UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

CARACTERIZACIÓN Y DETECCIÓN DE CONTRIBUYENTES QUE PRESENTAN FACTURAS FALSAS AL SII MEDIANTE TÉCNICAS DE DATA MINING

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PAMELA ANDREA CASTELLÓN GONZÁLEZ

PROFESOR GUÍA:
JUAN VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
SEBASTIÁN RÍOS PÉREZ
LUIS ABURTO LAFOURCADE
HUGO SÁNCHEZ RAMÍREZ

SANTIAGO DE CHILE
JULIO, 2012

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN GESTIÓN DE OPERACIONES Y AL TÍTULO
DE INGENIERO CIVIL INDUSTRIAL
POR: PAMELA CASTELLÓN
FECHA: 10/07/2012
PROF. GUÍA: JUAN VELÁSQUEZ SILVA

El presente trabajo de título tiene por objetivo identificar patrones de comportamiento de los contribuyentes declarantes de IVA, que transan facturas falsas para evadir impuestos. Para ello se utiliza el proceso KDD, el cual considera una serie de pasos y técnicas que permiten extraer conocimiento oculto a partir de un gran volumen de datos, para encontrar relaciones o patrones asociados a un cierto fenómeno.

La utilización y venta de facturas falsas tiene un impacto significativo en la recaudación que percibe el Estado, generando además efectos negativos que ponen en riesgo la competitividad de las empresas. Históricamente, la evasión por este concepto ha representado entre un 20% a un 30% de la evasión en el IVA, alcanzando la cifra de $450 millones de pesos durante la crisis económica de 2009. Adicionalmente, la detección, investigación, sanción y cobro de los impuestos adeudados, provoca un importante costo administrativo, debido a la cantidad de facturas transadas en el año y al tiempo requerido para su detección. En ese contexto, resulta necesario contar con procedimientos sistematizados y efectivos que gestionen la información disponible para detectar potenciales defraudadores de impuestos, focalizando los recursos en aquellos contribuyentes de mayor riesgo tributario.

Para la construcción del vector de características se utiliza la información de casos registrados con facturas falsas en el año 2006, considerando información del pago de impuestos en tal período, características particulares, comportamiento histórico en diferentes etapas de su ciclo de vida e indicadores del comportamiento de sus relacionados, entre otros. En una primera instancia, se aplican técnicas de SOM y Gas Neuronal, para analizar el potencial de contribuyentes que tienen un buen o mal comportamiento tributario e identificar sus características más relevantes. Posteriormente, se utilizan Árboles de Decisión, Redes Neuronales y Redes Bayesianas, para identificar aquellas variables que están relacionadas con un comportamiento de fraude y/o no fraude y detectar patrones de conducta, estableciendo en qué medida se pueden predecir estos casos con la información disponible.

El resultado indica que las variables que más discriminan entre fraude y no fraude en las micro y pequeñas empresas son el porcentaje de créditos generado por facturas, el resultado de las fiscalizaciones previas, la cantidad de facturas emitidas en el año y su relación con las facturas timbradas en los últimos dos años, el monto de IVA total declarado, la relación entre remanentes y créditos, los delitos e irregularidades históricas asociadas a facturas, y la participación en otras empresas. En las medianas y grandes empresas, en tanto, las variables más relevantes son la cantidad de remanente acumulado, el porcentaje de crédito asociado a facturas, el total de créditos, la relación entre gastos rechazados y activos, el capital efectivo, la cantidad de irregularidades previas asociadas a facturas, la cantidad de fiscalizaciones históricas, y el número de representantes legales.

En relación a los modelos predictivos, el mejor resultado se obtuvo con la red neuronal, donde el porcentaje de casos con fraude correctamente asignado fue de un 92% para las micro y pequeñas empresas, y de 89% para las empresas medianas y grandes. De acuerdo a esto y al potencial universo de usuarios de facturas falsas (120.768 empresas), se estima que con los modelos obtenidos se puede generar un potencial de recaudación de $101.446 millones de pesos al año, lo que permitiría reducir la evasión por concepto de IVA de manera significativa.

Finalmente, se concluye que es posible caracterizar y predecir contribuyentes que evaden impuestos a través de facturas falsas utilizando técnicas de Data Mining, y que los factores que inciden en la probabilidad que un contribuyente utilice facturas falsas dependen del tamaño o segmento del contribuyente, relación que hasta el momento se establecía sólo de manera intuitiva.

Se recomienda, para trabajos futuros, generar nuevas variables de comportamiento históricas relacionadas con fiscalizaciones y cobertura, explorar otros métodos para el preprocesamiento y selección de las variables, con los que eventualmente podrían obtenerse resultados diferentes. Igualmente, sería interesante explorar técnicas de validación cruzada y aplicar otras técnicas de data mining para mejorar la predicción de casos de fraude.

Dedicada a mis padres, María Mirtha y René

AGRADECIMIENTOS

En el camino recorrido hasta aquí tuve la oportunidad de conocer a muchas personas que contribuyeron en mayor o menor medida a la realización de este trabajo. Quiero agradecerles a todas ellas por el tiempo y dedicación prestados, por los momentos compartidos y el aprendizaje enseñado.

En particular, quisiera destacar a….

A mi Familia, por todo el amor, cariño y valores que me han entregado. Gracias a ustedes, hoy puedo cumplir una de las etapas importantes de mi vida.

A Daniel, por incentivarme a culminar este proceso y acompañarme en cada momento de este largo camino. Gracias por tu apoyo incondicional y el amor que me has entregado. A partir de hoy, comienza una nueva etapa.

A mi profesor guía Juan Velásquez, por la enseñanza, ayuda y paciencia al guiar mi trabajo. Gracias por sus consejos y las conversaciones sostenidas durante este proceso. Siempre estaré en deuda con usted.

A los profesores integrantes de mi comisión, por sus consejos y aportes al trabajo realizado. Especialmente a Hugo Sánchez por contribuir con su mirada experta y experiencia en temas tributarios y de fiscalización.

A mis amigos, Claudi, Marce, Patty, Enzo, Camilo, Joel y Saby, por su gran amistad y los maravillosos momentos que hemos pasado juntos a través de los años.

A Lorena, por darme la posibilidad de realizar este trabajo y sus aportes al inicio de este proceso.

A mi Universidad, por todas las experiencias vividas, y a sus grandes docentes, que siempre me orientaron con profesionalismo en la adquisición de nuevos conocimientos, afianzando mi formación académica.

A Julie Lagos, por responder todas mis consultas, siempre con buena disposición.

A toda la gente del Servicio de Impuestos Internos que contribuyó a este trabajo, ya sea aportando con su conocimiento, gestionando la realización de entrevistas y/o la entrega de información. Especialmente a Patricio Barra, Rodolfo Bravo, Bernardita Moraga, Lorena Cerda, Brandon Peña y Hugo Sánchez, quienes fueron parte fundamental de este trabajo.

ÍNDICE GENERAL

1. INTRODUCCIÓN

El Servicio de Impuestos Internos (SII) es la Institución responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario, propiciar la reducción de los costos de cumplimiento, y potenciar la Modernización del Estado y la Administración Tributaria en línea. Lo anterior en pos de fortalecer el nivel de cumplimiento tributario y el desarrollo económico de Chile y de su gente.

Un tema que ha sido una constante preocupación de todas las Administraciones Tributarias, en especial de aquellas pertenecientes a países en vías de desarrollo, es la "Evasión Tributaria". En efecto, la recaudación de impuestos es la principal fuente de ingresos de una nación, que le permite contar con los recursos necesarios para cumplir las funciones básicas de administración pública y entrega de servicios como salud, seguridad, educación, entre otras.

Los instrumentos para disminuir la evasión son variados. Algunos buscan aumentar la efectividad de la fiscalización y otros la colaboración de los contribuyentes. Entre los primeros, aquellos que aparecen más eficaces están relacionados con la recopilación, manejo y análisis de información, en cuyo caso el desafío es no afectar la privacidad de las personas y seleccionar casos para fiscalización de manera eficiente. Por otra parte, la colaboración de los contribuyentes crece, entre otros factores, con la calidad de los servicios tributarios, la simplicidad del sistema impositivo, el buen uso de la recaudación tributaria y la moderación en las tasas impositivas.

El fenómeno de las facturas falsas respecto del Impuesto al Valor Agregado (IVA) se explica por la mecánica de determinación del impuesto a pagar. Cuando una empresa recibe una factura falsa simula con ello una compra que nunca existió, con lo que aumenta fraudulentamente su crédito fiscal y disminuye su pago de IVA. Pero la factura falsa tiene un beneficio adicional para el evasor, pues incrementa también sus costos o gastos, disminuyendo su pago de Impuesto a la Renta. Es así como se ha ido generando una creciente industria de facturas falsas que se transan en el mercado, haciendo de la confección y distribución de documentos falsos un lucrativo negocio.

El SII utiliza actualmente diversos métodos para seleccionar contribuyentes para fiscalización, los cuales se basan en análisis de información parcial del contribuyente. Además, el SII maneja grandes volúmenes de datos, pues administra la información tributaria de más de 4 millones de contribuyentes, quienes realizan un sinfín de trámites durante su ciclo de vida, y por tanto, dispone de información valiosa para realizar análisis y detectar patrones de comportamiento.

Por otra parte, se requiere que los recursos invertidos en fiscalizar sean bien enfocados, lo que implica detectar a aquellos contribuyentes que realmente no cumplen con sus obligaciones tributarias, y no importunar ni desperdiciar tiempo y recursos en aquellos que cumplen de manera correcta. Para ello, resulta fundamental aprovechar los avances de la tecnología para mejorar el uso de los recursos, además de desarrollar nuevas metodologías que permitan descubrir otros patrones y formas de caracterizar el comportamiento de los contribuyentes.

1.1. ANTECEDENTES GENERALES

1.1.1. EL IMPUESTO A LAS VENTAS Y SERVICIOS

El Impuesto al Valor Agregado (IVA) es el principal impuesto al consumo que existe en Chile, que grava la venta de bienes corporales muebles e inmuebles; cuando son de propiedad de una empresa constructora construidos totalmente por ella o en parte por un tercero para ella; y los servicios que se presten o utilicen en el país que provengan de las actividades que la Ley señala[1]. A partir del 1 de octubre de 2003, dicho tributo se aplica con una tasa del 19% sobre la base imponible de ventas y servicios establecidos por la Ley, debiendo ser declarado y pagado mensualmente.

Son contribuyentes de este impuesto, las personas naturales o jurídicas, incluyendo las comunidades y las sociedades de hecho, que realicen ventas, presten servicios o efectúen alguna operación gravada de impuesto, entendiéndose por operación gravada aquellas situaciones que de acuerdo con la ley dan origen a la obligación de pagar impuesto. Es decir, contribuyentes que generen actividades económicas de primera categoría (industria, comercio, agricultura, otros) o actividades económicas de segunda categoría cuando corresponda.

El IVA afecta al consumidor final, pero se genera en cada etapa de la comercialización del bien. El monto a pagar surge de la diferencia entre el débito fiscal, que es la suma de los impuestos recargados en las ventas y servicios efectuados en el período de un mes y el crédito fiscal. Este último equivale al impuesto recargado en las facturas de compra y de utilización de servicios, y en el caso de las importaciones, al tributo pagado por la importación de especies.

Este impuesto se aplica en más de 130 países en diferentes etapas de desarrollo económico, convirtiéndose en un componente clave de la recaudación fiscal. Es así como para los países miembros de la OCDE[2], el IVA representa aproximadamente el 18% del total de los ingresos fiscales recaudados.

En el caso de Chile los ingresos tributarios proporcionan alrededor de un 75% de los recursos con que año a año el Estado sustenta sus gastos e inversiones, alcanzando durante el año 2011 los $21,5 billones de pesos[3]. El Impuesto al Valor Agregado representa el 45% de este monto, aportando un total de $9,7 billones de pesos, lo que da cuenta de la importancia que este impuesto tiene para el desarrollo económico del país.

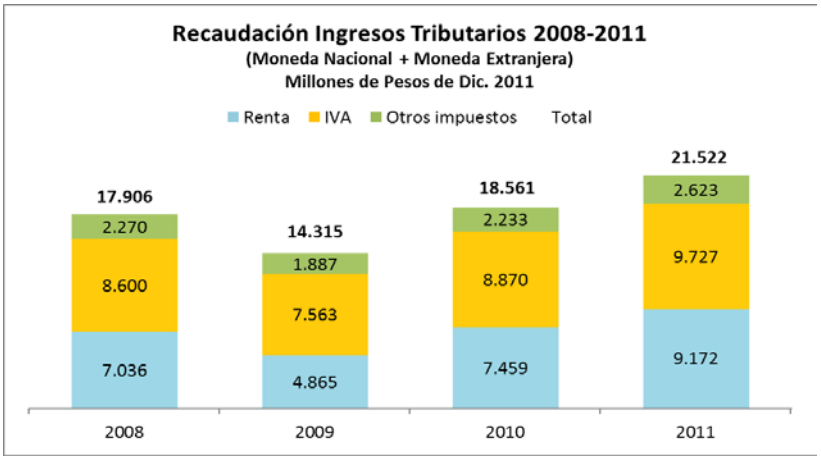
---

[1] Ley sobre el Impuesto a las Ventas y Servicios, D.L.N° 825.

[2] Los 30 países miembros de la OCDE representan a países que han alcanzado un nivel relativamente alto de desarrollo y comparten un compromiso con la economía de mercado y la democracia pluralista. Sus miembros representan el 60% del producto nacional bruto del mundo, las tres cuartas partes del comercio mundial y el 14% de la población mundial.

[3] Información publicada en la Cuenta Pública SII 2011 de mayo 2012, considerando los Ingresos Tributarios del Gobierno Central (sin incluir a Codelco, las Municipalidades y la Seguridad Social).

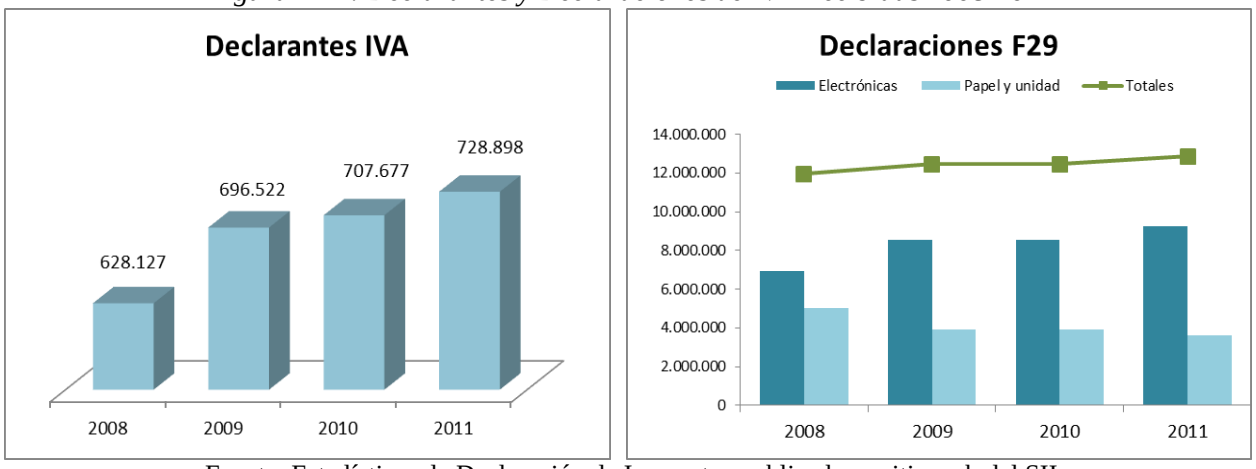
Figura N° 1: Recaudación Principales Impuestos 2008-2011



Fuente: Subdirección de Estudios SII en base a Informes de Ejecución Presupuestaria Sector Público DIPRES

Actualmente existen 729 mil contribuyentes que declaran IVA, quienes durante el año 2011 generaron 12,8 millones de declaraciones del F29, de las cuales un 72% se realiza en formato electrónico a través del sitio web del SII y un 28% se recibe en formato papel. Esto debido a la política adoptada por la institución para modernizar su gestión, la cual ha generado un aumento sostenido del uso de medios electrónicos para realizar la declaración y pago de impuestos.

Figura N° 2: Declarantes y Declaraciones de IVA Recibidas 2008-2011



Fuente: Estadísticas de Declaración de Impuestos publicada en sitio web del SII

1.1.2. LAS FACTURAS

Las facturas son documentos que deben emitir los contribuyentes del IVA en la enajenación de bienes corporales muebles y/o prestación de servicios, afectos o exentos, que efectúan con cualquier persona natural o jurídica que hubiese adquirido los bienes para su reventa, uso o consumo o que tengan la calidad de prestadores de servicios.

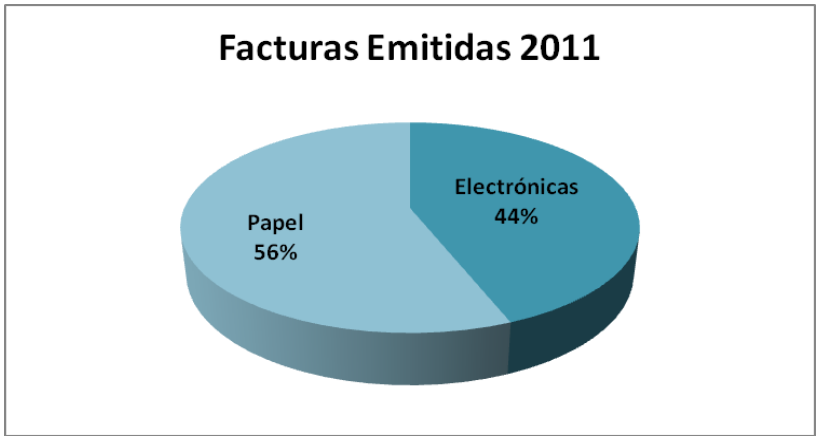
Deben emitir facturas los vendedores e importadores, en las operaciones que realicen con otros vendedores, importadores o prestadores de servicio, tratándose de ventas o promesa de venta de

inmuebles, contratos de instalación o confección de especialidades y contratos generales de construcción, y aquellos beneficiarios de la prestación afectos al mismo tributo, para acreditar que se le ha recargado separadamente el impuesto y hacer efectivo el crédito fiscal. No obstante, la Dirección Nacional del Servicio está facultada para eximir de la obligación de emitir facturas y boletas a determinadas actividades o grupos[4], como contribuyentes que presten servicios o vendan productos exentos y contribuyentes afectos a los impuestos establecidos en el D.L. N° 825, entre otros.

Estos documentos son de gran importancia probatoria en materia tributaria, tanto respecto del uso legítimo del crédito fiscal del IVA como de la justificación del cargo lícito de costos y gastos en la determinación del Impuesto sobre las Rentas. Se trata de un contrato escrito que justifica ingresos y egresos, que debe ser congruente principalmente con la caja y otros libros de contabilidad. Por otro lado, la factura es el documento más fiscalizado y reglamentado, ya que permite cometer grandes fraudes tributarios o fiscales a través de su no emisión, falsificación o simulación.

Actualmente, gran parte de las facturas se emiten en papel, sin embargo, como parte de la política adoptada para modernizar la gestión del SII y utilizar Internet como canal de comunicación con los contribuyentes, a partir del año 2003 se comienza a utilizar la factura electrónica (e-factura) como un medio para asegurar la autenticidad de sus emisores y cautelar la integridad de los documentos. Es así como de un total de 298 millones de facturas emitidas durante el año 2011, un 56% se confeccionan en formato papel y un 44% en formato electrónico[5].

Figura N° 3: Distribución Facturas Emitidas 2011



Fuente: Cuenta Pública SII 2011

La e-factura otorga la ventaja crucial de poder verificar su validez a través del sitio web del Servicio (www.sii.cl). Adicionalmente, constituye un golpe de timón para introducir una reforma micro-económica de alto impacto en la productividad y competitividad de las empresas. Por este motivo, el Servicio de Impuestos Internos se encuentra constantemente en un proceso de masificación de la e-factura a través de diversas iniciativas.

[4] Señalados taxativamente en el Inc. 1° del Art. N° 56 de la Ley de IVA.

[5] Información publicada en la Cuenta Pública SII 2011, mayo 2012.

1.1.3. LA EVASIÓN EN EL IVA

La evasión de impuestos corresponde a lo que el Estado deja de percibir a raíz del incumplimiento tributario, es decir, producto de que los contribuyentes no declaran lo que les corresponde de acuerdo con sus obligaciones tributarias, ya sea en forma voluntaria o involuntaria.

En el caso del IVA, los mecanismos utilizados para evadir pasan necesariamente por una ***subdeclaración de los débitos***, o bien, por un ***abultamiento de los créditos***. En términos simples, el evasor registra menos ventas, y por tanto menos débitos de IVA o, alternativamente, más compras y más créditos de IVA de los que en realidad realiza, y que según lo establece la ley deberían ser consignados por éste para determinar su obligación fiscal.

Las figuras evasoras de débitos más recurrentes son las ventas sin comprobante, en especial las ventas que se realizan a consumidor final[6]; el uso fraudulento de notas de crédito y la subdeclaración en los registros contables y en las declaraciones tributarias. Para abultar créditos, en tanto, los evasores recurren a mecanismos como las ***facturas falsas***, compras personales que se registran a nombre de la empresa; compras a contribuyentes ficticios y sobredeclaración en los registros contables y declaraciones tributarias.

Para cuantificar la evasión, las Administraciones Tributarias utilizan tanto métodos directos como indirectos. Los primeros parten de la realización de muestreos estadísticos, para obtener información de los contribuyentes. Los métodos indirectos, por su parte, buscan una relación causal entre el fraude y determinadas variables económicas observables. Dentro de estos métodos se encuentran los métodos de contraste, que dan una medida de la evasión fiscal basada en fuentes tributarias y/o en cifras de las cuentas nacionales, así como también en estadísticas oficiales nacionales.

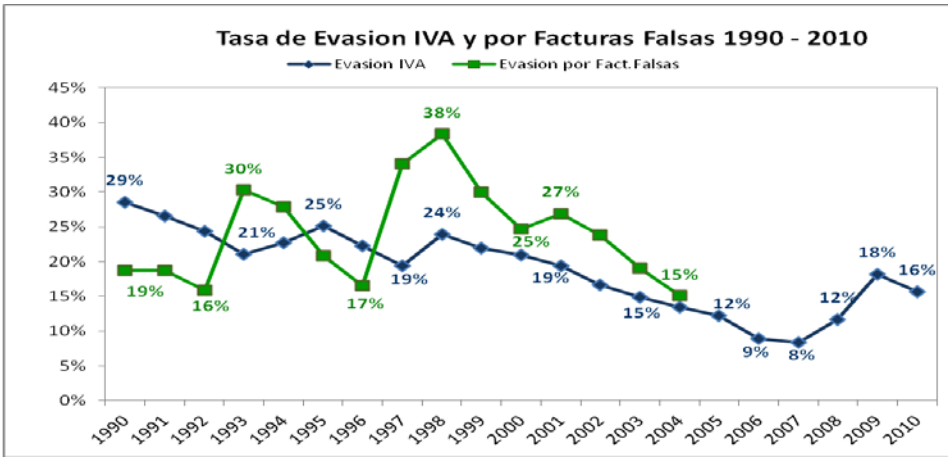
En el caso de Chile, la estimación de IVA se obtiene a través de la aplicación del método basado en cuentas nacionales, el cual construye un IVA teórico a partir de la información del consumo de hogares, que luego es comparado con la recaudación efectiva del período. De acuerdo con este método, la tasa de evasión del IVA alcanzaba un 29% en el año 1990, la cual fue disminuyendo desde el año 2000 en adelante, debido a la aplicación de la Ley de Lucha contra la Evasión, llegando en 2007 a un mínimo histórico del 8%. Sin embargo, producto de la crisis económica que afectó a Chile a fines del año 2008, se observa un alza considerable en el incumplimiento tributario, llegando a una tasa de evasión del 18% en 2009, por un monto de $1,5 billones de pesos[7], la cual disminuye al año siguiente a un 16%.

---

[6] En las transacciones a nivel de consumidor final el comprobante de venta es una ‘boleta de venta’, mientras que en las transacciones intermedias, el comprobante es una ‘factura’.

[7] Análisis posteriores indican que la evasión del IVA en el año 2009 podría haber llegado incluso a un 23% por un monto de $2 billones de pesos, disminuyendo a un 22% en el año 2010 y a un 20% en el año 2011, de acuerdo a las nuevas cifras publicadas de las cuentas nacionales 2012 y a la nueva Matriz Insumo Producto del año 2008. Sin embargo, a la fecha de este documento esas cifras sólo tienen carácter provisional.

Figura N° 4: Tasa de evasión estimada del IVA para el período 1990 - 2010

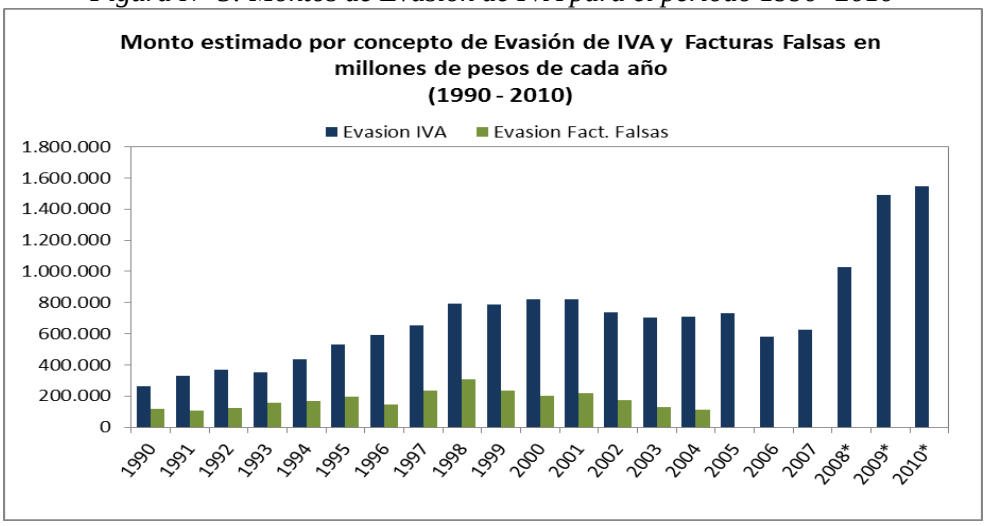


Fuente: Subdirección de Estudios del SII sobre la base de información del Banco Central

A su vez, el SII ha desarrollado un método para estimar la evasión del IVA por concepto de facturas falsas y otros abultamientos de créditos, aplicado en el período 1990-2004. De acuerdo a esto, en ese lapso, la evasión por facturas falsas ha representado entre un 15% y un 25% de la evasión total del IVA, aumentando considerablemente en años de crisis económicas. Es así como en el año 1992, el porcentaje de participación aumentó a un 30% y en la crisis de 1998-1999 alcanza su punto máximo con un 38% de participación, año en que alcanza una cifra cercana a los $317.000 millones de pesos.

Si bien no existe una cifra oficial que indique qué porcentaje representa la evasión por facturas falsas en los últimos años, se estima que la evasión por utilización y venta de facturas falsas en 2009 podría haber llegado hasta un 30% de la evasión en el IVA, de acuerdo con lo acontecido en las crisis anteriores, lo que equivale a $450.000 millones de pesos[8].

Figura N° 5: Montos de Evasión de IVA para el período 1990- 2010



(*) Cifras provisionales de acuerdo a actualización de cuentas nacionales del Banco Central
Fuente: Subdirección de Estudios del SII sobre la base de información del Banco Central

[8] Análisis posteriores indican que la evasión en el IVA durante el año 2009 puede haber llegado incluso a los $2 billones de pesos, lo que implica una evasión por facturas falsas del orden de los $615 millones de pesos (cifras de carácter provisional).

Por otra parte, las Administraciones Tributarias requieren gastar importantes recursos en auditorías y actividades de seguimiento para la detección de la evasión fiscal. Un reporte del año 1993 de la GAO[9], estimó que más del 70% anual de los costos administrativos de IVA están relacionados con la fiscalización de ese impuesto.

1.1.4. EL FENÓMENO DE LAS FACTURAS FALSAS

Todo empresario, comerciante y distribuidor, debe pagar IVA tanto por los insumos que compra para producir, como por las ventas que realiza. Al momento de pagar los impuestos, la ley otorga el derecho de recuperar el IVA que pagó por las materias primas adquiridas (llamado crédito fiscal) aliviando su carga impositiva. De esta manera, hay quienes pretenden engañar al fisco y recuperar más IVA, declarando compras de insumos con facturas que son falsas[10].

Conforme con las definiciones existentes en el SII, se entiende por **"factura falsa"** aquella que, en su materialidad o contenido, falta a la verdad. La falsedad del documento puede ser **"material"**, si en él se han adulterado los elementos físicos que conforman la factura. Esto ocurre, cuando se produce una imitación de un modelo verdadero y/o legítimo. Por ejemplo, cuando se hace una factura en una imprenta consignando un nombre, domicilio, RUT, numeración u otro dato o elemento esencial que no corresponda a la realidad. Así también, cuando estos antecedentes son verídicos, y se imitan los timbres y sellos del SII. La falsedad puede ser también **"ideológica"**, cuando la materialidad del documento no está alterada, pero las operaciones que en ella se consignan son engañosas o inexistentes.

En general la falsedad ideológica es más difícil de detectar, ya que implica transacciones no existentes o adulteradas que no se descubren a simple vista, y requiere una investigación o auditoría en la que se revisen los libros de compra, las rectificaciones y se realicen cruces de información con clientes y proveedores para su detección. Por otra parte, estos casos son más costosos para el SII, ya que requieren una mayor cantidad de tiempo destinado a la recopilación de antecedentes y pruebas, que a su vez son más difíciles de encontrar.

Entre los casos importantes que se han descubierto, se encuentra el fraude realizado por el ex dueño de Lozapenco, entre los años 1987 y 1990, el cual alcanzó los USD 46 millones de dólares producto de más de mil declaraciones, en las cuales se abultaba el valor de los productos que se enviaban al exterior (palos de escoba y lavatorios con pedestal), gracias al beneficio de reintegro del IVA por exportaciones, obteniendo una devolución tributaria mensual cercana a los USD 2 millones de dólares. Dentro de este último grupo también se encuentra el caso Publicam del año 2006, el cual ha sido el ejemplo de fraude más grande descubierto en los últimos 5 años, a raíz del cual el SII reconoció la existencia de una verdadera mafia de empresas dedicadas al comercio ilegal de facturas, al detectar 67 empresas que actuaron como emisoras de facturas falsas y 81 empresas receptoras de éstas.

---

[9] GAO Tax Policy, Value Added Tax: Administrative Costs Vary with Complexity and Number of Businesses (Washington DC, May 3, 1993).

[10] El artículo 21 del Código Tributario indica que "corresponde al contribuyente probar con los documentos, libros de contabilidad u otros medios que la ley establezca, en cuanto sean necesarios u obligatorios para él, la verdad de sus declaraciones o la naturaleza de los antecedentes y monto de las operaciones que deban servir para el cálculo del impuesto".

Otro caso que despierta la atención son las “empresas de papel” que tienen un calce casi perfecto entre el IVA de su ventas y el de sus compras, ya que no es normal que una persona compre para vender sin ganancias, durante un período prolongado de tiempo. También es común que algunos contadores utilicen cheques y facturas de sus clientes para cometer delitos en su nombre y paguen sus propias imposiciones, o de terceros, con dineros de sus empleadores. Esto es posible porque, según la legislación actual, un cheque nominativo a la Tesorería General de la República sirve para pagar las contribuciones de cualquier ciudadano.

Los “vendedores de IVA” saben que si van a mostrar papeles falsos, éstos deben estar dentro del rango de timbraje del cliente, o bien si se van a “colgar” del RUT de una empresa, ésta debe ser una que emita miles de facturas mensualmente para “colarse” subrepticiamente dentro de las ventas. Generalmente, los falsificadores tienen “palos blancos” o testaferros que concurren al SII para iniciar actividades profesionales, obtener facturas timbradas para venderlas, o arriendan oficinas que jamás van a funcionar para el rubro que declaran en sus documentos.

El SII al efectuar su labor de fiscalizar el correcto empleo de los créditos fiscales, investigará si se producen las situaciones descritas, para cuyo efecto deberá proceder a la verificación del emisor de la factura y de los elementos que conforman la operación, tales como: individualización de los contribuyentes, RUT, iniciación de actividades, timbraje de facturas, domicilios, medios y forma de pago de la mercadería, antecedentes contables (libros, documentación, inventarios, libro de existencias), efectividad de la operación (guías de despacho que acrediten el traslado de la mercadería, antecedentes de recepción en bodega), circulación de facturas de contribuyentes que han hecho término de giro, etc.

Igualmente hay que considerar que la falsedad que afecta a una factura es consecuencia del actuar doloso o malicioso de una o más personas: el emisor, el receptor o terceros que las confeccionan, venden o facilitan a cualquier título. Lo anterior implica que se está en presencia de un delito tributario sancionado penalmente, por lo que la calificación de falsedad de una factura es un elemento esencial que debe ser apropiadamente acreditado con los medios de prueba legales.

A partir de las entrevistas realizadas a fiscalizadores y a los documentos existentes en el SII, se identifican las siguientes situaciones asociadas a facturas falsas, las cuales se agrupan considerando su correspondencia con casos de falsedad material o ideológica u otros tipos de delitos relacionados con la utilización de facturas falsas para evadir impuestos.

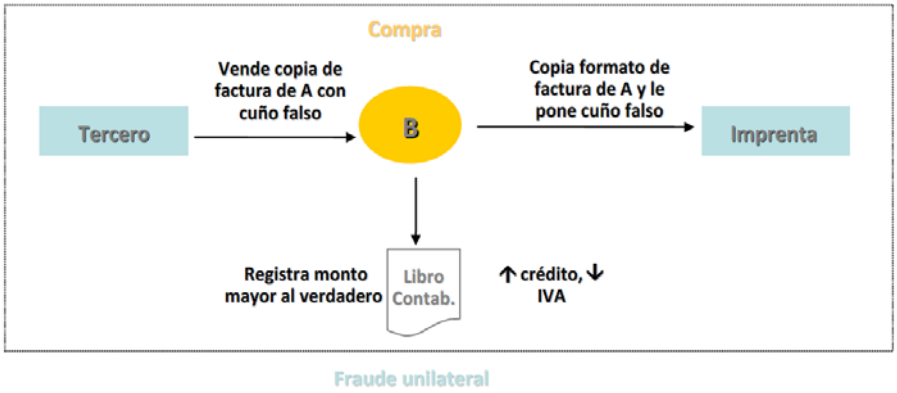
1.1.4.1. Facturas Materialmente Falsas

A continuación se describen los tipos de facturas falsas detectados que implican una adulteración material del documento tributario:

a) *Facturas Colgadas*: Esta figura se da cuando una persona toma los datos de una empresa real y manda a confeccionar las facturas a nombre de esa empresa, sin autorización ni conocimiento alguno por parte de ésta, para luego emitirla a algún comprador propio. Para ello se suplanta a contribuyentes de buen comportamiento, que tienen buena imagen comercial, utilizando la numeración y formato de facturas timbradas, pero con un timbre falso para darle

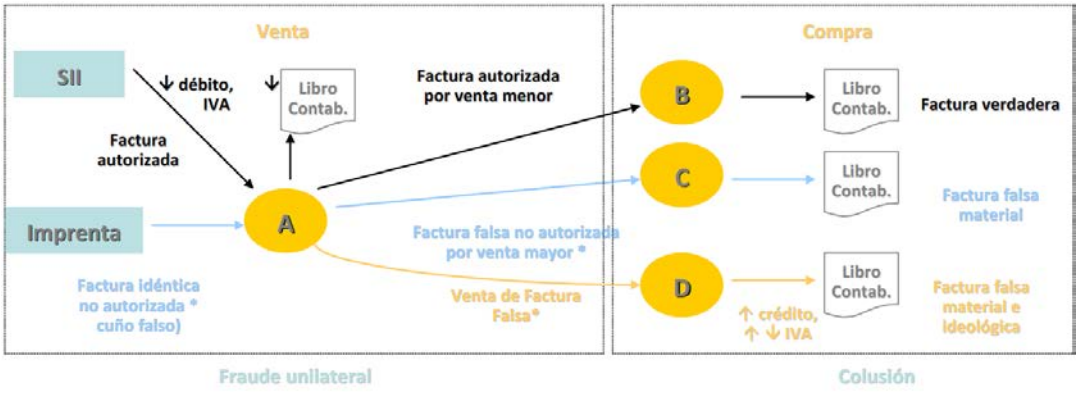
apariencia de verdaderas, para respaldar operaciones de compra. En algunos casos, se llega incluso a falsificar resoluciones del SII y mandatos de contribuyentes para perfeccionar el delito y realizar operaciones que no cumplen con el pago de los impuestos. Esta adulteración, afecta tanto al comprador de la factura como a la persona cuyos datos fueron usados para su confección, así como también a la imprenta que confecciona la factura sin requerir la identificación necesaria para elaborar el documento falso.

Figura N° 6: Esquema de operación de facturas colgadas



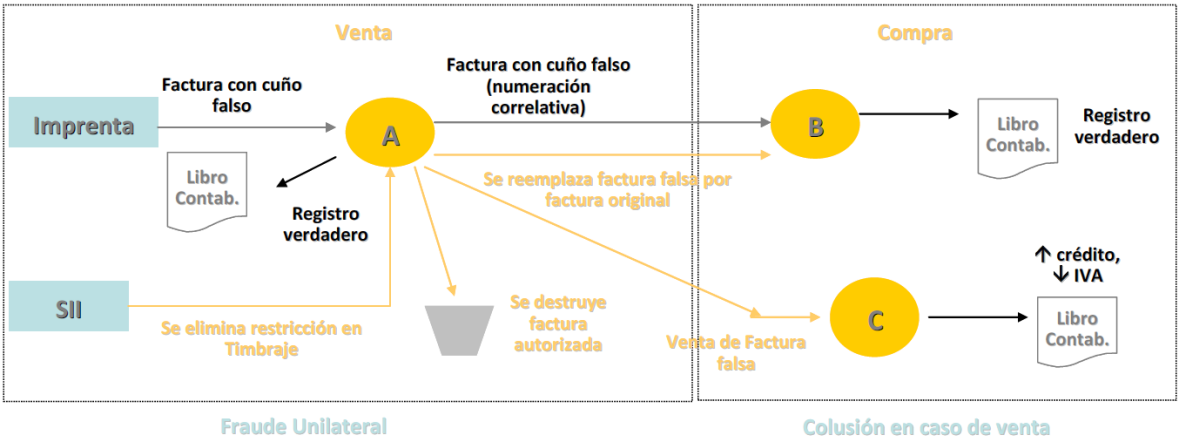
b) *Doble Juego de Facturas*: En este caso se opera con dos facturas, una timbrada y autorizada legalmente por el SII, y otra de formato y numeración idéntica, pero timbrada con un cuño falso. Con la documentación legítima se da cuenta sólo de algunas operaciones realizadas, con el propósito de reflejar un movimiento inferior al real. En tanto, con el juego de facturas irregulares, se registran operaciones más numerosas y por montos más elevados. Este caso corresponde a un ejemplo de factura falsa material, en la que se efectúa una venta con la finalidad de disminuir los débitos. En ocasiones, este mecanismo va acompañado de una venta de factura para respaldar una operación inexistente que requiere el concierto de dos o más personas: el que vende o proporciona las facturas y el que las adquiere para su uso fraudulento. Este es uno de los casos más comunes y difíciles de hallar, pues su detección sólo es posible mediante el cruce de información cliente-proveedor, que en algunos casos deja en evidencia la emisión de 2 ó 3 facturas con el mismo número correlativo, pero que amparan operaciones absolutamente distintas.

Figura N° 7: Esquema de operación de doble juego de facturas



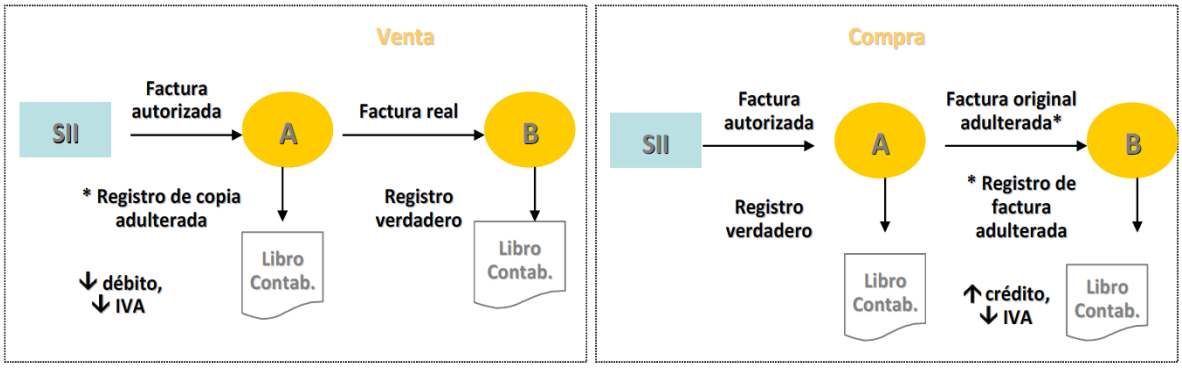
c) *Utilización de Factura Falsa Material por Timbraje Restringido*: También puede ocurrir que un contribuyente emita una factura falsa, debido a que se encuentra con timbraje restringido, sin el propósito de evadir impuestos, sino más bien de respaldar una operación existente. Una vez que se regulariza la situación y el Servicio autoriza el timbraje de documentos, pueden darse tres situaciones: (i) El contribuyente destruye la factura para no tener dos facturas con igual numeración. (ii) El contribuyente le solicita al cliente reemplazar la factura falsa por la factura autorizada, explicándole la situación. (iii) El contribuyente vende la factura autorizada por el Servicio a un comprador que la utiliza para abultar sus créditos. En este caso se entrega factura falsa con igual numeración (caso de doble juego de facturas).

Figura N° 8: Esquema de operación de facturas falsas por timbraje restringido



d) *Adulteración de Factura Original de Compra o Registro de Copia de Venta Adulterado*: Puede darse el caso de que un contribuyente adultera físicamente la factura entregada con el objeto de abultar el crédito fiscal de IVA y los costos o gastos del Impuesto a la Renta, en los libros de compra. Lo anterior corresponde a un caso de utilización de factura falsa material e ideológica, producto de que se altera físicamente el documento y se registra una operación adulterada. Para esta acción no se requiere participación de terceros, y corresponde a un caso de fraude unilateral. Sin embargo, al igual que el ejemplo anterior, no es una situación muy común, ya que la adulteración de la factura puede ser fácil de detectar al visualizarla. De la misma manera en el caso de las ventas, se puede adulterar la copia de la factura emitida para disminuir los débitos, caso que corresponde a una factura irregular.

Figura N° 9: Esquema de operación de adulteración del original de la compra

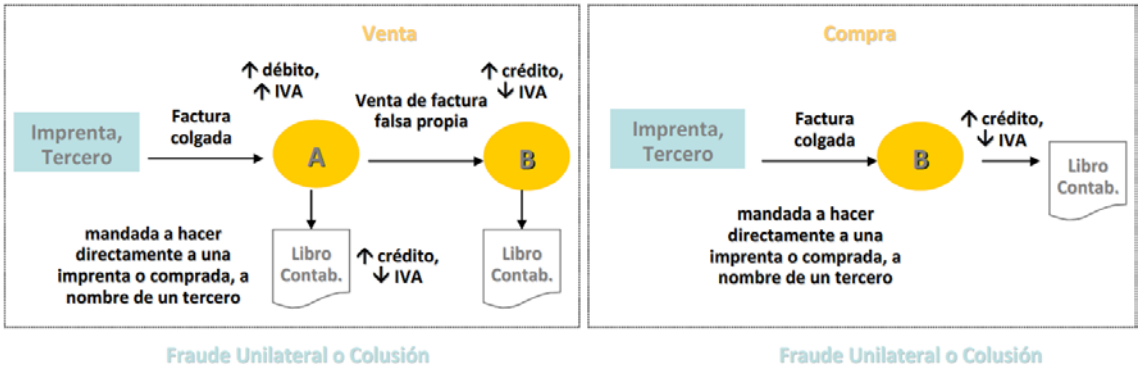


1.1.4.2. Facturas Ideológicamente Falsas

A continuación se describen los tipos de facturas falsas detectados que no implican necesariamente una adulteración material del documento tributario:

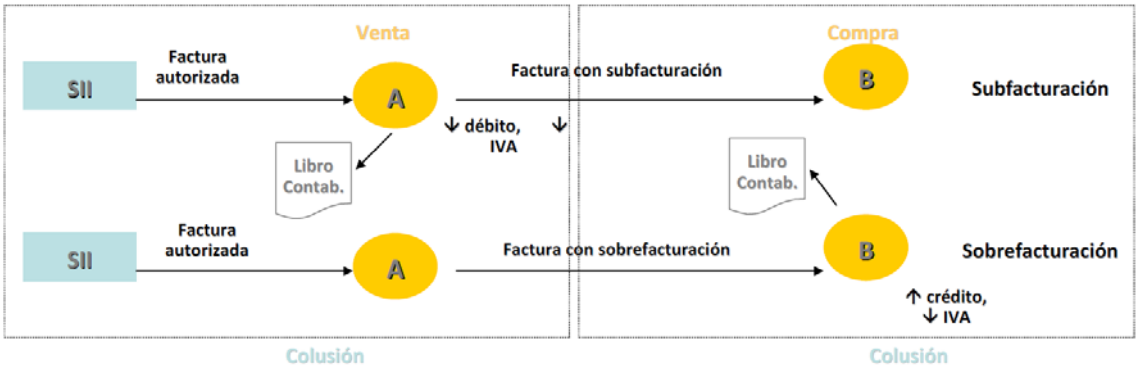
e) *Con Operación Inexistente*: El caso más común de este tipo de evasión, es la utilización de una factura falsa para registrar una operación inexistente, ya sea para disminuir el débito fiscal o abultar los créditos. En ambos casos, se requiere una factura colgada y/o la compra de una factura a nombre de un tercero. Si se utiliza una factura falsa para aumentar los créditos, probablemente el contribuyente mande a hacer una factura a una imprenta con un cuño falso o compre una factura a nombre de un tercero. En el caso de disminuir el débito, el contribuyente puede utilizar este mecanismo para compensar la venta de facturas falsas propias del contribuyente, con el objeto de aumentar el crédito fiscal. Esta forma de operar requiere de la colusión de la persona que vende o proporciona las facturas (imprenta, contador, tercero), y el contribuyente que las adquiere o recibe para su uso fraudulento.

Figura N° 10: Esquema de operación de facturas falsas por operación inexistente



f) *Subfacturación o Sobrefacturación de Ventas*: Otra forma de evasión es adulterar el contenido de una operación existente, ya sea registrando un monto mayor o menor al real, o registrando especies diferentes a las verdaderas. Este mecanismo se utiliza principalmente en los sectores primarios, emitiendo facturas que indican un producto de calidad inferior para registrar un valor menor, o sencillamente, se valoran a un precio inferior las operaciones. El vendedor entrega una factura a un menor valor del pagado por el comprador para pagar menos IVA. Para ello se requiere el concierto del vendedor y el comprador, que se coluden para obtener un beneficio ilegítimo en común en perjuicio del interés fiscal.

Figura N° 11: Esquema de operación por subfacturación o sobrefacturación de las ventas

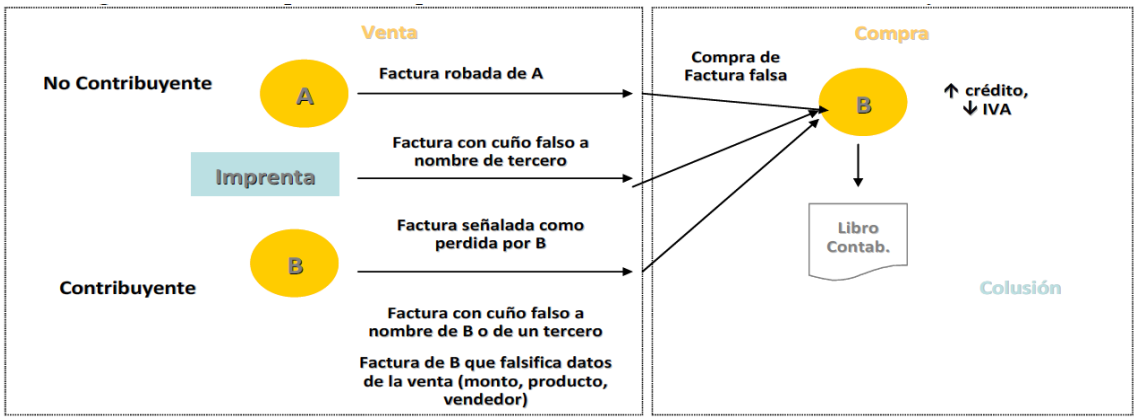


1.1.4.3. Otros Delitos Asociados

Existen también otros delitos asociados al uso de facturas falsas, los cuales pueden ser cometidos por personas no contribuyentes de IVA, entre los que se encuentran:

g) *Venta de Facturas*: Como se menciona en los ejemplos anteriores, para respaldar un operación inexistente y evadir impuestos, se requiere de la facilitación de una factura a nombre de un tercero, la cual puede ser fabricada por una imprenta, imitando el sello del SII a través de un cuño falso o hurtada a sus legítimos dueños por mandatarios, contadores, dependientes o terceros y vendidas a contribuyentes para respaldar costos o gastos del Impuesto a la Renta o crédito fiscal en el IVA. En ambos casos, se produce un negocio lucrativo, ya que al vender y entregar facturas que no le pertenecen, no se tiene que declarar ni pagar impuestos. Actualmente existen verdaderas mafias, las que se constituyen como empresas de papel, con el único fin de vender y facilitar documentos falsos a otros contribuyentes. Esta figura es compleja desde el punto de vista legal, ya que hace alusión a una organización criminal normalmente constituida para engañar al Fisco, con cierto patrón de conducta destinado a hacer muy complejo su accionar y dificultar su detección.

Figura N° 12: Esquema de operación del delito asociado a la venta de facturas



h) *Inicio de Actividades Fraudulento*: Respecto del punto anterior, se ha detectado que normalmente los falsificadores utilizan “palos blancos” para realizar el trámite de inicio de actividades, valiéndose de personas jóvenes o ancianas, de escasos recursos, con nulos conocimientos tributarios, quienes no tienen antecedentes en el Servicio, para que actúen como mandatarios y proporcionen sus antecedentes para dar inicio a la empresa con fines fraudulentos. Por otra parte, utilizan “domicilios virtuales” con el único fin de cumplir con el procedimiento de verificación en terreno que realiza el SII antes de autorizar el timbraje de documentos. También se da el caso de declaraciones de inicio falsas, cuando figura un tercero utilizando información de otra persona para crear una empresa a través de una cédula de identidad robada, sin consentimiento del tercero.

Cabe señalar que el hecho que un contribuyente tenga registrado en su libro de compras, como respaldo del crédito fiscal del Impuesto a las Ventas y Servicios, una o más facturas falsas, no necesariamente implica que éste tiene participación en la falsedad. Lo anterior, por cuanto se debe considerar la posibilidad de que no obstante haber actuado de buena fe, haya sido engañado por terceros inescrupulosos quienes le hayan otorgado una factura falsa por una operación real y

que el contribuyente puede no estar involucrado directamente en el delito, pudiendo éste ser cometido por el contador, un socio o un tercero. Por otro lado, que una factura contenga información errónea, no implica que haya intencionalidad de cometer fraude y por tanto, que sea falsa.

Es por esto, que el SII diferencia la factura falsa de otras irregularidades que se pueden encontrar en dichos documentos y que ameriten el rechazo del crédito fiscal, pero que no constituyen un delito, entre las que se encuentran:

- *Facturas no fidedignas*, que contienen irregularidades materiales que hacen presumir con fundamento que no se ajustan a la verdad, como tener una numeración y datación que no guarda correlatividad con el resto de la facturación; facturas con correcciones, enmendaduras o interlineaciones en su texto; facturas que no guardan armonía con los asientos contables que registran la operación de que dan cuenta; y casos en que no haya concordancia entre los distintos ejemplares de la misma factura.
- *Facturas que no cumplen con los requisitos legales o reglamentarios del SII*, como la omisión del número o timbre, información incompleta del emisor (nombre completo, rol único tributario, dirección del establecimiento, giro del negocio), información incompleta del comprador (nombre, rol único tributario, dirección), información incompleta de la operación (fecha de emisión, detalle de la operación, recargo separado del impuesto, condiciones de la venta), formato de la factura (dimensiones mínimas, color blanco del papel, fondo impreso, nombre del documento, recuadro rojo) o la unidad del Servicio en que debe efectuarse el timbraje, entre otros.
- *Facturas que han sido otorgadas por personas que no resulten ser contribuyentes del IVA*, ya sea por tratarse de personas que carecen de la calidad de vendedores habituales de bienes corporales muebles o inmuebles de su propiedad, construidos totalmente por ella o que en parte hayan sido construidos por un tercero para ella, o del carácter de prestadores de servicios afectos al IVA, o bien, por no ser sujetos de las demás operaciones que constituyen hechos gravados con el tributo.

Con relación a esta materia, es preciso señalar que la circunstancia de que el contribuyente no haya efectuado declaraciones de impuestos no lo priva de la calidad de contribuyente del IVA, sino que lo constituye en un contribuyente no declarante, circunstancia que el comprador no está en condiciones de verificar. Asimismo, que el emisor de la factura no sea ubicado con posterioridad al otorgamiento del documento en el domicilio indicado en éste, presente inconcurrencia a notificaciones del Servicio o se encuentre observado negativamente en los sistemas informáticos que mantiene la Administración Tributaria, no le quita su condición de contribuyente del IVA y no es posible fundar sólo en tales hechos el rechazo del crédito fiscal.

La detección, investigación, sanción y cobro de los impuestos adeudados como consecuencia del uso de estos documentos, generan un importante costo administrativo para el SII, que ha establecido planes específicos para la detección de estos casos. En el período 2001-2007 se han presentado más de 2.300 querellas por facturas falsas y otros delitos de defensa judicial, las cuales involucraron a más de 4.000 querellados, por un monto de perjuicio fiscal cercano a los $274.130 millones de pesos.

Cuadro N° 1: Estadísticas de acciones legales relacionadas con facturas falsas 2001-2007

ESTADÍSTICAS SCE	2001	2002	2003	2004	2005	2006	2007	Acumulado
Cantidad de Querellas	171	394	358	407	451	306	243	2.330
Cantidad de Querellados	371	835	667	839	801	537	386	4.436
Monto Perjuicio Fiscal (MM$)	29.370	36.407	49.751	58.812	47.856	21.620	30.314	274.130
Casos SCE[11]	830	2081	1794	1609	1553	1052	870	9.789

Fuente: Cuenta Pública SII 2005, 2006, 2007

En la práctica las querellas se efectúan por una combinación de las tipologías de fraude antes mencionadas. Es así como la mayor cantidad de querellas interpuestas considera la utilización de facturas material e ideológicamente falsas, junto a la venta de facturas. No menos importantes son los casos registrados por concepto de suplantación de identidad, usurpación, inicio de actividades falsa y fabricación de cuño falso, denominados "Otros delitos", el cual es el segundo en cantidad de casos registrados.

1.2. DEFINICIÓN Y JUSTIFICACIÓN DEL PROYECTO

Como se mencionaba anteriormente, en Chile los impuestos proporcionan alrededor de un 75% de los recursos con que el Estado financia los gastos e inversiones que realiza en beneficio de la comunidad. Durante el año 2011 se obtuvieron $21,5 billones de pesos por concepto de recaudación de impuestos[12], de los cuales el Impuesto al Valor Agregado tiene una participación cercana al 45% dentro del total de los ingresos tributarios.

En ese mismo año, la evasión en el IVA alcanzó los $2 billones de pesos[13], aumentando considerablemente a partir de la crisis económica mundial que impactó a Chile a fines de 2008 y mediados de 2009, llegando a tasas de evasión cercanas al 23%[14]. Por otra parte, se estima que la evasión por concepto de facturas falsas podría haber llegado a un 30% de la evasión total del IVA en ese año, con un potencial de evasión del orden de los $615 mil millones de pesos, comparando los créditos declarados por los contribuyentes en el IVA con los débitos por ventas intermedias.

Además de afectar la recaudación, la utilización de facturas falsas distorsiona el comportamiento de los contribuyentes, generando efectos económicos negativos, ya que las empresas cumplidoras deben enfrentarse a la competencia desleal de las incumplidoras. Al mismo tiempo, se pone en riesgo la existencia de empresas eficientes, que generan empleo y que aportan recursos a la financiación de las actividades públicas, pero que se ven incapaces de competir ante empresas que se valen del incumplimiento de sus obligaciones tributarias para desplazar a las primeras.

---

[11] El Sistema de Control de Expedientes (SCE) contiene los antecedentes de los casos asociados a delitos tributarios. Considera más casos que el número de querellados, pues comprende la información de todos los contribuyentes involucrados en la investigación del caso, incluyendo a mandatarios, proveedores, socios, contadores, entre otros.

[12] Información publicada en la Cuenta Pública SII 2011, mayo 2012, considerando ingresos tributarios del gobierno central.

[13] Estimación elaborada por la Subdirección de Estudios del SII, en base al método del potencial teórico de cuentas nacionales.

[14] Cifra provisoria, considerando las nuevas cifras publicadas de las cuentas nacionales en el año 2012 y la actualización de la Matriz Insumo Producto 2008, que generan un aumento de la tasa de evasión de un 18% a un 23% para el año 2009.

Por otra parte, la detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera un importante costo administrativo para las Unidades Operativas (áreas de Fiscalización y Jurídica). Esto se debe a la dedicación de fiscalizadores durante lapsos prolongados, para reunir las pruebas que permitan evaluar debidamente la participación del contribuyente, y determinar si procede la aplicación de la sanción civil en el rechazo del crédito fiscal o el inicio del procedimiento para la aplicación de sanciones por delito tributario. Actualmente se estima que el costo de recaudación de $100 es $0,91, es decir la recaudación cuesta cerca del 1% del valor recolectado.

El SII cuenta con 4.169 funcionarios, de los cuales el 31% corresponde a fiscalizadores que deben atender a más de 4 millones de contribuyentes, lo que equivale a aproximadamente 1 fiscalizador cada 3.000 contribuyentes. El 67% de la dotación se encuentra en las Direcciones Regionales, quienes tienen interacción directa con el contribuyente y realizan gran parte de las labores de fiscalización y auditoría del Servicio.

Debido a que no es posible fiscalizar a todos los contribuyentes que utilizan facturas falsas, en la práctica sólo es posible detectar un porcentaje de estos casos a través de las labores de fiscalización que el SII efectúa. Es así como durante el año 2011 se efectuaron 365.820 auditorías masivas de IVA, realizándose además 8.917 casos de fiscalización selectiva – emergente y 692.000 controles presenciales, para lo cual se destinaron más de 436.000 horas de trabajo, considerando horas normales, extras y nocturnas. La probabilidad de éxito de estas acciones durante ese mismo año fue de un 55% en las auditorías selectivas y emergentes, las que se basan en una hipótesis de evasión y requieren de mayor tiempo de investigación, mientras que las auditorías masivas de IVA tienen una probabilidad de éxito cercana al 22%, las cuales se realizan con la finalidad de generar una sensación de control y evitar ciertos tipos de comportamientos.

Respecto de las facturas falsas, el SII ha desarrollado planes de fiscalización selectivos para su detección, orientados a la verificación de facturas en terreno, y elaborado documentos de apoyo[15] que definen criterios de fiscalización específicos. Estos instrumentos han permitido mejorar la detección y fiscalización de facturas falsas y aumentar la recaudación por este concepto[16], generándose más de 2.500 querellas y 4.000 querellados junto a otros delitos de defensa judicial y comercio clandestino, por un monto de perjuicio fiscal de aproximadamente $280 mil millones de pesos, cifra bastante menor a la estimación de evasión por este tipo de fraude en cada año.

Lo anterior da cuenta de la relevancia que tiene focalizar los esfuerzos en la detección de casos de evasión y fraude fiscal. Para ello, el SII ha implementando recientemente un proyecto de segmentación de contribuyentes que permitirá focalizar los esfuerzos de fiscalización de los evasores, diferenciándolos según tamaño y tipo de contribuyentes, de acuerdo con sus necesidades y características específicas. No obstante, los procedimientos de selección actuales consideran sólo aspectos puntuales, como la relación débito/crédito o la utilización de ratios

[15] Circulares N° 60 y 93 del año 2001.

[16] A partir de diciembre de 2007, el Departamento de Diseño de Procedimientos de Fiscalización e Internacional de la Subdirección de Fiscalización, elabora una guía denominada “Procedimiento para la Fiscalización de las Facturas Falsas”, la cual considera la normativa vigente sobre la materia, la experiencia recogida en la ejecución de los programas de fiscalización, y los aportes de los funcionarios fiscalizadores.

tributarios para la selección de casos. Consecuentemente, no se han desarrollado herramientas de apoyo que contemplen un análisis global del contribuyente considerando toda la información disponible, contexto en el cual la Inteligencia Fiscal tiene mucho que aportar.

Considerando las potencialidades que ofrecen las técnicas de Data Mining para extraer y generar conocimiento de grandes volúmenes de datos, este estudio busca caracterizar a los contribuyentes que utilizan o venden facturas falsas en un año determinado en base a información histórica de su pago de impuesto y de comportamiento tributario, así como de sus relacionados, para detectar patrones de conductas asociadas con este tipo de fraude fiscal. De esta manera se espera identificar patrones de comportamiento que permitan detectar a aquellos contribuyentes que tienen mayor probabilidad de utilizar facturas falsas para evadir impuestos.

En una primera instancia, se identifican patrones de conducta del total de contribuyentes declarantes de IVA, considerando aquellas variables más relacionadas con el uso de facturas falsas, según el comportamiento de casos conocidos de fraude y no fraude. Posteriormente, se identifican patrones específicos de comportamiento de aquellos contribuyentes en los que el resultado de fraude y no fraude es conocido, obteniendo una probabilidad asociada al uso de facturas falsas.

De esta forma, se espera apoyar la acción fiscalizadora del SII, identificando aquellas características que definen un buen y mal comportamiento tributario, permitiendo así priorizar la selección de contribuyentes para la realización de auditorías futuras.

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Caracterizar a los contribuyentes declarantes de IVA que incurren en el uso y venta de facturas falsas para evadir impuestos y detectar patrones de comportamiento de fraude en base a su información tributaria, para contribuir al desarrollo de una metodología que permita apoyar la fiscalización de este tipo de casos en el Servicio de Impuestos Internos.

1.3.2. OBJETIVOS ESPECÍFICOS

Los objetivos específicos de esta tesis consideran:

- Determinar las características tributarias de los contribuyentes que incurren en el uso y comercialización de facturas falsas.
- Identificar patrones de fraudes por la exploración exhaustiva del conjunto de datos de manera de generar nuevos conocimientos.
- Seleccionar los mejores modelos de Data Mining para detectar los patrones de comportamiento asociados a este tipo de casos.
- Establecer una metodología que le permita al SII focalizar los esfuerzos y los recursos en la determinación de los contribuyentes que cometen este tipo de delito tributario.

1.4. ALCANCES

Para efectos de la caracterización se escoge el año 2006 como año de estudio, considerando a todos aquellos contribuyentes que hayan presentado al menos una declaración de IVA entre el año 2005 y 2007, correspondiente a 582.161 empresas.

En la caracterización de casos de fraude/no fraude se utiliza información de aquellas auditorías en las que existe certeza que se le revisaron sus facturas del año 2006, independiente de la oportunidad de esta revisión, generando un total de 1.692 empresas. Debido a que las auditorías se realizan hasta un período de 3 años atrás, se dificulta utilizar datos más recientes, pues durante 2010 aún se estaban generando casos que podrían haber utilizado facturas falsas desde el año 2007 hacia adelante.

Para explicar el comportamiento de estos contribuyentes se utiliza información que dé cuenta del pago de impuestos en esos períodos, considerando sus declaraciones mensuales de IVA y anual de Renta, indicadores relacionados con giros, denuncias y clausuras, así como otras características particulares como la antigüedad de la empresa, el nivel de actividad y la cantidad de asociados a la empresa e información de sus relacionados, entre otros.

Adicionalmente, se considera información histórica relacionada con diferentes etapas del ciclo de vida del contribuyente, como su inicio de actividades, verificación de actividades, modificaciones de información efectuadas, timbraje de documentos, fiscalizaciones realizadas y anotaciones formuladas durante su paso por el Servicio.

Debido a que existen diferencias entre el comportamiento de las empresas de menor tamaño respecto de las de mayor tamaño, ya sea en su pago de impuesto como en sus características particulares, se utilizan dos grupos para realizar el análisis, considerando por una parte a las micro y pequeñas empresas, y por otro, a las medianas y grandes empresas. Esto permitirá también focalizar las futuras acciones de detección de fraude de acuerdo con el segmento al que pertenecen los contribuyentes, conservando así la diferenciación utilizada actualmente en el SII.

1.5. RESULTADOS

Mediante el presente trabajo se obtiene:

- Una selección de las variables relevantes que permitan caracterizar y detectar a los contribuyentes que utilizan y/o comercializan facturas falsas.
- La selección del vector de características que mejor discrimine entre contribuyentes.
- La aplicación y posterior selección de diversas técnicas de Data Mining para la caracterización de contribuyentes que utilizan y/o comercializan facturas falsas.
- Un agrupamiento de contribuyentes de acuerdo a características similares.
- La aplicación y posterior selección de diversas técnicas de Data Mining de aprendizaje supervisado para la detección de contribuyentes que utilizan y/o comercializan facturas falsas.

- Una propuesta de metodología para seleccionar a contribuyentes a fiscalizar, considerando su probabilidad de evasión.

1.6. CONTRIBUCIONES Y PUBLICACIONES

Las principales contribuciones de esta tesis incluyen:

- Recopilación de antecedentes del fraude tributario por facturas falsas en Chile.
- Revisión del estado del arte en cuanto a las técnicas utilizadas en la detección del fraude tributario y otros tipos de fraude.
- Aplicación de distintas técnicas y variables para la caracterización y detección de fraude por facturas falsas.
- Identificación de patrones de fraude y no fraude por facturas falsas diferenciadas por segmento.
- Generación de conocimiento respecto de factores que inciden en este tipo de delito tributario.
- Aplicación de distintas técnicas y variables para la detección de fraude por facturas falsas.
- Generación de reglas que permitan seleccionar casos para fiscalización.

Dentro de las publicaciones se tiene:

- Publicación paper versión español en Revista Ingeniería de Sistemas de la Universidad de Chile, edición marzo 2012.
- Postulación paper versión inglés a una revista de prestigio internacional como el Expert Systems with Applications.

Ambas publicaciones se encuentran adjuntas en los anexos de este documento.

1.7. ESTRUCTURA DE LA TESIS

En el capítulo 1 se exponen antecedentes generales del sistema tributario chileno, particularmente aquellos relacionados con el impuesto a las ventas y servicios y las facturas, así como la problemática actual que enfrenta la Administración Tributaria Chilena respecto de la utilización y comercialización de facturas falsas como mecanismo para evadir impuestos, describiendo los objetivos y resultados que se esperan alcanzar en este trabajo.

En el capítulo 2 se presentan los aspectos teóricos necesarios para abordar el problema de detección de fraude mediante técnicas de inteligencia artificial. Para ello se señalan las características del fraude y la manera en que se ha resuelto el problema de su detección en distintos ámbitos, haciendo hincapié en el fraude tributario.

Posteriormente, en el capítulo 3, se describe el funcionamiento y las características de las técnicas de Data Mining a utilizar en esta tesis, así como también el procesamiento de datos requerido para realizar el análisis y aplicar las distintas técnicas.

En el capítulo 4 se describe la manera en la que se procesa la información, detallando los tipos de datos considerados y los resultados de su procesamiento y transformación. A continuación, se presentan los resultados de los experimentos realizados con cada técnica de Data Mining, y se propone una metodología de selección de casos para fiscalización, de acuerdo con los modelos generados, que permita detectar contribuyentes que son usuarios potenciales de facturas falsas según el segmento al que pertenecen.

En el capítulo 5 se presentan las conclusiones del trabajo realizado y las recomendaciones para la realización de futuros trabajos y análisis.

Finalmente, en los Anexos de este documento se presentan la estructura y los procedimientos efectuados por el SII para la detección de contribuyentes con facturas falsas, el detalle de la información utilizada para la construcción del vector de características, el resultado del análisis de componentes principales, los árboles de decisión generados junto con las reglas que definen si un caso es fraude o no fraude, y los papers para publicación a nivel nacional e internacional.

2. MARCO CONCEPTUAL

El fraude, en sus diversas manifestaciones, es un fenómeno del que no está libre ninguna sociedad moderna. Todas las Administraciones, con mayores o menores medios y resultados, se esfuerzan por convencer a la sociedad del valor que el cumplimiento de las leyes tiene para el buen funcionamiento de un Estado de derecho, y en combatir un fenómeno que atenta gravemente contra los principios de solidaridad y de igualdad de los ciudadanos ante la ley [BID, 2006].

Muchos problemas de detección de fraude involucran una gran cantidad de información. Procesar estos datos en búsqueda de transacciones fraudulentas, requiere un análisis estadístico y necesita algoritmos rápidos y eficientes, entre los cuales el Data Mining aporta técnicas relevantes que facilitan la interpretación de datos y mejoran la comprensión de los procesos. Las herramientas utilizadas son muy variadas, considerando que existen diversos tamaños y tipos de problemas.

En este capítulo, se presenta la teoría necesaria para poder abordar el problema de la detección de fraude utilizando una solución de inteligencia artificial. En las secciones 2.1 a 2.4 se entregan definiciones de fraude existentes en la literatura y se presentan las características del fraude tributario, señalando mecanismos mediante los cuales las Administraciones Tributarias han determinan los contribuyentes que tienen mayor riesgo fiscal. De manera complementaria, en la sección 2.5 se presenta el estado del arte de la detección de fraude en otros ámbitos de acción, como el sector bancario, el sector de telecomunicaciones, el ámbito de los seguros, del lavado de dinero, entre otros, en los cuales las técnicas de Data Mining han sido aplicadas exitosamente.

2.1. DEFINICIÓN DE FRAUDE

Existen muchas definiciones de fraude, según sea el punto de vista de consideración. La Real Academia Española[17] lo define como un "acto tendiente a eludir una disposición legal en perjuicio del Estado o de terceros", mientras que el Diccionario Inglés de Oxford[18] lo define como un "engaño injusto o criminal, con la finalidad de obtener un beneficio o ganancia personal".

Para articular un caso de fraude, [Davia et al, 2000] señala una serie de ítems que deben ser identificados: una víctima, los detalles del acto engañoso o fraudulento, el perjuicio a la víctima, el autor (es decir, un sospechoso), las pruebas de que el autor haya actuado con la intención de cometer fraude y que haya sido beneficiado por el acto cometido.

A modo general, es posible distinguir tres perfiles de individuos que cometen fraude, los cuales tienen una constante evolución en el tiempo en su modo de operar:

- El *"defraudador promedio"* (average offender) que comete fraude cuando existe la oportunidad, ya sea por tentación repentina o por sufrir de dificultades financieras en ese momento,
- El "*defraudador criminal*" (criminal offender), que comete fraude de manera continua y normalmente tiene registros de antecedentes penales por otros delitos cometidos en el pasado, y
- Las *"organizaciones criminales"* que se constituyen con el único fin de cometer fraude, destinando una mayor cantidad de tiempo, esfuerzo y recursos para perpetrarlo [Baldock, 1997].

El fraude cometido por el primer grupo se conoce como "soft fraud", siendo el más difícil de mitigar, debido a que el costo de investigación suele ser mayor que el costo del fraude mismo. Por otro lado, el fraude cometido por los defraudadores y organizaciones criminales se denomina "hard fraud", ya que involucra mayor cantidad de dinero y es más complejo de detectar.

Sea cual fuere la acepción que de él se tome, el fraude tiene como elemento definitorio su adaptación a la realidad económica y social. Si bien sus características son muy diversas, se pueden destacar como las más relevantes la generalidad, la mutabilidad, la internacionalización, la especialización y la adaptación a las nuevas tecnologías, sin desconocer la incidencia que la tipología de propios defraudadores tiene en su configuración [BID, 2006].

2.2. EL FRAUDE TRIBUTARIO

Conceptualizar el "fraude fiscal" no es tarea fácil, al tratarse de una calificación genérica carente de regulación legal, que puede ser analizada desde diversos ámbitos y afectada por nociones o

[17] http://www.rae.es/

[18] http://oxforddictionaries.com/

figuras afines que a veces lo delimitan y otras se solapan con él. En los diferentes documentos de trabajo de Organizaciones Internacionales sobre esta materia, se opta generalmente por abordar el fraude fiscal desde una visión práctica, con perfiles no siempre definidos, aludiendo por un lado a los inconvenientes del fraude, y por otro, a las medidas encaminadas a su erradicación. Así, se utiliza el término "fraude fiscal" en un sentido amplio, haciendo alusión tanto al conjunto de conductas que en muchos países se denomina "evasión fiscal", como a aquellas que se catalogan como "fraude fiscal" en los países que hacen esta diferenciación. Lo anterior resulta en que estas dos denominaciones genéricas sean las más utilizadas para referirse a las actividades fraudulentas o de incumplimiento fiscal[19].

Independiente que las denominaciones formales o técnicas utilizadas para delimitar y referirse a las conductas que suponen un incumplimiento tributario pueden ser variadas, todas tienen un rasgo en común: la materia u objeto al que se refieren, es decir, la conducta que se pretende perseguir y, en su caso, sancionar. Normalmente se habla de **"elusión fiscal"** cuando se hace referencia a conductas que, dentro de la ley, evitan o reducen el pago de impuestos, mientras que la **"evasión o fraude fiscal"** supone un quebrantamiento de la legalidad para obtener esos mismos resultados.

Aunque el fraude tributario requiere respuestas adecuadas, tanto preventivas como de control, no cabe desconocer que la gran diversidad de fraudes y la distinta valoración de los mismos, hacen necesario focalizarse por determinadas actuaciones en detrimento de otras. En el Cuadro Nº 2 se señalan los tipos de incumplimiento más característicos relacionados con el IVA.

Cuadro Nº 2: Tipos de Incumplimiento relacionados con el IVA

EVASIÓN POR SUBESTIMACIÓN DE IMPUESTOS RETENIDOS EN LAS VENTAS	
Fraude carrusel o missing trader fraud.	Una empresa es creada con la finalidad de capturar IVA en las ventas y luego desaparece sin remitir el IVA al gobierno.
Empresas fallidas	Una empresa fracasa o va a la quiebra antes de remitir al gobierno el IVA recaudado.
Transacciones menores	Una empresa (o negocio), puede cargar un precio más bajo, libre de IVA a las transacciones en efectivo, o puede reportar menores ventas en efectivo y retener el IVA captado.
Fraude de importación	Una empresa o individuo importa artículos para consumo personal y subestima su valor para efectos de IVA.
EVASIÓN POR SOBREDIMENSIONAMIENTO DE IMPUESTOS PAGADOS EN LAS COMPRAS	
Reembolsos fraudulentos	Un negocio o un defraudador informa retornos falsos, solicitando la devolución del IVA por parte del gobierno.
Tergiversación de compras	Una empresa solicita falsamente créditos fiscales de entrada, trastocando (o tergiversando) gastos personales de consumo como gastos comerciales.
Facturas falsas o adulteradas	Una empresa crea o altera facturas para aumentar la cantidad de crédito que puede solicitar por impuestos de entrada.
Fraude de exportación	Una empresa utiliza facturas falsas de exportación por bienes que no han sido exportados para solicitar crédito fiscal (de entrada)

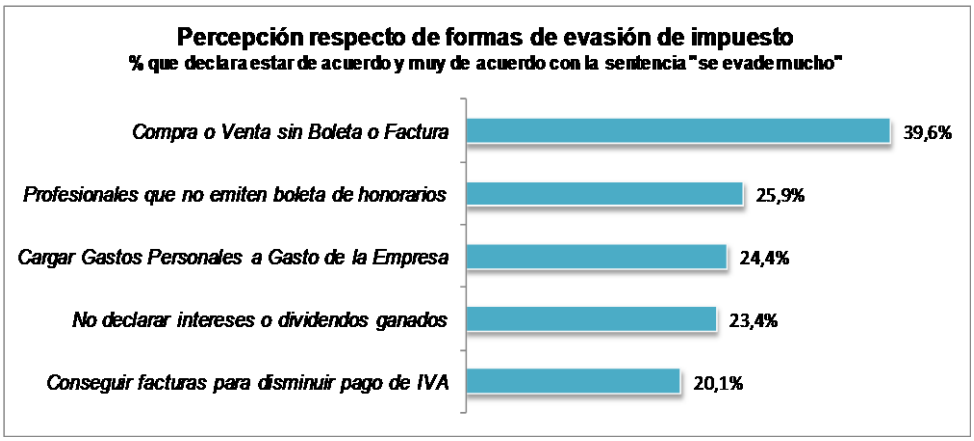
Fuente: GAO Analysis

[19] Si bien cada país utiliza una terminología diferente, lo más habitual es el uso de ambas denominaciones, incluso como terminología aceptada en las Organizaciones Internacionales. La OCDE, recoge ambas acepciones en multitud de documentos, así por ejemplo en "Fraude Fiscale. En savoir plus" (www.oecd.org.). Como ejemplos de países que utilizan estos términos con diferentes significados se puede señalar Uruguay (evasión y fraude), Venezuela (evasión y elusión), Argentina (evasión, fraude, contrabando), Perú (evasión, contrabando), México (fraude, evasión y contrabando), entre otros.

En particular, el fraude carrusel es de interés de la Unión Europea, donde las fronteras comerciales entre los países miembros ya no existen, y las empresas fraudulentas se aprovechan de la oportunidad de retener la totalidad del importe del IVA percibido sobre las ventas, aprovechando que las empresas no pagan IVA soportado por los bienes y servicios importados.

De acuerdo con una encuesta efectuada en el año 1996, en la cual se consulta por la percepción de incumplimiento asociado a distintas formas de evasión en Chile, se tiene que aproximadamente un 40% de los encuestados declaró estar de acuerdo con la sentencia "se evade mucho" a través de transacciones sin boleta o factura. Un 20%, en tanto, se declara de acuerdo con que una de las formas de evasión utilizada se asocia a conseguir facturas para disminuir el pago de IVA, que correspondería al uso de facturas falsas propiamente tal.

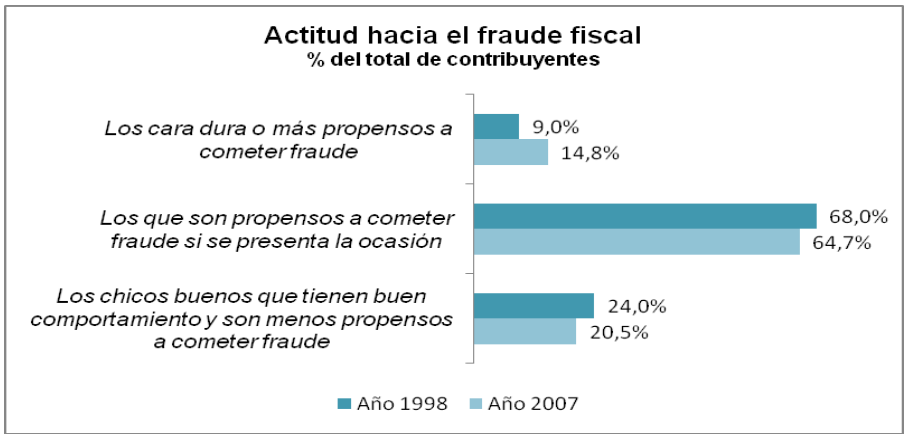
Figura Nº 13: Resultados Estudio "Imagen del SII" respecto de formas de evasión.



Fuente: "Imagen del SII", Informe 1996, Adimark

De igual manera, según estudios de percepción fiscal hacia las obligaciones tributarias realizados en los años 1998 y 2007, se aprecia una mayor tendencia de los contribuyentes a cometer fraude tributario y evadir impuestos en comparación a lo que se observaba una década atrás e implica que actualmente existe una mayor probabilidad de que los contribuyentes cometan este tipo de fraude si se presenta la ocasión.

Figura Nº 14: Resultados Estudio "Percepción Fiscal"



Fuente: Estudio de Percepción Fiscal en Chile, Mori 1998 e IPSOS 2007

2.2.1. FACTORES QUE INFLUYEN EN LA EVASIÓN TRIBUTARIA

Uno de los aspectos más destacados que debe tenerse en cuenta al diseñar un sistema de control tributario orientado a disminuir los índices de evasión y/o elusión tributaria, es conocer el origen de ese comportamiento por parte de los sujetos pasivos de las obligaciones fiscales. En ese sentido, y en particular, en países donde existen altas tasas de evasión motivadas por un comportamiento generalizado, el conocimiento de las causas de este fenómeno permite programar estrategias y desarrollar acciones tendientes a atacar tales enunciados, muchos de los cuales tienen un fuerte ingrediente social, más que estrictamente económico.

La teoría de la evasión tributaria señala una serie de variables que explican este fenómeno, entre los que se encuentran cuatro factores principales, en los que la Administración Tributaria tiene incidencia total o parcial.

Figura N° 15: Factores que influyen en la Evasión Tributaria

Fuente: Basado en Slemrod, Kagan, Tanzy y Shome, Serra y Toro.

La *Eficacia de la Fiscalización* se basa en la probabilidad de detección del incumplimiento tributario, de forma tal que aquellos sometidos a auditorías sean los que tienen una mayor probabilidad de ser evasores [Kagan, 1989]. En ella influye la disponibilidad de información del nivel de cumplimiento de los contribuyentes, las dificultades que enfrentan los fiscalizadores en la detección y prueba de las infracciones, y la optimización del uso de la información disponible para elevar la probabilidad de detección.

La *Simplicidad de la Estructura Tributaria* depende a su vez de la certidumbre con que las obligaciones tributarias están definidas claramente por las leyes, que puede llevar a un error involuntario o una interpretación equivocada o distinta de la ley; la complejidad para cumplir con las obligaciones tributarias, en el sentido de los costos en que debe incurrir para tal fin; y la manipulabilidad o mayores posibilidades de evasión y elusión que se generan a partir de esta complejidad [Slemrod, 1989].

La *Fiscalización* y las *Sanciones*, son dos herramientas principales que la Administración Tributaria posee para reducir la evasión. La combinación de ambas determina los incentivos económicos a evadir que tienen los contribuyentes. Sin embargo, no es claro que el aumento de las sanciones necesariamente conduzca a un mayor cumplimiento, pues un castigo excesivo puede conducir a la inaplicabilidad de las sanciones, sobre todo cuando la legislación tributaria tiene un alto grado de incertidumbre. Por otra parte, las sanciones elevadas pueden incentivar la corrupción, puesto que la negociación entre el contribuyente y el fiscalizador se hace más rentable para ambas partes. Otro factor importante es el tiempo que transcurre entre el delito y la aplicación del castigo [Tanzi y Schome, 1993], pues es frecuente que pasen varios meses e incluso años antes que las infracciones sean castigadas, con lo cual la sanción pierde todo efecto disuasivo.

También es importante la opinión que los contribuyentes tengan del destino de los impuestos, pues si un contribuyente siente que el sistema tributario es injusto, obviamente estará menos dispuesto a cumplir con sus obligaciones tributarias; más aún, desde su perspectiva la evasión podría ser un acto de justicia más que un delito. Por tanto, la *Aceptación del Sistema Tributario* es esencial para elevar el nivel de cumplimiento [Serra y Toro, 1994]. En la medida en que los trámites tributarios sean expeditos y los contribuyentes reciban un trato justo y digno, tendrán también una mejor disposición a pagar sus impuestos.

Figura N° 16: Factores que influyen en la actitud hacia el fraude fiscal

Fuente: Basado en Modelo BISEP, Valerie Braithwaite

De igual manera, el modelo BISEP [Braithwaite, 2009], indica un conjunto de factores (legislación, industria, negocio, económicos, sociológicos y psicológicos) que inciden en las actitudes y comportamiento de los contribuyentes respecto del cumplimiento tributario. Estas actitudes hacia el fraude fiscal son dinámicas, pues un contribuyente puede adoptar alguna de ellas en momentos diferentes y no equivalen a características de una persona o grupo, sino que son reflejo de la interacción entre la persona o grupo y quienes le imponen ciertas exigencias.

2.3. PREVENCIÓN vs DETECCIÓN

Es importante distinguir entre dos conceptos: la prevención y la detección del fraude. La *prevención del fraude* considera todas aquellas medidas utilizadas para detener el fraude antes de que éste ocurra. Por supuesto, ninguno de estos métodos es infalible, y su uso depende en gran medida de los inconvenientes generados para los usuarios y la relación entre el costo y la efectividad de los métodos utilizados. Esto implica contar con un sistema bien diseñado de procesos y procedimientos con el fin de prevenir y detener el fraude. Sin embargo, muchas veces esto no es suficiente.

En contraste, la *detección del fraude* se basa en identificar el fraude tan rápidamente como sea posible, una vez que éste ha sido perpetrado y las medidas de prevención han fallado. Si bien el uso de tecnología para prevenir este tipo de delito es la mejor vía para reducirlos, ésta es una disciplina en continua evolución, debido a que quienes cometen fraude normalmente van modificando y perfeccionando las técnicas utilizadas, encontrando nuevas formas para lograr sus objetivos.

Debido a la imposibilidad de revisar en línea que todas las transacciones del sistema sean correctas, o bien por el alto costo que esto tendría, la detección del fraude por lo general se hace a posteriori (Ej.: después de un mes que el cliente no hizo un determinado cargo en su tarjeta de crédito, después que hizo su declaración de impuestos, etc.). Es claro ver que cualquiera sea la empresa, las transacciones fraudulentas generan un costo importante en la operación del negocio. Por lo tanto, cualquier esfuerzo para detectar el fraude de manera concurrente (es decir, mientras se realiza la transacción), tiene el potencial de generar importantes ahorros a estas instituciones.

No obstante, sin una afinación y exhaustiva verificación, el sistema de detección puede costar más en términos de recursos humanos destinados a investigar las falsas alarmas, que la ganancia misma obtenida con su reducción. Por lo tanto, los datos de los ensayos deben ser adecuados y representativos para la evaluación de los sistemas de detección.

2.4. DETECCIÓN DEL FRAUDE TRIBUTARIO

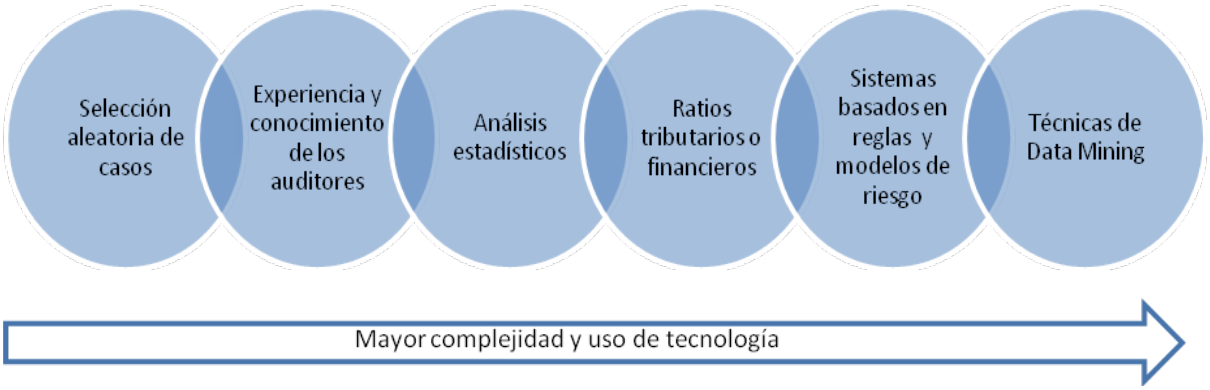
En términos generales, la mayor parte de los países que cuentan con una Administración Tributaria y Aduanera desarrollada, planifican su lucha contra el fraude fiscal. No obstante, existen importantes diferencias en los mecanismos, alcances, enfoque y énfasis puestos en dicha labor.

Uno de los caminos más utilizados para cuantificar el fraude consiste en realizar estudios de corte macroeconómico que vinculan la recaudación percibida con variables económicas como Cuentas Nacionales, Renta Nacional o el Producto Nacional Bruto para medir evasión [Tanzi, 1980]. Sin embargo, este tipo de estimaciones indirectas son limitadas, ya que sólo permiten construir una serie que indica la tendencia de la evasión o fraude en cada año.

De manera complementaria también son numerosos los estudios empíricos y experimentales que tratan de replicar el comportamiento de los individuos y sus correspondientes decisiones ante modificaciones en las variables consideradas clave para el cumplimiento fiscal [Torgler, 2002], [Sánchez y De Juan, 1994], [Alm et al, 1995] y [Rivas, 1997].

Como se aprecia en la Figura Nº 17, para detectar el fraude o evasión fiscal propiamente tal, las Administraciones Tributarias comenzaron aplicando *auditorías de selección aleatoria* para medir la evasión en diferentes grupos de contribuyentes, identificar tendencias de evasión emergente y evaluar los resultados de las políticas implementadas. Igualmente se utilizaba como criterio la selección de casos en los que no se habían realizado auditorías en períodos anteriores o por un cierto período de tiempo.

Figura N° 17: Técnicas utilizadas para seleccionar casos para fiscalización



Fuente: Elaboración propia

De igual forma, la selección de casos a partir de la *experiencia y conocimiento de los auditores*, se remonta a la época en la que había poco o ningún soporte de TI, cuando la disponibilidad de información se encontraba limitada [OCDE, 1999]. Su ventaja es la utilización en pleno del conocimiento adquirido por los fiscalizadores y el personal de primera línea, y su desventaja, el utilizar un conjunto limitado de información, sin el aprovechamiento de otros datos disponibles dentro de la administración. Por otro lado, al basarse en la experiencia del auditor, sólo se seleccionan casos con los que éste se encuentra familiarizado. Esto se ha visto potenciado con la utilización de sistemas computacionales que permiten registrar las auditorías efectuadas y sus resultados, de manera de compartir los conocimientos adquiridos con fiscalizadores localizados en distintos sectores geográficos.

Posteriormente, se utilizan metodologías basadas en *análisis estadísticos*, como las técnicas de a*nálisis de funciones discriminantes*, utilizado por el Internal Revenue Service[20], de forma tal que aquellos sometidos a auditorías sean los que tienen una mayor probabilidad de ser evasores, clasificándolos en categorías de alto o bajo riesgo y utilizando técnicas de matching para poner en relieve las disparidades encontradas en la información de declaraciones de impuestos.

[20] Administración Tributaria de los Estados Unidos.

También es habitual la construcción de *ratios tributarios o financieros* que proporcionan un medio de convertir la información de impuestos, en indicadores que permitan realizar seguimiento a una entidad durante un período de tiempo y compararlo con el comportamiento que presenta la industria o sector económico al que se vincula. Esta población se clasifica y se agrupa en cuartiles. El promedio de la industria debe encontrarse entre el percentil 25% y el percentil 75%. Los ratios pueden indicar que un individuo tiene un comportamiento inusual, pero son incapaces de proporcionar el detalle o las razones por las que esto ocurre para resolver el problema.

Actualmente, gran parte de las Administradoras Tributarias utilizan s*istemas basados en reglas* y *modelos de riesgo*, que transforman la información en indicadores que permitan rankear a los contribuyentes en términos del riesgo de su cumplimiento. Estos sistemas son una herramienta esencial para concentrar los esfuerzos en aquellos que tienen una mayor tendencia a la evasión fiscal. Los desafíos inherentes a este tipo de sistemas se relacionan con el hecho que muchas de las reglas utilizadas dependen de la comparación de ratios, los cuales pueden cambiar en el tiempo. Por otro lado, deben poder responder al conocimiento adquirido por el personal de primera línea que se encuentra trabajando en los casos de estudio. De la misma manera, deben considerar información proveniente de distintas fuentes, como declaraciones de impuestos, información de terceros y de dominio público disponible en la Web, para lo cual se requiere una inversión de recursos en TI.

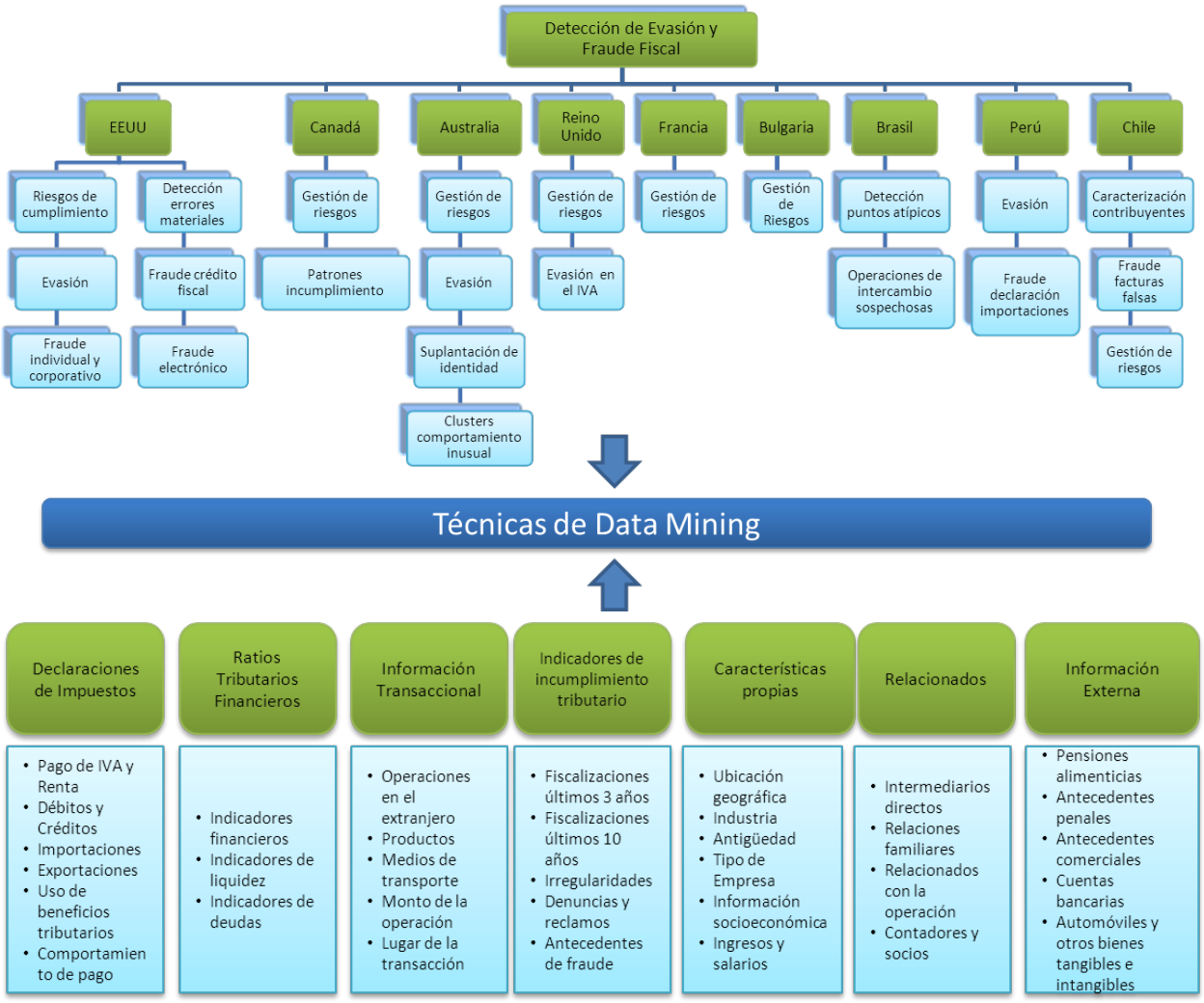
Para identificar los riesgos se requiere una acumulación continua de datos, que progresivamente se va transformando en inteligencia y en conocimiento. Hay dos elementos claves necesarios para el tratamiento de riesgos en la selección de casos para auditoría: precisión y oportunidad de la información utilizada, y conocimientos y técnicas que permitan analizar y priorizar la información[21].

Durante los últimos años, las *técnicas de data mining e inteligencia artificial*, han sido incorporadas en las actividades de planificación de auditorías, principalmente para detectar patrones de fraude o de evasión fiscal, siendo utilizadas por instituciones de administración tributarias con fines específicos [OECD, Compliance Risk Management, 2004]. Entre ellas se encuentran las Administraciones Tributarias de EE.UU., Canadá, Australia, Reino Unido, Bulgaria, Nueva Zelanda y a nivel latinoamericano, Brasil, Chile y Perú.

En el caso de las Administraciones Tributarias éstas han sido empleadas principalmente para la detección de evasión tributaria, construcción de sistemas de gestión de riesgos, detección de fraude tributario, detección de patrones de comportamiento y agrupación en conjuntos de características similares, utilizando para ello información de las declaraciones de impuestos, ratios tributarios y financieros, información de las transacciones efectuadas, indicadores de incumplimiento tributario, características propias de los contribuyentes, comportamiento de sus relacionados e información disponible en otras instituciones, como se muestra en la Figura N° 18.

[21] Estos sistemas son utilizados en las administraciones tributarias de EE.UU., Canadá, Reino Unido, Suiza, Austria, Francia, Nueva Zelanda, Suiza, Irlanda, Tanzania y Brasil.

Figura N° 18: Tipos de fraude detectados mediante técnicas de Data Mining por las Administraciones Tributarias e información utilizada



Fuente: Elaboración propia en base a revisión de la literatura

En términos generales, las Administraciones Tributarias tienen varias similitudes respecto del tipo de información utilizado para la detección de comportamientos de riesgo mediante la aplicación de técnicas de data mining, diferenciándose en función del tipo de comportamiento que se espera detectar, y la posibilidad de contar con información en fuentes externas, que permitan complementar la caracterización e identificación de los contribuyentes que se encuentran en situación de riesgo de cumplimiento.

Particularmente, en Estados Unidos las técnicas de data mining han sido utilizadas en más de 128 agencias y departamentos federales para mejorar el funcionamiento de sus servicios, el análisis científico, el manejo de recursos humanos, la detección de fraude, detección de actividades criminales y actividades terroristas. En el caso de la detección de fraude, estas técnicas han sido utilizadas en 24 agencias federales en diferentes ámbitos como seguros, salud, compras del Estado, educación, seguridad social, agricultura, energía, en el ámbito financiero y policial, por nombrar algunos ejemplos.

Como se muestra en el Cuadro Nº 3, las técnicas de data mining más utilizadas corresponden a técnicas de clasificación y predicción como las redes neuronales, los árboles de decisión, la regresión logística y el Support Vector Machines, y en menor medida, técnicas de clusterización como SOM y K-means, aplicando en todos los casos distintas combinaciones de ellas.

Cuadro N°3: Técnicas de data mining utilizadas por las adm. tributarias para detección del fraude fiscal

Técnica Aplicada	EEUU	Canadá	Australia	Reino Unido	Bulgaria	Brasil	Perú	Chile
Redes Neuronales	✓	✓		✓	✓		✓	✓
Árboles de Decisión	✓	✓	✓				✓	✓
Regresión Logística	✓		✓	✓	✓			
SOM			✓					✓
K-means			✓					✓
Support Vector Machines	✓		✓					
Técnicas de Visualización	✓					✓		
Redes Bayesianas			✓					
K-Nearest Neighbour			✓					
Reglas de Asociación							✓	
Reglas Difusas							✓	
Cadenas de Markov						✓		
Series de Tiempo		✓						
Regresión				✓				
Simulaciones	✓							

Fuente: Elaboración propia en base a revisión de la literatura

El Internal Revenue Service, institución a cargo de administrar los impuestos en Estados Unidos, posee un modelo de cumplimiento del riesgo (RBC)[22], que se compone de varios módulos: Filing Analysis Module, que analiza los rendimientos de los pagos y declaraciones de impuestos, asignando puntajes al incumplimiento de reglas; Post Filing Analysis Module, que clasifica los casos en categorías de alto y bajo riesgo dependiendo de las reglas violadas; Workload Analysis Module, que equilibra la clasificación de riesgo con el personal disponible para asignar trabajos de auditoría, considerando la experiencia del auditor, las tasas de respuesta y los tiempos requeridos, entre otros factores, y el Post Treatment Analysis Module, que reúne información de los resultados de los pagos involucrados, para la planificación y seguimiento del caso.

En particular, el sistema *Reveal* fue desarrollado para detectar patrones de actividades criminales y actividades terroristas. El sistema provee la capacidad de consultar datos de múltiples fuentes para identificar relaciones, los que incluyen información del Bank Secrecy Act, información de impuestos y de actividades contra el terrorismo. Adicionalmente, permite crear un perfil de las acciones y personas asociadas al objeto investigado, ya sea un individuo o una corporación, entregando información de las transacciones financieras realizadas entre ellos a través de una representación gráfica, y permite la búsqueda por nombres, números de seguro social, y otros tipos de datos para ayudar a reducir su investigación. Posteriormente pueden utilizar el componente de visualización, el cual muestra las relaciones entre los datos consultados, y facilita el descubrimiento de las relaciones entre las entidades, patrones y tendencias.

[22] Risk Based Compliance Methodologies

Por su parte, la Administración Tributaria de Canadá, utiliza un sistema de gestión de riesgos, combinado con un análisis estadístico a nivel macro, para establecer tendencias de cumplimiento utilizando series de tiempo económicas. Esto permite comparar las ventas de un sector económico con la recaudación de impuesto asociada, visualizar cambios en los patrones de cumplimiento, y marcar áreas potenciales de no cumplimiento para priorizar los análisis posteriores. Para ello posee un herramienta de software denominada Compliance Measurement Profiling and Assessment System (COMPASS), que permite analizar riesgos por industria, ubicación geográfica, entre otros, a partir de la definición de riesgos individuales [OECD, Compliance Risk Management, 2004]. Adicionalmente utiliza redes neuronales y árboles de decisión para distinguir las características de los contribuyentes que evaden o cometen fraude, en base a los resultados de auditorías pasadas, y detectar los patrones de incumplimiento o evasión.

El “Compliance Program” es uno de los cinco subplanes que integran el “Tax Office Plan" desarrollado por Australia, el cual es uno de los mejores ejemplos de cómo articular y razonar los planes de lucha contra el fraude. Se trata de procedimientos de carácter específico, dirigidos a los diversos asuntos implicados en el cumplimiento voluntario, de prevención y corrección del fraude. Estos planes basan las actividades de control en un modelo de riesgos, que se relaciona con la naturaleza de los contribuyentes. Más recientemente, la oficina de impuestos ha desarrollado proyectos basados en estadística y Data Mining, con el objetivo de realizar comparaciones, encontrar asociaciones y patrones, utilizando modelos de regresión logística, árboles de decisión y Support Vector Machines (SVM).

Un caso de interés ha sido el enfoque utilizado por Denny, Williams y Christen, de descubrimiento de pequeños clusters o subpoblaciones inusuales, denominadas “Hot Spots”, utilizando técnicas como el Self-Organizing Map (SOM) para explorar sus características y algoritmos de agrupación como K-means, utilizando representaciones visuales que son fáciles de entender para usuarios no técnicos. Por otra parte, al utilizar este enfoque, los usuarios no tienen que determinar el número de grupos ni las definiciones de agrupaciones de distancia de antemano. Los usuarios pueden seleccionar las regiones para ver los detalles, y seleccionar regiones o grupos sobre la base de sus necesidades.

De igual modo, las Administraciones Tributarias del Reino Unido, Austria y Nueva Zelanda, poseen un sistema de análisis de riesgo para determinar evasión en el IVA [OECD, Compliance Risk Management, 2004]. En el caso de Austria se identifican 35 factores de riesgo relacionados con la facturación, el comportamiento de pagos irregulares, formularios de pago, umbrales, riesgos producidos por cambios en el sistema, situaciones históricas, liquidez e información de reclamos, utilizando un sistema de puntaje acumulado en el tiempo. El puntaje asignado depende de la antigüedad del contribuyente, razón por la cual se realizan auditorías especiales para las empresas nuevas con menos de 2 años de antigüedad.

El Reino Unido utiliza una herramienta de modelamiento de riesgo multivariable que se vale de un análisis de regresión para encontrar las variables de riesgo más relevantes, denominado Central Risk Analysis. En el caso de Nueva Zelanda, el modelo que asocia atención del control y grado de cumplimiento coincide con el utilizado por la Australian Taxation Office. El Plan incluye un análisis del entorno económico, internacional, poblacional, de diversidad étnica, de estructura familiar, entre otros.

En Francia, el análisis de riesgos se lleva a cabo a través de dos principales herramientas de software, OASIS y SYNFONE. El primero es un programa de software que permite realizar estudios y comparaciones de declaraciones de impuestos de IVA, impuesto de sociedades o actividades por cuenta propia. Permite el cálculo de un conjunto de ratios entre los diferentes ítems de los balances y estados financieros. También es posible hacer comparaciones y relaciones entre las anteriores declaraciones de impuestos del mismo contribuyente, por un período de tres o cuatro años, y entre los contribuyentes con la misma actividad y el mismo nivel de volumen de negocios, a nivel local, regional o nacional. El segundo, es un software dedicado principalmente a la evaluación del riesgo. Utiliza los recursos de las principales bases de datos de impuestos, bases europeas (intra transacciones) e información de las actividades de auditoría del contribuyente durante los últimos diez años para dar una clasificación de riesgos para cada individuo.

A nivel latinoamericano, Perú fue uno de los primeros en aplicar estas técnicas para detectar evasión tributaria, incorporando al sistema de selección en la Aduana Marítima del Callao, una herramienta de inteligencia artificial basada en redes neuronales [SUNAT, 2006]. Esta herramienta utiliza una historia de hechos y las relaciona para identificar la correspondencia entre un acto determinado -como por ejemplo, el fraude- con los valores de variables asociados a él. Una característica importante es su capacidad de identificar patrones y eventos no típicos, que la hacen una herramienta poderosa para aumentar significativamente el nivel de incidencia de las Declaraciones Únicas de Aduanas (DUA) que son seleccionadas para el reconocimiento físico y detectar el comportamiento ilícito. Posteriormente, este modelo fue mejorado a través de la aplicación de reglas difusas y de asociación para el pre-procesamiento de las variables y árboles de clasificación y regresión (CART) para seleccionar las variables más relevantes.

Por su parte, Brasil desarrolló el proyecto HARPIA (Análise de Risco Aduaneiro e Inteligência Artificial Aplicada) de manera conjunta entre la Receita Federal do Brasil y las universidades de ese país [Digiampetri, 2008]. Este proyecto consistía en desarrollar un sistema de detección de puntos atípicos (CARANCHO y ANACOM) que ayude a los fiscalizadores a identificar operaciones sospechosas y un sistema de información de exportación de productos (PFEIS) para ayudar a los importadores en el registro y clasificación de sus productos, utilizando atributos para evitar posibles duplicidades. CARANCHO se basa en la visualización gráfica de información de importaciones y exportaciones históricas y PFEIS se basa en cadenas de Markov, para calcular la probabilidad de que una palabra cadena, sea válida en un determinado dominio.

En el caso de Chile, la primera experiencia fue desarrollada en el año 2007, utilizando SOM y K-means para segmentar contribuyentes de IVA de acuerdo con sus declaraciones de F29 y características particulares [Luckeheide, 2007]. Posteriormente, siguiendo la tendencia internacional, en el año 2009 se construyen modelos de riesgos en distintas etapas del ciclo de vida del contribuyente, en los que se aplican técnicas de redes neuronales, árboles de decisión y regresión logística. Adicionalmente se desarrolla la primera experiencia para detectar potenciales usuarios de facturas falsas a través de redes neuronales artificiales y árboles de decisión, utilizando principalmente información de su declaración de IVA y Renta en micro y pequeñas empresas y algunos indicadores relacionados al incumplimiento tributario [Bravo, 2009].

2.5. DETECCIÓN DE OTROS TIPOS DE FRAUDE

La problemática de la detección de fraude abarca diversos ámbitos de acción, afectando empresas del sector bancario, telecomunicaciones, seguros, informática, salud y educación, entre otros. De acuerdo con un estudio realizado por Ernst&Young en el año 2006 en el cual se encuestó a 150 empresas chilenas, medianas y grandes, un 41% de ellas declaró haber sido víctima de algún tipo de fraude en los dos últimos años. Esto plantea grandes desafíos, considerando que el fraude normalmente es mayor que lo declarado, debido a que se resiente la imagen de la compañía y en muchos casos, hay empresas que no están en conocimiento de que han sido víctimas de fraude.

A continuación se describen las características del fraude en diversos sectores en los que la aplicación de herramientas de Data Mining para detectar transacciones fraudulentas ha sido exitosa.

2.5.1. FRAUDE EN TARJETAS DE CRÉDITO

Una de las aplicaciones más interesantes para detectar fraude se produce en las tarjetas de crédito a raíz de la cantidad de operaciones que se realizan por día y los montos involucrados. En este tipo de industria interesa tanto prevenir el fraude como detectarlo lo antes posible, de lo contrario la confianza de los consumidores se pierde y se reducen los ingresos, lo que se suma a las pérdidas directas ocasionadas por las acciones de fraude.

Los tipos de fraude más comunes son el robo simple, la falsificación de tarjetas, la clonación, la transferencia de fondos por robo de claves, el phishing o envío de email falso para capturar los datos de la tarjeta, el robo de bases de datos y el hackeo del sitio web. Normalmente, el estafador utiliza una tarjeta física para perpetrar el fraude, aunque la posesión física no es esencial. En el caso de una tarjeta robada, el defraudador gasta tanto como le es posible en el menor lapso posible, antes de que se detecte el robo de la tarjeta y ésta sea bloqueada. La falsificación de tarjetas y la clonación requieren obtener los detalles de la tarjeta sin el conocimiento del dueño, lo cual se hace de varias formas, incluyendo el skimming donde los empleados copian ilegalmente la tira magnética de la tarjeta deslizándola en un lector manual, ingresando los detalles de la tarjeta en un teléfono móvil mientras se espera en una cola, o haciéndose pasar por un empleado de la compañía y tomando los datos a través del teléfono.

Las transacciones realizadas por estafadores haciendo compras sin tarjeta y con tarjeta falsificada, pueden ser detectadas a través de métodos que buscan cambios en los patrones de las transacciones, así como a través de la revisión de patrones específicos que son conocidos en estas estafas. Las bases de datos de tarjetas de crédito contienen información de cada transacción, la cual incluye variables como el código del comercio, el número de la cuenta, el tipo de tarjeta de crédito, el tipo de compra, el nombre del cliente, el monto y la fecha de la transacción. Algunos de los datos son numéricos, como el monto de la transacción, y otros son nominales, como el código del comercio, que puede tener cientos de categorías o valores simbólicos. Los tipos de datos mezclados han conducido a que las aplicaciones utilicen una amplia variedad de herramientas estadísticas, aprendizaje de máquinas y técnicas de Data Mining.

Para detectar si una cuenta puede estar comprometida en un fraude, el análisis se puede basar en modelos de patrones previos de uso del cliente, patrones de uso estándar esperados o en patrones particulares que son conocidos por estar normalmente asociados con fraude. Un ejemplo simple es aquel que señala que el gasto acumulado en el tiempo es lineal. Saltos imprevistos en estas curvas o cambios repentinos de pendiente pueden ameritar una investigación. Adicionalmente el uso de una tarjeta para realizar un tipo de compra inusual puede gatillar una alarma de generación de un patrón no esperado. Finalmente se encuentran los patrones de transacciones conocidos, como las compras repentinas de muchos bienes que permiten una fácil venta en el mercado negro, como dispositivos electrónicos pequeños o joyas, así como el uso inmediato de una nueva tarjeta en ubicaciones geográficas muy diferentes.

En la práctica se ha visto que la herramienta utilizada para la detección de este tipo de fraude depende del contexto y del caso particular. Debido a que los métodos de detección de anomalías dependen del contexto, mucha de la literatura publicada se concentra en métodos de clasificación supervisada. En particular las alertas basadas en reglas y redes neuronales han atraído interés y son predominantes [Gosh y Reilly, 1994], [Alexkerov et all, 1997], [Dorronsoro, Ginel, Sánchez y Cruz, 1997], y [Brause et al, 1999]. Entre ellos se encuentra el software HNC desarrollado por el Falcon Fraud Manager, que descansa principalmente en tecnología de redes neuronales para detectar fraudes en tarjetas de crédito. También se ha aplicado un procedimiento denominado Análisis Discriminante No Lineal (NLDA), que construye clasificadores que combinan redes neuronales con técnicas de análisis discriminante [Dorronsoro y Santa Cruz, 1997].

También son utilizados distintos métodos de clasificación. [Grosh, 1998] recomienda una clasificación donde hay un atributo de fraude y un enfoque de agrupación. [Chan, 1999] indica que la partición de un gran conjunto de datos en subconjuntos más pequeños para generar clasificadores con diferentes algoritmos mejora significativamente el ahorro de costos y [Stolfo et al, 1997] entrega un sistema meta-clasificador basado en la idea de detectar fraude localmente y ofrecer servicios de detección de intrusiones en un único sistema de información empresarial, fusionando los resultados para conducir a una herramienta global más precisa.

El modelo de fraude de crédito desarrollado por [Brause et al, 1999] utiliza reglas de asociación para minar datos simbólicos y una función de base radial en redes neuronales para minar a lo largo de los datos, comparando cada transacción de fraude con el resto para encontrar pares de similares características. Cada par se fusiona en una regla generalizada en la que se sustituye un atributo no repetido por un símbolo específico generando una nueva regla, formando un árbol y reglas por nivel. El uso de redes neuronales supervisadas, para comprobar los resultados de las reglas de asociación, aumenta la exactitud predictiva.

[Bolton y Hand, 2001] utilizan métodos de aprendizaje no supervisado desarrollando el Peer Group Analysis, herramienta que utiliza modelos locales de comportamiento del gasto en el tiempo para detectar cambios en las cuentas y un análisis de punto de quiebre para detectar cuándo los objetos comienzan a comportarse de manera distinta. Simultáneamente, se emplean pruebas estadísticas para evaluar si las recientes operaciones siguen un patrón diferente al comportamiento anterior. [Aggarwal y Yu, 2001], en tanto, utilizan algoritmos genéticos para detectar puntos atípicos (outlier detection) en conjuntos de datos de grandes dimensiones.

[Maes et al, 2002] utilizan dos técnicas de aprendizaje para el razonamiento bajo incertidumbre: redes neuronales artificiales (ANN) y redes bayesianas (BBN), introduciendo una medida de rendimiento independiente del problema de aprendizaje. Los resultados muestran que los BBN fueron más precisos y más rápidos en el entrenamiento, pero son más lentos cuando se aplican a nuevos casos, mientras que la detección del fraude es más rápido al utilizar ANN.

Finalmente, [Fan, 2004] propone un algoritmo sobre la base de árboles de decisión para combinar la utilización de datos antiguos y nuevos en problemas de detección de fraude, revisando si los datos tienen consistencia y eligiendo la mejor hipótesis en diferentes situaciones, determinando así si los nuevos datos son suficientes para obtener la predicción.

2.5.2. FRAUDE EN TELECOMUNICACIONES

En el sector de telecomunicaciones la diversidad de tipos de fraude es aún mayor, considerando el desarrollo tecnológico y la expansión de los teléfonos móviles en los últimos años, generando pérdidas estimadas entre un 4% [Telecom and Network Security Review, 1997] hasta un 20% [Cahill, Lambert, Pinheiro y Sun, 2002].

Este tipo de fraude tiene diversos objetivos y ámbitos de aplicación. Por ejemplo, se distingue entre el fraude destinado a los proveedores de servicios del cometido a los dueños de la cuenta telefónica. También se puede distinguir entre el fraude que tiene por finalidad la obtención de dinero, de aquel que busca la obtención de un servicio gratis. Los que más prevalecen son el fraude de suscripción y el fraude de navegación [Shawe-Taylor et al, 2000]. El primero ocurre cuando el estafador obtiene una suscripción con identificación falsa, sin pagar por las llamadas. El segundo es aquel en que se usa el servicio sin tener autorización, haciendo que otra persona pague por las llamadas. Usualmente se detecta por la aparición de una llamada fantasma en la cuenta telefónica, para lo cual se utiliza la clonación de los teléfonos móviles o la obtención de permisos para efectuar los llamados, ya sea a través del uso de tecnología para intervenir la red o del abuso de información privilegiada accediendo a información de la compañía.

Las redes de telecomunicaciones generan una vasta cantidad de datos, algunas veces del orden de varios gigabytes por día, donde las técnicas de Data Mining son particularmente importantes. Al igual que otros tipos de fraudes, los métodos de detección utilizados son preferentemente modelos supervisados, empleando métodos basados en reglas o sobre la comparación de resultados estadísticamente catalogados como sospecha. Los sistemas de detección basados en reglas, usan pautas tales como el aparente uso del mismo teléfono en dos lugares geográficamente distantes de manera sucesiva y rápida, y llamadas que se traslapan en el tiempo, con un alto valor y duración. En un nivel mayor, se elabora una distribución estadística de las llamadas (a menudo llamada profiles o signatures) los cuales son comparados con perfiles determinados por expertos o por métodos de aprendizaje supervisado con casos de fraude y no fraude conocidos.

[Murad y Pinkas, 1999] y [Rosset et al, 1999] distinguen entre funcionamiento a nivel de llamadas individuales, patrones de llamadas diarias y patrones de llamadas totales, y describen cuáles son los efectos de los métodos de detección de puntos atípicos para detectar comportamientos anómalos. [Cortés y Pregibon, 1998] y [Rogers, 2000] utilizan programas

basados en perfiles concernientes a la duración promedio de las llamadas, la llamada más larga y al número de llamadas a regiones particulares en el día. Técnicas de clasificación son utilizadas por [Fawcett y Provost, 1997 a.b. 1999] y [Moreau, Verrelst y Vandewalle, 1997], [Burge y Taylor, 2001], focalizándose en detectar cambios en el comportamiento de los usuarios. Mientras que [Pinheiro y Evsukoff, 2006] aplican un método de segmentación a través de mapas de Kohonen y luego de clasificación a través de redes neuronales en Brasil Telecom.

El software más conocido basado en redes neuronales fue desarrollado por el Fraud Solutions Unit of Nortel Networks [Nortel, 2000], el cual usa una combinación de éstas técnicas. Del mismo modo, se encuentra el software desarrollado por [Moreau et al, 1996] y [Shawe Taylor et al, 2000] para la Comisión Europea. [Tanigushi, Haft, Hollmen y Tresp, 1998] utilizan redes neuronales, modelos mixtos y redes bayesianas basados en los registros históricos de llamadas facturadas. El grupo de investigación [Wheatherford, 2002] de seguridad avanzada para las tecnologías de comunicaciones personales (ASPECT), se focaliza en las redes neuronales, en particular sin supervisión, para entrenar la legalidad actual de los perfiles de usuario, utilizando información reciente e histórica para definir patrones de normalidad. El fraude es muy probable cuando hay una diferencia entre un teléfono móvil del usuario y el perfil histórico.

[Burge y Shawe-Taylor, 1996-1997] se centran en técnicas de aprendizaje sin supervisión, considerando los perfiles de usuarios más las secuencias de registros de llamadas, calculando una distancia medida entre dos perfiles. En [Moreau y Vandewalle, 1997] y [Moreau, Verrelst, Vandewalle, 1997] se aplican redes neuronales supervisadas, etiquetando manualmente los perfiles de usuario recientes y antiguos, similares a las de [Burge y Shawe-Taylor, 1997], en casos fraudulentos y no fraudulentos. [Howard y Gosset, 1998], utilizan una combinación de herramientas no supervisadas y de supervisión de perfiles de usuario con ayuda de regresión logística, obteniendo mejores resultados, sobre todo en la región baja de falsos positivos.

[Fawcett y Provost, 1996-1997] desarrollaron un método de detección de fraude adaptativo basado en reglas de aprendizaje de comportamiento fraudulento, que generan umbrales de forma automática, para determinar el perfil del fraude de cuentas individuales de telefonía celular. Durante el año 1999, los mismos autores desarrollan un procedimiento denominado “Activity monitoring” mediante el cual se vigila el comportamiento de los usuarios, analizando los flujos de datos, con el fin de detectar la aparición de un comportamiento fraudulento o interesante.

Los métodos de visualización también han sido aplicados [Cox et al, 1997], en los cuales se utiliza un despliegue computacional gráfico que muestra la cantidad de llamadas entre diferentes suscriptores en varias ubicaciones geográficas junto con patrones de detección humana. El trabajo de [Cahill et al, 2002] se basa en la detección de fraudes orientado al manejo de eventos, asignando puntajes de fraude para detectar cuándo ocurre el fraude, donde la ponderación de las llamadas recientes de teléfono móvil pesan más que las previas. Estas marcas permiten identificar cuentas que pueden tener actividades fraudulentas y llamadas que pueden ser sospechosas. Este marco se ha aplicado tanto a sistemas inalámbricos como a sistemas de cables de línea.

De manera complementaria, [Lundin y Kvarnstrom, 2003] desarrollan un método para construir datos sintéticos para las pruebas realizadas en los sistemas de detección de fraude. El método identifica las propiedades estadísticas importantes, preservando los parámetros relevantes para la

formación y detección, como el usuario y el comportamiento del servicio. Se aplica un método de generación de datos sobre IP basado en Video-On-Demand (VOD), corriendo en un entorno de prueba piloto con clientes reales para generar archivos de registro de síntesis, que incluya el comportamiento de los usuarios fraudulentos y de los usuarios normales.

[Mazhelis y Puuronen, 2004] abordan el problema de sustitución de usuarios en telefonía móvil, planteándolo mediante la detección de cambios atípicos en el comportamiento del usuario. Para ello, se valen del supuesto que el comportamiento de un usuario y un impostor se diferencian en algunos detalles, y que esas diferencias pueden ser detectadas automáticamente.

Actualmente, la extensión del fraude se mide considerando factores tales como la duración de las llamadas y la tarifa, pero la nueva generación de teléfonos móviles puede necesitar también tomar en cuenta aspectos como el contenido de la llamada, porque la tecnología usada permite largas transmisiones de datos que pueden contener varios números diferentes de datos.

2.5.3. FRAUDE EN SEGUROS

El fraude de seguros es una de las áreas críticas, después del fraude de tarjetas de crédito, falsificación de cheques y fraude de telecomunicaciones. A pesar que los estudios realizados se han centrado en el sector seguros de manera global, es frecuente encontrar aplicaciones en áreas específicas. El seguro de automóviles ha acaparado la mayor atención junto al de viajes, casas y compensación de trabajadores, por las pérdidas que normalmente genera a las compañías que las ofertan. Implementar una correcta política de selección de riesgos es, sin duda, uno de los aspectos claves en el funcionamiento de una empresa aseguradora. El contrato de seguro se fundamenta en el principio de buena fe de las partes contratantes, de forma que la entidad se compromete a cubrir al asegurado de la posible materialización de un riesgo, a cambio del cobro de una prima.

La modelación del comportamiento ilícito de los asegurados ha derivado en la aparición de estudios relacionados con el diseño óptimo de la póliza y cobertura ofertadas, y con la influencia que la investigación de siniestros puede tener en la estructura de costos de la entidad [Picard, 1996-1997]; [Crocker y Morgan, 1998]. En el ámbito de aplicación de técnicas cuantitativas, la tendencia ha sido determinar las principales señales de fraude, seleccionando indicadores directamente relacionados con la aparición de comportamientos fraudulentos [Weisberg y Derrig, 1993] y [Artis, Ayuso y Guillén, 1999].

Cronológicamente hablando, el primer estudio cuantitativo fue realizado por [Weisberg y Derrig, 1993], donde los autores persiguen como objetivo identificar señales de fraude en los siniestros utilizando un modelo de regresión lineal. En su trabajo distinguen entre siniestros legítimos, planeados, con fraude oportunista y con un incremento injustificado de los daños producidos.

Trabajos posteriores se centran en modelos de elección discretos [Brockett, Xia y Derrig, 1993], [Derrig y Ostaszewski, 1995], los cuales permiten determinar las variables que justifican un determinado comportamiento y cuantificar la probabilidad de aparición del mismo. Conocidas las características de un siniestro, el modelo permite tomar decisiones sobre la conveniencia de

investigarlo o no, atendiendo a la probabilidad estimada de fraude. La aplicación de modelación logística anidada, permite tener en cuenta las etapas de decisión seguidas por el individuo a la hora de decidir si actúa o no en forma fraudulenta [Ayuso, 1998]. Posteriormente, [Ayuso y Guillén, 2000] desarrollan un modelo de elección binaria, incorporando la existencia de errores de clasificación, que permita obtener una aproximación al porcentaje de casos fraudulentos no detectados.

[Von Altrock, 1995] sugirió un sistema de lógica difusa usando valores de umbrales óptimos, que entrega la probabilidad de un fraude y las razones del por qué un reclamo puede ser fraudulento. [Cox, 1995] propuso otros sistemas de lógica difusa: el primero usa una red neuronal no supervisada para aprender las relaciones naturales en los datos y encontrar clusters significativos; el segundo, en tanto, emplea el algoritmo de Wang-Mendel para buscar y explicar las razones del por qué los proveedores del cuidado de la salud cometen fraude a las compañías de seguros. El sistema EFD [Major y Riedinger, 1995], es un sistema experto, en el cual el conocimiento especializado es integrado con información estadística, apoyando la identificación de proveedores cuyo comportamiento no está dentro de las normas. Este sistema ha sido usado para detectar fraude de seguros en 12 ciudades de EE.UU.

La metodología Hot Spot [Williams y Huang, 1997] aplica un proceso de tres pasos: el algoritmo K-means para la detección de clusters, el algoritmo C4.5 de árboles de decisiones con inducción de reglas y dominio de conocimientos, y resúmenes estadísticos y herramientas de visualización de la reglas de evaluación. Éstas han sido aplicadas al área de la salud para la detección de fraude por parte de médicos y en la Comisión de Seguros de Enfermedad. [Williams, 1999] ha ampliado el foco de la arquitectura a utilizar para generar reglas, posibilitando que el usuario especialista en el fraude explore las normas y les permita evolucionar, de acuerdo a cuán interesante sea el descubrimiento. [Brockett et al, 1998] presentó una metodología similar usando SOM para la detección de clusters antes de las redes neuronal con propagación hacia atrás (backpropagation) en reclamos de fraudes.

[He et al, 1998] utiliza redes neuronales backpropagation y SOM para analizar los resultados de la clasificación. Los resultados del cluster muestran que fuera de las cuatro clasificaciones de categorías usadas por la tasa de perfiles de la práctica médica, sólo dos de ellas son importantes. [Ormerod et al, 2003] recomienda también el uso de Redes Bayesianas dinámicas en tiempo real (BBNs) para la detección temprana de potenciales engaños fraudulentos.

2.5.4. FRAUDE EN INFORMÁTICA

El fraude en sistemas computacionales, es un gran negocio y tiene una extensiva área de investigación. Los hackers pueden encontrar passwords, leer y modificar archivos, alterar recursos de códigos, leer emails, entre otros [Denning, 1997]. Si este tipo de fraude puede ser prevenido o detectado tempranamente, entonces podría ser virtualmente eliminado. A menudo, este tipo de ataque es adaptativo, y una vez que un fraude de este tipo es reconocido, los hackers pueden buscar una ruta diferente, sobre todo si la recompensa es alta.

Debido a su importancia, se han desplegado grandes esfuerzos para desarrollar métodos de detección, generándose varios productos comerciales, entre ellos, el sistema de detección de intrusos de Cisco [CSIDS, 1999] y el sistema experto de detección de próxima generación [NIDES, Anderson, Frivold y Valdes, 1995]. Dado que el único registro de las actividades de un hacker es la secuencia de comandos que son usados cuando se compromete el sistema, los analistas de datos de fraude computacional usan preferentemente técnicas de análisis de las secuencias, utilizando tanto métodos no supervisados como supervisados.

[Forrest, 1994-1997] propuso un algoritmo de selección negativa para varios problemas de detección de anomalías. Este algoritmo genera patrones aleatorios que son comparados con cada patrón normal definido. Si alguno de estos coincide con un patrón normal, este patrón no se convierte en detector y es removido. De lo contrario, se convierte en un patrón detector y por tanto se requiere monitorearlo y hacerle seguimiento. Este algoritmo de selección negativa ha sido utilizado con éxito para detectar virus informáticos [Forrest et al, 1994], herramientas de detección de roturas, series temporales de detección de anomalías [Dasgupta, 1998], y redes de detección de intrusos [Hofmeyr, 1999] y [Hofmeyr y Forrest, 2000]. Siguiendo esta línea [Kim y Bentley, 2001] también utilizan un algoritmo de selección negativa para construir un detector de anomalías. [Shieh y Gligord, 1991- 1997] describen un método de matching y sostienen que es más efectivo que los métodos estadísticos para determinar tipos conocidos de intromisión, pero es incapaz de detectar nuevos tipos de patrones.

Las redes neuronales también han sido utilizadas para estos propósitos [Ryan, Lin y Miikkulainen, 1997]. [Lee y Stolfo, 1998] se valieron de técnicas de clasificación de un usuario o programa que ha sido identificado como normal o anormal utilizando datos, las que son aplicadas en tiempo real. Debido a que los patrones normales son a menudo muy numerosos, se desarrolla una codificación para convertir estos patrones en un número que se pueda visualizar fácilmente, comparar y entender. [Lippman et al, 2000] concluyó que el énfasis debería estar puesto en métodos de desarrollo para detectar nuevos patrones, más que en patrones antiguos, pero [Kummar y Spafford, 1994] resaltan que la mayoría de las brechas son el resultado de un pequeño número de ataques conocidos, como lo evidencian los equipos de respuesta.

Dado que la intromisión representa un comportamiento, y el objetivo es distinguir entre comportamientos de intromisión y comportamientos normales en secuencias, los modelos de Markov se han aplicado naturalmente [Ju y Vardi, 1991]. [Qu et al, 1998] también utilizan probabilidades de eventos para determinar los perfiles de fraude. [Forrest, Hofmeyr, Somayaji y Longstaff, 1996], describen un método basado en cómo los sistemas inmunes naturales distinguen entre comportamiento propio y ajeno. Tal como en los datos de telecomunicaciones, los patrones de individuos cambian a lo largo del tiempo, de manera que un sistema de detección debe ser capaz de adaptarse a los cambios, pero no tan rápidamente, de manera que acepte intrusiones como cambios legítimos. [Lanne y Brodley, 1998] y [Kosoresow y Hofmeyer, 1997] usaron la semejanza de secuencias que pueden ser interpretadas en un marco probabilístico.

También se han desarrollado incursiones para detectar intrusiones en correos electrónicos personales. [Stolfo, Hershkop, Wang y HU, 2003] utilizan modelos supervisados bayesianos para construir un clasificador que señale si el correo es benigno o malicioso. Por otro lado, [Airoldi y Malin, 2004] diseñaron un mensaje de filtro que discrimina entre los mensajes que contienen

patrones de intención fraudulenta, el cual está capacitado para hacer una decisión booleana sobre un conjunto de datos etiquetados, donde las etiquetas son “fraude” y “no fraude”. Después de que el filtro se ha entrenado, puede ser aplicado a los mensajes entrantes a un servidor de correo en tiempo real. Se trata de un problema nuevo, pero sigue una tendencia reciente de investigación en text mining, denominado *semantic learning*, utilizando árboles de decisión temporales.

2.5.5. LAVADO DE DINERO

El lavado de dinero, proceso de oscurecer la fuente, propiedad o uso de los fondos usualmente efectivos que constituyen las ganancias de actividades ilícitas, es uno de los tipos de fraude más complejos de identificar. Mientras que los fraudes con tarjetas de crédito ocurren en períodos cortos de tiempo, en el lavado de dinero pueden pasar años antes que las transferencias individuales o las cuentas sean identificadas como parte de un proceso de este tipo. Esto se dificulta por el hecho que los bancos almacenan información de manera distinta y no suelen compartirla. Por otra parte, los bancos no son las únicas entidades que transfieren dinero electrónicamente y otros negocios se han establecido precisamente para este propósito. Debido a las grandes sumas involucradas, quienes lavan dinero son altamente profesionales y a menudo tienen contactos en los bancos que pueden indicarles los detalles de las estrategias de detección que están siendo aplicadas.

La detección del lavado de dinero trabaja mano a mano con la prevención. En Estados Unidos, el acta de secreto bancario requirió que los bancos reportaran todas las transacciones superiores a 10.000 dólares a las autoridades en 1970. Sin embargo, los operadores adaptaron su modus operandi dividiendo el monto transferido en cantidades menores y realizando depósitos en diferentes bancos, lo que se denomina “smurfing” o “structuring”. En los Estados Unidos, ahora esto es ilegal, pero por la forma en que quienes incurren en este tipo de fraude se adaptan a los métodos de detección actuales, es posible aproximarse a la perspectiva pesimista, en que sólo los defraudadores incompetentes en lavado de dinero son detectados. Claramente esto limita el valor de los métodos de detección supervisada, pues los patrones detectados serán aquellos donde había características de fraude en el pasado, pero que pueden ya no serlo actualmente.

Las transferencias electrónicas de fondos contienen ítems tales como la fecha de transferencia, la identidad del remitente, el número de ruta y banco de origen, la identidad del receptor, el número de ruta del banco receptor y el monto transferido. Algunas veces ciertos campos no son requeridos para estas transferencias y pueden dejarse en blanco, los que inevitablemente da pie a errores. Se han desarrollado software que detectan y corrigen automáticamente estos errores basados en restricciones semánticas y sintácticas entre contenidos posibles, pero esto nunca puede ser una solución completa. Además, el tema es complicado por el hecho que, como ya se ha comentado, los bancos no comparten su información y, por otra parte, los bancos no son las únicas entidades que transfieren dinero electrónicamente y otros negocios se han establecido precisamente para este propósito. Es por esto que las transacciones electrónicas de fondos proveen un dominio natural para el lavado de dinero.

En general, el lavado de dinero involucra 3 pasos: (1) La introducción del efectivo en los sistemas bancarios o en negocios legítimos, por ejemplo, transfiriendo las notas bancarias obtenidas de transacciones de venta por drogas en un cheque al portador; (2) layering, que se describe como

múltiples transacciones a través de múltiples cuentas con diferentes propietarios a diferentes instituciones financieras en el sistema financiero legítimo; (3) integración, que consiste en fusionar los fondos con dineros obtenidos en actividades legítimas.

La técnica de detección más utilizada se denomina *Link Analysis* (Andrews y Peterson, 1990) que busca relaciones entre elementos de información y permite identificar los participantes involucrados en las transacciones, tanto a nivel de transacción individual, como a nivel de cuentas y a nivel de negocios. Existen sistemas que utilizan reglas basadas en la experiencia y en el desarrollo de estadísticos descriptivos directos como la distribución de Benford.

Es común también el uso de métodos para detectar cambios en el comportamiento basados en análisis de grupos de pares [Bolton y Hand, 2001] y detección de quiebres [Goldberg y Senador, 1997], así como en puntajes de sospecha. El Advanced Detection System [Kirkeland et al 1998 y Senator, 2000] usa un patrón de reglas de asociación y un patrón de secuencia de tiempo que pone énfasis en herramientas de visualización. Un acercamiento similar es el desarrollado por SearchSpace Ltd. que combina algoritmos genéticos, lógica difusa y redes neuronales para detectar tratos internos y manipulación de mercados. [Chartier y Spillane, 2000] también describen una aplicación de redes neuronales para detectar lavado de dinero.

2.5.6. OTROS ÁMBITOS

El fraude puede ocurrir también en un contexto comercial: por ejemplo, en prescripción de recetas médicas, en personas que no existen o están muertas, y donde un médico realiza un procedimiento, pero cobra al asegurador por otro servicio que es más caro, o incluso no realiza ninguno. [He, Wang, Graco y Hawkins, 1997] y [He, Graco y Yao, 1999] describieron el uso de redes neuronales, algoritmos genéticos y el método del vecino más cercano para clasificar los perfiles prácticos de médicos en Australia en clases de normal y anormal.

El fraude médico está a menudo asociado a un fraude en el seguro. [Major y Riedinger, 1992] crearon un sistema basado en estadísticas y conocimientos para detectar fraudes de salud al comparar las observaciones con aquellas que deberían ser similares. [Brockett, Xia y Derrig, 1998] utilizaron redes neuronales para clasificar cobros fraudulentos y no fraudulentos, para lesiones corporales automovilísticas en cobros de seguros de salud.

Por supuesto, la medicina no es la única área científica donde los datos han sido falsificados para soportar una teoría. Los problemas de fraude en ciencias están atrayendo una atención en aumento, pero han estado siempre presentes: científicos errantes han sido conocidos por arreglar números de experimentos para empujar el desarrollo de un producto o alcanzar un nivel de significancia para una publicación. [Press y Tanur, 2001] presentan una discusión del rol de la subjetividad en el proceso científico.

El plagio es también otro tipo de fraude. Con la evolución de internet es extremadamente simple para los estudiantes plagiar artículos y entregarlos como su propio trabajo en cursos escolares y

de universidad, para los cuales se han desarrollados sistemas de detección, como el caso del sistema DOCODE[23] en Chile [Oberreuter, L'Huillier, Rios y Velásquez, 2011].

Otras áreas en las que también se ha usado este tipo de herramientas para detectar irregularidades es en el ámbito financiero, que pueden ser usadas para detectar fraude contable y administrativo en contextos más amplios que en los del lavado de dinero. Métodos de sampling estadísticos son importantes en este tipo de auditorías y herramientas de screening se aplican para decidir cuáles devoluciones de impuestos ameritan una investigación detallada.

Herramientas estadísticas para la detección de fraude también han sido aplicadas en eventos deportivos. Por ejemplo, [Robinson y Tawn, 1995], [Smith, 1997] y [Barao y Tawn, 1999] examinaron los resultados de velocistas para evaluar si algunos tiempos excepcionales estaban fuera de línea con lo que se esperaba. Las herramientas estadísticas se están utilizando también en métodos biométricos de detección de fraude, los que se están expandiendo gradualmente. Estos incluyen huellas digitales computarizadas e identificación de retina y reconocimiento facial.

En muchas de las aplicaciones discutidas, la velocidad de procesamiento es esencial. Este es particularmente el caso en el procesamiento de transacciones con datos de telecomunicaciones e intrusión, donde un alto número de registros es procesado diariamente, pero también aplica en tarjetas de crédito, bancos y tiendas de retail.

3. MARCO TEÓRICO

Muchos problemas de detección de fraudes involucran una gran cantidad de datos. Procesarlos, requiere un análisis estadístico y necesita rápidos y eficientes algoritmos, entre los cuales el Data Mining aporta técnicas relevantes que facilitan la interpretación de los datos y ayuda a mejorar la comprensión de los procesos. Las herramientas utilizadas son muy variadas, considerando que existen diversos tamaños y tipos de problemas de detección de fraudes. A partir de estas herramientas y la metodología KDD, se pueden encontrar los patrones más relevantes asociados a las transacciones fraudulentas, con los cuales el negocio puede generar políticas y tomar acciones para controlarlas.

La principal característica de este tipo de problemas, es lo desbalanceada que se encuentra la base de datos, comparando el porcentaje de casos con fraude y sin fraude. Por lo general, el patrón de fraude es bastante poco frecuente, lo que dificulta identificar o diferenciar el patrón entre la gran masa de transacciones correctas. Esto significa que para predecir todos los ejemplos legítimos, se requiere una gran cantidad de información, ya que se puede tener una alta tasa de éxito, sin detección de ningún fraude.

Adicionalmente, la detección de fraude posee algunos problemas técnicos y prácticos debido a la limitación producida por la pobre calidad de los datos, ya que éstos son usualmente recogidos para otras tareas, diferentes al propósito de la detección misma. Aunque se han definido

[23] Document Copy Detector (www.docode.cl).

estándares para la forma de recoger este tipo de datos [National Association of Insurance Commissioners, 2003], no todos los datos son relevantes para realizar predicciones y muchos de ellos tienen comúnmente errores. Otro problema involucra encontrar la mejor vía para hacer predicciones más comprensibles para el análisis. Uno de los problemas frecuentes es la falta de dominio de conocimiento, o conocimiento previo que revela datos, como atributos importantes, las relaciones y los patrones probablemente existentes.

Finalmente, un punto trascendental son los tipos de errores que se manejan al predecir que una determinada transacción será o no será fraudulenta. Por ejemplo, el costo de inculpar a un cliente de fraude siendo que no lo es (Error Tipo I o Falso Positivo) puede ser tremendamente alto (cliente terrorista, se va de la empresa y anuncia a otras 20 personas su malestar) comparado con el error inverso: dejar pasar a un cliente que está cometiendo fraude (Error Tipo II o Falso Negativo). De esta forma, dependiendo de los costos, el sistema preferirá equivocarse de una forma que de otra. Claramente que estos costos dependen de la industria en que uno se encuentra.

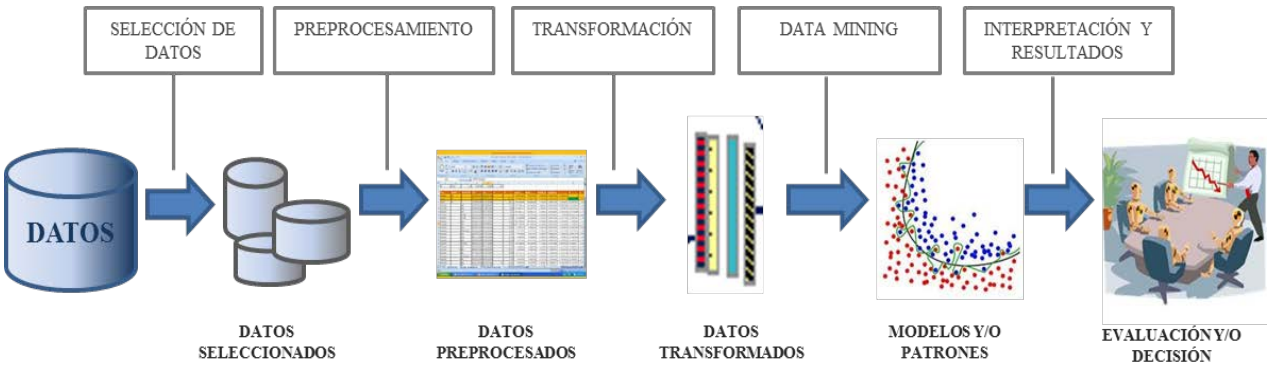
En este capítulo se describe el proceso de KDD y el funcionamiento de las técnicas de Data Mining que se utilizaran en esta tesis, considerando técnicas de aprendizaje no supervisado y supervisado, para la caracterización de los contribuyentes que utilizan y comercializan facturas falsas así como para su detección.

3.1. EL PROCESO KDD

Los algoritmos de Data Mining se enmarcan en el proceso de extracción de información conocido como KDD (Knowledge Discovery in Databases), el cual considera la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita (aunque no se conoce previamente), para encontrar relaciones o patrones y extraer conocimiento.

El proceso de KDD está dividido en una serie de pasos, desde la selección y limpieza de datos hasta la evaluación e interpretación de los resultados, como se muestra en la Figura Nº 19.

Figura N° 19: Etapas del Proceso KDD según Fayyad



Fuente: Basado en Fayyad

Las primeras etapas del proceso KDD se abocan a limpiar, preparar, seleccionar y formatear los datos de acuerdo a los patrones a buscar y al algoritmo que se utilizará. A esta etapa se le conoce como *preprocesamiento*. Posteriormente aparece la etapa de minería de datos, en la cual se buscan o descubren los patrones ocultos en los datos, los cuales pasan a una etapa de *evaluación*, en donde se determina la validez y la confiabilidad de dichos patrones. Al final de todo este proceso, se obtiene una serie de patrones llamados *conocimiento*.

El desarrollo del descubrimiento del conocimiento es además iterativo e interactivo, por lo que las fases del proceso pueden ser en cualquier momento interrumpidas para volver a comenzar en alguno de los pasos anteriores, siendo este proceso de iteración muchas veces necesario para poder lograr un descubrimiento de conocimiento de alta calidad.

A continuación se describen los aspectos más relevantes del procesamiento de datos y de las técnicas que se aplicarán en esta tesis:

3.1.1. LIMPIEZA Y PROCESAMIENTO DE DATOS

En la mayoría de los proyectos de Data Mining, la preparación de los datos es un paso fundamental para obtener buenos resultados, ya que la información puede provenir de muchas fuentes diferentes, tener errores, ambigüedades o ser redundante, consumiendo gran parte del tiempo del proyecto. Por otra parte, los datos deben ser transformados de manera apropiada para realizar el análisis.

3.1.1.1. Limpieza

La calidad de los datos tiene una incidencia directa en los resultados, ya que si los datos no son de calidad los resultados tampoco lo serán. Por otra parte, los datos deben ser confiables de manera de representar al grupo objetivo. Para ello, es importante identificar los puntos anómalos o atípicos que se presentan en los datos, ya sea por un error en su registro o digitación para proceder a su eliminación.

La detección de un punto atípico requiere de la experiencia de los involucrados en el negocio para determinar si es un error, y por ende debe ser eliminado del análisis, o si corresponde a un dato verdadero. Para su identificación se pueden utilizar histogramas, análisis de clusters (detectando aquellos casos que quedan fuera del cluster) o regresión (de manera de encontrar la curva que mejor se ajusta a los datos existentes, considerando como outliers aquellos puntos relativamente lejanos a la curva).

En el caso de no tener celdas nulas o sin información, existen métodos que tratan de estimar un valor para los datos faltantes. Por ejemplo, se puede usar la media del atributo, o la media de la variable para todas las muestras que pertenecen a una misma clase, o el valor más probable basado en los datos existentes[24]. En general, este proceso es beneficioso para entender la

[24] Utilizando árboles de decisión o Bayes, entre otros métodos.

exactitud con que se recopila la información, así como corregir cualquier error o incoherencia. Sin embargo, hay que tener en cuenta que al ser sólo estimaciones, pueden conllevar a resultados inválidos.

3.1.1.2. Integración

La extracción de datos de múltiples fuentes es una situación común en estudios de Data Mining. Los datos deben ser cuidadosamente preparados antes de realizar cualquier análisis para asegurarse que puedan responder a las hipótesis de trabajo y representar lo más fielmente posible la población objetivo. A menudo, éstos han sido recogidos con un fin totalmente distinto que el objetivo de ejercer la minería de datos. Esto implica que pueden tener inconsistencias y redundancias, generando tuplas repetidas en los distintos orígenes de la base de datos. Para eliminar estas inconsistencias, se integra la información en un mismo "depósito", usando metadata o realizando un análisis de correlación que permita medir en qué magnitud un atributo implica en el otro.

3.1.1.3. Reducción

Un aspecto importante a considerar es la representatividad de los datos. Al mismo tiempo, no es recomendable utilizar un gran número de variables para realizar el análisis, ya que los sistemas tienen capacidades limitadas de procesamiento. En ese sentido, se busca contar con el menor número de variables posible que permita representar el comportamiento del grupo de estudio. Para solucionar este problema, se realiza una reducción de la data, de manera de obtener resultados analíticos iguales o muy similares.

Entre los métodos más utilizados se encuentran:

- Agregación de los datos, por ejemplo, se puede utilizar valores promedios o totales en vez de información desagregada por período.
- Discretización, mediante la transformación de datos numéricos o continuos en valores categóricos.
- Reducción de dimensiones, en el cual se seleccionan atributos de manera que la distribución de probabilidad de las diferentes clases dados los valores de esos atributos sea lo más parecida posible a la distribución original. Entre ellos podemos mencionar métodos heurísticos de selección de atributos, Decision Tree Induction o Análisis de Componentes Principales.

Una de las opciones para eliminar variables del estudio es no considerar aquellas variables que tienen un nivel bajo de información, y que tengan muchos campos con información incompleta o nula. De la misma forma, se recomienda eliminar aquellas variables que tienen una baja correlación con la variable que se desea explicar, ya que no aportan información para el análisis.

3.1.1.4. Transformación

Muchas técnicas de análisis de datos tienen restricciones sobre los tipos de variables que están en condiciones de procesar. Como resultado, estas técnicas implican que los datos deben ser transformados en una forma adecuada para el análisis. Además, ciertas características de las variables tienen implicaciones en términos de cómo los resultados del análisis se interpretarán.

Una transformación común es la normalización, que permite tratar por igual las variables, a pesar de contar con rangos diferentes. Los métodos más utilizados se describen a continuación:

- **Min- Max:** Transforma la variable en un rango 0-1, mediante la fórmula

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}(X'_{\max} - X'_{\min}) + X'_{\min}$$

  donde X' es el nuevo valor de la variable normalizada, X es el valor original de la variable, X_{min} es el mínimo valor posible de la variable, X_{max} es el máximo valor posible, X'_{min} es el valor mínimo para el rango normalizado, y X'_{max} es el máximo valor de la gama normalizada. Esta es una fórmula útil que se utiliza ampliamente, para la cual se requiere los valores mínimos y máximos de las variables originales. Si estos no contienen todo el rango de valores posibles, el uso de la fórmula debe ser restringido para su uso en el rango especificado.

- **Z-score:** Este método normaliza los valores de la variable en torno a su valor promedio o media, utilizando la fórmula

$$X' = \frac{X - \bar{x}}{s}$$

  donde $\bar{x}$ es la media o valor medio de la variable y s es la desviación estándar. Esta fórmula es útil cuando los valores mínimos y/o máximos son desconocidos, o el valor mínimo y máximo de la variable es muy influenciado por valores fuera de rango. El objetivo es que la mayor cantidad de datos permanezca entre el origen y la desviación estándar.

- **Decimal Scaling:** Esta transformación considera un rango de valores entre -1 y 1, donde n es el número de dígitos del máximo valor absoluto y cuya fórmula es de la forma

$$X' = \frac{X}{10^n}$$

- **Value Mapping**: Esta transformación consiste en convertir datos no numéricos en información numérica.

- **Discretización**: También denominado "suavización de los datos", se aplica para transformar información numérica en información discreta, lo cual es conveniente en ciertas situaciones.

Por ejemplo cuando un valor se define en un intervalo de valores posibles, pero el conocimiento acerca de cómo se recopilan los datos sugiere que la exactitud de los datos no garantiza estas escalas. En segundo lugar, algunas técnicas solo pueden usar datos categóricos y por tanto, la conversión de los datos continuos en valores discretos es absolutamente necesaria. La discretización también puede aplicarse a variables nominales.

3.1.1.5. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, que permite reducir la dimensión del problema, y expresar los datos de forma tal que destaquen sus similitudes y diferencias, perdiendo la menor cantidad de información posible. Una de sus ventajas es que retiene aquellas características que contribuyen más a la varianza de los datos.

El ACP construye nuevos componentes principales o factores que son una combinación lineal de las variables originales del problema, donde la varianza de mayor tamaño del conjunto de datos es capturada en el primer eje, llamado el primer componente principal, la segunda varianza más grande corresponde al segundo eje y así sucesivamente, generando de esta forma variables independientes entre sí.

Para construir esta transformación lineal debe generarse primero la matriz de covarianza o matriz de coeficientes de correlación. Debido a la simetría de esta matriz, existe una base completa de vectores propios de la misma. La transformación que permite llevar las antiguas coordenadas a la nueva base, es precisamente la transformación lineal necesaria para reducir la dimensionalidad de los datos. Las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales.

Un aspecto clave a considerar en el ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los nuevos factores con las variables iniciales. Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

3.2. TÉCNICAS DE DATA MINING

La mayoría de los algoritmos de aprendizaje usados en Data Mining, se dividen en no supervisado y supervisado. Los primeros permiten detectar relaciones a través de técnicas de agrupación, asociación y de clasificación, lo que permite realizar un análisis descriptivo. Los segundos, en tanto, se utilizan con mayor frecuencia para la predicción.

Las técnicas de aprendizaje supervisado requieren un conjunto de elementos de entrada para los cuales se conoce la respuesta de salida (fraude y/o no fraude), de manera que los algoritmos se entrenen y aprendan a distinguir los patrones de cada respuesta de salida. Para esto, es indispensable contar con registros históricos de fraude en los que se tenga certeza de la clasificación resultante. Las predicciones se basan en la muestra de entrenamiento de casos,

donde los valores conjuntos de todas las variables son conocidos. Esto se llama aprendizaje supervisado o “aprender con un maestro”. Bajo esta metáfora, el “estudiante” presenta una respuesta para cada variable en la muestra de entrenamiento, y el supervisor o “maestro” provee la respuesta correcta y/o un error asociado a la respuesta generada, lo cual es caracterizado por alguna función de costo.

Con el aprendizaje supervisado, se tiene una clara medida de éxito o falta del mismo, que puede ser usado para juzgar el nivel de adecuación y comparar la efectividad de diferentes métodos en variadas situaciones. La falta de éxito es medida directamente por la pérdida o costo esperado a través de la distribución conjunta, y puede ser estimada de varias formas, incluyendo la validación cruzada (cross-validation).

En contraste, los métodos de aprendizaje no supervisado se utilizan cuando no existe un conjunto legítimo de observaciones fraudulentas con que comparar la salida. En estos casos, se enseña al algoritmo a descubrir por sí mismo las correlaciones y similitudes entre los patrones de entrada del conjunto de datos y agruparlos en diferentes categorías. De esta forma, el análisis de componentes principales, el escalado multidimensional (Multidimensional Scaling o MDS), los mapas auto-organizativos (Self-Organizing Maps o SOM), y las curvas principales, intentan identificar variedades de pocas dimensiones en el espacio de valores de entrada que representa data de muchas dimensiones. Esto provee información acerca de las asociaciones entre las variables y si pueden o no ser consideradas como funciones de conjuntos más pequeños de variables “latentes”.

A su vez, el análisis de clusters intenta encontrar regiones convexas múltiples del espacio de entrada que contienen modas de la densidad de probabilidad conjunta. Esto puede indicar si esta densidad de probabilidad conjunta puede o no ser representada por una mezcla de densidades más simples que representan distintos tipos de clases de observaciones. Las reglas de asociación, por su parte, intentan construir descripciones simples (reglas conjuntivas) que describen regiones de alta densidad, para el caso especial de data de valores binarios de muy alta dimensionalidad.

En general, el enfoque de técnicas de aprendizaje supervisado es más popular, sin embargo, la utilización de aprendizaje no supervisado puede reducir la dimensionalidad y organizar el espacio de entrada acelerando el proceso de aprendizaje [Silipo, 2003]. Muchas de las investigaciones indican que los primeros estudios utilizaban aprendizaje no supervisado para derivar clusters y luego se usaba aprendizaje supervisado para obtener puntajes o reglas de cada cluster [Berry y Linoff, 2000] o viceversa.

En nuestro caso, se utilizará una combinación de ambas técnicas, aplicando en una primera instancia técnicas de aprendizaje no supervisado para comprender las características de los contribuyentes que utilizan facturas falsas y en una segunda instancia, técnicas de aprendizaje supervisado para detectar los contribuyentes que tienen patrones similares, y en consecuencia, tengan una alta probabilidad de haber utilizado facturas falsas. A su vez, se aplicarán varias técnicas para comparar los resultados y elegir el mejor modelo de predicción.

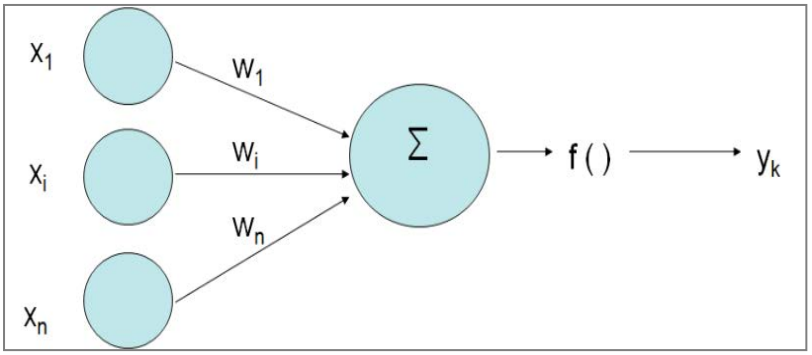
A continuación se describen las técnicas que se utilizarán en esta tesis.

3.2.1. REDES NEURONALES ARTIFICIALES

Una red neuronal artificial es un procesamiento distribuido masivamente en paralelo, que representa el funcionamiento biológico de las neuronas del cerebro, las que tienen una tendencia natural para almacenar conocimiento empírico y disponibilizarlo para su uso posterior.

Al igual que las estructuras neuronales biológicas, las redes neuronales artificiales suelen estar organizadas en capas y poseen un algoritmo de aprendizaje. Estos algoritmos están formados por un conjunto de reglas que permiten a la red neuronal aprender a partir de los datos que se le suministran, mediante la modificación de los pesos sinápticos de las conexiones entre las neuronas (el umbral de cada neurona se modifica como si fuese un peso sináptico más). Generalmente los datos que se usan para entrenar la red se suministran de manera aleatoria y secuencial. Un modelo simple de neurona se muestra en la Figura Nº 20.

Figura N° 20: Neurona natural y su modelación matemática basada en McCulloch y Pitts[25]



Fuente: Elaboración propia

Las entradas x_i representan las señales que provienen de otras neuronas y que son transmitidas a través de las dendritas. Los pesos w_i son la intensidad de la sinapsis que conecta dos neuronas, donde x_i y w_i son valores reales. Mientras que f es la función umbral (función de transferencia) que la neurona debe sobrepasar para activarse, para producir las salidas y_k. Este proceso ocurre biológicamente en el cuerpo de la célula.

Los pesos w_i pueden ser identificados siguiendo el criterio de minimización del error cuadrático, o error total definido como:

$$E = \sum_i (y_i - f(\sum_k w_k x_{ik}))^2$$

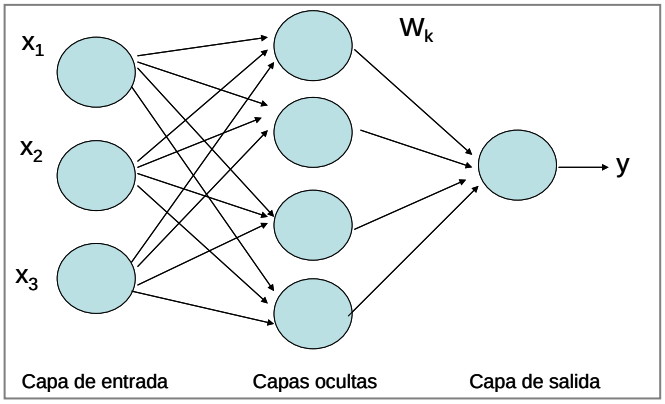
Existen diferentes funciones no lineales que pueden ser usadas para representar la función de activación f, la más utilizada es la función sigmoidal dada por la ecuación:

[25] McCulloch y Pitts realizaron el primer esfuerzo para construir una neurona artificial en 1943.

$$f(x) = \frac{1}{1+e^{-kx}}$$

La red neuronal de capa simple fue utilizada para introducir redes más complejas de múltiples capas, como se muestra en la Figura Nº 21, la cual corresponde a una red que considera tres neuronas de entrada, una capa oculta con cuatro neuronas y una neurona de salida. El modelo es más complejo en la medida que posee más capas ocultas y más neuronas de entrada o salida.

Figura N° 21: Red Neuronal Artificial con Múltiples Capas



Fuente: Elaboración propia

En este caso, es necesario reducir al mínimo el error de red, comenzando con las neuronas de salida, mediante la aplicación de la siguiente relación:

$$y = f(\sum_j W_j f(\sum_k w_{jk} x_k))$$

Para lo cual es necesario encontrar el conjunto de pesos W_j y w_{jk} que minimicen la función del error dada por:

$$E = \sum_i (y_i - f(\sum_j W_j f(\sum_k w_k x_{ik})))^2$$

Las redes neuronales artificiales de múltiples capas son principalmente usadas para:

- *Clasificación*: Mediante el entrenamiento de la red, las salidas pueden ser usadas como un vector de clasificación, por ejemplo, una categorización de riesgo de un banco o empresa aseguradora (aprendizaje no supervisado).

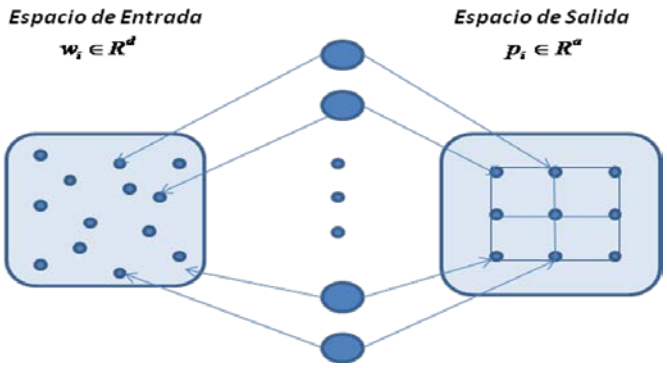
- *Predicción*: La red puede ser entrenada para representar una situación, y entregar una predicción basada en el aprendizaje de los casos de entrada (aprendizaje supervisado).

3.2.2. SELF-ORGANIZING MAP

Self-Organizing Maps (SOM) es un modelo de redes neuronales utilizado para el análisis y visualización de datos de alta dimensión, también conocido como Mapas de Kohonen, debido al nombre de su autor, el académico Teuvo Kohonen (1981). Este modelo es uno los modelos más populares de las Redes Neuronales Artificiales, basado en aprendizaje competitivo no supervisado, el cual tiene la capacidad de formar *mapas de características*, de manera similar a como ocurre en el cerebro.

La Red de Kohonen consiste en un conjunto de neuronas $i \in \{1...N\}$ dispuestas en una grilla o reticulado *d*-dimensional, que genera un espacio de salida a-dimensional, donde $a \leq d$, sobre la cual se definen relaciones de vecindad. Cada neurona tiene asociado un vector de pesos $w_i \in R^d$ en el espacio de entrada y un vector de posición $p_i \in R^a$ en la grilla de salida, como se muestra en la Figura Nº 22.

Figura N° 22: Arquitectura de la Red SOM

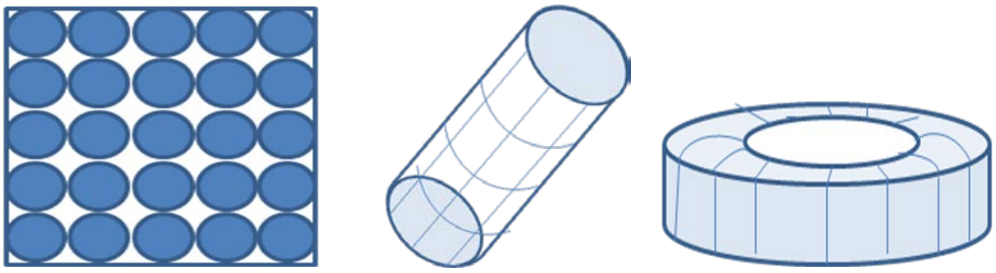


Fuente: Elaboración propia

Si bien la dimensión a de la grilla no posee restricción alguna, por lo general, es escogida menor o igual a 3, permitiendo una visualización directa. Cada neurona posee vecinos de distintos niveles dependiendo de la distancia que se encuentre de ésta. Los vecinos del primer nivel son los más cercanos y están conectados directamente a la neurona y el número de vecinos en cada nivel depende del tipo de grilla. Los tipos de grilla más utilizadas son la rectangular y la hexagonal, sin embargo, estas podrían tener cualquier topología como se muestra en la Figura Nº 23.

Los vectores prototipos en el espacio de entrada son ajustados a los datos durante el entrenamiento de la red, de manera que estos se asemejen a los puntos de la data lo mejor posible. Después de inicializar el prototipo de vectores, las neuronas de la red generan cierta actividad ante el estímulo de los datos de entrada, lo que permite determinar qué zonas o más específicamente qué neuronas han aprendido a representar ciertos patrones de la entrada. Las neuronas de mayor actividad, deben ser capaces de ajustarse más fácilmente a los ejemplos que intentan representar. De este modo se logra generar un mapa cuyas zonas de actividad van cambiando a medida que se presentan distintas relaciones.

Figura N° 23: Tipos de grilla utilizados en una Red SOM



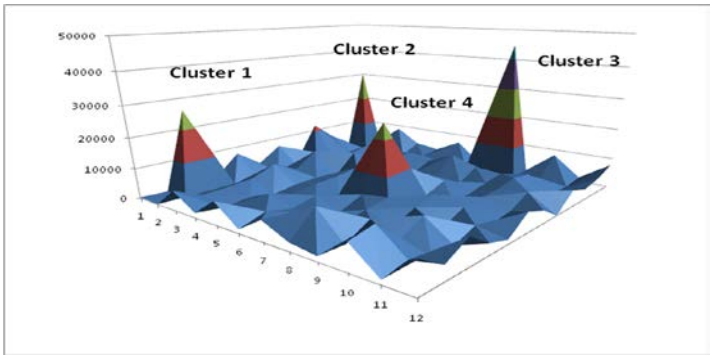
(a) Abierta (b) Cilíndrica (c) Toroidal

Fuente: Elaboración propia

Los patrones que generan actividad en la misma zona poseen características similares y pueden ser agrupados dentro de una misma categoría o cluster, basándose en una medida de distancia, normalmente Euclidiana. Dado que los vectores prototipo tienen posiciones bien definidas en el mapa, el SOM pasa a ser una especie de algoritmo de proyección de vectores. El ordenamiento topológico de las unidades del mapa depende principalmente de la vecindad local, dado que hay más unidades donde la densidad de la data es alta, la vecindad en esas áreas se vuelve más pequeña, medida en el espacio de entrada, por lo que la proyección sintoniza con la densidad local de la data. Al ser la proyección discreta (pues la cantidad de valores equivale al número de unidades del mapa), muchos vectores pueden ser proyectados al mismo punto.

La Figura Nº 24 muestra la visualización de un resultado tipo de un mapa de Kohonen, en el que se identifican 5 grupos principales o clusters, para los cuales se conoce el valor de los atributos que los caracteriza, identificando así los patrones de comportamiento de cada uno de ellos.

Figura Nº 24: Ejemplo de visualización de una Red de Kohonen



Fuente: Elaboración propia

DESCRIPCIÓN DEL ALGORITMO

- Inicialización

Inicializar los pesos w_{ij}, con valores aleatorios pequeños, fijando la zona de vecindad de las neuronas de salida. Esta inicialización por lo general es aleatoria y toma valores pequeños pero también puede realizarse a partir de las muestras, seleccionando al azar distintos puntos del conjunto de entrada para inicializar los vectores prototipos, o de manera ordenada, a partir del subespacio lineal generado por los vectores de entrada.

- Aprendizaje

Presentar a la red una información de entrada en forma de vector $x(t) \in R^d$ cuyas componentes sean valores continuos y calcular la distancia de todos los elementos de la red a cada neurona de salida. En el caso de la distancia euclidiana

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2$$

La neurona vencedora de la capa de salida o "Best Matching Unit" (BMU) es aquella cuyo vector de pesos $w_{ij}(t)$ es el más parecido a la información de entrada $x_i(t)$. Una vez localizada la neurona vencedora i*, se actualizan los pesos de sus conexiones de entrada y también los de las neuronas vecinas las que pertenecen a su zona de vecindad, de acuerdo al esquema $w_{ij}(t+1) = w_{ij}(t) + \alpha(t) h_{i,i^*}(t)(x_i(t) - w_{i^*j}(t))$,donde el término $\alpha(t)$ es un parámetro de ganancia o coeficiente de aprendizaje, que toma valores entre 0 a 1, decreciente con el número de iteraciones del proceso de entrenamiento y $h_{i,i^*}(t)$ es una función de densidad gaussiana centrada en la unidad ganadora, que permite controlar la adaptación de toda la red, de acuerdo a la distancia de cada neurona a la BMU.

Por lo general, la función de vecindad se define como:

$$h_{i,i^*}(t) = e^{-\frac{\|r_{i^*} - r_i\|^2}{2\sigma^2(t)}}$$

donde r_{i^*} y r_i son la posición de la neurona ganadora y la posición de la i-ésima neurona en el arreglo. El parámetro $\sigma(t)$ define el ancho de la vecindad y es de la forma:

$$\sigma(t) = \sigma_i * \left(\frac{\sigma_f}{\sigma_i} \right)^{t/T}$$

donde σ_i y σ_f son escogidos al inicio del entrenamiento y corresponden al tamaño inicial y final de la vecindad y T es el número máximo de iteraciones, con $\sigma_i > \sigma_f$. Por otra parte, la tasa de aprendizaje $\alpha(t)$ asegura la convergencia del algoritmo y es de la forma:

$$\alpha(t) = \alpha_i * \left(\frac{\alpha_f}{\alpha_i} \right)^{t/T}$$

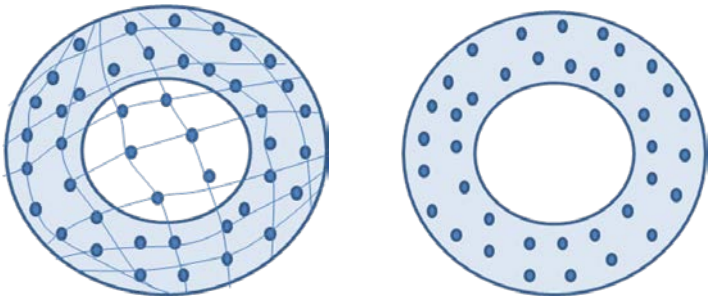
donde los valores α_i y α_f son elegidos al comienzo del entrenamiento y T es el número máximo de iteraciones. Mientras más grande sea el número de neuronas mejor es la resolución del mapa, sin embargo un número excesivo de neuronas puede hacer que el clustering sea inexistente y que cada patrón de entrada sea representado por una neurona diferente.

La mayoría de las técnicas de proyección de vectores, como los mapas de Sammon, tienen salidas continuas. Por lo tanto, en algunas tareas de visualización, la proyección definida por SOM es a menudo complementada con otros métodos. Las ventajas de este tipo de mapas es que permiten proyectar datos altamente dimensionales a un esquema de representación de baja dimensión, así como encontrar similitud en los datos, para encontrar grupos con características similares que sean representativos de un problema de estudio. Entre sus desventajas se encuentra la necesidad de definir una tipología de red a priori y predeterminar el tamaño de la red.

3.2.3. GAS NEURONAL

El Gas Neuronal (NG: Neural Gas), es un algoritmo relativamente nuevo de redes neuronales no supervisada, orientada a la cuantización vectorial de estructuras arbitrarias [Henríquez, 2008]. Este método es más general, comparado con los Mapas de Kohonen. La mayor diferencia entre ambos modelos es que el Gas Neuronal no define una grilla que impone relaciones topológicas entre unidades de la red y cada neurona puede moverse libremente a través del espacio de datos. Esta libertad permite al algoritmo una mejor capacidad para aproximar la distribución de los datos en el espacio de entrada, ya que las neuronas no están obligadas a tener que mantener ciertas relaciones de vecindad, como se aprecia en la Figura Nº 25.

Figura Nº 25: Visualización de una Red de Kohonen (a) y una Red de Gas Neuronal (b) tipo anillo



(a) (b)

Fuente: Elaboración propia

La Figura N° 25 (a) muestra una cuantización mediante SOM, mientras que la Figura N° 25 (b) muestra una cuantización mediante Gas Neuronal. Se puede observar que debido a las restricciones topológicas, existen neuronas en el SOM que no están correctamente ubicadas (se encuentran fuera de la distribución de los datos).

Las coordenadas de las neuronas son llamadas tradicionalmente “pesos”. Si consideramos una colección de M neuronas, cada neurona puede ser imaginada como un punto en el espacio de datos. Las coordenadas de la i-ésima neurona será denotada como $w_i = (w_{i1}, w_{i2} ... w_{ip})$. Al inicio, la colección de neuronas es distribuida aleatoriamente sobre el espacio de datos. En las siguientes iteraciones, las neuronas cambian su posición y se adaptan ellas mismas a la nube de datos. El proceso de adaptación es llamado *aprendizaje* o *entrenamiento*.

En este algoritmo, cada patrón de entrada, genera una excitación sobre cada unidad de la red. En cada iteración se presenta un vector de datos aleatorio $x(t)$ al conjunto de neuronas. Para cada vector de datos $x(t)$ se encuentra la neurona más cercana, de acuerdo a la distancia euclidiana. Esta neurona es llamada ganadora y obtiene el índice i*.

El vector de pesos de la neurona ganadora satisface la relación $d(x, w_{i*}) = \min d(x, w_i)$ con $1 \le i \le m$. En el paso siguiente se establece el vecindario de la neurona ganadora. La magnitud del vecindario (diámetro) decrece exponencialmente con el número de iteraciones. Para cada iteración t, todas las neuronas pertenecientes al vecindario de la neurona ganadora cambian su posición para ubicarse más cerca del vector $x(t)$ actualmente expuesto. El cambio es descrito mediante la regla $\Delta w_i = \alpha(t) h_\lambda(t)(x(t) - w_i)$, donde la función $h_\lambda(t)$ describe el vecindario de la neurona ganadora, de acuerdo a la expresión:

$$h_\lambda(t) = e^{\frac{-d^2(x, w_i)}{2\sigma^2(t)}}$$

y la tasa de aprendizaje $\alpha(t)$ determina qué tan grande puede ser el cambio de posición. Esta tasa usualmente decrece con el número de iteraciones, con un valor de inicio α_0, el cual va decreciendo gradualmente a un valor final $\alpha_{\min}$ alcanzado al final de todas las iteraciones T, descrito por la función:

$$\alpha(t) = \alpha_0 \left(\frac{\alpha_{\min}}{\alpha_0} \right)^{t/T}$$

De la misma manera tenemos que $\sigma(t)$ define el diámetro del área de vecindad en la iteración t. Normalmente decrece con el número de iteraciones, el cual va disminuyendo desde un valor inicial σ_0 hasta llegar a un valor final $\sigma_{\min}$ en el número máximo de iteraciones T, de acuerdo a la fórmula:

$$\sigma(t) = \sigma_0 \left(\frac{\sigma_{\min}}{\sigma_0} \right)^{t/T}$$

En el algoritmo Gas Neuronal Creciente (GNG: Growing Neural Gas), a diferencia de los previamente descritos, el número de unidades crece durante el proceso auto-organizativo. Comenzando con 2 unidades, para luego ir agregando neuronas a la red, de manera sucesiva con vectores prototipos w_1 y w_2 seleccionados aleatoriamente. Presentando un valor de entrada $x(t)$ de acuerdo a alguna distribución. Para determinar donde insertar nuevas unidades, se calculan algunas medidas de error locales durante el proceso de adaptación. Una nueva unidad se inserta cerca de aquella neurona que posee el mayor error acumulado.

En este caso, el algoritmo encuentra el BMU tal que los nodos i* y j* sean los más cercanos a $x(t)$, con vectores prototipos w_{i*} y w_{j*} tales que $\|w_{i*} - x(t)\|^2$ es el valor más pequeño y

$\|w_{j^*} - x(t)\|^2$ es el segundo valor más pequeño para todos los nodos. Si no existe una conexión entre i* y j* entonces ésta es creada, configurando la edad del borde entre i* y j* con $age_{(i^*, j^*)} = 0$. A continuación se incrementa la edad de todos los bordes que emanan de i* de acuerdo a $age_{(i^*,i)} = age_{(i^*,i)} + 1, \ \forall i \in N_{i^*}$ donde N_{i^*} es el conjunto de vecinos topológicos directos en los que existe un borde entre i y i*.

Finalmente se remueven aquellos bordes con una edad mayor que la edad máxima $T(t)$, definida como:

$$T(t) = T_i \left(\frac{T_f}{T_i} \right)^{t / t \max}$$

Existe también el algoritmo Gas Neuronal de Mergel (MNG: Merge Neural Gas), el cual es más rápido e intuitivo que los anteriores, orientado al procesamiento de secuencias. Este modelo combina una arquitectura tolerante al ruido, basada en las unidades ganadoras de las instancias anteriores con el cuantizador del Neural Gas (NG), para ir construyéndose de manera recursiva. Una de las grandes cualidades de esta arquitectura es que puede ser combinada con grillas arbitrarias.

3.2.4. ÁRBOLES DE CLASIFICACIÓN

Los árboles de clasificación, también llamados árboles de decisión, son uno de los métodos de aprendizaje supervisado no paramétrico más utilizado, ya que destacan por su sencillez y su aplicabilidad a diversas áreas e intereses. En general, los algoritmos de construcción de árboles se diferencian en las estrategias utilizadas para particionar nodos y podar el árbol. Uno de los primeros fue el AID (Automatic Interaction Detection) [Sonquist, Baker y Morgan, 1971], el cual se basa en un algoritmo recursivo con sucesivas particiones de las observaciones originales en otros subgrupos menores y más homogéneos mediante secuencias binarias. Una adaptación posterior se conocido como CART (Classification And Regression Trees, o árboles de clasificación y regresión) propuesto por [Breiman et al, 1984]. También existen algoritmos de clasificación no binaria como el CHAID (Chi-square automatic interaction detection) introducido por [Kass, 1980], el algoritmo C5 desarrollado por [Quinlan ,1993], el algoritmo ID3 (Interactive Dichotomizer) introducido por [Quinlan, 1986], los Arboles Bayesianos basados en la aplicación de métodos bayesianos a arboles de decisión y el MARS (Multivariate Adaptive Regression Splines), propuesto por [Friedman, 1991].

En esta tesis, se utilizarán los árboles basados en la metodología CHAID, que permite generar un número distinto de ramas a partir de un nodo considerando tanto variables continuas como categóricas. Por otra parte, tiene la ventaja de ser muy intuitivo en su forma de presentarse como árbol. Para utilizar esta técnica es necesario disponer de tamaños de muestra significativos, ya que al dividirse en múltiples subgrupos, cabe el riesgo de encontrar grupos vacíos o poco representativos si no se dispone de suficientes casos en cada combinación de categorías.

Básicamente el algoritmo consiste en:

- Formar inicialmente todos los pares posibles y combinaciones de categorías. Para cada nodo padre potencial, el algoritmo evalúa todas las combinaciones de las posibles variables explicativas, agrupando las categorías que se comportan homogéneamente con respecto a la variable respuesta en un grupo y manteniendo separadas aquellas categorías que se comportan de forma heterogénea.
- Para cada posible par se calcula el estadístico correspondiente a su cruce con la variable dependiente. El par con valor más bajo de este indicador formará una nueva categoría de dos valores fusionados, siempre que no sea significativo. Esta última condición es importante, pues si fuese significativa, implica que las dos categorías que se pretenden fusionar no lo pueden hacer porque son heterogéneas entre sí con la variable dependiente, y lo que se busca es justamente lo contrario, relacionar categorías con comportamiento similar.
- Para las categorías fusionadas se procede a realizar nuevas fusiones de los valores del pronosticador, pero esta vez con una categoría menos, pues dos de las antiguas categorías fueron reducidas a una sola.
- El proceso se acaba cuando ya no pueden realizarse más fusiones porque los estadísticos entregan resultados significativos.

La prueba estadística utilizada depende de la medición nivel del campo de destino. Si el campo de destino es una variable categórica, se utiliza una prueba estadística chi-cuadrado y si es una variable continua, se utiliza una prueba estadística F.

- Prueba estadística Chi-cuadrado

Si el campo de destino Y es una variable con salidas categóricas, se realiza un test que mide la independencia entre la variable de destino y alguna otra variable X. Para ello se construye una tabla de contingencia que contiene la frecuencia observada del número de casos que se encuentra en cada casilla, considerando una categorización de la variable X, en caso que ésta sea continua.

Para determinar la independencia se construye el estadístico Chi-cuadrado de Pearson $X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$, donde $n_{ij} = \sum_n f_n I(x_n = i \wedge y_n = j)$ es la frecuencia observada y $\hat{m}_{ij}$ es la frecuencia esperada de cada celda ij si se cumple la relación de independencia, calculada como: $\hat{m}_{ij} = \frac{n_{i*} n_{*j}}{n_{**}}$, donde $n_{i*} = \sum_{j=1}^{J} n_{ij}$, $n_{*j} = \sum_{i=1}^{I} n_{ij}$ y $n_{**} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$.

El correspondiente p-valor se calcula como $p = \Pr ob(x_d^2 > X^2)$ donde x_d^2 sigue una distribución chi-cuadrado con $d = (J-1)(I-1)$ grados de libertad.

Adicionalmente, para realizar el análisis de segmentación, normalmente se ponen filtros que imponen distintos criterios para detener el proceso de clasificación. De lo contrario, la clasificación podría considerar una gran cantidad de grupos terminales de tamaño muy pequeño que serían difíciles de interpretar. En el otro extremo, con un número elevado de variables y sin restricción alguna, esta clasificación produciría tantos grupos como individuos tuviese la muestra.

Los filtros o criterios antes mencionados se clasifican en:

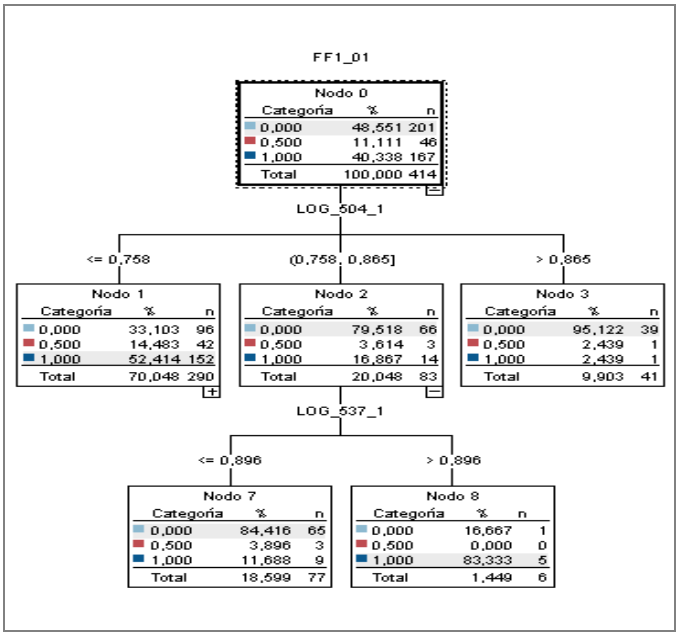
- *Filtros de significación*: Son los más utilizados en la técnica de segmentación CHAID. Su criterio consiste en no permitir segmentaciones que no sean estadísticamente significativas. Por defecto, se utiliza un nivel de significación del 5% que indica que los resultados tienen un 95% de nivel de confianza.
- *Filtros de asociación*: En este caso el criterio de detención ocurre porque el coeficiente de asociación elegido no alcance un determinado nivel. Este criterio es más permisivo que el anterior, porque el filtro de significación es más sensible al número de casos de cada grupo, mientras que el de asociación elimina este efecto.
- *Filtros de tamaño*: Su principal objetivo es evitar que se formen grupos muy pequeños durante el proceso de segmentación, dado el problema que supone la generalización de estos casos. Estos filtros pueden aplicarse antes o después de la segmentación.
- *Filtros de nivel*: Este criterio supone poner como condición un número máximo de niveles de segmentación. Este filtro evita que se formen múltiples segmentaciones en segmentos desproporcionadamente grandes de la muestra. Por otra parte, simplifica los resultados al reducir el número de variables requeridos para explicar la variable dependiente.

De esta forma, se evaluará la utilización de un CHAID exhaustivo, el cual es una modificación del algoritmo que busca hacer frente a algunas de sus debilidades [Biggs, de Ville, y Suen, 1991]. En particular, a veces no es posible encontrar la separación óptima para una variable, ya que la fusión de categorías se detiene tan pronto como se encuentra que todas las categorías restantes son estadísticamente diferentes. El CHAID exhaustivo remedia esto, combinando las categorías de la variable predictora hasta dejar sólo dos categorías. Posteriormente, examina la serie de fusiones para la predicción y encuentra el conjunto de categorías que entrega el nivel de asociación más fuerte con la variable de destino, calculando un p-valor ajustado. Este método utiliza las mismas pruebas estadísticas y trata de la misma forma los valores perdidos. Debido a su método de combinar las categorías de las variables es más completo, pero normalmente requiere más tiempo para realizar los cálculos.

Entre las ventajas de este método se encuentra la posibilidad de generar reglas que permitan distinguir entre casos de fraude y no fraude, considerando las variables más relevantes y significativas respecto de la variable de estudio. Por otro lado, su representación a través de grafos permite tener una mejor comprensión del problema, representando situaciones complejas mediante una estructura de fácil entendimiento como el árbol. Su desventaja principal es su dificultad cuando se presentan muchas alternativas, lo cual puede ocurrir si el modelo busca aproximarse a la realidad, donde el número de cálculos puede crecer en forma desproporcionada.

En la Figura Nº 26 se muestra un ejemplo de formato de salida para este tipo de árboles, considerando una salida con tres categorías. Como se aprecia en la figura este árbol en particular contiene 3 niveles y está compuesto por 6 nodos. Para cada nodo se obtiene el número de casos que cumplen la condición que lo caracteriza y el porcentaje que representa del tamaño total de la muestra.

Figura Nº 26: Visualización de un árbol con salida categórica



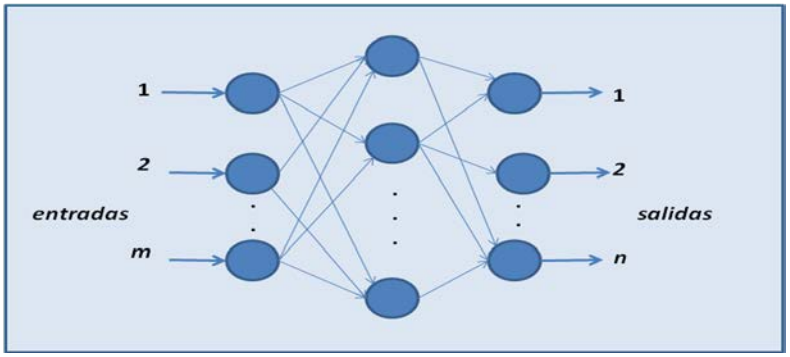
Fuente: Salida tipo utilizando el software Clementine de SPSS

Esta técnica será utilizada principalmente con una finalidad exploratoria, para conocer las variables explicativas más significativas respecto de la variable que indica el resultado de la auditoría. De igual forma se evaluará esta técnica como herramienta predictiva para la detección de contribuyentes que utilizan facturas falsas, de acuerdo a información de casos de fraude conocidos.

3.2.5. REDES NEURONALES ARTIFICIALES CON BACKPROPAGATION

El modelo perceptrón de multicapas (MLP), es una red neuronal artificial de varias capas (Figura Nº 27) utilizado para el aprendizaje supervisado. Para esto, se cuenta con un conjunto de ejemplos (observaciones) con atributos de entrada para los cuales se conoce la salida deseada. La red debe encontrar la relación existente entre los atributos de entrada y la salida deseada para cada ejemplo. Esto lo realiza a través de un método de aprendizaje llamado “Backpropagation” o “Retropropagación del error”, que minimiza el error de predicción mediante un ajuste a los pesos de la red.

Figura Nº 27: Esquema de una Red Neuronal Artificial con 2 capas (MLP)



Fuente: Elaboración propia

Este método posee dos etapas: una fase de propagación (forward pass phase) y una fase de retropropagación (backward pass phase):

- En la fase de propagación, se calculan las salidas basado en las entradas y los pesos asignados a la red inicial, para la cual se calcula el error de la predicción.
- En la fase de retropropagación, se calcula el error hacia atrás a través de la red, desde las unidades de salida hacia las unidades de entrada, obteniendo un error en cada unidad. De esta forma se actualizan los pesos de modo de minimizar el error a través de un método de descenso por gradiente.

Este proceso es iterativo, por lo que tras realizar varias veces el algoritmo, la red va convergiendo hacia un estado que permita clasificar todos los patrones de entrenamiento, que minimiza el error[26].

Suponiendo que nuestra red consta de p neuronas de entrada y j neuronas de salida, el error en la neurona de salida j en la iteración t se define como $e_j(t) = d_j(t) - y_j(t)$ donde j es un nodo de salida, $d_j(t)$ corresponde a la respuesta deseada del nodo j y $y_j(t)$ corresponde a la señal de salida de la neurona j. Por otra parte, la suma de errores cuadráticos de la red se define como:

$$E(t) = \frac{1}{2}\sum_{j \in C} e_j^2(t)$$

donde el conjunto C incluye todos los nodos en la capa de salida de la red. El número de patrones o ejemplos forman el conjunto de entrenamiento o *Training Set, el* cual denotaremos como N. Así el Error Cuadrático Medio es obtenido sumando $E(t)$ para todas las iteraciones y normalizando respecto N es

$$E_{av} = \frac{1}{N}\sum_{n=1}^{N} E(t)$$

[26] Normalmente se calcula el error cuadrático medio.

El nivel de activación producido en la entrada asociada a la neurona j es

$$v_j(t) = \sum_{i=0}^{p} w_{ij}(t) y_i(t)$$

donde p es el número total de entradas (excluyendo la umbral) aplicadas al nodo j. Por ello, la señal que nos aparece a la salida de la neurona j en la iteración t será:

$$y_j = \varphi_j(v_j(t))$$

Aplicando la regla de la cadena podemos expresar el gradiente de la suma cuadrática de los errores respecto a los pesos asociados de la forma

$$\Delta w_{ji}(t) = \eta \delta_j(t) y_i(t)$$

donde η es una constante que determina la velocidad de aprendizaje, llamado parámetro de aprendizaje del algoritmo de backpropagation y $\delta_j(t)$ viene dado por

$$\delta_j(t) = e_j(t)\varphi_j'(v_j(t))$$

En el caso que la neurona j pertenezca a una capa oculta de la red, podemos reescribir el Error Cuadrático Medio considerando que la neurona k es un nodo de salida como

$$E(t) = \frac{1}{2}\sum_{k \in C} e_k^2(t)$$

En este caso el nivel de activación para la neurona k vendrá dado por

$$v_k(t) = \sum_{j=0}^{q} w_{kj}(t) y_j(t)$$

donde q nos indica el número total de entradas (excluyendo la umbral) aplicadas a la neurona k. En este caso el gradiente local $\delta_j(t)$ para una neurona oculta j queda de la forma:

$$\delta_j(t) = \varphi_j'(v_j(t)) \sum_{k \in C} \delta_k(t) w_{kj}(t)$$

Varias investigaciones han demostrado que, durante el proceso de entrenamiento, la red neuronal artificial backpropagation tiende a desarrollar relaciones internas entre neuronas con el fin de organizar los datos de entrenamiento en clases. Esto hace que el modelo se degenere, pues memoriza los datos con los que está siendo entrenada, reproduciendo bien los datos para el grupo de entrenamiento, pero prediciendo de mala forma si se aplica sobre un conjunto nuevo de datos. Para evitar lo anterior, existen diversas soluciones, como el decaimiento exponencial de los

pesos, en donde los pesos se van “congelando” a medida que el entrenamiento avanza, evitando así el sobreajuste.

El método más utilizado para evitar el sobreajuste, es la detección temprana. Para ello los datos se dividen en dos conjuntos: un conjunto de entrenamiento y un conjunto de testeo. Con el conjunto de entrenamiento, la red realiza el aprendizaje y el ajuste de sus pesos, minimizando el error sobre este conjunto. Paralelamente, la red neuronal con los pesos obtenidos, se evalúa sobre otro conjunto que no conoce (conjunto de testeo).

Otro factor a considerar es el excesivo uso de parámetros o grados de libertad en el modelo, correspondiente a los pesos que hay entre las capas. Las variables de entrada y de salida están dadas por el problema, sin embargo, no hay ninguna regla que defina el número de unidades en las capas intermedias. Mientras más unidades se usen en esta capa, más pesos tendrá la red, aumentando el riesgo de sobreajuste.

Las redes neuronales tienen la ventaja que debido a su capacidad de establecer cualquier relación entre la entrada y la salida mediante el entrenamiento, son capaces de resolver problemas que, de otra manera, serían muy complejos de atacar, como por ejemplo:

- *No linealidad*: En general las neuronas no tienen una respuesta lineal, y por tanto, la interconexión de neuronas es no lineal. Además, la arquitectura de la red se puede seleccionar en función de la aplicación, lo que les dota de gran versatilidad.
- *Relación entrada – salida*: Las redes neuronales aprenden de un conjunto de muestras etiquetadas donde a cada entrada le corresponde una única salida, que termina por construir una relación entre la entrada y la salida del problema que está tratando.
- *Adaptabilidad*: Debido a su capacidad aprender de manera continua, son capaces de adaptarse a problemas no estacionarios en el tiempo, adecuando sus pesos al entorno que le rodea.
- *Tolerancia a fallos*: El alto grado de paralelismo les otorga una gran rapidez de funcionamiento y una gran tolerancia a fallos y robustez, ya que son sistemas distribuidos en los que el fallo se ve amortiguado por la presencia de neuronas y pesos vecinos.

Con esta técnica se espera aprender los patrones de comportamiento de los contribuyentes de IVA auditados catalogados con y sin fraude de facturas falsas en el año 2006, para posteriormente testear este aprendizaje en un grupo de testeo, que permita evaluar en qué medida se pueden generalizar los patrones a nuevos conjuntos de datos, de manera de detectar quiénes son usuarios potenciales de facturas falsas.

3.2.6. REDES BAYESIANAS

Una red bayesiana es un grafo dirigido acíclico, utilizado para predecir la probabilidad de ocurrencia de diferentes resultados, sobre la base de un conjunto de hechos. La red consta de un conjunto de nodos que representan las variables del problema que se desea resolver, y de arcos dirigidos, que conectan los nodos e indican una relación de dependencia existente entre los

atributos de los datos. Las redes bayesianas describen la distribución de probabilidad que gobierna un grupo de variables, especificando suposiciones de independencia condicional junto con probabilidades condicionales.

Sea $X = \{X_1, X_2 ... X_n\}$ un conjunto de variables aleatorias. Formalmente, una red bayesiana para X es un par B = (G, P) en el que:

- G es un grafo acíclico dirigido en el que cada nodo representa una de las variables $\{X_1, X_2 ... X_n\}$ y cada arco representa relaciones de dependencia directas entre las variables. La dirección de los arcos indica que la variable 'apuntada' por el arco depende de la variable situada en su origen.
- P es un conjunto de parámetros que cuantifica la red. Contiene las probabilidades $P(x_i / pa_i)$ para cada posible valor x_i de la variable X_i y cada valor posible pa_i de Pa_i, donde éste último denota al conjunto de padres de X_i en G.

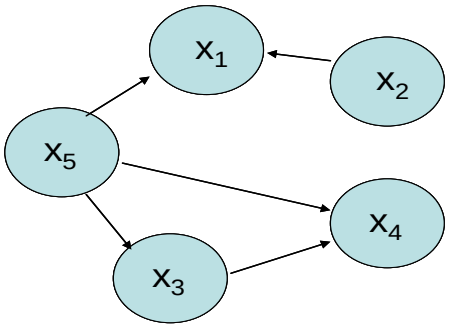
Así, una red bayesiana B define una distribución de probabilidad conjunta única sobre X dada por

$$P(X) = P(X_1, X_2 ... X_n) = \prod_{i=1}^{n} P(x_i / pa_i)$$

Un ejemplo de red bayesiana se presenta en la Figura Nº 28. La función de probabilidad conjunta de esta red es:

$$P(X_1, X_2 ... X_5) = P(X_1 / X_2, X_5) P(X_2) P(X_3 / X_5) P(X_4 / X_3, X_5) P(X_5)$$

Figura Nº 28: Ejemplo de una red bayesiana básica



Fuente: Elaboración propia

La topología o estructura de la red no sólo proporciona información sobre las dependencias probabilísticas entre las variables, sino también sobre las independencias condicionales existentes entre ellas. Cada variable es independiente de aquellas variables que no son descendientes suyas en el grafo, dado el estado de sus variables padre. Esta inclusión de las relaciones de independencia en la propia estructura del grafo hace de las redes bayesianas una buena herramienta para representar conocimiento de forma compacta, reduciendo el número de parámetros necesarios. Además, proporcionan métodos flexibles de razonamiento basados en la

propagación de las probabilidades a lo largo de la red de acuerdo con las leyes de la teoría de la probabilidad.

El problema del aprendizaje bayesiano puede describirse informalmente como: dado un conjunto de entrenamiento $D = \{u_1, u_2 ... u_n\}$ de instancias de X, encuentre la red B que se ajuste mejor a D, el cual se divide en dos partes:

a) Aprendizaje estructural, que consiste en obtener la estructura de la red.
b) Aprendizaje paramétrico, en el que conocida la estructura del grafo, se obtienen las probabilidades correspondientes a cada nodo.

Teniendo en cuenta que el tamaño de las tablas de parámetros crece exponencialmente con el número de padres de un nodo, es conveniente observar distintas técnicas para reducir el número de parámetros necesarios.

a) Aprendizaje Estructural

Existen diferentes técnicas que permiten construir redes bayesianas. La primera de ellas reúne métodos que exploran las *relaciones de dependencia* existentes entre subconjuntos de variables, para elegir la forma en que deben conectarse. El estudio de esas relaciones requiere establecer un criterio cuantitativo para medir la dependencia entre variables, y es dicho criterio el que guía la construcción de la red. La otra aproximación habitual al aprendizaje de redes, consiste en realizar una *búsqueda guiada por una medida global de calidad*. En esta otra aproximación, la operación general consiste en generar distintos grafos mediante un algoritmo de búsqueda, y aplicar a cada uno de ellos una función de medida de calidad para decidir qué grafo conservar en cada paso. Generalmente, la segunda metodología tiene menor tiempo de complejidad pero puede no encontrar la mejor solución debido a su naturaleza heurística, mientras que la primera es generalmente asintóticamente correcta cuando la función de probabilidad cumple ciertas condiciones.

b) Aprendizaje Paramétrico

El aprendizaje paramétrico es simple cuando todas las variables son completamente observables en el conjunto del entrenamiento. El método más común es el llamado *estimador de máxima verosimilitud,* que consiste en estimar las probabilidades deseadas a partir de la frecuencia de los valores de los datos de entrenamiento. Con esta técnica se espera calcular la probabilidad de ocurrencia de un determinado evento de fraude dado el cumplimiento de ciertas condiciones en sus atributos.

La calidad de estas estimaciones dependerá de que exista un número suficiente de datos en la muestra. Cuando esto no es posible se puede cuantificar la incertidumbre existente representándola mediante una distribución de probabilidad, para así considerarla explícitamente en la definición de las probabilidades. Habitualmente se emplean distribuciones Beta en el caso

de variables binarias, y distribuciones Dirichlet[27] para variables multivaluadas. Esta aproximación es útil cuando se cuenta con el apoyo de expertos en el dominio de la aplicación, para concretar los valores de los parámetros de las distribuciones.

Suponiendo que la función de distribución de probabilidad conjunta de X sobre la estructura G depende de parámetros $\vartheta_G = (\vartheta_1, \vartheta_2 ... \vartheta_n)$, esta función queda de la forma:

$$P(X / \vartheta_G, G) = \prod_{i=1}^{n} P(x_i / pa_i, \vartheta_i, G)$$

En consecuencia, el problema de estimación de parámetros se reduce a calcular la función de probabilidad posterior $P(\vartheta_G / D, G)$, donde D representa el conjunto de datos de entrenamiento. Suponiendo que los parámetros ϑ_i son mutuamente independientes, tenemos que:

$$P(\vartheta_G / D, G) = \prod_{i=1}^{n} P(\vartheta_i / D, G)$$

En general, podemos suponer que todas las variables X_i en la red Bayesiana G tienen distribución multinomial, ya que en caso que la variable sea continua ésta se discretiza. Cada función de probabilidad local P_i asociada con la variable X_i es una colección de distribuciones multinomiales P_{ij} que corresponde a la distribución para cada configuración de los padres Pa_i.

Luego $P(x_i^k / pa_i^j, \vartheta_i, G) = \vartheta_{ijk}$ corresponde a la probabilidad de que la variable X_i se encuentre en su configuración *k* y sus padres en la configuración *j*. El número de configuraciones de la variable X_i es r_i y el número de configuraciones de sus padres es q_i. El vector de parámetros asociados con cada función de probabilidad local P_{ij} se denota por $\vartheta_{ijk} = (\vartheta_{ij1}, \vartheta_{ij2} ... \vartheta_{ijr_i})$ con una distribución a priori Dirichlet. Se puede actualizar cada vector de parámetros de manera independiente, de acuerdo a

$$P(\vartheta / D, G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\vartheta_{ij} / D, G)$$

Usando el supuesto de que la distribución Dirichlet es la distribución a priori conjugada y que los parámetros pueden estimarse usando el valor esperado de esa distribución, se tiene que la probabilidad de un caso no observado x_{n+1} es de la forma

$$P(x_{n+1} / D, G) = \prod_{i=1}^{n} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}$$

[27] La distribución Dirichlet corresponde a la versión multivariante de la distribución Beta.

donde N_{ijk} es el número de casos en D en el que la variable X_i está en la configuración k y los padres de X_i están en la configuración j, es decir, $X_i = x_i^k$ y $Pa_i = pa_i^j$ con $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Por otra parte, α_{ijk} es el parámetro de la distribución Dirichlet con $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ y $\alpha_{ijk} > 0$.

Específicamente, en este trabajo se utiliza el Algoritmo TAN de Friedman y Goldzmith para construir la red bayesiana que se encuentra disponible en el software Clementine de SPSS, el que se describe a continuación:

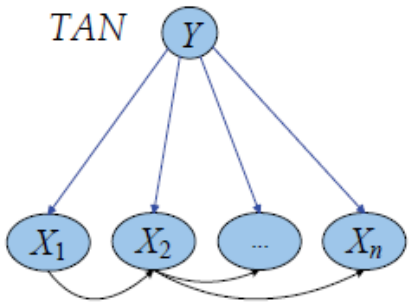
Algoritmo Estructural de TAN de Friedman y Goldzmith

Este algoritmo crea una red con una topología restringida y corresponde a una mejora del Modelo Naive Bayes. El principio que guía la construcción de la red, es una medida de información mutua que cuantifica la relación entre las variables. Por otra parte, utiliza como condición que cada una de ellas dependa de otro factor además de la variable que indica el resultado final.

Sus principales ventajas son su precisión en la clasificación y sus resultados favorables en comparación con otros modelos bayesianos generales. Su desventaja es que debido a su sencillez, se imponen mayores restricciones sobre la estructura de dependencia no cubierto entre sus nodos.

Sea $X = (X_1, X_2, ...X_n)$ el vector que representa las variables predictivas de la variable objetivo Y . El algoritmo TAN primero aprende una estructura de árbol sobre X utilizando la información mutua condicionada a Y . Luego agrega un arco desde el nodo Y resultante a cada nodo predictor X_i . Este clasificador define las siguientes condiciones: (i) Cada variable predictora X_i tiene como padre la variable objetivo resultante Y . (ii) Cada variable predictora X_i puede tener otras variables predictoras X_j como padres, tal como se aprecia en la Figura Nº 29.

Figura Nº 29: Ejemplo de una estructura de red bayesiana algoritmo TAN



Fuente: Elaboración propia

La información mutua entre dos nodos X_i, X_j dada una variable objetivo Y , dada por Friedman, se define como:

$$I(X_i, X_j / Y) = \sum_{x_i, x_j, y_k} P(x_i, x_j, y_k) \log(\frac{P(x_i, x_j / y_k)}{P(x_i / y_k) P(x_j / y_k)})$$

La red se construye a partir de los siguientes pasos:

1) Se calcula $I(X_i, X_j / Y), i = 1...n, j = 1...n, i \neq j$ para cada par de variables.
2) Se utiliza el algoritmo de Prim [Prim et al., 1957] para construir un árbol recubridor máximo ponderado con el peso de una unión de borde entre X_i, X_j dado por $I(X_i, X_j / Y)$. Este algoritmo funciona de la siguiente manera: comienza con un árbol sin bordes y marca una variable al azar como entrada. Luego se encuentra una variable cuyo peso con una de las variables marcadas sea máximo, a continuación, marca esta variable y agrega el borde al árbol. Este proceso se repite hasta que todas las variables están marcadas.
3) Se transforma el árbol resultante no dirigido a uno dirigido escogiendo X_1 como un nodo raíz y ajustando la dirección de todos los bordes para ser externos a éste.

Una de las ventajas principales de las redes bayesianas, es que permiten representar tanto el aspecto cuantitativo de un problema como su aspecto cualitativo. Esto, debido a que posibilita conocer la estructura causal de un conjunto de datos. Por otra parte, permite trabajar con datos perdidos de una manera eficiente, lo que en la práctica es deseable. Una desventaja de este último punto es que si un estado de la variable no aparece en el grupo analizado, la probabilidad de que ocurra ese evento es cero. Adicionalmente, al presentar toda la información en un único formato (probabilístico y gráfico), facilita las interpretaciones posteriores, entregando una visión general del problema.

3.2.7. HERRAMIENTA TECNOLÓGICA

Para esta tesis se utilizará una combinación de dos herramientas. Una de ellas corresponde a la herramienta de open source denominada R, la cual consiste en un lenguaje de programación y sistema de gráficos, que provee una gran variedad de técnicas estadísticas y de data mining, distribuido gratuitamente bajo los términos de la GNU[28]. De manera complementaria se utilizará la herramienta de minería de datos del software estadístico SPSS, Clementine.

R está disponible en varias formas: el código fuente escrito principalmente en C (y algunas rutinas en Fortran), esencialmente para máquinas Unix y Linux, o como archivos binarios pre compilados para Windows, Linux [Debian, Mandrake, RedHat, SuSe], Macintosh y Alpha Unix.

Una de sus principales ventajas, es que tiene una gran colección de funciones que se actualizan constantemente, y posee grandes capacidades gráficas, que superan a otros softwares, con la posibilidad de producir outputs de gráficos en distintos formatos. Por otra parte, R trabaja bien con muchas otras herramientas, importando data de distintas fuentes. Además, es un software gratuito, lo que permite a cualquier usuario descargarlo e instalarlo sin mayor problema. Entre sus desventajas se encuentran el uso de comandos y rutinas que pueden ser complicadas de

[28] Para mayor información: http://www.gnu.org/

utilizar y comprender para usuarios principiantes, y el tiempo requerido para aprender a usar el programa y encontrar los comandos requeridos. Por otra parte, si bien existe documentación de los comandos, no existen muchos ejemplos de su aplicación en la web de manera libre.

En el caso de Clementine, su ventaja principal viene dada por su mayor facilidad de uso, ya que es mucho más sencillo de manejar que R, al poseer una interfaz gráfica que permite seleccionar los parámetros de los modelos utilizados de una manera mucho más simple. Esto implica también que se requiere menor tiempo para implementar las técnicas utilizadas. Entre sus desventajas se encuentran la dependencia de Windows y el elevado costo de su licencia. Por otra parte, sólo se pueden utilizar las rutinas y funciones que vienen incorporadas, a diferencia de R que permite desarrollar nuevas funciones, mediante la programación de éstas.

4. APLICACIÓN TÉCNICAS DE DATA MINING

Tal como se menciona en el capítulo 1, el objetivo de esta tesis es identificar los patrones de comportamiento asociados a la utilización y venta de facturas falsas, extrapolando quiénes están evadiendo impuestos de esta forma, mediante la aplicación de técnicas de Data Mining. Para ello, en una primera instancia se utilizan herramientas de aprendizaje no supervisado, que permitan entender de mejor forma el comportamiento del universo de contribuyentes en relación a variables que implican un comportamiento irregular; y en una segunda instancia, herramientas de aprendizaje supervisado para aprender del comportamiento de los contribuyentes que han utilizado facturas falsas, y detectar quiénes podrían estar en esa condición, a partir de resultados de auditorías pasadas en las que el resultado final es conocido (fraude y/o no fraude).

En este capítulo se describe el procedimiento efectuado para seleccionar los datos que serán utilizados en el análisis, describiendo las fuentes de información y las transformaciones aplicadas, así como el tratamiento de los outliers y los casos nulos. Posteriormente se indican las variables seleccionadas para el modelamiento, presentando los experimentos realizados con cada una de las técnicas de Data Mining, y los resultados obtenidos en cada uno de ellos, para caracterizar a los contribuyentes y detectar quiénes son usuarios potenciales de facturas falsas.

4.1. DESCRIPCIÓN DE LOS DATOS

Para efecto de obtener patrones de comportamiento se considera como universo del estudio, a todos aquellos contribuyentes que hayan presentado al menos una declaración de IVA entre enero de 2005 y diciembre de 2007, es decir, empresas que se presenten algún tipo de movimiento en ese período, distinguiendo entre dos grupos:

- Micro y Pequeñas Empresas: Se considera como micro empresa a toda entidad que ejerce una actividad económica de forma regular, ya sea artesanal u otra, a título individual, familiar o como sociedad, y cuyas ventas anuales son inferiores a 2.400 UF. Las pequeñas empresas consideran ventas superiores a las 2.400 UF, pero inferiores a las 25.000 UF al año.

- Medianas y Grandes Empresas: Este segmento tiene una gran importancia por el volumen de ventas que tiene en relación al total de ventas en Chile. Las medianas empresas venden más de 25.000 UF al año pero menos que 100.000 UF y los grandes contribuyentes son aquellos que venden por sobre las 100.000 UF.

Se utiliza esta clasificación debido a que, según las entrevistas realizadas, existe la hipótesis que el comportamiento de evasión depende del tamaño de la empresa. A modo de ejemplo, las empresas grandes utilizan sistemas de contabilidad más complejos que podrían dificultar evadir a través del uso fraudulento de facturas, mientras que en las empresas pequeñas los sistemas de contabilidad son más informales. De igual forma, la magnitud de sus declaraciones y pagos de impuestos son muy diferentes. Por otra parte, las estrategias de fiscalización del Servicio se encuentran segmentadas por tipo de contribuyente, existiendo departamentos diferenciados que generan programas de fiscalización específicos en cada grupo, por lo que es útil para el SII conocer los patrones de comportamiento de cada segmento por separado.

La información relativa a los casos de facturas falsas históricos fue obtenida del Sistema de Control de Expedientes (SCE), administrado por el Departamento de Jurídica, el cual contiene información del resultado de las principales investigaciones de facturas falsas realizadas por el SII. Se selecciona el año 2006 como año de estudio, debido a que las auditorías consideran la revisión de hasta tres períodos hacia atrás, y por tanto, en el año 2010 todavía se están registrando en el sistema facturas falsas del año 2007, generando el universo señalado en el Cuadro Nº 4.

Cuadro Nº 4: Número de contribuyentes utilizados en el análisis

CONTRIBUYENTES DEL ANÁLISIS	MI y PE	ME y GR	TOTAL
Empresas activas en el período 2005-2007	558.319 (96%)	23.842 (4%)	582.161 (100%)
Empresas auditadas por facturas en 2006 con resultado de fraude o no fraude conocido	1.280 (76%)	412 (24%)	1.692 (100%)

La recopilación de información de casos con fraude y no fraude presentó algunos inconvenientes debido la forma en que ésta se almacena, pues se registra la fecha de inicio y término de la auditoría, los períodos tributarios revisados y el resultado obtenido, pero la información de los períodos en los que ocurren las diferencias no está automatizada. Por lo tanto, para saber si la factura falsa detectada correspondía al año 2006 específicamente, hubo que revisar las anotaciones y comentarios efectuados por el auditor y las rectificatorias efectuadas en códigos relacionados con facturas de ese año.

Debido a que existen empresas que se les revisó sus facturas del año 2006, en los que se detecta fraude en otros períodos cercanos y no necesariamente en ese mismo año, los casos de fraude y no fraude se categorizaron en tres tipos: “0” indica que el contribuyente fue auditado y no se encontraron facturas falsas en ninguno de los períodos revisados, “1” indica que el contribuyente no utilizó facturas falsas en el año de análisis pero sí en otros períodos revisados (normalmente el año anterior o siguiente), y “2” indica que el contribuyente utilizó facturas falsas en el año de estudio. A priori no existe claridad si los casos con categoría “1” tienen un comportamiento similar a la categoría “2” o a la categoría “0”.

Para construir el vector de características, se investigó cuáles eran los tipos de irregularidades y comportamientos comunes de los actores involucrados en la emisión y el uso de facturas falsas, así como los procedimientos de fiscalización utilizados para detectarlos, de manera de considerar el juicio experto de los fiscalizadores y el aprendizaje ya adquirido para detectar estos casos[29]. También se toma en consideración la información utilizada por otras Administraciones Tributarias para detectar fraude tributario mencionado en el capítulo 2. No obstante, debido a que no es posible representar todos los comportamientos irregulares con la información que maneja el Servicio, se descartan aquellas revisiones para las cuales no se encuentra la información disponible, ya sea porque sólo es posible detectarlo con una revisión en terreno o se requiera realizar un cruce con información de otras instituciones.

De esta forma, se seleccionaron 20 códigos del Formulario de Declaración y Pago Simultáneo de Impuestos (F29) relacionados con la operatoria del pago de IVA; 31 códigos del Formulario de Impuestos Anuales a la Renta (F22) asociados a la generación de la base imponible de primera categoría y datos contables de la empresa, y 31 ratios tributarios que relacionan la información de IVA y Renta y la rentabilidad de la empresa con su liquidez, entre otros. Adicionalmente se generan 14 variables relativas a características propias de los contribuyentes, 42 indicadores relacionados con su comportamiento histórico y su comportamiento en el año de análisis, 15 indicadores relacionados con las distintas etapas del ciclo del vida del contribuyente y 14 indicadores relacionados al comportamiento de sus relacionados, que pueden dar indicios de un buen o mal comportamiento tributario en el tiempo, como se muestra en el Cuadro Nº 5.

Cuadro Nº 5: Tipo de información utilizada para construir el vector de características[30]

CONCEPTO	# DATOS	TIPO DE INFORMACIÓN
Pago de Impuestos	82	Declaraciones de IVA (F29), Declaración de Renta (F22), Ratios Tributarios de IVA/Renta
Características Propias	14	Edad, Antigüedad Empresa, Cobertura, Facturador electrónico, Contabilidad computacional, Actividades económicas, Cambio sujeto, Declara por internet, Tiene domicilio y sucursales propias
Comportamiento Histórico y en el año de análisis	42	Fiscalizaciones selectivas, Delitos Previos, Problemas con el domicilio, Inconcurrencias, Denuncias y Clausuras, Pérdidas de RUT, Destrucción de documentos, Deuda regularizada, Pérdida de Facturas, Facturas observadas y/o bloqueos, Marcas Preventivas
Ciclo de Vida	15	Inicio de actividades, Verificación de actividades, Timbraje de documentos, Modificaciones de información, Términos de giro previos
Relacionados	14	Mandatarios, Representantes Legales, Socios, Familiares, Proveedores, Contadores, Sociedades y Representaciones (activos, antecedentes de delito, investigados, bloqueados)

Cabe señalar que las variables de comportamiento requirieron un mayor tiempo de preparación, ya que en la mayoría de los casos no se encontraban en forma directa en los sistemas. Por otra parte, se construyen variables numéricas que puedan ser posteriormente interpretadas.

[29] Se realizan entrevistas a profesionales del Depto. de Mediana y Grandes Empresas de la Subdirección de Fiscalización, profesionales del Depto. Fiscalización Selectiva y RIAC de la Dirección Regional Centro y profesionales del Depto. de Resoluciones, de Timbraje y de Operación IVA de la Dirección Regional Poniente.

[30] El detalle de los indicadores generados se encuentra en el Anexo E de este documento.

4.2. SELECCIÓN Y PROCESAMIENTO

La preparación de los datos es una parte fundamental del proceso KDD, ya que muchas veces éstos contienen errores en su almacenamiento que pueden llevar a la extracción de patrones y reglas poco útiles, de baja calidad, e incluso erróneos. Por otra parte, debido a limitaciones de tiempo y restricciones de procesamiento de los computadores, en la mayoría de los casos se debe limitar el número de variables a examinar, por lo que adquiere relevancia seleccionar aquellos atributos que sean los más representativos y expliquen mejor el comportamiento que se quiere detectar o caracterizar. A continuación se describe el procedimiento de limpieza, transformación y normalización utilizado para la selección de las variables de cada modelo.

4.2.1. LIMPIEZA DE DATOS

Para llevar a cabo la limpieza de datos, se detectan los casos inconsistentes y los outliers de cada grupo, los cuales son eliminados para no introducir ruido en la muestra. Particularmente esto adquiere relevancia en la información de impuestos, ya que en ocasiones la información contiene errores por el hecho de transcribir las declaraciones recibidas en papel a los sistemas.

- Datos outliers

Para la detección de los outliers, en la información de impuestos, se eliminaron aquellos casos que cumplen la condición:

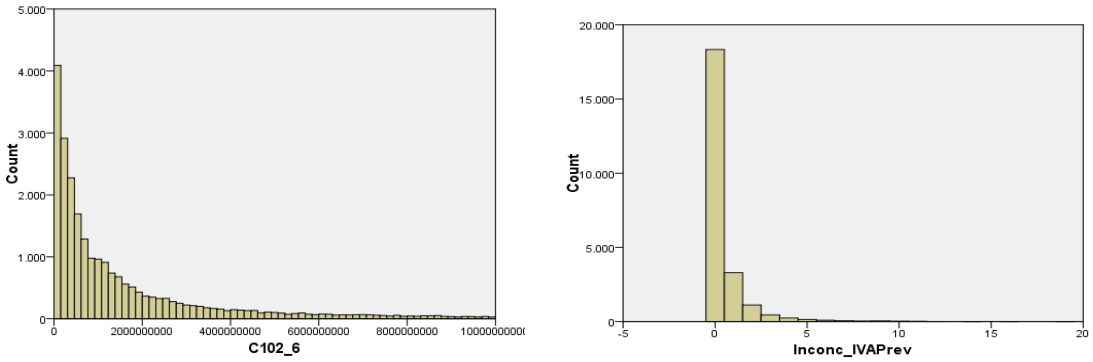
$$x_i > media(x / x > 0) + 5 * desvest(x / x > 0)$$

donde x_i es el valor de la suma en el código x, para el contribuyente i. Por lo tanto, se eliminan aquellos casos en que se supera la media más cinco veces la desviación estándar de alguna variable, considerando únicamente aquellos con valor positivo en ese código, lo que indica que el dato corresponde a un valor extremo.

En la mayoría de los casos, la distribución de estas variables es decreciente, debido a que un gran porcentaje de contribuyentes paga montos bajos, y sólo un pequeño porcentaje paga montos altos de impuestos, por lo que la eliminación de casos se hizo de manera cuidadosa, complementado con histogramas y diagramas de caja de cada variable, de manera de no eliminar casos que estuvieran correctos pero alejados del promedio, cómo se grafica en la Figura N° 30, caso (a).

En el caso de las variables de comportamiento, se observa que sólo un pequeño grupo posee algunas conductas irregulares, provocando que los promedios sean cercanos a cero o bajos en relación al máximo valor observado. Por lo tanto, al eliminar los casos con valores más altos, se elimina a aquellos contribuyentes con comportamiento irregular, los que son el grupo de interés del estudio, por lo que en esos casos se aplicó un criterio experto, consultándole a fiscalizadores si le parecía irregular los valores observados, como se grafica en la Figura N° 30, caso (b).

Figura Nº 30: Ejemplo de distribución decreciente de variables de impuestos y de comportamiento



(a) Código de impuestos tipo (b) Variable de comportamiento tipo

- Datos inconsistentes

En general las variables de comportamiento no tenían inconsistencias, debido a que fueron construidas en forma manual. Sin embargo, se presentaron algunas inconsistencias en los códigos del F29. Por ejemplo, existe información de boletas emitidas (cód.111) pero no se declara el débito generado por ese número de boletas (cód.110) o viceversa. Dado que estos casos no eran muchos, se determina eliminarlos de la base. El mismo criterio se utilizó para el resto de los códigos de débitos y créditos.

Luego de quitar los outliers y los datos considerados inconsistentes, el conjunto de datos final queda compuesto por 532.755 contribuyentes que son micro y pequeñas empresas y 22.609 contribuyentes que son medianas y grandes empresas, eliminando un 4.6% del primer grupo y un 3.4% del segundo grupo.

4.2.2. TRANSFORMACIÓN DE VARIABLES

Debido a que la declaración del pago de IVA se realiza mensualmente y la declaración del Impuesto a la Renta se realiza en forma anual, la primera transformación consiste en agrupar la información de IVA, de manera que sea comparable con la declaración de Renta. Para ello se genera la suma de los montos mensuales de cada código en el año[31].

Inicialmente se consideró también los promedios y desviación estándar de cada código en el año, pero posteriormente fueron descartados, ya que están muy correlacionados con las sumas y tienen menos variabilidad. Por otra parte, existía un estudio previo que indicaba que los promedios y desviaciones estándar no generaban un gran aporte a los modelos y que era mejor considerar los totales[32].

[31] El detalle de las variables agregadas de pago de impuesto generadas se encuentra en el Anexo E de este documento.

[32] "Segmentación de los contribuyentes que declaran IVA utilizando técnicas de Data Mining", Sandra Luckeheide (abril 2007)

Adicionalmente, se definen ratios tributarios que relacionen la información presentada en los Formularios de IVA y Renta, y ratios financiero-contable que relacionen la rentabilidad de la empresa y su liquidez, así como otros ratios normalmente utilizados para seleccionar casos de fiscalización como la razón débito/crédito. Por otra parte, se proponen algunos ratios como la razón entre facturas emitidas/facturas timbradas, que puede dar indicios de empresas que venden o compran facturas falsas, la desviación de la razón débito/crédito que puede indicar si una empresa vende o compra más de lo normal en algún período, y otros, que dan cuenta de la importancia que tienen las facturas en el respaldo de los créditos, como el credfact/créditototal, ya que una empresa ficticia que vende o utiliza facturas falsas, normalmente no realiza ventas a través de boletas[33].

Una tercera transformación se utiliza para la completitud de los datos nulos, colocando el valor cero en los códigos de impuestos, cuando el contribuyente presenta una declaración. Esto debido a que lo normal es que el contribuyente sólo complete los códigos con valor distinto de cero al momento de presentar la declaración. No ocurre lo mismo en el caso que el contribuyente no presenta la declaración, ya que no implica que el contribuyente no haya tenido movimiento nulo en el período.

En general, la información de IVA es más completa que la de renta, porque los códigos del reverso del formulario F22, sólo deben ser presentados por contribuyentes que llevan contabilidad completa. Por lo tanto, se utiliza información de débitos y créditos de IVA para completar datos de ingresos y costos del período, debido a la relación directa existente entre ambos. Para el resto de los campos de renta se utiliza la mediana del código para el mismo tramo de venta[34] como técnica de reemplazo.

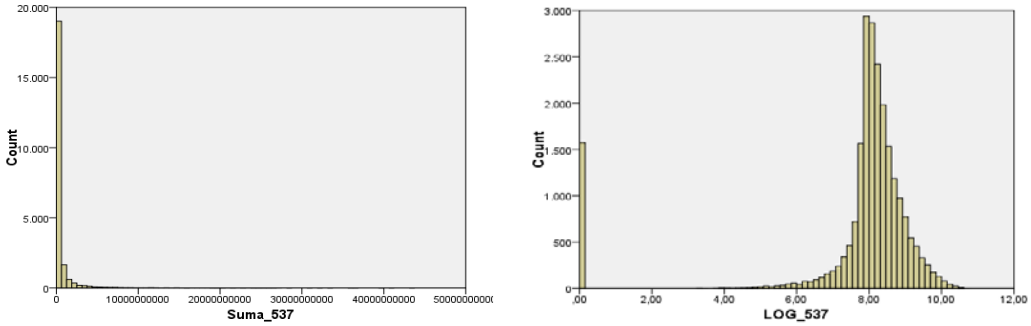
En el caso de variables de comportamiento, también se procede a reemplazar valores nulos por valor cero, aunque no son muchas las variables en esta situación. Principalmente hay problemas con las variables edad del contribuyente, antigüedad de la empresa, domicilio propio y meses desde el último timbraje, que posteriormente son agrupadas por análisis de componentes principales, por lo que se les debe asignar un valor que no afecte su resultado.

Finalmente, producto de la distribución decreciente de las variables de impuestos, fue necesario aplicar una transformación logarítmica de manera de disminuir el efecto de los casos extremos, como se muestra en la Figura N° 31. Esto debido a que al aplicar los modelos de clusterización con las variables originales, se tendía a formar un solo grupo donde estaban concentrados los valores bajos – medios de la variable. Ésta transformación fue utilizada para los códigos de IVA en el caso de las medianas y grandes empresas, y para los códigos de IVA y Renta en el caso de las micro y pequeñas empresas.

[33] El detalle de los ratios tributarios financieros – contables utilizados se encuentra en el Anexo E de este documento.

[34] El Servicio establece una clasificación de los contribuyentes de acuerdo a tramos de venta que indica dentro de un mismo segmento (micro, pequeñas, medianas, grandes) quiénes son empresas con monto de ventas bajo, intermedio y alto.

Figura Nº 31: Ejemplo de distribución original y transformada de códigos de impuestos



(a) Variable original x (b) Variable transformada Log(x+1)

Como se aprecia en la Figura N° 31, la transformación utilizada resalta los casos con valor cero diferenciándolos del resto aunque sean valores bajos. Por otra parte, esta transformación no se puede aplicar a variables con valores negativos, por lo que fue necesario eliminar del análisis aquellas variables que tenían valores menores a cero, como la renta líquida imponible.

4.2.3. NORMALIZACIÓN DE VARIABLES

Para evitar que las variables con un mayor rango de valores le quiten importancia a otras con un rango menor, se procede a normalizar todas las variables de manera que sean comparables una de la otra. Para ello se utiliza la normalización "Min-Max" en el rango [0,1], de acuerdo a:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

donde X' es el nuevo valor de la variable normalizada, X es el valor original de la variable , X_{min} es el mínimo valor posible de la variable, X_{max} es el máximo valor posible[35].

4.2.4. ANÁLISIS DE COMPONENTES PRINCIPALES

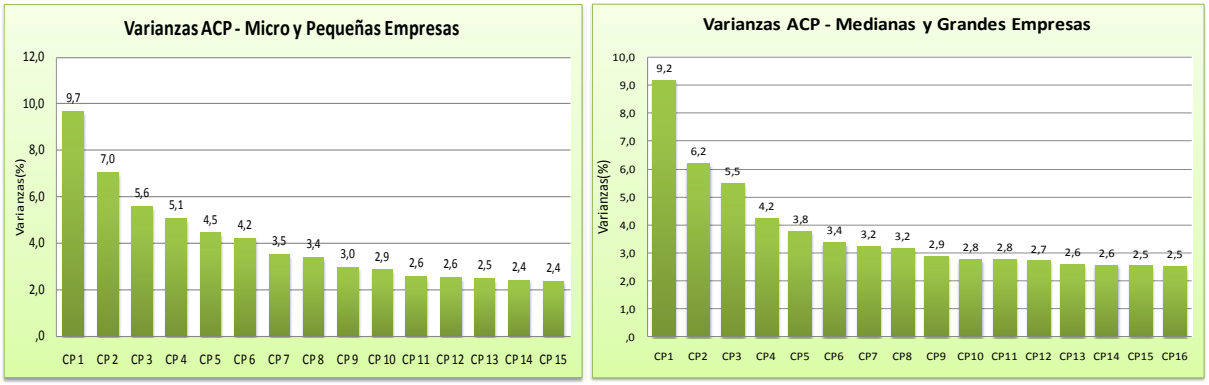
Debido a la gran cantidad de variables de comportamiento generadas, se procede a reducirlas a través del análisis de componentes principales (ACP). Esto permite construir indicadores de comportamiento del contribuyente que den señales de buenos o malas conductas, considerando un amplio esquema de características en el tiempo para cada segmento. Dado que nuestro interés es generar variables de comportamiento que estén relacionados al uso y venta de facturas falsas, previo a la aplicación de las componentes principales, se seleccionan aquellas variables que tienen una correlación mediana-alta con la variable de uso de Facturas Falsas en el período, seleccionando aquellas que tienen un sigma (bilateral) menor a 0,5 de los coeficientes de correlación de Pearson y eliminando aquellas con sigma (bilateral) mayor a 0,5[36].

[35] En este caso no se recomienda normalizar de acuerdo al promedio y la desviación estándar, ya que las distribuciones de las variables son decrecientes y no centradas en torno al promedio.

[36] El sigma bilateral indica la probabilidad de que el coeficiente de correlación lineal de Pearson sea cero. Mientras menor sea este sigma, mayor será la probabilidad de que las variables estén correlacionadas.

Como resultado se generan 15 componentes principales para el grupo de las micro y pequeñas empresas, que explican un 61,3% de la varianza de los datos. Del mismo modo, se generan 16 componentes principales para las medianas y grandes empresas, que explican un 59,9% de la varianza de los datos, como se indica en la Figura Nº 32.

Figura Nº 32: Varianzas del Análisis de Componentes Principales



Fuente: Elaboración propia

En el grupo de las micro y pequeñas empresas se aplica el análisis de componentes principales sobre un conjunto de 42 variables que presentan una relación con la variable que indica si existe fraude por facturas falsas en el año de estudio, generando 15 componentes principales como se muestra en el Cuadro N° 6.

Cuadro Nº 6: Conceptos asociados a cada componente principal grupo micro y pequeñas empresas

NOMBRE COMPONENTE PRINCIPAL Y CONCEPTO ASOCIADO	VARIABLES QUE LA COMPONEN[37]
ACP1 Nivel de facturas timbradas en los últimos años	Fact0406, Factmax0406, Factprom0406, Facturas06
ACP2 Delitos e irregularidades de facturas previos	FactObsHist, FactObsRcte, DelitoPrev
ACP3 Fiscalizaciones previas con resultado positivo	FiscSelecPos, RendTotal, FiscSelecPrev, NotifFiscPrev
ACP4 Frecuencia de Timbraje	FrecTim06, FrecProm, MesesUltimb
ACP5 Participación en otras empresas	NRepresentac, NSociedades, RelacionadoBloq
ACP6 Problemas de localización	InconTotPrev, GirosPrev, InconTot06, NoubicadoRcte
ACP7 Antigüedad	Edad, Antig, Espropia
ACP8 Clausuras y denuncios históricos	ClausurasHist, DenunciosHist
ACP9 Cobertura de la empresa	Nsucursales, Ncomunas, ActecoDF06
ACP10 Fiscalizaciones previas con resultado negativo	FiscSelecNeg
ACP11 Verificaciones de actividad	VIANegativa, VerificaAct
ACP12 Delitos de relacionados indirectos	DelitoMand, DelitoConta
ACP13 Irregularidades previas (pérdida facturas)	PerdFactHIst, NoubicadoHist
ACP14 Nivel de formalidad de la empresa	FactElect, ContabCompleta
ACP15 Delitos de relacionados directos	DelitosRL, DelitosFam, Cambiosujeto, AlertaNDF06

Fuente: Elaboración propia

De éstas, la que aporta un mayor nivel de varianza se relaciona con el nivel de facturas timbradas en los últimos dos años (ACP1) que está compuesta por la cantidad total de facturas timbradas en

[37] La descripción de cada variable se encuentra en el Anexo E de este documento.

los dos últimos años, el promedio de facturas timbradas en ese mismo período, el máximo de facturas timbradas en ese período y la cantidad de facturas timbradas en el año de análisis, aportando un 9,7% de la varianza. En segundo lugar se encuentran los delitos e irregularidades de facturas previas (ACP2), compuesta por la cantidad de veces que registra antecedentes de delito asociado a facturas en el pasado, la cantidad de veces que registra anotación de facturas observadas recientemente e históricamente, aportando un 7,0% de la varianza.

En el grupo de las medianas y grandes empresas se aplica el análisis de componentes principales sobre un conjunto de 40 variables que presentan una relación con la variable que indica si existe fraude por facturas falsas en el año de estudio, generando 16 componentes principales como se muestra en el Cuadro N° 7.

Cuadro Nº 7: Conceptos asociados a cada componente principal grupo medianas y grandes empresas

NOMBRE COMPONENTE PRINCIPAL Y CONCEPTO ASOCIADO	VARIABLES QUE LA COMPONEN[38]
ACP1 Cobertura de la empresa	Ncomunas, Nsucursales, Nregiones
ACP2 Fiscalizaciones previas	FiscSelecPrev, FiscSelecNeg, FiscSelecPos, NotifFiscPrev
ACP3 Nº Actividades económicas	Actecoact, ActecoDF06, DestruccDoc
ACP4 Nivel de formalidad de la empresa y antigüedad	DeclaInternet, ContabCompleta, Antig
ACP5 Clausuras y denuncios históricos	ClausuraHist, DenunHist
ACP6 Verificaciones de actividad	VIANegativa, VerificaAct
ACP7 Giros e inconcurrencias	GirosPrev, InconIVA06, InconTotPrev
ACP8 Representantes legales	NReprInactivo, NRepActivo
ACP9 Delitos de los relacionados	DelitoMandPrev, DelitoRLPrev, DelitoContaPrev
ACP10 Irregularidades de facturas y nivel de timbraje	IRRPrev, Factprom0406
ACP11 Rendimiento de fiscalizaciones previas	RendTotal, Mesultimb
ACP12 Irregularidades recientes	InconNotif06, NoubicadoRcte, FactObsRcte
ACP13 Cambio de sujeto	AlertaNDF06, Cambiosujeto06
ACP14 Antecedentes de término de giro y no ubicado	NoubicadoHist, Tgiroprev, MarcaPrevent
ACP15 Antecedentes de timbraje restringido	TimbrajeRestr
ACP16 Regularización de deudas y pérdidas de rut.	DeudaRegul, PerdidaRut

Fuente: Elaboración propia

En este caso, la componente principal que aporta un mayor nivel de varianza se relaciona con la cobertura de la empresa (ACP1) que está compuesta por el número de comunas, sucursales y regiones en las cuales tiene locales el contribuyente, aportando un 9,2% de la varianza.

En segundo lugar, se encuentra un indicador de las fiscalizaciones previas (ACP2), que está compuesta por la cantidad de fiscalizaciones selectivas previas, la cantidad de fiscalizaciones selectivas con resultado positivo y negativo, y la cantidad de anotaciones que indican notificaciones de fiscalización previas, aportando un 6,2% de la varianza.

Las componentes principales obtenidas indican que existen variables comunes que pueden estar relacionadas al uso de facturas falsas en ambos grupos, como los delitos de los relacionados, la

[38] La descripción de cada variable se encuentra en el Anexo E de este documento.

antigüedad, la cobertura de la empresa, el nivel de formalización de la contabilidad, los antecedentes de verificación de actividades, la frecuencia del timbraje, el nivel de fiscalizaciones previas, las clausuras y denuncios históricos y las irregularidades previas asociadas a las facturas.

Sin embargo, como se esperaba, también se observan diferencias. En el caso de las micro y pequeñas empresas, se incluye la participación en otras empresas, los delitos de los familiares y las facturas timbradas en el último período, mientras que en las medianas y grandes empresas aparecen variables relacionadas a las actividades económicas de la empresa, antecedentes de timbraje restringido, de término de giro, de regularización de deudas y pérdida de RUT.

4.2.5. SELECCIÓN DE VARIABLES

Dado que interesaba generar variables de comportamiento relacionadas al uso y venta de facturas falsas y no a otros comportamientos, se seleccionan sólo aquellas variables que tienen una correlación mediana-alta con la variable de uso de facturas falsas en el año 2006 (FF06), eliminando aquellas que tienen más de un 10% de probabilidad que el coeficiente de Pearson sea cero, exceptuando algunos códigos de interés como el total de débitos, total de créditos y pago de IVA, entre otros. Igualmente se descartan aquellas variables que tienen un gran porcentaje de valores nulos como Débboletas/boletas, Crédactfijo/Factactfijo, Debnotascréd/Notasdecréd, y la variable Edad del contribuyente en el caso de las medianas y grandes empresas.

De esta forma, se seleccionan 42 variables en el segmento micro y pequeñas y 36 variables en el segmento medianas y grandes para el análisis. En el primer grupo, un 35% de las variables corresponde a códigos de la declaración de IVA, un 35% a códigos relacionados con renta y un 30% a variables relacionadas al comportamiento, las que se presentan en el Cuadro Nº 8.

Cuadro Nº 8: Correlaciones de variables seleccionadas con FF06, Grupo Micro y Pequeñas

Variable	Pearson Correlation	Sig. (2-tailed)	Variable	Pearson Correlation	Sig. (2-tailed)	Variable	Pearson Correlation	Sig. (2-tailed)
LOG_CFTOT	,496**	,000	ACP10	-,129**	,000	LOG_GTORECHACT	,099**	,003
LOG_REMCRED	-,469**	,000	ACP4	,148**	,000	LOG_INGCOST	-,102**	,003
LOG_SUMA89	,398**	,000	LOG_129	,106**	,000	LOG_101	,084**	,003
LOG_IVAFACT	,372**	,000	ACP6	-,148**	,000	LOG_OTGTOACT	,095**	,005
LOG_SUMA91	,307**	,000	LOG_SUMA509	,100**	,000	LOG_DFACT	-,077**	,007
LOG_SUMA503	,302**	,000	LOG_ACTPAS	,121**	,000	ACP14	-,072*	,022
LOG_INGACT	,307**	,000	ACP5	-,112**	,000	LOG_CAPEFACT	-,069*	,037
LOG_SUMA502	,244**	,000	LOG_SUMA111	,100**	,000	LOG_SUMA537	-,057*	,041
LOG_CTOACT	,280**	,000	ACP3	-,110**	,000	ACP2	0,06	,051
LOG_DEBCRED	,238**	,000	LOG_628	,097**	,001	ACP15	-0,05	,097
LOG_CFACT	,223**	,000	ACP7	-,164**	,001	LOG_122	0,04	,115
LOG_SUMA520	,191**	,000	LOG_FEMTIM	,095**	,001	ACP11	0,05	,125
LOG_SUMA538	,190**	,000	LOG_REMUACT	,113**	,001	LOG_102	0,04	,129
LOG_630	,153**	,000	ACP13	-,102**	,001	ACP12	0,05	,144
ACP1	,130**	,000	LOG_639	,087**	,002	LOG_DESVDC	0,03	,300

Notas: ** La correlación es significativa al nivel 0,01 (bilateral), * La correlación es significativa al nivel 0,05 (bilateral)

Asimismo, de las 36 variables seleccionadas en el segundo grupo, un 31% corresponde a códigos relacionados con el IVA, un 38% a códigos relacionados con Renta y un 31% a variables

relacionadas con el comportamiento, con mayor preponderancia de variables relacionadas con la renta, como se muestra en el Cuadro Nº 9.

Al analizar las correlaciones, se observa que las más relacionadas con la variable FF06, son las mismas que se plasman en el grupo de micro y pequeñas empresas, correspondientes al porcentaje que representa el crédito por facturas respecto de crédito total y la relación entre el remanente de los créditos y el crédito total promedio. Por otra parte, se observa mayor relevancia de variables de comportamiento, como el número de representantes legales activos e inactivos de la empresa y el nivel de fiscalizaciones previas, y mayor relevancia de variables de renta como activos, y pasivos, entre otros.

Cuadro Nº 9: Correlaciones de Variables seleccionadas con FF06, Grupo Medianas y Grandes

Variable	Pearson Correlation	Sig. (2-tailed)	Variable	Pearson Correlation	Sig. (2-tailed)	Variable	Pearson Correlation	Sig. (2-tailed)
Rem_Cred	-,408**	,000	C779_6	-,195**	,000	LOG_91_1	-,102*	,037
CredFact_CredTot	,384**	,000	ACP4_1	-,191**	,000	DebFact_DebTot	,097*	,049
ACP8_1	-,347**	,000	Rem_Activos	,186**	,000	Ing_Costos	-,098	,051
ACP2_1	-,316**	,000	Ingl_Activos	,183**	,000	C646_6	-,092	,063
C123_6	-,314**	,000	CostoI_Activos	,170**	,001	ACP14_1	,088	,074
LOG_89	,279**	,000	ACP10_1	,152**	,002	LOG_111	-,078	,206
C122_6	-,279**	,000	DevExp_Cred	-,150**	,002	ACP5_1	-,062	,206
C634_6	-,271**	,000	ACP9_1	,149**	,002	Factem_Facttim	,060	,224
LOG_504	-,264**	,000	Deb_Cred	,147**	,003	ACP15_1	-,053	,279
LOG_525	-,240**	,000	ACP7_1	-,138**	,005	C630_6	,049	,325
LOG_537	-,237**	,000	Suma_538	-,115*	,020	LOG_519	-,040	,414
ACP12_1	,209**	,000	ACP1_1	-,110*	,025	IVA_Ing	-,024	,623
OtGastos_Activos	,213**	,000	C304_6	-,104*	,035	LOG_509	-,008	,871

Notas: ** La correlación es significativa al nivel 0,01 (bilateral), * La correlación es significativa al nivel 0,05 (bilateral)

4.2.6. MEDIDA DE DISTANCIA

La medida de distancia que se utilizará en la aplicación de los algoritmos de clustering en el vector de características generado, es la distancia Euclidiana, que por ser la más comúnmente utilizada, viene por defecto en la mayoría de los algoritmos en R (como en gran parte de las herramientas de data mining).

4.3. MODELAMIENTO

Para efectos de caracterización e identificación de patrones, en una primera instancia se aplican las técnicas de data mining al universo de empresas, con el objetivo de identificar relaciones entre su pago de impuestos (IVA y Renta) y variables de comportamiento asociadas a la utilización de facturas falsas. Posteriormente se aplican técnicas de clasificación para aquellos casos en los que la condición de fraude y no fraude es conocida, de manera de identificar patrones específicos de este conjunto de contribuyentes. Finalmente se aplican herramientas de clasificación para detectar casos de fraude y no fraude con la información generada.

4.3.1. CARACTERIZACIÓN DEL UNIVERSO DE EMPRESAS

Para caracterizar al universo de empresas se aplican técnicas de aprendizaje no supervisado como el SOM y el Gas Neuronal. Estas técnicas permiten detectar relaciones, enseñando al algoritmo a descubrir por sí mismo las correlaciones y similitudes entre los patrones de entrada del conjunto de datos, de manera de agruparlos en diferentes categorías.

La hipótesis de trabajo supone que al considerar sólo las variables de comportamiento relacionadas al uso de facturas falsas combinadas con variables de impuestos, es posible detectar grupos de contribuyentes que tienen un buen o mal comportamiento tributario y conocer cómo realizan su pago de impuesto.

Debido a que el software utilizado (R) solicita no dejar valores nulos en la matriz, se generaron algunos inconvenientes al momento de trabajar con los ratios tributarios cuando el denominador era cero, impactando principalmente al grupo de micro y pequeñas empresas que tiene más casos y variables con valor cero en sus códigos de IVA y Renta. En el caso de las medianas y grandes empresas, la mayoría de ellas contiene información de todos sus ratios, por lo que sólo se eliminan los casos en los que se tiene un valor nulo en alguno de sus campos.

En una primera etapa se aplica el método SOM para identificar clusters o grupos de empresas de comportamiento similar. Para ello se utiliza el paquete “Som” de R, considerando una topología de red rectangular, con 3 neuronas de entrada y 24x24 neuronas de salida en el grupo de las micro y pequeñas empresas y 36x36 neuronas de salida en el grupo de las medianas y grandes empresas, con un número máximo de 100 iteraciones. Posteriormente se aplica el Gas Neuronal, considerando el mismo número de clusters que el Mapa de Kohonen, utilizando el paquete “Clust” de R, el cual genera una matriz con las características de los centroides de cada variable y un vector de clasificación que señala el grupo al que pertenece cada contribuyente.

Debido a la cantidad de variables consideradas en el análisis y al tamaño del universo de micro y pequeñas empresas, los experimentos de ese grupo se realizan en base a muestras aleatorias de 50.000 empresas[39] manteniendo todos los casos conocidos de FF06, debido a que constituyen un porcentaje pequeño del universo.

En el Cuadro Nº 10 se presentan los experimentos realizados para la caracterización del universo de empresas utilizada en el estudio y a continuación se presentan los resultados obtenidos en cada uno de ellos.

Cuadro Nº 10: Experimentos para caracterizar al universo de empresas

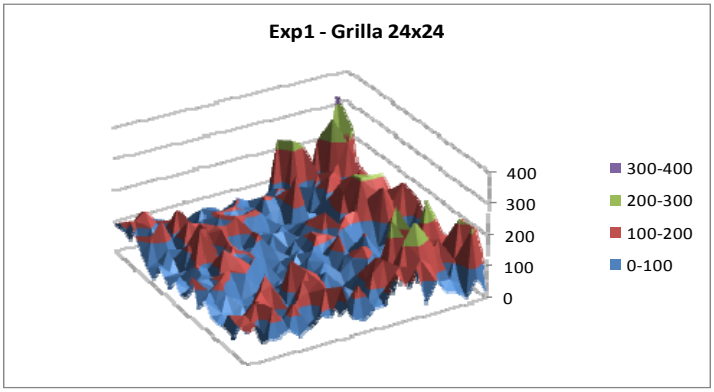
EXP.Nº	TÉCNICA DM	GRUPO OBJETIVO	Nº CONTRIBUYENTES	Nº CASOS CON FF	Nº VARIABLES
1	SOM	MI y PE	50.000	705	30
2	SOM	ME y GR	18.671	245	24
3	Gas Neuronal	MI y PE	50.000	705	30
4	Gas Neuronal	ME y GR	18.671	245	24

[39] Este tamaño de muestra fue seleccionado debido a las restricciones de procesamiento del programa R.

Experimento 1: SOM – Micro y Pequeñas Empresas – Vector de Características compuesto por 30 variables (considera ratios tributarios de IVA y Renta)

Este experimento contempla como universo a los contribuyentes que tienen valor no nulo en sus ratios tributarios de IVA y Renta, el cual representa aproximadamente un 40% del total de micro y pequeñas empresas. Como resultado se identifican 6 clusters principales, cuyos peaks representan grupos que en términos de comportamiento no presentan grandes diferencias; sin embargo, varían en la composición de su pago de impuesto, en aspectos tales como el nivel de existencias declarado, el monto de débitos con boletas, uso de facturas e IVA determinado.

Figura Nº 33: SOM de 24 x 24 rectangular – muestra experimento 1



Fuente: Elaboración propia

El Cuadro N° 11 indica los promedios del vector de características de cada cluster, con los cuales se pueden identificar las características de cada uno de ellos, como se presenta a continuación:

- **Cluster 1 (5,8%):** Tiene un nivel de actividad bajo (no genera IVA, no emite boletas, tiene ingresos bajos, el mayor nivel de remanentes y menos créditos por facturas recibidas). En términos de comportamiento, tiene la menor frecuencia de timbraje, delitos e irregularidades previas.
- **Cluster 2 (10,9%):** También tiene niveles bajos de movimiento pero declara niveles de existencia altos; tiene menos facturas emitidas, pero la mayor cantidad de créditos y crédito promedio por factura. En términos de comportamiento, es el que tiene mayor cantidad de irregularidades y delitos previos asociados a facturas, mayor cantidad de fiscalizaciones previas con resultado positivo y negativo, y delitos de relacionados directos.
- **Cluster 3 (17,8%):** Tiene niveles intermedios de uso de facturas y pagos intermedio-altos de IVA (no emite boletas y tiene nivel bajo de remanentes). En cuanto a su comportamiento tiene la menor cantidad de fiscalizaciones positivas, irregularidades previas por facturas y problemas de localización, además tiene la mayor cantidad de verificaciones de actividad.
- **Cluster 4 (24,2%):** Tiene niveles altos de movimiento producto del uso de boletas, tiene uso intermedio de facturas, el mayor valor de crédito promedio por factura y los mayores costos. En relación a su comportamiento tiene la mayor cantidad de delitos e irregularidades previas y valor intermedio del resto de los indicadores.

- **Cluster 5 (25,8%):** Tiene nivel intermedio de movimiento (pago intermedio de IVA, débitos por facturas intermedio, no emite boletas), pero tiene la menor cantidad de créditos y créditos promedio por facturas, tiene además el mayor valor de débitos/créditos en términos promedio y de desviación. En cuanto al comportamiento tiene valor intermedio de sus indicadores.
- **Cluster 6 (15,5%):** Tiene el nivel de actividad más alto (mayor pago de IVA, mayor cantidad de débitos y créditos asociados a facturas, no emite boletas). En términos de comportamiento presenta irregularidades previas, problemas de localización y delitos de los relacionados indirectos, tales como mandatarios y contadores.

En términos generales, los cluster se diferencian principalmente por el nivel de movimiento de IVA, donde los cluster 1 y 2 tienen bajo movimiento, los cluster 3 y 6 tienen un nivel intermedio, y los cluster 4 y 6 tienen un movimiento alto de IVA, ya sea por boletas o por combinación de boletas y facturas. En cuanto a su comportamiento, los cluster 2 y 4 son los que tienen un comportamiento más irregular, representando el 35% de los contribuyentes, mientras que los cluster 1 y 3 tienen el mejor comportamiento tributario, representando el 24% de los contribuyentes. El resto (41%), tiene un comportamiento más bien intermedio, con algunos problemas de localización.

Cuadro Nº 11: Centroides de cada cluster – muestra experimento 1

Cluster	FF01	LOG_129	LOG_INGACT	LOG_CTOACT	LOG_SUMA503	LOG_SUMA111	LOG_SUMA538	LOG_SUMA520	LOG_SUMA537	LOG_SUMA89
1	0,000	0,000	0,066	0,016	0,240	0,000	0,740	0,662	0,768	0,000
2	0,001	0,768	0,077	0,042	0,195	0,000	0,754	0,759	0,791	0,000
3	0,000	0,000	0,203	0,071	0,242	0,000	0,790	0,746	0,664	0,738
4	0,000	0,783	0,238	0,172	0,380	0,844	0,857	0,829	0,734	0,755
5	0,000	0,001	0,196	0,058	0,365	0,000	0,779	0,706	0,626	0,758
6	0,000	0,667	0,179	0,068	0,585	0,000	0,899	0,834	0,742	0,865

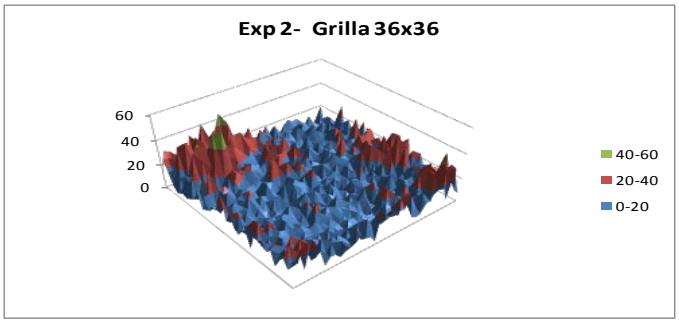
Cluster	LOG_SUMA91	LOG_CFACT	LOG_CFTOT	LOG_DEBCRED	LOG_REMCRED	LOG_IVAFACT	LOG_DESVDC	LOG_IVAING	LOG_PIVAING	LOG_FEMTIM
1	0,644	0,569	0,033	0,020	0,330	0,000	0,014	0,000	0,029	0,107
2	0,664	0,636	0,098	0,022	0,328	0,000	0,017	0,000	0,034	0,083
3	0,743	0,624	0,464	0,201	0,037	0,650	0,107	0,062	0,061	0,110
4	0,774	0,581	0,490	0,168	0,012	0,518	0,023	0,027	0,032	0,085
5	0,764	0,543	0,486	0,289	0,014	0,638	0,168	0,106	0,100	0,099
6	0,874	0,610	0,467	0,250	0,017	0,680	0,092	0,094	0,097	0,093

Cluster	ACP1	ACP2	ACP3	ACP4	ACP6	ACP10	ACP11	ACP12	ACP13	ACP15
1	0,012	0,000	0,004	0,363	0,142	0,036	0,044	0,001	0,009	0,487
2	0,012	0,001	0,009	0,365	0,159	0,041	0,077	0,000	0,023	0,505
3	0,015	0,000	0,001	0,381	0,104	0,032	0,085	0,002	0,006	0,497
4	0,022	0,003	0,006	0,370	0,167	0,036	0,029	0,002	0,012	0,486
5	0,023	0,000	0,002	0,386	0,136	0,032	0,033	0,001	0,010	0,485
6	0,066	0,002	0,005	0,417	0,192	0,035	0,028	0,003	0,016	0,487

Experimento 2: SOM – Medianas y Grandes Empresas – Vector de Características compuesto por 24 variables (considera ratios tributarios de IVA y Renta)

Este experimento contempla como universo a los contribuyentes que tienen valor no nulo en sus ratios tributarios de IVA y Renta, considerando un total de 24 variables relacionadas con el uso de facturas falsas. Como se aprecia en la Figura Nº 34, más que peaks se identifican 5 zonas de mayor densidad. Las principales diferencias se producen por el nivel de costos, la cantidad de boletas emitidas, el porcentaje de los débitos correspondiente a la emisión de facturas, las notas de crédito emitidas por facturas, y en menor medida por el nivel de cobertura de la empresa, la cantidad de fiscalizaciones previas y las irregularidades de facturas y nivel de timbraje promedio.

Figura Nº 34: SOM de 36 x 36 rectangular – muestra experimento 2



Fuente: Elaboración propia

El Cuadro N° 12 indica los promedios del vector de características de cada cluster, con los cuales se pueden identificar las características de cada uno de ellos:

Cuadro Nº 12: Centroides de cada cluster – muestra experimento 2

Cluster	FF01	C630	C122	C779	C123	ING_COSTOS	ING_ACTIVOS	COSTO_ACTIVOS
1	1,000	0,762	0,070	0,006	0,063	0,000	0,008	0,014
2	1,000	0,503	0,027	0,002	0,023	0,000	0,030	0,027
3	1,000	0,431	0,041	0,009	0,036	0,000	0,020	0,016
4	1,000	0,130	0,046	0,003	0,036	0,000	0,008	0,005
5	1,000	0,457	0,060	0,003	0,058	0,000	0,013	0,011

Cluster	REM_ACTIVOS	OTGASTOS_ACTIVOS	DF_DTOTAL	CF_CTOT	REM_CRED	LOG_519	LOG_509	LOG_111
1	0,014	0,010	0,354	0,498	0,013	0,410	0,282	0,000
2	0,038	0,027	0,032	0,512	0,012	0,446	0,012	0,819
3	0,036	0,026	0,315	0,517	0,010	0,422	0,340	0,714
4	0,023	0,018	0,347	0,504	0,020	0,393	0,007	0,000
5	0,018	0,014	0,357	0,508	0,014	0,416	0,404	0,000

Cluster	LOG_DEBCRED	ACP1	ACP2	ACP8	ACP10	ACP11	ACP12	ACP13
1	0,113	0,007	0,111	0,241	0,193	0,258	0,010	0,235
2	0,101	0,012	0,100	0,216	0,196	0,263	0,016	0,213
3	0,108	0,013	0,096	0,253	0,203	0,255	0,015	0,220
4	0,112	0,003	0,062	0,251	0,191	0,253	0,022	0,225
5	0,117	0,011	0,084	0,255	0,198	0,253	0,014	0,225

Las características principales de cada cluster son:

- **Cluster 1 (30,9%):** Presenta el nivel más alto de costos, activos y pasivos, niveles intermedios de movimientos de IVA y no presenta boletas. En términos de comportamiento tiene baja cobertura, menor cantidad de irregularidades recientes y mayor cantidad de fiscalizaciones previas. Además tiene mayor cantidad de actecos de cambio de sujeto y de difícil fiscalización.
- **Cluster 2 (14,6%):** Presenta el nivel más bajo de activos, pasivos y cuentas por pagar a empresas relacionadas y la relación más alta de ingresos y costos respecto de sus activos. Tiene nivel alto de movimiento de IVA con la mayor cantidad de débitos por boletas y cantidad de facturas recibidas. En términos de comportamiento, tiene el valor más alto de rendimiento por fiscalizaciones previas, nivel de cobertura alto.
- **Cluster 3 (11,1%):** Registra el valor más alto de cuentas por pagar a empresas relacionadas. En cuanto a IVA tiene nivel intermedio-alto de movimiento por uso de boletas, además tiene el mayor valor de crédito promedio por factura y la menor cantidad de remanente. Respecto de su comportamiento, tiene la cobertura más alta y la mayor cantidad de irregularidades previas por facturas y nivel de timbraje.

- **Cluster 4 (16,9%):** Registra los costos más bajos y de ingresos respecto de sus activos. Presenta la mayor cantidad de remanentes y la menor cantidad de notas de crédito por facturas y de facturas de compra recibidas, no emite boletas. En relación a su comportamiento, tiene la menor cobertura y cantidad de fiscalizaciones previas y el más alto de irregularidades recientes y menor nivel de timbraje e irregularidades por facturas.
- **Cluster 5 (27,8%):** Presenta niveles intermedios de costos, activos y pasivos. Tiene la mayor cantidad de notas de crédito por facturas y nivel alto de cantidad de facturas recibidas, además, tiene la mayor razón entre débitos y créditos y porcentaje de débitos asociados a facturas. En cuanto al comportamiento tiene la mayor cantidad de representantes legales y nivel de cobertura intermedia, junto a los otros indicadores.

En términos generales, el cluster 1 es el que ha sido más fiscalizado, presentando además un comportamiento tributario positivo. El cluster 2 es el que históricamente ha generado más rendimiento, probablemente por su nivel de cobertura alto y movimiento de impuestos. El cluster 3 tiene la mayor cobertura, nivel de timbraje e irregularidades asociadas a facturas. El cluster 4 tiene la menor cobertura y cantidad de fiscalizaciones previas con irregularidades recientes. El cluster 5 tiene en general un comportamiento intermedio.

Experimento 3: Gas Neuronal – Micro y Pequeñas Empresas – Vector de Características compuesto por 30 variables (considera ratios tributarios de IVA y Renta)

Para comparar los resultados obtenidos con el SOM, se utiliza la misma cantidad de clusters que los obtenidos con esa técnica en el Gas Neuronal, identificando las características de cada uno de ellos. En particular este experimento considera sólo los contribuyentes que tienen información de sus ratios de IVA y renta no nulos, generándose 6 cluster. Las características principales de cada grupo se describen en el Cuadro Nº 12.

Cuadro Nº 13: Centroides de cada cluster – muestra experimento 3

Centro	FF01	LOG_129	LOG_INGACT	LOG_CTOACT	LOG_SUMA503	LOG_SUMA111	LOG_SUMA538	LOG_SUMA520	LOG_SUMA537	LOG_SUMA89
1	0,018	0,000	0,216	0,101	0,312	0,000	0,810	0,736	0,668	0,759
2	0,008	0,062	0,229	0,125	0,335	0,698	0,792	0,738	0,663	0,729
3	0,021	0,768	0,296	0,204	0,528	0,815	0,895	0,857	0,763	0,804
4	0,003	0,316	0,134	0,085	0,216	0,248	0,712	0,723	0,741	0,012
5	0,000	0,675	0,232	0,154	0,240	0,740	0,780	0,744	0,666	0,694
6	0,010	0,728	0,225	0,136	0,394	0,001	0,843	0,795	0,716	0,763

Centro	LOG_SUMA91	LOG_CFACT	LOG_CFTOT	LOG_DEBCRED	LOG_REMCRED	LOG_IVAFACT	LOG_DESVDC	LOG_IVAING	LOG_PIVAING	LOG_FEMTIM
1	0,775	0,598	0,410	0,263	0,078	0,662	0,180	0,079	0,079	0,098
2	0,749	0,572	0,437	0,221	0,058	0,592	0,115	0,068	0,071	0,094
3	0,824	0,622	0,463	0,178	0,028	0,589	0,047	0,042	0,046	0,106
4	0,599	0,606	0,144	0,038	0,298	0,010	0,019	0,000	0,029	0,091
5	0,718	0,560	0,453	0,179	0,046	0,528	0,070	0,052	0,060	0,081
6	0,784	0,622	0,424	0,192	0,066	0,616	0,104	0,053	0,058	0,092

Centro	ACP1	ACP2	ACP3	ACP4	ACP6	ACP10	ACP11	ACP12	ACP13	ACP15
1	0,023	0,000	0,006	0,389	0,166	0,036	0,054	0,001	0,016	0,490
2	0,031	0,000	0,010	0,385	0,173	0,039	0,043	0,002	0,021	0,486
3	0,076	0,003	0,010	0,425	0,186	0,041	0,044	0,001	0,019	0,488
4	0,014	0,000	0,008	0,365	0,144	0,039	0,061	0,001	0,015	0,493
5	0,015	0,000	0,004	0,354	0,157	0,034	0,041	0,000	0,016	0,487
6	0,038	0,002	0,009	0,398	0,176	0,039	0,053	0,001	0,021	0,491

Las características de cada cluster son:

- **Cluster 1 (34,4%):** Tiene el mayor nivel de pago de IVA respecto de sus ingresos, mayor valor del ratio entre débitos y créditos, con bajo nivel de boletas emitidas y el menor nivel de existencias. En cuanto a su comportamiento tiene bajo número de fiscalizaciones previas con resultado negativo y niveles intermedios de los otros comportamientos.
- **Cluster 2 (13,0%):** Similar al anterior, pero con mayor cantidad de débitos con boletas y menor cantidad de créditos, Presenta la mayor cantidad de delitos de los relacionados indirectos y mayor cantidad de fiscalizaciones previas con resultado negativo.
- **Cluster 3 (15,2%):** Tiene la mayor cantidad de existencias, nivel de ingresos y costos respecto a sus activos, emite más facturas y boletas, registra más créditos por facturas emitidas y mayor pago de IVA. En relación a su comportamiento, tiene la mayor cantidad y frecuencia de facturas timbradas, y mayor ratio entre facturas emitidas y timbradas. Además presenta la mayor cantidad de problemas de localización, delitos e irregularidades previas de facturas y fiscalizaciones con resultado positivo y negativo.
- **Cluster 4 (9,4%):** Registra el menor pago de IVA, con nivel intermedio de uso de boletas y de créditos, con el menor ratio entre débitos y créditos e IVA por factura emitida. Presenta bajo nivel de timbraje, menos problemas de localización, mayores verificaciones de actividad en el domicilio y menos irregularidades previas.
- **Cluster 5 (15,4%):** Tiene nivel intermedio de pago de IVA y del resto de los comportamientos. Presenta la menor frecuencia de timbraje, menor cantidad de fiscalizaciones previas con resultado positivo y negativo, menor cantidad de verificaciones de actividad y delitos de los relacionados indirectos.
- **Cluster 6 (12,7%):** Presenta nivel intermedio de pago de IVA y del resto de los comportamientos, con poca emisión de boletas y alto número de existencias. Tiene algunas irregularidades previas y problemas de localización.

De acuerdo a lo anterior, los clusters 2 y 4 son los que tienen un mejor comportamiento tributario, mientras que los clusters 3 y 6 tienen más irregularidades. Al analizar la distribución de los casos conocidos de fraude por facturas falsas en ese año, se observa además que los clusters 2 y 4 tienen preponderancia de casos con buen comportamiento, mientras que el cluster 1 tiene preponderancia de casos con fraude en período cercano, y los clusters 3 y 6 tienen preponderancia de casos con fraude por facturas falsas en el año de estudio. En cuanto al tamaño, el cluster 1 es el que presenta mayor tamaño con un 34,4% de la muestra, mientras que el más pequeño corresponde al cluster 4 con un 9,4% de la muestra.

Cuadro Nº 14: Tamaño del cluster y distribución de casos de Facturas Falsas en cada uno de ellos

	Tamaño	Resultado		
Cluster	Nº casos	0	1	2
1	17.200	35%	59%	39%
2	6.500	13%	6%	8%
3	7.600	26%	12%	31%
4	4.700	11%	6%	2%
5	7.700	2%	2%	4%
6	6.350	13%	15%	15%
Total	50.000	100%	100%	100%

A modo general, se observa que los grupos que se forman al aplicar el gas neuronal también se encuentran influenciados por el pago de impuestos, pero en menor medida que en los mapas de Kohonen. Por otra parte, se observan mayores diferencias en términos del comportamiento, aunque sólo podemos concluir si un grupo tiene mejor o peor comportamiento que otro, pero existe una alta probabilidad de encontrar casos con y buen comportamiento en un mismo grupo, como se muestra en la distribución de casos conocidos de fraude por facturas falsas en cada cluster.

Experimento 4: Gas Neuronal – Medianas y Grandes Empresas – Vector de Características compuesto por 24 variables (considera ratios tributarios de IVA y Renta)

Este experimento considera 5 clusters, debido a que el SOM arroja para este grupo la generación de ese número de grupos. El cluster de mayor tamaño contiene el 41,9% de la muestra, mientras que el de menor tamaño contiene un 9,1% de los casos.

Cuadro Nº 15: Centroides de cada cluster – muestra experimento 4

Centro	FF_01	C630	C122	C779	C123	ING_COSTOS	ING_ACTIVOS	COSTO_ACTIVOS
1	0,0122	0,4003	0,0874	0,0356	0,0782	0,00003	0,0177	0,0149
2	0,0058	0,3663	0,5523	0,1672	0,5645	0,00003	0,0009	0,0009
3	0,0074	0,3657	0,1635	0,6669	0,1545	0,00068	0,0081	0,0062
4	0,0189	0,3505	0,3649	0,1820	0,3608	0,00282	0,0015	0,0015
5	0,0111	0,4300	0,0678	0,0507	0,0631	0,00002	0,0230	0,0222

Centro	REM_ACTIVOS	OTGASTOS_ACTIVOS	DF_DTOTAL	CF_CTOTAL	REM_CRED	LOG_519	LOG_509	LOG_111
1	0,0270	0,0216	0,3543	0,4116	0,1151	0,3863	0,1808	0,0022
2	0,0022	0,0022	0,3459	0,2955	0,2545	0,4098	0,2596	0,0224
3	0,0147	0,0138	0,3376	0,3479	0,1667	0,3869	0,2252	0,1055
4	0,0043	0,0037	0,2944	0,3428	0,1302	0,4579	0,5149	0,7262
5	0,0412	0,0298	0,2196	0,4691	0,0371	0,4140	0,2322	0,7284

Centro	LOG_DEBCRED	ACP1	ACP2	ACP8	ACP10	ACP11	ACP12	ACP13
1	0,1128	0,0121	0,1015	0,2383	0,0940	0,0404	0,1248	0,7336
2	0,0844	0,0435	0,1649	0,2927	0,1022	0,0372	0,1432	0,7578
3	0,1059	0,0236	0,1190	0,2717	0,0964	0,0341	0,1313	0,7453
4	0,0920	0,0702	0,1610	0,2961	0,1150	0,0421	0,1416	0,7689
5	0,1052	0,0254	0,1028	0,2168	0,1011	0,0278	0,1290	0,7610

Las características de cada cluster son:

- **Cluster 1 (34,4%):** Presenta un bajo nivel de emisión de boletas y el mayor ratio entre débitos y créditos. Además presenta el mayor porcentaje de débitos asociados a facturas y la menor cantidad de facturas recibidas, notas de crédito asociadas a facturas y cuentas por pagar a empresas relacionadas. En cuanto a su comportamiento, tiene la mayor cobertura y menor cantidad de fiscalizaciones previas, de irregularidades de facturas y nivel de timbraje, irregularidades recientes y actividades económicas de difícil fiscalización.
- **Cluster 2 (13,0%):** Registra el nivel más alto de activos y pasivos, pero el nivel más bajo de ingresos y costos en relación a esos activos, y de créditos asociados a facturas. Además tiene el mayor nivel de remanentes de créditos y poca emisión de boletas. Registra un nivel de cobertura intermedio con más fiscalizaciones previas e irregularidades recientes.

- **Cluster 3 (15,2%):** Presenta niveles intermedios de sus indicadores de pago de impuesto, con la mayor cantidad de cuentas por pagar a empresas relacionadas. Tiene pocas irregularidades previas y nivel de cobertura intermedio.
- **Cluster 4 (9,4%):** Presenta el mayor valor del ratio entre ingresos y costos, de notas de compra emitidas por facturas y de facturas de compra emitidas, con alto uso de boletas, presentando los menores costos. Registra un nivel de cobertura alta, mayor cantidad de representantes legales, de irregularidades con facturas y nivel de timbraje, y rendimientos por fiscalizaciones previas, lo que puede estar relacionado a su mayor cobertura. Además tiene la mayor cantidad de actecos de cambio de sujeto y difícil fiscalización.
- **Cluster 5 (15,4%):** Presenta el mayor nivel de costos y el valor más bajo del ratio entre ingresos y costos, de activos y pasivos, además registra un monto mayor de otro tipo de gastos. Presenta además el menor porcentaje de débitos asociados a facturas y el nivel más alto de créditos asociados a facturas, con alto uso de boletas. En cuanto al comportamiento, registra niveles intermedios-bajos de cobertura, de representantes legales y de fiscalizaciones previas. Tiene además el menor monto asociado a rendimientos de fiscalización.

De acuerdo a lo anterior, los clusters 2 y 3 son los que tienen un mejor comportamiento tributario, mientras que los clusters 1, 4 y 5 tienen más irregularidades. De los anteriores, los más fiscalizados con controles selectivos son los clusters 2 y 4.

Respecto de la distribución de casos conocidos de fraude por facturas falsas se tiene que los casos con fraude en período cercano se encuentran principalmente en el cluster 1, el cual tiene también gran porcentaje de casos con fraude y en menor medida casos sin fraude. Los clusters 2 y 3 tienen pocos casos con fraude. Mientras que el cluster 4 tiene mayor porcentaje de casos sin fraude aunque con un número significativo de casos con fraude. El cluster 5 en tanto tiene mayor porcentaje de casos con fraude, aunque con un número significativo de casos sin fraude.

Cuadro Nº 16: Tamaño del cluster y distribución de casos de Facturas Falsas en cada uno de ellos

	Tamaño	Resultado		
Cluster	Nº casos	0	1	2
1	7.818	24%	61%	40%
2	1.847	23%	0%	5%
3	1.697	11%	7%	3%
4	2.036	21%	16%	18%
5	5.273	21%	16%	35%
Total	18.671	100%	100%	100%

Lo anterior, corrobora el hecho que los grupos formados se diferencian principalmente por su pago de impuestos más que por su comportamiento asociado al fraude. Por lo tanto, no se puede concluir si alguno de ellos está asociado específicamente a un patrón de fraude, sino más bien cuáles de ellos tienen un mejor comportamiento en relación a otros.

4.3.2. CARACTERIZACIÓN Y DETECCIÓN DE USUARIOS DE FACTURAS FALSAS

Como se mencionaba en los primeros capítulos de esta tesis, el objetivo principal es identificar las características de los contribuyentes que utilizan y/o venden facturas falsas, así como la detección de aquellos contribuyentes qué estarían cometiendo fraude a través de este mecanismo. Esto permitiría generar planes de control específicos, aumentando la probabilidad de encontrar casos con rendimiento que permitan disminuir la evasión de impuestos, aumentar la recaudación y mejorar la productividad de la acción de fiscalización.

Comúnmente, las técnicas de aprendizaje supervisado son más populares que las de aprendizaje no supervisado para la detección de fraude, ya que si bien es de interés conocer cuáles son las características y los clusters que se forman en el universo de empresas, lo relevante y principal, es determinar quiénes son los contribuyentes propensos a ser defraudadores. Para ello se utiliza información en los que se conoce la respuesta de salida (fraude y/o no fraude), de manera que los algoritmos se entrenen y aprendan a distinguir los patrones de cada tipo de respuesta.

Considerando que los histogramas y diagramas de caja señalan que los casos con y sin fraude se encuentran en los casos extremos de las variables que componen el vector de características, se utiliza en una primera instancia árboles de decisión para caracterizar a este grupo, dado que permite conocer el umbral de estas variables a partir de los cuales se puede diferenciar entre un caso de fraude y no fraude. Esta técnica se utiliza tanto para caracterizar como para detectar fraude por facturas falsas, considerando todas aquellas que son de interés para el análisis, como el total de débitos y créditos y el pago de IVA del período. Para la detección del fraude tributario, se aplica también una red neuronal con backpropagation y una red bayesiana. La primera es una de las herramientas más utilizadas en la resolución de problemas en el campo de la Inteligencia Artificial. La segunda permite obtener inferencias a través de un aprendizaje causal, en la que se evalúan todas las posibilidades de los sucesos.

Para presentar los resultados de cada experimento se utiliza la matriz de confusión, la cual es una herramienta que se emplea en aprendizaje supervisado para visualizar rápidamente en cuántas ocasiones fue exacta la predicción del modelo. Las columnas de cada matriz representan los valores de predicción del modelo, mientras que las filas representan los valores reales. Esta matriz de clasificación se crea ordenando todos los casos en categorías que indican si el valor de predicción coincide con el valor real, y si el valor de predicción es correcto o incorrecto. Estas categorías se conocen como Falso Positivo, Verdadero Positivo, Falso Negativo y Verdadero Negativo y es de la forma:

Cuadro Nº 17: Forma de una Matriz de Confusión

		PREDICCIÓN DEL MODELO	
		FF=1	**FF=0**
VALOR REAL	**FF=1**	Verdadero Positivo (VP)	Falso Negativo (FN)
	FF=0	Falso Positivo (FP)	Verdadero Negativo (VN)

Adicionalmente se utilizan los siguientes indicadores:

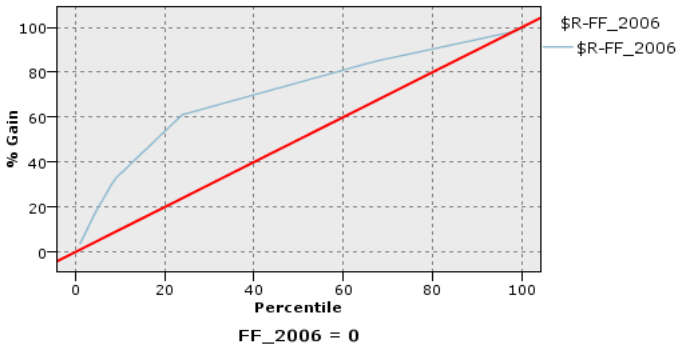
$$\text{Sensibilidad} = \frac{VP}{VP + FN} \; ; \; \text{Especificidad} = \frac{VN}{VN + FP}$$

$$\text{Concordancia} = \frac{VP + VN}{VP + VN + FP + FN} \; ; \; \text{Tasa de Error} = \frac{FP + FN}{VP + VN + FP + FN}$$

La “Sensibilidad” indica la proporción de casos con fraude en que la predicción fue correcta y la “Especificidad” la proporción de casos sin fraude en los que la predicción fue correcta. Por otro lado, la “Tasa de Concordancia” representa la proporción de casos con y sin fraude en las que la predicción fue realizada correctamente y la “Tasa de Error” los casos con y sin fraude que fueron asignados a una clase incorrecta.

Adicionalmente se utiliza la curva ROC[40], que es una representación gráfica de la sensibilidad frente a la especificidad del modelo, donde el eje de las abscisas está determinado por la especificidad y el eje de las ordenadas por la sensibilidad. La mayor exactitud de la prueba está determinada por un desplazamiento hacia “arriba y a la izquierda” de la curva, lo que se puede utilizar como un índice de la exactitud global de la prueba. La exactitud máxima corresponde al valor 1, y la diagonal es el valor de la indiferencia o aleatorio, es decir, es equivalente al experimento de lanzar una moneda para asignar un caso de fraude.

Figura Nº 35: Forma de una Curva ROC



En el Cuadro Nº 18 se describen los experimentos realizados para la caracterización y detección del fraude por facturas falsas, y posteriormente, se presentan los resultados obtenidos en cada uno de ellos.

[40] Receiver Operating Characteristic, o Característica Operativa del Receptor.

Cuadro Nº 18: Experimentos para caracterizar y detectar contribuyentes con facturas falsas

EXP. Nº	TÉCNICA DM	GRUPO OBJETIVO	Nº CASOS CON FF CONOCIDO	Nº VARIABLES
1	Árbol de Decisión	Micro y Pequeñas	1.280	40
2	Árbol de Decisión	Medianas y Grandes	414	42
3	Red Neuronal	Micro y Pequeñas	1.280	40
4	Red Neuronal	Medianas y Grandes	414	42
5	Red Bayesiana	Micro y Pequeñas	1.280	40
6	Red Bayesiana	Medianas y Grandes	414	42

4.3.2.1. Árboles de Decisión

Esta técnica se utiliza tanto para la caracterización como para la detección de fraude por facturas falsas, utilizando para ello los casos de fraude y no fraude conocidos en el año de análisis, que comprende a 1.280 contribuyentes del segmento micro y pequeños y 414 contribuyentes del segmento medianas y grandes empresas.

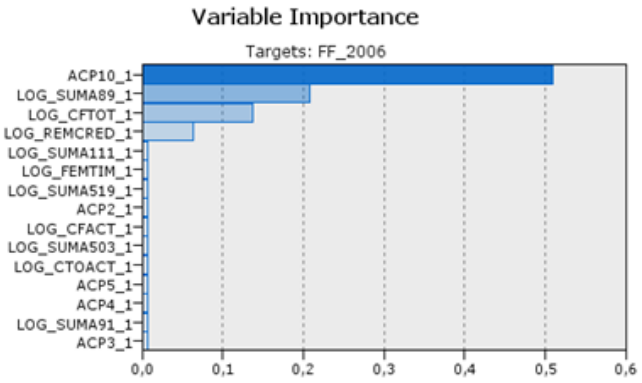
A priori no era claro si el comportamiento de los casos categorizados como "1" en la variable FF06, que indica que no le encontraron facturas falsas en el año de estudio pero sí en otros años revisados, es más parecido al grupo categorizado como "0" que indica que no registra facturas falsas, o más parecido al grupo "2" que utilizan facturas falsas en el año de análisis. Por lo tanto, se determina mantener estas categorías a modo exploratorio para conocer a cuál de estos dos grupos se asemeja más en términos de comportamiento.

Experimento 1: Árbol de Decisión – Micro y Pequeñas Empresas – Vector de Características compuesto por 40 variables (salida categórica de FF06)

Para este segmento se tiene información de 1.280 empresas auditadas en el que se revisa el período tributario del año 2006, de los cuales 316 tuvo resultado negativo en todos los períodos revisados (FF = 0), 371 tuvo resultado negativo en el año de estudio pero positivo en otros períodos cercanos (FF = 1) y 593 tuvo resultado positivo en el año de estudio (FF=2).

Como se observa en la Figura Nº 36, del total de 40 variables consideradas en el árbol, sólo 15 de ellas tienen un nivel de importancia significativo, siendo las más relevantes el número de fiscalizaciones previas con resultado negativo (ACP10), el total del IVA determinado en el año (Cód. 89 del F29), el porcentaje de crédito por facturas respecto del crédito total (Cftot) y el promedio de remanentes respecto del crédito total del año (Remcred). Otras variables de comportamiento relevantes son los delitos e irregularidades de facturas previos (ACP2), el nivel de participación en otras empresas (ACP5), las fiscalizaciones previas con resultado positivo (ACP3) y la frecuencia de timbraje (ACP4).

Figura Nº 36: Nivel de importancia de las variables del experimento 1 – árbol de decisión



Fuente: Resultado obtenido a partir del software Clementine de SPSS

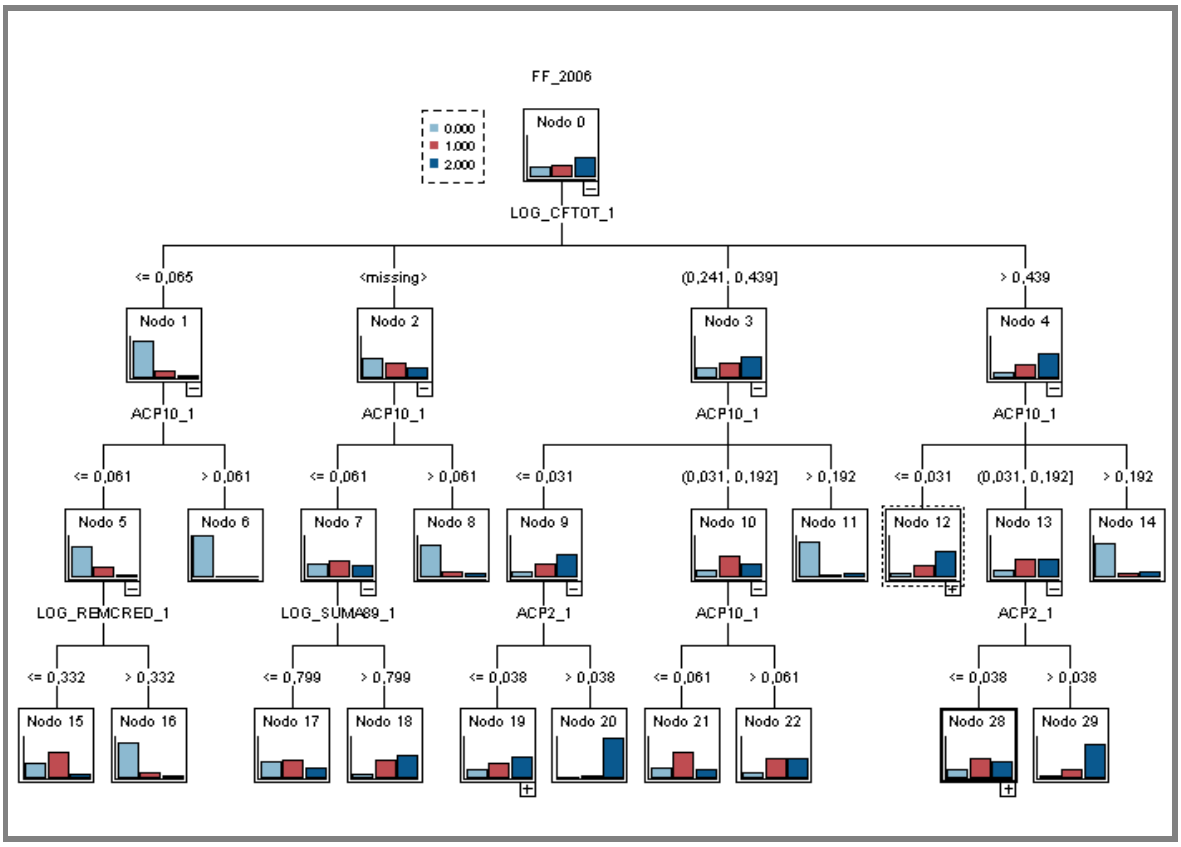
Al analizar la distribución de casos en el árbol en la Figura Nº 37, vemos que a medida que la empresa tiene un menor porcentaje de créditos asociados a facturas (Cftot), tiene menos posibilidades de estar utilizando facturas falsas. También influye el número de fiscalizaciones realizadas con resultado negativo (sin rendimiento). Esto hace sentido, pues si ha sido fiscalizado varias veces y nunca se le ha encontrado nada, es menos probable que esté utilizando facturas falsas respecto de otra empresa que tiene un bajo número de fiscalizaciones sin rendimiento. Estas dos variable por sí solas, determinan varios nodos finales con preponderancia de casos sin fraude correspondientes al nodo 6 (61 casos), nodo 8 (42 casos), nodo 11 (21 casos) y nodo 14 (27 casos).

Por otra parte, la variable que indica una mayor preponderancia de delitos e irregularidades asociadas a facturas (ACP2) genera nodos finales con preponderancia de casos con fraude por facturas falsas en el año de estudio, como sucede con el nodo 20 (26 casos) y el nodo 29 (22 casos), el cual combinado con la frecuencia del timbraje (ACP4) genera grupos asociados a fraude.

Particularmente el nodo 12 contiene casi la mitad de los casos (46%) y se descompone en varias ramas en función del valor que toma el crédito promedio por factura emitida (mientras mayor sea el indicador CFact, más posibilidad hay de que cometa fraude). Adicionalmente, en cada rama generada se tiene un gran número de casos con fraude, el cual depende a su vez del número de facturas emitidas y el IVA pagado, el total de débitos por boletas, el nivel de participación en otras empresas y la relación entre costos y activos.

Al analizar la composición de cada nodo final, se observa que los patrones generados son un buen descriptor de los casos auditados sin fraude en el año de estudio. Por otro lado, se generan reglas que permiten encontrar casos con fraude, no sólo en el año de estudio (FF=2) sino que también en años previos (FF=1). Esto debido a que los contribuyentes con fraude en años previos se encuentran generalmente en los mismos nodos donde se encuentran los casos de fraude en el año de análisis. Por este motivo, para efectos de predicción se agruparán ambos casos (FF=1 y FF=2) en un mismo grupo, de manera que “1” indica que el contribuyente utiliza facturas falsas y “0” que no.

Figura Nº 37: Árbol de decisión, ramas más representativas del experimento 1 [41]



Fuente: Resultado obtenido a partir del software Clementine de SPSS

En el Cuadro N° 19, se presentan los resultados de la predicción. De acuerdo a la matriz de confusión, los casos con fraude correctamente asignados fue cercano al 89%, mientras que los casos sin fraude correctamente asignados fue un 79%. De esta forma, el porcentaje de casos con y sin fraude correctamente asignado fue de un 87% con un 13% de error.

Cuadro N° 19: Matriz de confusión resultante, árbol de decisión – experimento 1

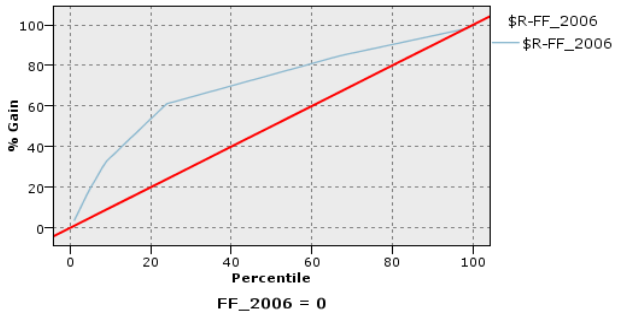
	FF=1	FF=0	Total
FF=1	910 (88,9%)	54 (21,1%)	964
FF=0	113 (11,1%)	203 (78,9%)	316
Total	1.023	257	1.280

Sensibilidad	88,9%
Especificidad	78,9%
Concordancia	86,9%
Tasa Error	13,1%

[41] El árbol completo generado en este experimento se encuentra en el Anexo G.

La Figura Nº 38 relaciona la sensibilidad con la especificidad del modelo, donde el eje de las abscisas está determinado por la especificidad y el eje de las ordenadas es la sensibilidad. La curva generada en este experimento se encuentra desplazada hacia arriba y a la izquierda de la diagonal, indicando que el modelo entrega mejores resultados que el caso de selección aleatoria. De acuerdo a este gráfico, con un 20% de las fiscalizaciones se puede detectar al 55% de contribuyentes con fraude, mientras que con un 10% de las fiscalizaciones se puede descubrir al 35% de los contribuyentes que utilizan facturas falsas.

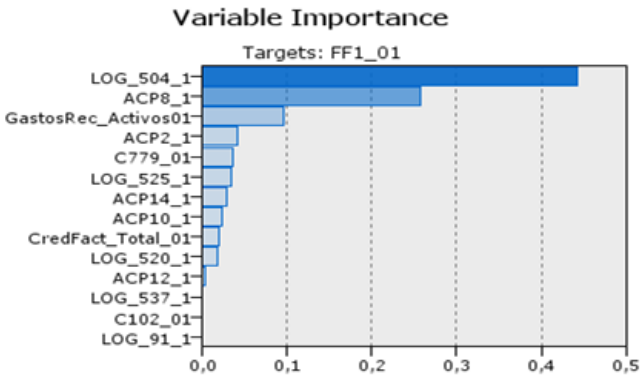
Figura Nº 38: Curva ROC, árbol de decisión – experimento 1



Experimento 2: Árbol de Decisión – Medianas y Grandes Empresas – Vector de Características compuesto por 42 variables (salida categórica de FF06)

Este experimento contempla un universo de 414 contribuyentes, en los cuales el resultado de fraude y no fraude es conocido, considerando un total de 42 variables. De éstos, 201 tuvo resultado negativo en los períodos auditados (FF06=0), 167 tuvo resultado positivo (FF06=2) y 46 tuvo resultado negativo en el año de estudio pero utilizó facturas falsas en otros períodos cercanos (FF06=1).

Figura Nº 39: Nivel de importancia de las variables del experimento 2 – árbol de decisión



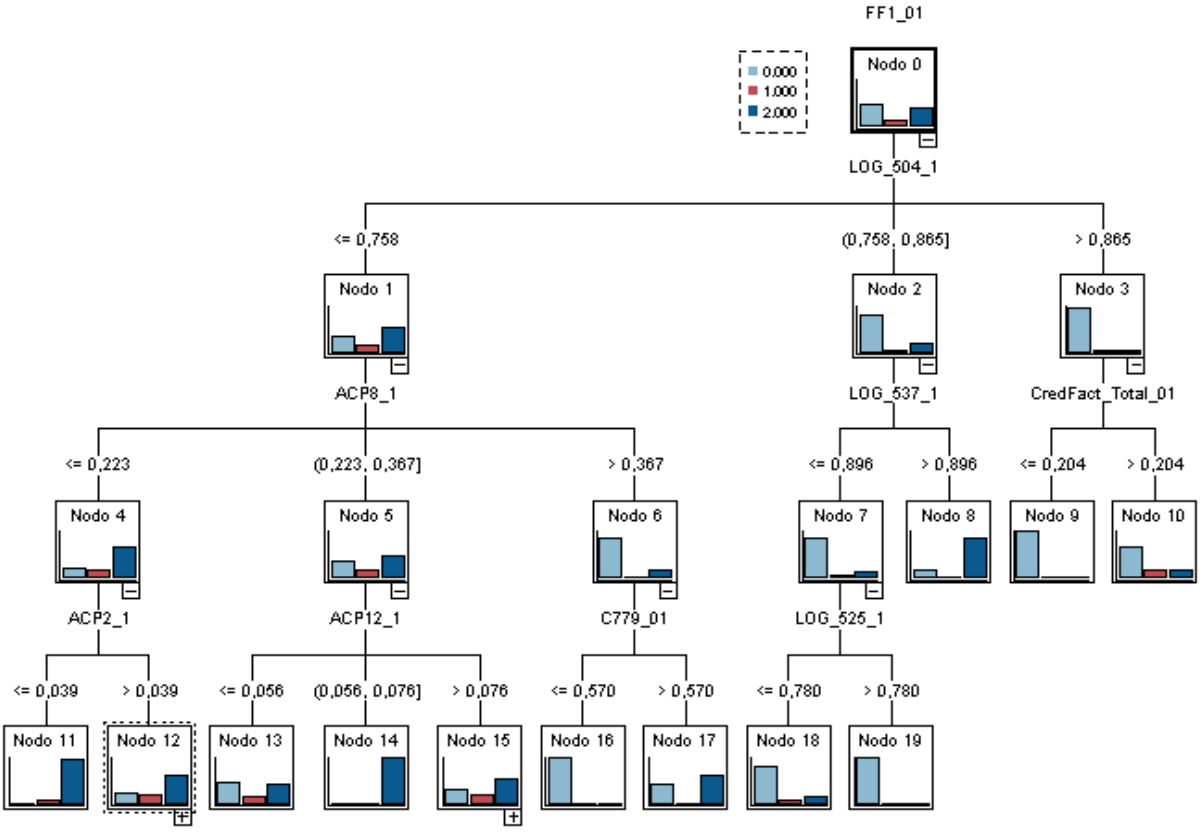
Fuente: Resultado obtenido a partir del software Clementine de SPSS

Como se observa en la Figura Nº 39, del total de 42 variables consideradas en el árbol, sólo 14 de ellas tienen un nivel de importancia significativo, siendo las más relevantes la cantidad de

remanente del período anterior (Cód. 504 del F29), las variables asociadas al número de representantes legales (ACP8), la proporción de los gastos rechazados en relación al total de activos (Gtosrec/Activos), el nivel de fiscalizaciones previas (ACP2), las cuentas por pagar a empresas relacionadas (Cód. 779 del F22), si registra antecedentes de término de giro y de no ubicado (ACP14) y si presenta irregularidades previas asociadas a facturas y el nivel de timbraje promedio (ACP10), entre otros.

Al analizar el árbol generado, se observa que a medida que la empresa registra mayor cantidad de remanente de períodos anteriores (Cód. 504 del F29) menor es la probabilidad de encontrar fraude por facturas falsas, lo que se acentúa con un nivel bajo de créditos por facturas respecto del crédito total (Credfact/Total) ubicados principalmente en el Nodo 9. De igual forma, un nivel bajo de créditos totales (Cód. 537 del F29) y un nivel alto de créditos por facturas de activo fijo (Cód. 525 del F29) generan casos sin fraude ubicados en los Nodos 18 y 19. Lo mismo sucede en empresas con niveles bajos de remanentes de crédito que tengan mayor cantidad de representantes legales (ACP8) y pocas cuentas por pagar a empresas relacionadas (Cód. 779 del F22) ubicados en el Nodo 16.

Figura Nº 40: Árbol de decisión, ramas más representativas del experimento 2 [42]



Fuente: Resultado obtenido a partir del software Clementine de SPSS

[42] El árbol completo generado en este experimento se encuentra en el Anexo H.

Los casos con mayor probabilidad de fraude, en cambio, se encuentran en empresas que tienen nivel intermedio-alto de remanente de créditos (Cód. 504 del F29) y nivel alto de crédito total (Cód. 537 del F29) ubicados en el Nodo 8, en empresas que tienen bajo nivel de remanente de crédito (Cód. 504 del F29) que tienen pocos representantes legales (ACP8) y niveles bajos de fiscalizaciones previas (ACP2) ubicados en el Nodo 11, en empresas que tienen bajo nivel de remanente de crédito (Cód. 504 del F29) que tienen nivel intermedio de representantes legales (ACP8) e intermedio de irregularidades recientes (ACP12) ubicados en el Nodo 14.

Adicionalmente existen otras reglas que determinan fraude considerando un valor bajo de pago de IVA (Cód. 91 del F29), mayor cantidad de irregularidades de facturas y de timbraje promedio y menor cantidad de antecedentes de término de giro y no ubicados, los que pueden apreciarse de mejor forma en el árbol completo.

Al igual que en el experimento 1, se concluye que las empresas de la categoría “1” en la variable de salida FF06, tienen un comportamiento más parecido a la categoría “2”. Por lo tanto, las reglas generadas permiten descubrir también algunos casos de fraude que se producen en períodos cercanos al año de estudio. Debido a que el número de casos en esa categoría es reducido, se determina agrupar ambas categorías en una sola para la aplicación de los modelos predictivos de fraude.

De acuerdo con la matriz de confusión resultante del experimento realizado para la detección, se tiene que los casos con fraude correctamente asignados fue cercano al 78%, mientras que los casos sin fraude correctamente asignados fue un 85%. De esta forma, el porcentaje de casos con y sin fraude correctamente asignado fue de un 81% con un 19% de error.

Cuadro N° 20: Matriz de confusión resultante, árbol de decisión – experimento 2

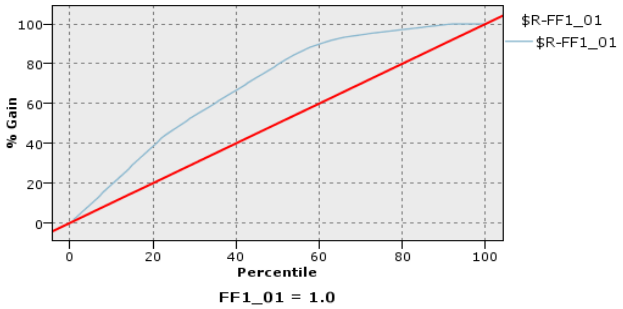
	FF=1	FF=0	Total
FF=1	187 (78,3%)	26 (14,9%)	213
FF=0	52 (21,7%)	149 (85,1%)	201
Total	239	175	414

Sensibilidad	78,3%
Especificidad	85,1%
Concordancia	81,1%
Tasa Error	18,9%

A diferencia del experimento anterior, este modelo predice mejor los casos sin fraude que con fraude, obteniéndose resultados de menor calidad que el experimento anterior, lo cual se ve reflejado en el gráfico comparativo de sensibilidad y especificidad respecto del caso aleatorio, pues la curva del modelo se encuentra menos desplazada hacia “arriba y a la izquierda” que el experimento anterior. En este caso con un 40% de los fiscalizados se puede detectar al 65% de

los contribuyentes con fraude, mientras que con 20% de las fiscalizaciones se puede descubrir al 40% de los contribuyentes que utilizan facturas falsas.

Figura Nº 41: Curva ROC, árbol de decisión – experimento 2



4.3.2.2. Redes Neuronales con Backpropagation

Este método analiza las relaciones existentes entre los atributos de entrada y el resultado conocido de la auditoría (fraude y/o no fraude), a través de una técnica de aprendizaje llamado "Backpropagation" o "Retropropagación del error" que minimiza el error de predicción mediante un ajuste a los pesos de la red. Este proceso es iterativo, por lo que tras realizar varias veces el algoritmo, la red va convergiendo hacia un estado que permita clasificar todos los patrones de entrenamiento que minimiza el error.

Para evitar el sobreajuste de la red, los datos se dividen en dos conjuntos: uno de entrenamiento y uno de testeo. Con el primero, la red realiza el aprendizaje y el ajuste de sus pesos, minimizando el error, el cual es posteriormente evaluado con los pesos obtenidos en el segundo conjunto. Para distribuir el número de casos conocidos en cada grupo (entrenamiento y testeo) se utiliza la regla 70/30, lo cual se realiza de manera aleatoria. Considerando además que los grupos de fraude y no fraude no son homogéneos en cuanto a su tamaño, se realizó un balance previo de la red, para que considere la misma cantidad de casos con fraude y no fraude en el entrenamiento.

En este tipo de problemas lo más difícil de determinar es el número de capas y nodos ocultos, así como la cantidad de épocas o iteraciones a utilizar. El error del modelo disminuye mientras más iteraciones se realicen, sin embargo, se corre el riesgo de estar sobreentrenando la red, y por tanto, se obtendría una red de entrenamiento muy ajustada, pero que al momento de ser aplicado en otro grupo entrega una predicción de baja calidad. A esto último se le llama capacidad de generalización del modelo.

Para considerar estos efectos se realizan experimentos considerando distintos números de ciclos y nodos en las capas ocultas, de manera de determinar a través de ensayo y error, los parámetros más adecuados. En el caso de las iteraciones se utilizan los valores: 1.000, 5.000, 10.000 y 20.000. Para determinar los nodos de la capa oculta, en tanto, se utiliza la cantidad de nodos que calcula el programa por defecto en función de los datos del modelo y otra correspondiente a la mitad del número de nodos de entrada, equivalente a 3 y 20 nodos respectivamente.

De estas combinaciones se obtuvo que con un mayor número de nodos en la capa oculta, los errores del entrenamiento disminuían, mientras que al generalizar los resultados, éstos bajaban su calidad. Lo mismo sucedía al aumentar el número de iteraciones. En consecuencia se determina utilizar la combinación que mostró mejores resultados en la generalización, correspondiente a 1.000 iteraciones y 3 nodos en la capa oculta.

Experimento 3: Red Neuronal Backpropagation – Micro y Pequeñas Empresas – Vector de Características compuesto por 42 variables (considera agrupación de casos con valor "1" y "2" de la variable FF)

Al igual que con el árbol de decisión, en este segmento se utiliza información de 1.280 empresas auditadas, de los cuales 316 tuvo resultado negativo en todos los períodos revisados (FF = 0), 371 tuvo resultado negativo en el año de estudio pero positivo en otros períodos cercanos (FF = 1), y 593 tuvo resultado positivo en el año de estudio (FF=2). Por otra parte, se agrupan los contribuyentes que se les detecta fraude por facturas falsas, independiente de si las utiliza en el año de estudio o en períodos cercanos.

En el Cuadro Nº 21 se presentan las matrices de confusión resultantes del entrenamiento y el testeo, las cuales recogen los porcentajes de éxitos y fracasos clasificados en forma correcta e incorrecta por la correspondiente red. Como se aprecia en ambas matrices, la tasa de concordancia, que representa cuántos casos de fraude y no fraude son correctamente asignados es cercana al 88% en el entrenamiento y a un 87% en el testeo, con un error de un 12% y 13% respectivamente.

Cuadro N° 21: Resultados entrenamiento y testeo, red neuronal – experimento 3

Entrenamiento

	FF=1	FF=0	Total
FF=1	638 (93,7%)	43 (6,3%)	681
FF=0	64 (30,6%)	145 (69,4%)	209
Total	702	188	890

Sensibilidad	93,69%
Especificidad	69,38%
Concordancia	87,98%
Tasa Error	12,02%

Testeo

	FF=1	FF=0	Total
FF=1	262 (92,6%)	21 (7,4%)	283
FF=0	29 (27,1%)	78 (72,9%)	107
Total	291	99	390

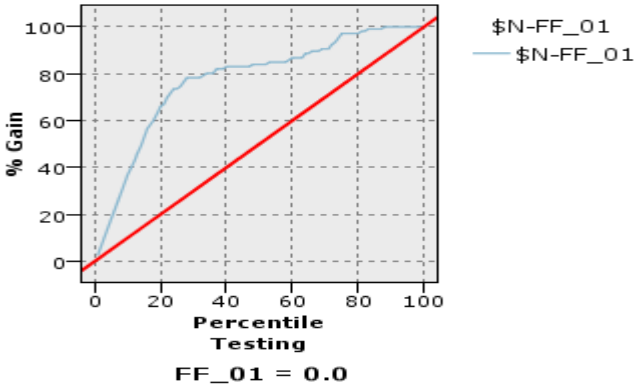
Sensibilidad	92,58%
Especificidad	72,90%
Concordancia	87,18%
Tasa Error	12,82%

La sensibilidad del modelo, que indica la probabilidad de obtener un resultado positivo cuando el contribuyente registra facturas falsas, y que representa los casos con fraude correctamente asignados, es del orden del 93,7% en el entrenamiento y del 92,6% en el testeo. Adicionalmente el porcentaje de casos sin fraude correctamente asignados fue de un 69,4% en el entrenamiento y

un 72,9% en el testeo. Ambas matrices son muy similares, por lo que el poder de generalización de la red se considera bueno, con un pequeño aumento de la asignación correcta de casos sin fraude en el testeo.

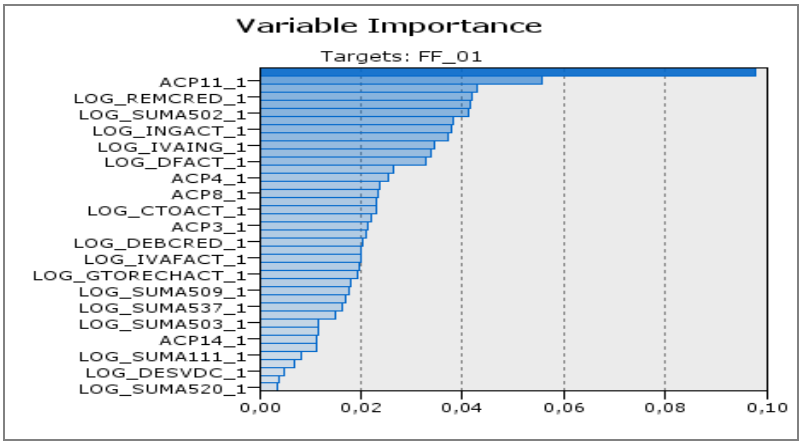
La Figura Nº 42 relaciona la sensibilidad con la especificidad del modelo, donde el eje de las abscisas está determinado por la especificidad y el eje de las ordenadas por la sensibilidad. La curva generada se encuentra desplaza hacia arriba y a la izquierda, lo cual es un buen indicador del modelo. De acuerdo con ella, se tiene que con un 20% de las fiscalizaciones se puede detectar al 65% con peor comportamiento, mientras que con 10% de las fiscalizaciones se puede descubrir al 40% más irregular.

Figura Nº 42: Curva ROC, red neuronal – experimento 3



Al revisar el nivel de importancia de las variables consideradas en este experimento, presentadas en la Figura Nº 43, se observa que las variables más relevantes se modifican bastante respecto de los resultados del árbol de decisión, ya que las dos variables consideradas como más relevantes en ese experimento (ACP10 y Cód. 89) no aparecen como relevantes en la red neuronal. En este caso se tiene que la variable que representa el nivel y resultados de las verificaciones de actividad (ACP11) adquiere mayor relevancia, manteniéndose como variable importante el ratio entre Remanentes y Crédito Total promedio del año (RemCred). Sin embargo, se obtienen muy buenos resultados en la predicción de casos con fraude.

Figura Nº 43: Nivel de importancia de las variables del experimento 3 – red neuronal



Fuente: Resultado obtenido a partir del software Clementine de SPSS

Por otra parte, el número de variables consideradas como relevantes en la red neuronal es mayor que el caso obtenido con el árbol de decisión, aunque el grado de relevancia es menor. En particular llama la atención que el ratio que representa qué porcentaje de los créditos se asocia a facturas en este caso pierde relevancia. Esto puede ocurrir porque la red neuronal se basa en otro tipo de relaciones no necesariamente lineales entre las variables, como sucede con las reglas de los árboles de decisión, ya que los códigos que componen ese ratio (Cód. 520 y Cód. 537) aparecen como significativos en la red neuronal.

Experimento 4: Red Neuronal Backpropagation – Medianas y Grandes Empresas – Vector de Características compuesto por 42 variables (considera agrupación de casos con valor "1" y "2" de la variable FF)

En el caso de las medianas y grandes empresas, existe información de 412 empresas auditadas con resultado conocido de fraude y/o no fraude relativo al uso o comercialización de facturas falsas en el año de estudio. De éstas, 212 tuvo resultado negativo en los períodos auditados (FF06=0), 167 tuvo resultado positivo (FF06=2), y 46 tuvo resultado negativo en el año de estudio pero utilizó facturas falsas en otros períodos cercanos (FF06=1). Al igual que los otros experimentos, se determina agrupar los contribuyentes que se les detecta fraude por facturas falsas, independiente si las utiliza en el año de estudio o en períodos cercanos, debido a que tienen un comportamiento similar.

Los resultados del entrenamiento y el testeo se presentan en el Cuadro Nº 22. Ambas matrices indican que la tasa de concordancia, que representa cuántos casos de fraude y no fraude son correctamente asignados, es cercana al 73%, con un error cercano al 27%. En este caso la sensibilidad, que indica el porcentaje de casos con fraude correctamente asignado, es superior al 87% en el grupo de entrenamiento y testeo, mientras que el porcentaje de casos sin fraude correctamente asignados tiene una especificidad cercana al 57% en el grupo de entrenamiento y 59% en el grupo de testeo.

Cuadro N° 22: Resultados entrenamiento y testeo, red neuronal – experimento 4

Entrenamiento

	FF=1	FF=0	Total
FF=1	138 (87,3%)	20 (12,7%)	158
FF=0	61 (43,0%)	81 (57,0%)	142
Total	199	101	300

Sensibilidad	87,34%
Especificidad	57,04%
Concordancia	73,00%
Tasa Error	27,00%

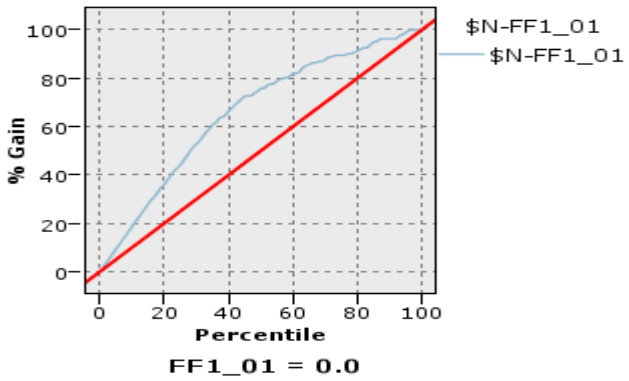
Testeo

	FF=1	FF=0	Total
FF=1	48 (88,8%)	6 (11,1%)	54
FF=0	27 (40,9%)	39 (59,1%)	66
Total	75	45	120

Sensibilidad	88,88%
Especificidad	59,09%
Concordancia	72,50%
Tasa Error	27,50%

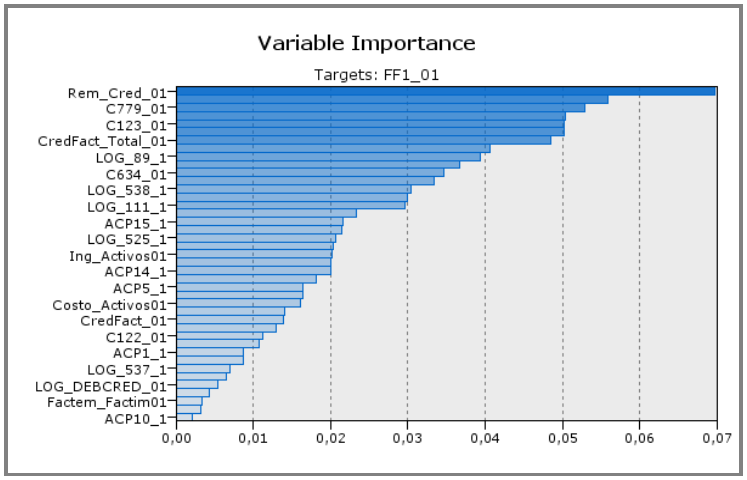
Si bien esta agrupación predice bastante bien los casos con fraude, en términos globales los resultados obtenidos son de menor calidad que en el experimento anterior. Esto se ve reflejado en el gráfico comparativo de sensibilidad y especificidad respecto del caso aleatorio, pues la curva del modelo se encuentra menos desplazada hacia "arriba y a la izquierda" que el experimento anterior, indicando que con un 40% de los fiscalizados se puede detectar al 65% de los contribuyentes con comportamiento más fraudulento, mientras que con un 20% de los fiscalizados se puede descubrir casi al 40% más irregular.

Figura N° 44: Curva ROC, red neuronal - experimento 4



Como se aprecia en la Figura Nº 45, la red neuronal de este experimento le da mayor importancia a variables relacionadas al comportamiento, pero mantiene como más relevantes los campos relacionados al pago de IVA y al impuesto a la Renta, principalmente información de ratios. En el caso del comportamiento considera relevante las anotaciones de timbraje restringido (ACP15), si tiene antecedentes de término de giro y no ubicado (ACP14), registra clausuras y denuncios históricos (ACP5), el nivel de cobertura de la empresa (ACP1) y si registra irregularidades con facturas y el nivel de timbraje promedio (ACP10). La variable más relevante corresponde al ratio entre remanentes y créditos promedio en el año (Remcred).

Figura N° 45: Nivel de importancia de las variables del experimento 4 – red neuronal



Fuente: Resultado obtenido a partir del software Clementine de SPSS

4.3.2.3. Redes Bayesianas

Las redes bayesianas son utilizadas para realizar pronósticos en diferentes situaciones. Su principal ventaja es que permite obtener información acerca de las relaciones causales entre distintos eventos, obteniendo la probabilidad de ocurrencia de un determinado suceso en función de un conjunto de acciones, entregando una vista clara de las relaciones mediante un gráfico de red. Al igual que en la aplicación de la red neuronal, para evitar el sobreajuste de la red, los datos se dividen en dos conjuntos: un conjunto de entrenamiento y un conjunto de testeo. Con el conjunto de entrenamiento, la red realiza el aprendizaje y el ajuste de sus pesos, minimizando el error, el cual es posteriormente evaluado con los pesos obtenidos en el segundo conjunto. En nuestro caso se utiliza la regla 70/30 para distribuir el número de casos conocidos en cada grupo (entrenamiento y testeo), lo cual se realiza de manera aleatoria.

En una primera instancia, se determina la estructura de la red adecuada para el conjunto de datos de la muestra, denominado aprendizaje estructural, y posteriormente se estiman las densidades de probabilidad conjunta de cada nodo hijo dado sus nodos padres, lo que se conoce como aprendizaje paramétrico. Específicamente, se evalúan dos métodos para construir la red bayesiana: el algoritmo TAN y el algoritmo de estimación Markov Blanket disponibles en el software Clementine de SPSS, utilizando un preprocesamiento previo de las variables para identificar cuáles son más relevantes, de manera de mejorar el tiempo de procesamiento y el rendimiento del algoritmo[43]. De igual forma se considera un test de independencia de máxima verosimilitud y chi-cuadrado, para el aprendizaje paramétrico.

Al igual que en los experimentos obtenidos con los árboles de decisión y red neuronal, se agrupan los casos que se detecta utilización de facturas falsas, ya sea en el año de estudio como en los períodos cercanos. Considerando que la condición anterior implica generar grupos no homogéneos en cuanto a su tamaño, se realizó un balance previo de la red, para que considere la misma cantidad de casos con fraude y no fraude en el entrenamiento.

Experimento 5: Red Bayesiana – Micro y Pequeñas Empresas – Vector de Características compuesto por 42 variables (considera agrupación de casos con valor "1" y "2" de la variable FF)

De los dos métodos antes mencionados, el algoritmo TAM fue el que tuvo un mejor comportamiento para formar la estructura de la red, por lo que sólo se presentan los resultados obtenidos con este algoritmo. En el caso del aprendizaje paramétrico el test de máxima verosimilitud tuvo un mejor desempeño, considerando un nivel de significancia del 5%.

Los resultados del entrenamiento y el testeo de la red bayesiana se presentan en el Cuadro Nº 23. La tasa de concordancia, que representa cuántos casos de fraude y no fraude son correctamente asignados es de un 82% en el grupo de entrenamiento y de 78% en el grupo de testeo, con un error aproximado al 18% y 22% respectivamente. Lo anterior implica que la predicción fue mejor en el grupo de testeo que en el grupo de entrenamiento.

[43] Este preprocesamiento selecciona los atributos más relevantes, basado en test estadísticos de independencia.

Cuadro N° 23: Resultados entrenamiento y testeo, red bayesiana – experimento 5

Entrenamiento

	FF=1	FF=0	Total
FF=1	414 (85,5%)	70 (14,5%)	484
FF=0	107 (20,9%)	403 (79,1%)	510
Total	521	473	994

Sensibilidad	85,53%
Especificidad	79,02%
Concordancia	82,19%
Tasa Error	17,81%

Testeo

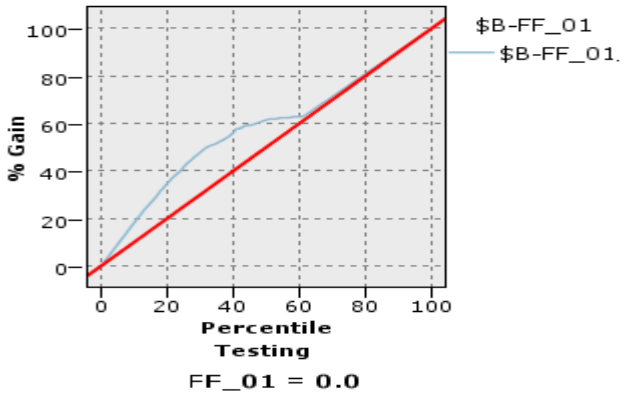
	FF=1	FF=0	Total
FF=1	163 (82,3%)	35 (17,7%)	198
FF=0	23 (35,9%)	41 (64,1%)	64
Total	186	76	262

Sensibilidad	82,32%
Especificidad	64,06%
Concordancia	77,86%
Tasa Error	22,13%

Analizando de manera desagregada por tipo de fraude, se tiene que la predicción de casos con fraude fue de un 85% en el grupo del entrenamiento y de un 82% en el grupo del testeo. Por otra parte, la predicción de casos sin fraude disminuye, obteniendo un 79% de aciertos en el caso del entrenamiento y un 64% en el caso del testeo. Esto indica que el modelo predice mejor los casos con fraude, aunque con tasas de error bastante significativas. En cuanto al poder de generalización de la red, se observa que es bastante bueno para los casos con fraude, pero disminuye el poder de predicción de los casos sin fraude, ya sea porque aprendió de mejor forma el comportamiento de las empresas con fraude o porque se sobre-entrenó con estos casos.

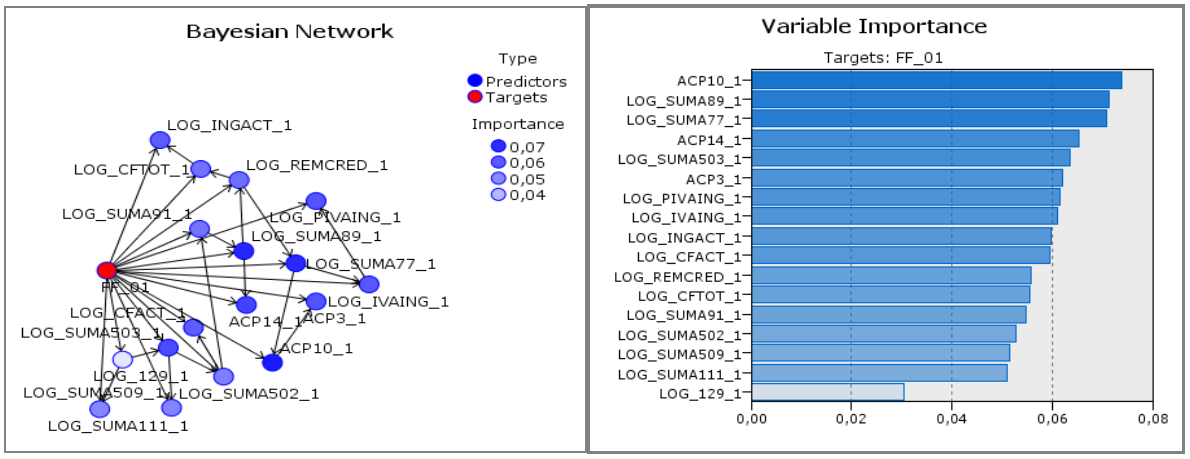
Al analizar la curva ROC, se observa que en una primera instancia la red realiza estimaciones en forma regular (percentil 60), pero después comienza a disminuir su poder de predicción, sin obtener ganancias respecto del caso aleatorio. Antes de ese percentil, la curva indica que con un 40% de los fiscalizados es posible detectar un 55% de los contribuyentes de mal comportamiento, mientras que con un 10% de los fiscalizados es posible encontrar un 20% de los contribuyentes más fraudulentos.

Figura Nº 46: Curva ROC, red bayesiana – experimento 5



La cantidad de variables relevantes es similar a la red neuronal, la que contiene principalmente información de IVA y de comportamiento, y en menor medida ratios que relacionan el pago de IVA con información de Renta. La variable que indica el número de fiscalizaciones previas con resultado negativo aparece como la variable más relevante (ACP10), incorporándose también el nivel de formalización de la contabilidad (ACP14) y las fiscalizaciones previas con resultado positivo (ACP3). En el caso del IVA, las más relevantes son el IVA determinado (Suma89), el remanente del período siguiente (Suma77) y la cantidad de facturas totales emitidas en el año (Suma503). Por otra parte, son relevantes el ratio que relaciona el Pago del IVA y el IVA determinado con los ingresos del giro (PivaIng e IvaIng), así como la relación entre ingresos del giro y activos (IngAct).

Figura Nº 47: Nivel de importancia de las variables del experimento 5 – red bayesiana



Fuente: Resultado obtenido a partir del software Clementine de SPSS

Al analizar las relaciones de causalidad entre las variables del grafo, se observan varias conexiones. Por ejemplo, existe una relación entre el IVA determinado (Suma89), el Pago del IVA (Suma91) y el ratio entre remanentes y créditos promedio (RemCred). También existe relación entre el porcentaje de créditos por facturas (CfTot) y el ratio que relaciona ingresos y activos (IngAct).

Otra relación está dada por los remanentes del período siguiente (Suma77), el ratio entre IVA Determinado e ingresos (Ivaing), el ratio entre Pago de IVA e ingresos (PivaIng) y las fiscalizaciones previas con resultado negativo (ACP10). A su vez, ésta última variable se relaciona con las fiscalizaciones previas con resultado positivo (ACP3), ya que puede darse el caso que una persona tenga sólo fiscalizaciones con resultado positivo o negativo, o tener combinaciones de ambas.

Experimento 6: Red Bayesiana – Medianas y Grandes Empresas - Vector de Características compuesto por 42 variables (considera agrupación de casos con FF06 igual a "1" y "2")

Al igual que en el grupo de micro y pequeñas empresas, el algoritmo TAM fue el que tuvo un mejor comportamiento para formar la estructura de la red, y se obtuvieron mejores resultados con el test de máxima verosimilitud en el caso del aprendizaje paramétrico.

En este caso las matrices de confusión del entrenamiento y la red indican que la predicción de casos con y sin fraude se acercan bastante, con una tasa de concordancia de 79% en el entrenamiento y de 70% en el testeo, con un error de 21% y 30% respectivamente, prediciendo con mayor exactitud los casos con fraude. La predicción de los casos fraudulentos fue de un 81% en el grupo de entrenamiento y de un 73% en el grupo del testeo. Mientras que la predicción correcta de casos sin fraude fue de un 77% en el caso del entrenamiento y de un 66% en el caso del testeo.

Cuadro Nº 24: Resultados entrenamiento y testeo, red bayesiana – experimento 6

Entrenamiento

	FF=1	FF=0	Total
FF=1	117 (81,3%)	27 (18,7%)	144
FF=0	36 (22,6%)	123 (77,4%)	159
Total	153	150	303

Sensibilidad	81,25%
Especificidad	77,35%
Concordancia	79,20%
Tasa Error	20,80%

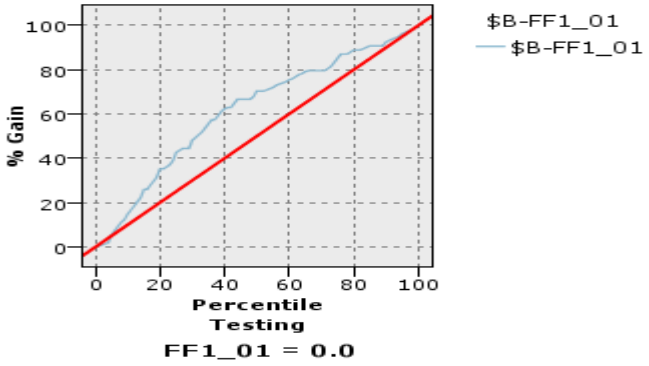
Testeo

	FF=1	FF=0	Total
FF=1	44 (73,3%)	16 (26,7%)	60
FF=0	17 (33,3%)	34 (66,7%)	51
Total	61	50	111

Sensibilidad	73,33%
Especificidad	66,66%
Concordancia	70,27%
Tasa Error	29,73%

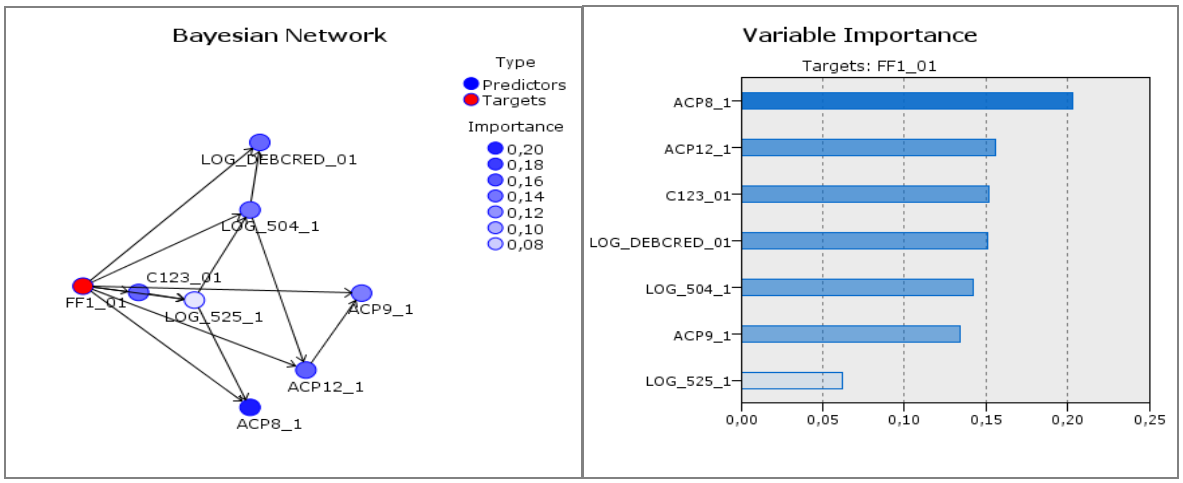
Sin embargo, al observar ambas matrices se tiene que el poder de generalización no es muy bueno, ya que los resultados obtenidos en el testeo fueron más bajos que los obtenidos en el entrenamiento. El gráfico comparativo de sensibilidad y especificidad respecto del caso aleatorio, indica que con aproximadamente un 40% de las fiscalizaciones se puede detectar a poco más del 60% de casos más irregulares, encontrándose la curva del modelo desplazada hacia “arriba y a la izquierda” respecto del caso aleatorio, mientras que con un 20% de las fiscalizaciones es posible encontrar el 35% de los casos fraudulentos.

Figura Nº 48: Curva ROC, red bayesiana – experimento 6



Como se aprecia en la Figura Nº 49, la red bayesiana de este experimento mantiene como variables relevantes la cantidad de representantes legales (ACP8), las irregularidades recientes (ACP12), los delitos de los relacionados (ACP9), la relación entre débitos y créditos (DebCred) y la cantidad de remanente del período anterior. Por otra parte, incorpora las variables relacionadas al nivel de pasivos de la empresa (Cód. 123 del F22) y el crédito obtenido por facturas de activo fijo (Cód. 525 del F29).

Figura Nº 49: Nivel de importancia de las variables del experimento 6 – red bayesiana



Fuente: Resultado obtenido a partir del software Clementine de SPSS

En términos de causalidad, se observa una relación entre las variables que indican el nivel de pasivos de la empresa (Cód.123 del F22), el crédito por facturas de activo fijo (Cód.525 del F29) y la cantidad de representantes legales (ACP8). Por otra parte existe relación entre las irregularidades recientes (ACP12) y los delitos de los relacionados (ACP9), así como también entre el nivel de remanentes (Cód.504 del F29) y la relación entre débitos y créditos (DebCred del F29).

4.4. RESULTADOS

4.4.1. CARACTERIZACIÓN DEL UNIVERSO DE EMPRESAS

La caracterización del universo de empresas efectuada con el SOM y el Gas Neuronal, y el vector de características compuesto por datos del pago de impuesto de IVA y Renta e indicadores de comportamiento tributario asociados al uso de facturas falsas, señalan la existencia de 6 clusters en el segmento de las micro y pequeñas empresas, y de 5 clusters en el segmento de las medianas y grandes empresas. En el caso del SOM, los clusters encontrados no tenían una alta densidad respecto del total de empresas, razón por la cual, las empresas circundantes fueron asignadas al cluster más cercano, y de menor distancia.

En particular, los clusters 2, 4 y 5, que representan un 45,5% del total de contribuyentes micro y pequeñas empresas de acuerdo al SOM y un 37,8% del total de contribuyentes de acuerdo al Gas

Neuronal, corresponden a contribuyentes que tienen un buen comportamiento tributario. Mientras que los clusters 3 y 6, que representan un 28,7% de los contribuyentes de acuerdo al SOM y un 27,9% de los contribuyentes de acuerdo al Gas Neuronal, corresponden a contribuyentes que tienen un comportamiento tributario irregular y potencialmente asociado al uso de facturas falsas.

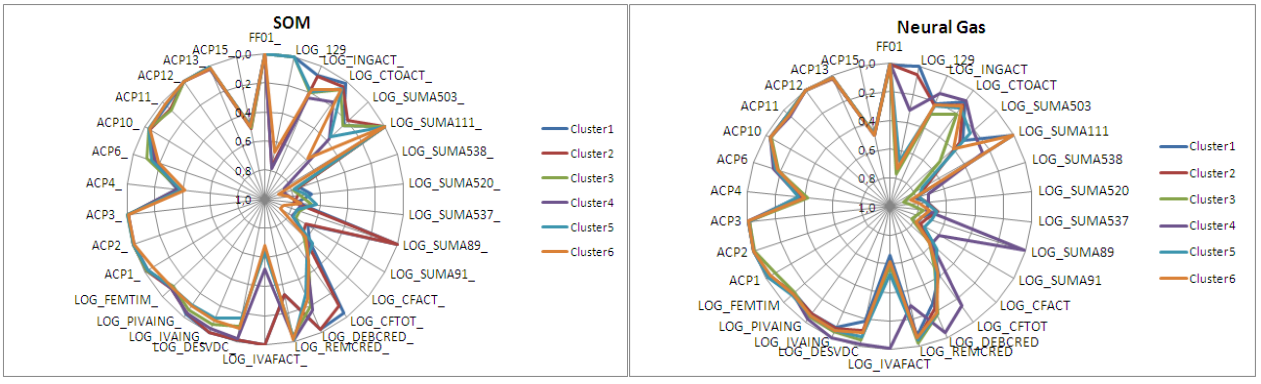
Cuadro N° 25: Tamaño Clusters Caracterización Micro y Pequeñas empresas

CLUSTER	SOM	NEURAL GAS
1	25,8%	34,4%
2	15,5%	13,0%
3	17,8%	15,2%
4	5,8%	9,4%
5	24,2%	15,4%
6	10,9%	12,7%

La Figura N° 50 compara el valor promedio del vector de características en los clusters resultantes con cada método en el segmento de las micro y pequeñas empresas. De acuerdo a este esquema, las variables que más inciden en la clusterización del SOM son aquellas que tienen un gran porcentaje de casos con valor cero, como los códigos asociados a emisión de boletas, ingresos, costos, nivel de existencias e IVA, mientras que las variables asociadas al comportamiento no presentan una gran variabilidad entre los grupos. Para identificar los clusters que tenían comportamiento más irregular se analizó también la distribución de casos con resultado de fraude conocido en el mapa. Si bien era posible identificar algunas zonas, éstas no se consideran muy confiables debido a que los casos con fraude representan un porcentaje bajo del total de empresas, y por lo tanto, su interpretación puede inducir a conclusiones erróneas.

Algo similar ocurre con la agrupación obtenida con el Gas Neuronal, donde los grupos formados también se encuentran influenciados por el pago de impuestos, generándose mayores diferencias en los indicadores de comportamiento entre un grupo y otro. Por otra parte, con este método se tiene información de cuántos casos de fraude y no fraude por facturas falsas se encontraban en cada cluster, lo que permitió caracterizar de mejor forma los grupos resultantes, en términos de su comportamiento.

Figura Nº 50: Comparación centros de cada cluster – Micro y Pequeñas empresas



Si bien se identifican algunos clusters que tienen antecedentes relacionados con un buen y mal comportamiento tributario, existe una alta probabilidad de encontrar casos con y sin fraude en un mismo grupo. Por lo tanto, más que generar clusters de buen y mal comportamiento, con los resultados obtenidos se puede identificar algunas características que señalan aquellos contribuyentes que tienen un mejor o peor comportamiento tributario respecto de otro, los que se presentan en el Cuadro N° 26.

Cuadro N° 26: Variables asociadas a un buen y mal comportamiento de las micro y pequeñas empresas

MICRO Y PEQUEÑAS EMPRESAS				
VARIABLE	**PERÍODO**	**CONCEPTO**	**BUEN COMPORT.**	**MAL COMPORT.**
Remanentes	t	IVA	↓	
Crédito Total			↓	↑
Pago de IVA			↓	↑
Uso de boletas				↑
Uso de facturas			↓	↑
Costos directos		Renta		↑
Existencias finales				↑
Ratio débito/crédito		Ratio IVA	↓	
Ratio crédito facturas/cantidad de facturas			↓	↑
Ratio ingresos del giro/activos		Ratio Renta		↑
Ratio costos directos/activos				↑
Cantidad de facturas timbradas	t-2	Nivel de timbraje		↑
Ratio facturas emitidas/facturas timbradas				↑
Frecuencia de timbraje			↓	↑
Fiscalizaciones con resultado positivo	< t	Comportamiento histórico	↓	↑
Fiscalizaciones con resultado negativo			↑	↑
Delitos e irregularidades por facturas			↓	↑
Delitos de los relacionados directos				↑
Problemas de localización			↓	↑
Verificaciones de actividad			↑	

En el segmento de las medianas y grandes empresas, los clusters 2 y 3, que representan un 25,7% del total de contribuyentes de acuerdo al SOM y un 19% de acuerdo al Gas Neuronal, corresponden a contribuyentes que tienen un buen comportamiento tributario. Mientras que los clusters 4 y 5, que representan un 44,7% de los contribuyentes de acuerdo al SOM y un 39,1% de de acuerdo al Gas Neuronal, corresponden a contribuyentes que tienen un comportamiento tributario irregular y potencialmente asociado al uso de facturas falsas.

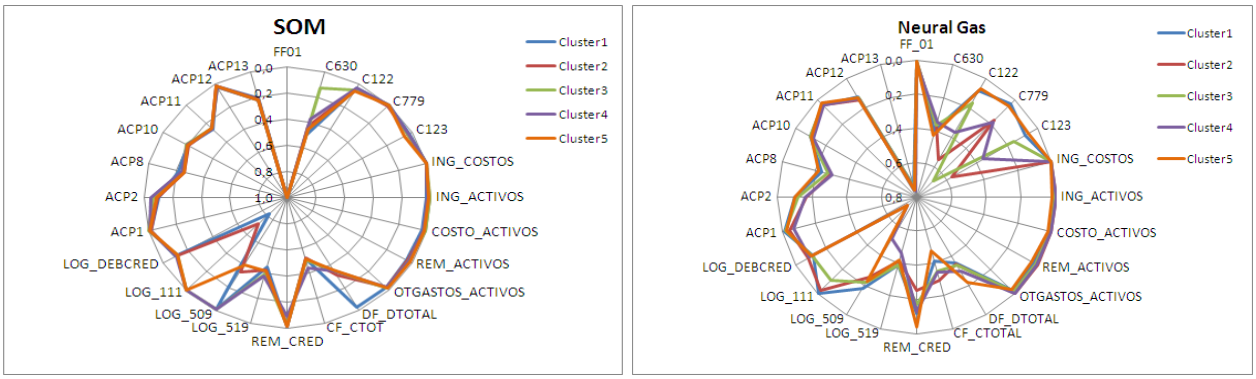
Cuadro N° 27: Tamaño Clusters Caracterización Medianas y Grandes empresas

CLUSTER	SOM	NEURAL GAS
1	30,9%	41,9%
2	14,6%	9,9%
3	11,1%	9,1%
4	16,9%	10,9%
5	27,8%	28,2%

Al igual que en el segmento de las micro y pequeñas empresas, los clusters resultantes en las medianas y grandes empresas están fuertemente influenciados por las variables que tienen un gran porcentaje de casos con valor cero, como los códigos asociados a la emisión de boletas, las notas de crédito por facturas y el nivel de costos de la empresa.

Como se aprecia en la Figura Nº 51, el SOM genera clusters diferenciados principalmente por la emisión de boletas, las notas de crédito emitidas por facturas, el porcentaje de débito por facturas respecto de los débitos totales y los costos de la empresa. El Gas Neuronal, en tanto, considera adicionalmente el porcentaje de créditos por facturas respecto de los créditos totales, los activos, los pasivos y las cuentas por pagar a empresas relacionadas.

Figura Nº 51: Comparación centros de cada cluster – Medianas y Grandes empresas



Si bien en este grupo se observan diferencias de comportamiento entre un grupo y otro, al parecer estos están influenciados en parte por el nivel de cobertura y tamaño de la empresa. En el Cuadro Nº 28 se presentan las características asociadas a un buen y mal comportamiento tributario, considerando su pago de impuestos e indicadores del comportamiento.

Cuadro N° 28: Variables asociadas a un buen y mal comportamiento de las medianas y grandes empresas

MEDIANAS Y GRANDES EMPRESAS				
VARIABLE	**PERÍODO**	**CONCEPTO**	**BUEN COMPORT.**	**MAL COMPORT.**
Remanentes	t	IVA	↑	↓
Boletas			↓	↑
Facturas de compra recibidas			↓	↑
Notas de crédito asociada a facturas				↑
Activos		Renta	↑	↓
Pasivos			↑	↓
Ratio crédito facturas/cantidad de facturas		Ratio IVA	↓	↑
Ratio ingresos del giro/activos		Ratio Renta	↓	↑
Ratio costos directos/activos			↓	↑
Nivel de cobertura	t	Características	↓	↑
Cantidad de representantes legales				↑
Actecos cambio sujeto y dificil fiscalización				↑
Irregularidades recientes	< t	Comportamiento histórico	↓	
Irregularidades por facturas y nivel de timbraje			↓	↑
Cantidad de fiscalizaciones			↑	↑
Rendimiento por fiscalizaciones				↑

4.4.2. CARACTERIZACIÓN Y DETECCIÓN DE USUARIOS DE FACTURAS FALSAS

Considerando los patrones y reglas generadas en las ramas del árbol de decisión para diferenciar entre casos de fraude y no fraude, en el Cuadro N° 29 se presentan los comportamientos principales observados en cada segmento, que resume las relaciones que generan nodos con y sin utilización de facturas falsas en el año de estudio, el período a que hace referencia cada comportamiento y el concepto asociado.

Cuadro N° 29: Comportamiento asociado a utilización de facturas falsas en las MIPE

VARIABLE		PERÍODO	CONCEPTO	NFF	FF
Remcred	Remanentes/ crédito prom.	t	Ratio IVA	↑	
CFTot	Créditos fact/ total créditos			↓	↑
CFact	Créditos fact/ fact emitidas				↑
Ctoact	Costos/activos		Ratio Renta		↑
Suma503	Facturas emitidas		IVA	↓	↑
Suma89	IVA determinado			↓	↑
Sum91	Pago de IVA			↑	↓
Suma111	Débitos boletas			↓	↑
Femtim	Facturas emitidas/ timbradas	t-2	Nivel de Timbraje	↓	↑
ACP4	Frecuencia timbraje			↓	↑
ACP10	Fiscalizaciones negativas	< t	Comportamiento histórico	↑	↓
ACP3	Fiscalizaciones positivas			↓	↑
ACP2	Delitos e irregularidades				↑
ACP5	Participación otras empresas				↑

Como se indica en el Cuadro N° 29, las micro y pequeñas empresas con resultado de fraude conocido se caracterizan por tener un mayor porcentaje de créditos asociados a facturas y mayor valor promedio de crédito por factura. Además tienen más débitos con boletas, cantidad de facturas emitidas y registran valores más altos del indicador costos/activos y del indicador facturas emitidas/facturas timbradas. Asimismo, tienen montos más altos de IVA determinado, registran menos fiscalizaciones previas con resultado negativo (improductivas) y más fiscalizaciones previas con resultado positivo (productivas), registran una mayor preponderancia de delitos e irregularidades históricas asociadas a facturas y mayor frecuencia de timbraje en los últimos dos años.

Los casos sin fraude, en cambio, se caracterizan por tener un menor porcentaje de créditos asociados a facturas, registrar más fiscalizaciones previas con resultado negativo y menor cantidad de facturas emitidas. Además, tienen un valor más bajo del indicador facturas emitidas/facturas timbradas y registran una mayor relación entre remanentes y créditos promedio.

Los resultados obtenidos para las medianas y grandes empresas se presentan en el Cuadro N° 30. En este segmento, los contribuyentes con fraude en el año de análisis se caracterizan por declarar

un monto menor de remanente acumulado en el mes anterior, declarar un mayor porcentaje de crédito asociado a facturas y crédito total, y presentar valores más altos del indicador gastos rechazados/activos y capital efectivo. . Además, tienen un mayor nivel de informalidad en su contabilidad, mayor cantidad de irregularidades previas asociadas a facturas y timbraje, mayor cantidad de irregularidades recientes y menor cantidad de fiscalizaciones previas y representantes legales.

Los casos sin fraude, en cambio, se caracterizan por declarar un monto mayor de remanente acumulado del período anterior, declarar un porcentaje menor de crédito asociado a facturas, y valores más bajos del indicador gastos rechazados/activos. Además, tienen una menor cantidad de irregularidades recientes, mayor número de representantes legales y realizan un mayor pago de IVA.

Cuadro N° 30: Comportamiento asociado a utilización de facturas falsas en las ME y GR

VARIABLE		PERÍODO	CONCEPTO	NFF	FF
Suma504	Remanentes mes anterior	t	IVA	↑	↓
Suma537	Total créditos			↓	↑
Suma520	Créditos fact. recibidas			↓	↑
Suma525	Créditos fact. activo fijo			↑	
Suma91	Pago de IVA			↑	↓
C779	Cuentas pagar emp. relacionadas		Renta	↑↓	↑↓
C102	Capital efectivo				↑
CredFactT	Créditos fact/total créditos		Ratio IVA	↓	
GtosRecAc	Gastos rechazados/ activos		Ratio Renta		↑
ACP8	Representantes legales		Características	↑	↓
ACP2	Fiscalizaciones	< t	Comportamiento histórico		↓
ACP10	Irregular. facturas y timbraje				↑
ACP12	Irregular. recientes	t-2		↓	↑

Además de caracterizar a los usuarios de facturas falsas, un aspecto esencial de este trabajo es detectar quiénes tienen una alta probabilidad de estar cometiendo fraude y evadir impuestos, a través de este mecanismo. Para ello se aplicaron tres técnicas de Data Mining de aprendizaje supervisado: árboles de decisión, redes neuronales y redes bayesianas.

Para determinar cuál de estos modelos es mejor, se comparan los resultados obtenidos en la detección de casos con y sin fraude, considerando la sensibilidad, especificidad y tasa de error de cada modelo de los grupos de testeo, que se presentan en el Cuadro N° 31. Como se observa en este Cuadro, los resultados de la detección de fraude fueron mejores en el grupo de las micro y pequeñas empresas, obteniéndose tasas de error en el rango [13% – 22%]. En el segmento de las medianas y grandes empresas, en tanto, la tasa de error se encuentra entre [18% – 30%]. Esto puede ocurrir porque en las empresas más grandes, la forma de evadir es más compleja y por lo tanto, más difícil de detectar.

En ambos segmentos, los mejores resultados de la detección de usuarios de facturas falsas se obtuvieron con el método de red neuronal. En el grupo de las micro y pequeñas empresas, el experimento Nº 2, que contempla la agrupación de casos con fraude en el año 2006 y con fraude en años previos, arrojó una detección del 92,6% para los casos con fraude y de un 72,9% para los casos sin fraude, con un error de 12,8%. Además, el poder de generalización del modelo es bastante bueno, ya que los resultados del testeo son similares a los obtenidos en el entrenamiento de la red. Por otra parte, este modelo considera un rango amplio de variables, abarcando tanto información de IVA como de ratios tributarios de Renta, además de información del comportamiento histórico de las empresas analizadas, lo que le da mayor confiabilidad al modelo.

En términos de resultados, el árbol de decisión presenta resultados similares a la red neuronal, con una asignación correcta de casos con fraude de 89% y de casos sin fraude del 79%, y un error de detección del 13%. La red bayesiana, en tanto, entrega resultados de menor calidad, con una detección del 82% de los casos con fraude y de 64% de los casos sin fraude, con un error de 22%.

Cuadro N° 31: Resultados experimentos de aprendizaje supervisado

EXP. N°	SEGMENTO	MÉTODO	SENSIBILIDAD	ESPECIFICIDAD	TASA DE ERROR
1	Micro y Peq.	Árbol de Decisión	89,0%	79,0%	13,0%
2	Micro y Peq.	Red Neuronal	92,6%	72,9%	12,8%
3	Micro y Peq.	Red Bayesiana	82,3%	64,1%	22,1%
4	Med. y Grandes	Árbol de Decisión	79,0%	85,0%	18,0%
5	Med. y Grandes	Red Neuronal	88,8%	59,1%	27,5%
6	Med. y Grandes	Red Bayesiana	73,3%	66,7%	29,7%

En el grupo de las medianas y grandes empresas, que constituye el segmento de interés principal, si bien el menor error se obtuvo con el árbol de decisión en el experimento Nº 4, el experimento Nº 5 de la red neuronal presenta una mejor detección de los casos con fraude, con una sensibilidad del 88,8%. Esto, a pesar de que presenta una detección de menor calidad de los casos sin fraude, con una especificidad de 59,1%.

Si, en cambio, interesa tener certeza de ambos grupos, vale decir empresas con y sin fraude, el árbol de decisión también entrega buenos resultados, con mayor certeza de los casos sin fraude tributario. Por otra parte, en el experimento Nº 6 que aplica una red bayesiana, se obtuvieron resultados más equitativos entre la predicción de casos con y sin fraude, correspondientes a 73% y 66% respectivamente, con una tasa de error similar al experimento Nº 5. En este caso, las variables más relevantes para la predicción consideran tanto información de comportamiento como de delitos de los relacionados y cantidad de representantes legales de la empresa, así como variables relacionadas al pago de IVA y ratios tributarios de renta.

Al analizar el grado de relevancia de las variables seleccionadas con cada modelo utilizado para la detección de utilización de facturas falsas, se observan diferencias respecto del número de variables clasificadas como relevantes. En parte, esto se explica por el tipo de relaciones que se generan en cada uno de ellos, lo que produce que algunos le den más importancia al ratio entre dos variables, mientras que otros trabajan directamente con las variables que componen los ratios.

En el Cuadro Nº 32 se presenta un resumen de las variables más relevantes determinadas por cada modelo, presentando sólo aquellas que tienen algún grado de coincidencia, es decir, que fueron determinadas como relevantes en más de un modelo.

Cuadro N° 32: Coincidencia de variables- modelos de uso facturas falsas en las MI y PE empresas

Variable	Concepto	Período	Árbol Decisión	Red Neuronal	Red Bayesiana
Log_Remcred	Relación entre remanentes y créditos	t	x	x	x
Log_Suma503	Cantidad facturas emitidas	t	x	x	x
Log_Suma111	Débitos por boletas	t	x	x	x
ACP3	Fiscalizaciones previas con resultado positivo	< t	x	x	x
Log_Cfact	Crédito por factura	t	x		x
Log_Cftot	Crédito por facturas respecto del crédito total	t	x		x
Log_Ctoact	Relación entre costos y activos	t	x	x	
Log_Ingact	Relación entre ingresos y activos	t		x	x
Log_IvaIng	Relación entre IVA e ingresos	t		x	x
Log_Suma502	Débito por facturas	t		x	x
Log_Suma509	Cantidad notas de crédito por facturas	t		x	x
Log_Suma89	IVA determinado	t	x		x
Log_Suma91	Total IVA a pagar	t	x		x
ACP14	Nivel de formalidad de la empresa	t		x	x
ACP4	Frecuencia de Timbraje	t-2	x	x	
ACP10	Fiscalizaciones previas con resultado negativo	< t	x		x

De acuerdo a lo anterior, las variables que dan cuenta de la relación entre remanentes y créditos, la cantidad de facturas emitidas, los débitos por boletas y las fiscalizaciones previas con resultado positivo (casos productivos para el SII) son catalogados como relevantes en los tres modelos generados para el segmento de las micro y pequeñas empresas. Otras variables relevantes fueron el crédito promedio por factura, el porcentaje que representan los créditos por facturas respecto del crédito total, los débitos por facturas, el pago de IVA, y su relación con los ingresos de la empresa y los activos. En términos de comportamiento, son relevantes también el nivel de formalidad de la empresa, la frecuencia del timbraje y las fiscalizaciones previas con resultado negativo.

En el segmento de las medianas y grandes empresas, la cantidad de variables catalogadas como relevantes y el grado de coincidencia es menor. Es así como sólo la variable que indica la cantidad de créditos por facturas de activo fijo es clasificada como relevante en los tres modelos, como se indica en el Cuadro Nº 33. En menor grado, son también relevantes las variables de comportamiento que indican que el contribuyente tiene antecedentes previos de término de giro y de no ubicado, irregularidades previas asociadas a facturas e irregularidades recientes y la cantidad de representantes legales activos e inactivos. Respecto del pago de impuestos, son relevantes la relación entre débitos y créditos, el porcentaje de créditos asociados a facturas respecto del total de créditos, los remanentes del mes anterior, las cuentas por pagar a empresas relacionadas y el total de pasivos, que indican el grado de endeudamiento de la empresa.

Cuadro N° 33: Coincidencia de variables- modelos de uso facturas falsas en las ME y GR empresas

Variable	Concepto	Período	Árbol Decisión	Red Neuronal	Red Bayesiana
Log_525	Crédito por facturas de activo fijo	t	x	x	x
ACP14	Antecedentes de término de giro y no ubicado	< t	x	x	
ACP10	Irregularidades de facturas y nivel de timbraje	< t	x	x	
ACP12	Irregularidades recientes	t-2	x		x
ACP8	Representantes legales	t	x		x
CredFact_Tot	Crédito por facturas respecto del crédito total	t	x	x	
Debcred	Relación entre débitos y créditos	t		x	x
Log_Suma504	Remanente de créditos del mes anterior	t	x		x
Log_Suma537	Total crédito	t	x	x	
Cod_779	Cuentas por pagar a empresas relacionadas	t	x	x	
Cod_123	Total pasivos	t		x	x

De lo anterior, se puede concluir que las variables de comportamiento son relevantes para determinar quiénes son usuarios potenciales de facturas falsas en un año determinado, ya que contienen la historia del comportamiento del contribuyente desde sus inicios. Por otro lado, la información del pago de IVA también es fundamental para encontrar usuarios potenciales de facturas falsas, principalmente en las micro y pequeñas empresas. En las medianas y grandes empresas, en tanto, la información del impuesto a la renta adquiere mayor relevancia.

En relación al ingreso potencial obtenido con las auditorías de los contribuyentes correctamente detectados como fraude o evasores de impuesto, y los costos asociados a sus auditorías, el beneficio neto por contribuyente se puede estimar como:

Cuadro N° 34: Matriz de Beneficio Neto por contribuyente

		VALOR PREDICCIÓN	
		FF=1	FF=0
VALOR REAL	FF=1	(Rendimiento de la Auditoría – Costo de la Fiscalización)	(Estimación Rendimiento no Percibido + Deterioro de la Imagen)
	FF=0	(Costo de la Fiscalización)	Mejora Percepción de la Imagen

El rendimiento de la auditoría se obtiene considerando el promedio del impuesto evadido por los contribuyentes de acuerdo a la tasa de evasión del IVA en el año 2006, o considerando el rendimiento promedio de las auditorías selectivas realizadas en ese mismo año que fueron productivas. Para ello hay que tener en cuenta que en las micro y pequeñas empresas es común la aplicación de programas masivos de IVA, los cuales buscan abarcar un gran número de contribuyentes en poco tiempo. Por otra parte, en las empresas medianas y grandes es más común la aplicación de programas selectivos, que poseen una mayor profundidad, debido a la magnitud de documentos y la complejidad de éstos casos.

De acuerdo a lo anterior, se estima que el rendimiento de una auditoría masiva realizada a una micro y pequeña empresa es cercano a los $200.000, mientras que el rendimiento promedio de una fiscalización selectiva aplicada a una mediana empresa es de $3.600.000, monto que aumenta a $20.000.000[44] en el caso de las grandes empresas.

El costo de la fiscalización, en tanto, se puede estimar directamente como el costo asociado a las horas requeridas por el fiscalizador para realizar la auditoría, y el costo ponderado de los estamentos y grados asociados a esta función. Un auditor con escalafón de fiscalizador grado 14 ó 15, tiene un valor hora aproximado de $8.750, mientras que un jefe de grupo con escalafón de fiscalizador grado 12 ó 13, tiene un valor hora de $10.929[45]. De acuerdo a esto, el costo promedio de una fiscalización a una micro y pequeña empresa es de $50.000, mientras que para una mediana empresa este costo aumenta a $900.000[46] y a $5.000.000[47] para una gran empresa.

Respecto del deterioro o mejora de la imagen, no existen estimaciones que den cuenta de cómo afecta el resultado de la auditoría a la percepción que el contribuyente tiene del Servicio, razón por la cual, sólo se expresará a modo teórico, asignándole un valor igual a cero para efecto de la estimación del beneficio neto.

De esta forma, la matriz de beneficio neto por unidad de contribuyente para una micro y pequeña empresa queda de la siguiente forma:

Cuadro N° 35: Matriz de Beneficio Neto por contribuyente cuantificada para una mipyme

		VALOR PREDICCIÓN	
		FF=1	**FF=0**
VALOR REAL	**FF=1**	150.000 = 200.000 – 50.000	-200.000
	FF=0	-50.000	0

De acuerdo a esta matriz y a los resultados obtenidos en experimento Nº 2 con la red neuronal, el beneficio promedio por contribuyente a quien se detecta correctamente fraude tributario hubiera sido de $86.282. Considerando además que la capacidad predictiva del modelo es del orden del 90%, en el grupo de testeo se genera un beneficio de $22,6 millones de pesos.

En el grupo de las medianas y grandes empresas, se tiene que aproximadamente un 65% son empresas medianas y el 35% restante corresponde a grandes empresas. Considerando los ingresos y costos de las auditorías realizadas a cada segmento y los porcentajes de participación de cada grupo, la matriz de beneficio neto promedio para las medianas y grandes empresas queda de la forma:

[44] Estimaciones proporcionadas por el Área de Riesgo de la Subdirección de Fiscalización del Servicio de Impuestos Internos.
[45] Dato extraído de los sueldos brutos publicados en el sitio web del SII, www.sii.cl.
[46] Considera que un auditor destina un promedio de 100 horas y un jefe de grupo 10 horas selectivas.
[47] Considera que un auditor destina un promedio de 450 horas, un abogado 50 horas y un jefe de grupo 50 horas.

Cuadro N° 36: Matriz de Beneficio Neto por contribuyente para una mediana y gran empresa

		VALOR PREDICCIÓN	
		FF=1	FF=0
VALOR REAL	FF=1	7.005.000	-9.340.000
	FF=0	-2.335.000	0

De acuerdo a esta matriz y a los resultados obtenidos en experimento Nº 2 con la red neuronal, el beneficio promedio por contribuyente a quien se detecta correctamente fraude tributario es de $1.245.333. Considerando además que la capacidad predictiva del modelo es del orden del 84%, en el grupo de testeo se genera un beneficio de $53,549 millones de pesos.

No existen estimaciones del porcentaje ni del monto de evasión diferenciado por tipo de contribuyente, sin embargo, se han realizado algunos estudios de imagen institucional que indican que aproximadamente un 20% de los contribuyentes han utilizado facturas para evadir impuesto[48]. Por otra parte, un estudio realizado el año 2007, indica que aproximadamente un 14,8% de los contribuyentes son "caradura" o propensos a cometer fraude, mientras que un 64,7% es propenso a cometer fraude si se presenta la ocasión, y un 20,5% tiene buen comportamiento y es menos propenso a cometer fraude.

Considerando que el universo de micro y pequeñas empresas activas en el año 2006 es del orden de las 580.000 empresas y que existen aproximadamente 23.842 medianas y grandes empresas activas en ese mismo año, existe un universo potencial de 120.000 contribuyentes que utilizan facturas para evadir impuestos. Esto demuestra la necesidad de mejorar los sistemas de selección de contribuyentes para fiscalización, ya que debido a la disponibilidad limitada de recursos, no es posible fiscalizarlos a todos.

4.5. PROPUESTA DE METODOLOGÍA PARA FISCALIZACIÓN

Con la finalidad de probar la capacidad predictiva real de los modelos desarrollados, que determinan el conjunto de empresas que tienen una alta probabilidad de estar utilizando facturas falsas en un cierto año, se considera absolutamente necesario aplicar un programa de fiscalización piloto, que permita determinar en terreno el nivel de acierto en la clasificación de los contribuyentes catalogados como potenciales defraudadores de impuestos.

Para lo anterior, se propone aplicar una combinación de los modelos predictivos generados, para detectar quiénes tienen una mayor probabilidad de estar utilizando facturas falsas en un año determinado. Esto, con la finalidad de priorizar un conjunto de empresas a fiscalizar en cada segmento, seleccionando sólo los casos que tienen mayor probabilidad de estar cometiendo fraude, sin necesidad de fiscalizar a todos los contribuyentes que aparezcan como usuarios de facturas falsas, tarea que puede resultar infactible en la práctica, ya sea por restricciones de presupuesto, tiempo y/o horas de trabajo disponibles de los fiscalizadores.

[48] Estudio "Imagen del SII" respecto de las formas más comunes de evasión, realizado en 1996 por la empresa Mori.

En la realización de este programa piloto, debe tenerse en cuenta al menos cuatros aspectos:

i. Cómo se determina a quiénes fiscalizar
ii. A quiénes seleccionar para fiscalización.
iii. Cuál es la hipótesis de evasión y qué ítems revisar.
iv. El período de impuestos a revisar.

El primer punto se relaciona con la *determinación de la probabilidad de fraude*, que está dada por la combinación de los modelos desarrollados, los que pueden ser aplicados en cada segmento de manera independiente. Como primer paso, se debe establecer el año de análisis en el que se desea encontrar contribuyentes que utilizan o venden facturas falsas. Posteriormente, es necesario actualizar el vector de características utilizado en los modelos. Para ello, debe realizarse lo siguiente:

- Reunir el conjunto de contribuyentes que presentan al menos una declaración de IVA en el año e identificar a qué segmento pertenecen.
- Construir las variables de impuestos y ratios tributarios, la cual es directa, pues en el caso de la Renta y el IVA se utiliza información de los códigos declarados en los formularios de cada impuesto en el año. En el caso de las variables de comportamiento, existen algunas que se determinan según el proceder del año y otras que consideran información histórica, para las cuales se recomienda actualizar las variables ya construidas, incorporando sólo la información de los últimos años. En el caso de las variables de comportamiento se requiere calcular además las componentes principales asociadas.
- Realizar una limpieza de los datos, extrayendo los outliers (valores que superan la media más 5 veces la desviación estándar) y los datos inconsistentes (casos en que las sumas de determinados códigos no coincidan con el código final).
- Aplicar la transformación logaritmo a las variables que lo requieran.
- Reemplazar valores nulos de ingresos y costos de renta, con información de ventas y compras de IVA.
- Normalizar las variables mediante la norma Min-Max (entre 0 y 1).
- Aplicar los modelos y obtener la predicción de casos con Facturas Falsas en cada uno de ellos para cada segmento.
- Combinar las probabilidades obtenidas en cada modelo, ya sea en forma lineal o asignado mayor preponderancia a los modelos más asertivos (red neuronal y árbol de decisión).

Los siguientes puntos se relacionan con el *impacto de la respuesta esperada*, el cual depende de los intereses del SII. Por ejemplo, se propone utilizar alguno(s) de los siguientes criterios: fiscalizar a aquellos que pueden generar un mayor rendimiento, que tienen un menor costo (por ejemplo, de transporte), que tienen mayor crédito, mayor cantidad de facturas de compra, que pertenecen a determinados sectores económicos o que en el tiempo han tenido menor cobertura de fiscalización. Lo anterior, considerando que la efectividad de un plan de fiscalización depende tanto de la probabilidad de éxito (casos productivos/casos totales) como del rendimiento promedio por fiscalizado productivo.

La selección de los casos para fiscalización, dependerá también del alcance del piloto. Si tendrá un alcance nacional o se elegirá un determinado lugar geográfico para realizarlo, así como de la cantidad de recursos disponibles en ese determinado lugar, principalmente las horas de trabajo disponibles de los fiscalizadores. Independiente de los criterios utilizados, se recomienda elegir los casos de mayor probabilidad de fraude diferenciado por segmento, debido a que generalmente la mayor respuesta se tiene para los contribuyentes de mayor tamaño, dejando de lado a contribuyentes fraudulentos de menor tamaño.

Para tener más información de lo anterior, se puede hacer un análisis del rendimiento obtenido en los casos catalogados como fraude en el análisis, de manera de determinar en cuáles de ellos se obtuvo el mejor resultado en términos de recaudación, información que no se tenía disponible al momento de realizar el estudio. Debido a que el rendimiento depende del tamaño del contribuyente, se puede hacer un análisis del rendimiento obtenido en relación a su nivel de ingresos o su pago de impuestos, y/o desagregado por sector económico o ubicación geográfica.

El tercer punto, no menor, es la *hipótesis de evasión* y la información que se debe revisar en la auditoría, que es fundamental para la revisión que deben realizar los fiscalizadores durante la visita al contribuyente y la generación de los documentos de apoyo. Debido a que los comportamientos detectados se focalizan en la detección de contribuyentes con facturas falsas, la auditoría debe enfocarse en lo que se revisa en un programa normal de revisión de facturas. No obstante, se pueden entregar algunos indicadores adicionales al fiscalizador, como el ratio anual facturas emitidas/facturas timbradas, el porcentaje de créditos asociados a facturas respecto de los créditos totales o el crédito promedio por factura. De igual forma, se puede indicar si alguno de sus relacionados, directos o indirectos, ha tenido irregularidades asociadas a facturas, para focalizar la revisión en facturas de esos relacionados.

En principio, el *período de revisión* de los documentos corresponde a un año como mínimo, dependiendo del tipo de reglas que se utilicen. Si, por ejemplo, se utilizan las reglas generadas con el árbol de decisión, se puede obtener la probabilidad de fraude en el año (FF=2) y/o la probabilidad de fraude en el año anterior (FF=1). Por otra parte, la red neuronal y la red bayesiana agrupan ambos comportamientos en uno, razón por la cual puede ser conveniente revisar los documentos del año de análisis y el año anterior.

Finalmente, al término de la aplicación del piloto es importante calcular el porcentaje de aciertos del total de casos fiscalizados en los que se encuentran efectivamente Facturas Falsas, para determinar la efectividad de los modelos. De igual forma, es importante continuar recopilando información de contribuyentes fiscalizados en los que se detectan o no facturas falsas, ya que esto permitirá ir actualizando los parámetros de los modelos, considerando cambios de comportamiento en el tiempo, para lo cual resulta fundamental registrar información del período en el cual se detecta el fraude.

5. CONCLUSIONES Y TRABAJO FUTURO

La utilización y venta de facturas falsas tiene un impacto significativo en la recaudación que percibe el Estado para financiar sus proyectos, además de los efectos económicos negativos que genera al poner en riesgo la competitividad de las empresas. En el año 2009, se estima que la evasión por concepto de IVA fue de $1,5 billones de pesos, de los cuales aproximadamente un 30% corresponde a evasión mediante utilización de facturas falsas. Esto equivale a $450 millones de pesos, cifra que puede aumentar a $600 millones de pesos de acuerdo a las últimas estimaciones de evasión realizadas[49]. Adicionalmente, la detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, provoca un importante costo administrativo para las Unidades Operativas, lo que da cuenta de la relevancia que tiene focalizar los esfuerzos en la detección de casos de evasión y fraude fiscal.

Los métodos utilizados para caracterizar a los contribuyentes que tienen buen o mal comportamiento tributario de acuerdo con información de su pago de impuestos y variables asociadas a la utilización de facturas falsas, indican que es posible identificar algunas características diferenciadoras de buen y mal comportamiento tributario. Particularmente, el "método de Kohonen" obtiene patrones de comportamiento relacionados con su pago de impuestos, donde las variables con mayor cantidad de ceros resultaron ser las que más impacto tuvieron en la conformación de los grupos, dándole menor importancia a las variables de comportamiento tributario. En el caso del "método de gas neuronal" los grupos formados también se encuentran influenciados por el pago de impuestos, pero con mayores diferencias en las variables de comportamiento.

Es así como en el segmento de las micro y pequeñas empresas se detectan 6 clusters, estimándose que aproximadamente un 45% tiene un buen comportamiento tributario y un 28% tiene un comportamiento irregular. En el segmento de las medianas y grandes empresas, en tanto, se detectan 5 clusters, de los cuales un 26% tiene buen comportamiento y un 39% tiene un comportamiento de tipo irregular. Al analizar la distribución de los casos conocidos de fraude por facturas falsas en cada cluster, se observa que existe una alta probabilidad de encontrar casos con y sin fraude en un mismo grupo, razón por la cual sólo es posible distinguir algunas características que señalan a los contribuyentes que tienen un mejor o peor comportamiento tributario respecto de otro, más que identificar clusters de buen y mal comportamiento tributario.

Posteriormente, el "método de árboles de decisión" aplicado a los casos en el que el resultado de fraude y no fraude era conocido resulta ser una buena técnica para detectar variables que permiten distinguir entre casos de fraude y no fraude. Esto, debido que al analizar la distribución de las variables en cada grupo, se observa que los casos con fraude tienden a tomar valores más extremos de las variables, lo que permite distinguir rangos a partir de los cuales existe una probabilidad de tener o no tener facturas falsas. Por otro lado, los resultados obtenidos fueron coherentes con lo observado en la realidad, de acuerdo a la vista experta.

[49] La evasión en el IVA estimada para el año 2009 es de $1,5 billones de pesos con una tasa de evasión del 18%. Sin embargo, este monto aumenta a $2 billones de pesos con una tasa de evasión el 23%, de acuerdo a la actualización de la tasa de evasión realizada en 2012, que considera la actualización de la Matriz Insumo Producto de 2008, cifra que tiene carácter provisional.

En el caso de las micro y pequeñas empresas, las variables que permiten distinguir entre fraude y no fraude se relacionan principalmente con el porcentaje de créditos generado por facturas respecto del crédito total, y las fiscalizaciones previas con resultado negativo. En la medida que el contribuyente fue fiscalizado más veces en el pasado y no se encontró nada, es más probable que no tenga fraude en el futuro. Por otro lado, mientras su crédito esté más asociado a otros ítems distintos a las facturas (activo fijo u otros), es menos probable que utilice facturas para respaldar sus créditos. Otras variables relevantes son la cantidad de facturas emitidas en el año y su relación con las facturas timbradas en los últimos dos años, el monto de IVA total declarado en el año, la relación entre remanentes y créditos promedio, las fiscalizaciones previas con resultado positivo, los delitos e irregularidades históricos asociadas a facturas, y la participación en otras empresas. Por otra parte, en las medianas y grandes empresas, las variables más relevantes para distinguir entre casos de fraude y no fraude son la cantidad de remanente acumulado en los períodos anteriores, el total de créditos, el porcentaje de crédito asociado a facturas y el pago de IVA, la relación entre gastos rechazados y activos, el capital efectivo, así como la cantidad de irregularidades recientes e irregularidades previas asociadas a facturas, la cantidad de fiscalizaciones históricas, y la cantidad de representantes legales.

En relación a los modelos de detección, los que tienen mejor desempeño son los modelos de red neuronal de perceptrón multicapa, que para efectos del estudio cuentan con una capa de entrada que contiene las variables explicativas, una capa intermedia de procesamiento y una capa de salida (FF=1, FF=0). En el caso de las micro y pequeñas empresas la sensibilidad del modelo es de un 92%, mientras que la especificidad es de un 72%. En el segmento de las medianas y grandes empresas, en tanto, el nivel de detección disminuye, obteniendo redes que predicen bien en un 89% de los casos con fraude, pero no tan bien los casos sin fraude, alcanzando sólo un 59% de certeza.

Los resultados obtenidos con el árbol de decisión arrojan resultados similares, con una sensibilidad del 89% en el caso de las micro y pequeñas empresas y de un 79% en el caso de las medianas y grandes empresas, y una especificidad del 79% y 85% en el primer y segundo grupo respectivamente, detectando con mayor nivel de certeza los casos sin fraude respecto de la red neuronal. Los resultados de la red bayesiana, en tanto, son de menor calidad, con una sensibilidad del 82% en el segmento de las micro y pequeñas empresas y de 73% en el segmento de las medianas y grandes empresas, con una especificidad del 64% y el 67% en cada grupo.

Dado lo anterior, y considerando que en la práctica sólo es posible fiscalizar a un grupo más bien reducido de empresas en un año, se recomienda realizar una combinación de los tres modelos predictivos generados, o en su defecto, utilizar las reglas generadas por el árbol de decisión, debido a su nivel de certeza y simplicidad, de manera de seleccionar para fiscalización a aquellos que tienen las probabilidades más altas de cometer fraude por facturas falsas.

En términos de recaudación, la predicción de un caso de fraude en una micro y pequeña empresa aporta un beneficio neto de $86.282, mientras que para una mediana y gran empresa, esta cifra aumenta a $1.245.333, lo que permitiría reducir la evasión por concepto de IVA, si se consideran el total de casos auditados en un año.

A modo general, se observa que las variables de comportamiento que contienen la historia del contribuyente son relevantes para determinar empresas que potencialmente utilizan facturas falsas para evadir sus impuestos. De igual manera, los ratios tributarios que relacionan información de IVA y Renta o dos códigos del mismo impuesto, son más efectivos que la información que proporciona cada código por separado. Por otro lado, la información relacionada con el pago de IVA es relevante al momento de determinar la probabilidad de evasión en el año de estudio para cualquier tipo de empresa, mientras que la de Renta impacta principalmente a las empresas de mayor tamaño, que en general disponen de mecanismos de evasión más sofisticados.

A partir de los resultados de esta investigación, y con el objeto de profundizar aún más en la problemática planteada, se hace oportuno e importante mencionar que resulta necesaria la formulación de estándares en la recopilación de información, que permitan continuar la senda de investigación de este tipo de evasión u otro. Esto, debido a la complejidad de recopilar información de casos catalogados como fraude o sin fraude en el año de estudio.

De igual forma, para probar la capacidad predictiva real del modelo desarrollado, resulta vital su aplicación en actividades que permitan determinar en terreno el nivel de acierto en la clasificación de los contribuyentes seleccionados en la muestra, para lo cual se recomienda la implementación de un programa piloto que esté dirigido a los dos segmentos económicos estudiados, que será concluyente en términos de la efectividad real del modelo.

Para trabajos futuros, se recomienda generar nuevas variables de comportamiento históricas relacionadas con fiscalizaciones y cobertura, así como explorar otros métodos para el preprocesamiento y selección de las variables, con los que eventualmente podrían obtenerse resultados diferentes. Igualmente, sería interesante explorar técnicas de validación cruzada y aplicar otras técnicas de data mining para mejorar la predicción de casos de fraude.

Finalmente, se concluye que mediante los métodos de Data Mining y el vector de características utilizado, es posible caracterizar y detectar contribuyentes que evaden impuestos a través de facturas falsas en un año dado. De esta manera, se contribuye a desarrollar nuevas formas de selección de casos para auditoría, generando una nueva visión para potenciar la efectividad de la acción fiscalizadora del Servicio de Impuestos Internos.

6. BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN

1. AIROLDI E. y MALIN B., “Data Mining Challenges for Electronic Safety: The Case of Fraudulent Intent Detection in E-Mails”. Proceedings of the Workshop on Privacy and Security Aspects of Data Mining, 2004.

2. AUSTRALIAN NATIONAL AUDIT OFFICE, “The Australian Taxation Office`s Use of Data Matching and Analytics in Tax Administration”. Audit Report Nº 30 2007-2008, Performance Audit, Abril 2008.

3. BANCO INTERAMERICANO DE DESARROLLO (BID). ARNÁIZ T., GARCÍA J.A., LÓPEZ J.M., “Los Planes Integrales para la Prevención y Corrección del Fraude Fiscal”, Marzo 2006.

4. AYUSO M. y GUILLÉN M., “Errores de Respuesta en la Clasificación de Siniestros Fraudulentos en el Seguro de Automóviles”. Universidad de Barcelona, Junio 2000.

5. BALDOCK T, “Insurance Fraud”. Trends and Issues in Crime and Criminal Justice. Australian Institute of Criminology”, Febrero 1997.

6. BOLTON R. y HAND D., “Unsupervised Profiling Methods for Fraud Detection”. Department of Mathematics Imperial College, London, 2001.

7. BOLTON R. y HAND D., “Statistical Fraud Detection: A Review”. Statistical Science, Vol. 17- N° 3, 2002.

8. BRAUSE R., LANGSDORF T. y HEPP M., “Neural Data Mining for Credit Card Fraud Detection”, Interner Bericht, 1999.

9. BRAVO J. 2009, “Aplicación de Redes Neuronales Artificiales en el Proceso de Selección de Contribuyentes a Fiscalizar por Utilización de Facturas Falsas en su Contabilidad”. Tesis para optar al título de Ingeniero Civil Industrial, Universidad de Santiago de Chile, 2009.

10. BRAITHWAITE V. “Tax Evasion”, in Michael Tonry (Ed.), The Oxford Handbook of Crime and Public Policy, Oxford: Oxford University Press, pp. 381-405. 2009.

11. BURGE P. y TAYLOR J, “Frameworks for Fraud Detection in Mobile Telecommunications Networks”, 1997.

12. BURGE P. y TAYLOR J., “An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users to use in Fraud Detection”. Journal of Parallel and Distributed Computing, 2001.

13. CAHILL M., LAMBERT D., PINHEIRO J. y SUN D., “Detecting Fraud in the Real World”, 2000.

14. CHAN P., FAN W., PRODROMIDIS A. y STOLFO S., “Distributed Data Mining in Credit Card Fraud Detection”, Computer Science, Florida Institute of Technology, IEEE Intelligent Systems’ Special Issue on Data Mining, 1999.

15. CLIFTON P. y CHUN W., “Investigative Data Mining in Fraud Detection”. School of Business Systems, Monash University, Noviembre 2003.

16. DAVIA H. R., COGGINS J.W. y KASTANTIN J., “Accountant’s Guide to Fraud Detection and Control (2da edición)”, 2000.

17. DE MOYA M.E. y NIÑO VASQUEZ L.F., “Representación y Clasificación de Datos Geoespaciales usando Redes Neuronales”, Universidad Nacional de Colombia, 2005.

18. DERRIG R., “Insurance Fraud”. The Journal of Risk and Insurance, 2002, Vol.69. N° 3.

19. DIGIMPIETRI L., TREVISAN N., MEIRA L., JAMBEIRO J., FERREIRA C. y KONDO A., “Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System”. Proceedings of the 9th Annual International Digital Government Research Conference, 2008.

20. DORRONSORO J. y SANTA CRUZ C., “Discrimination of Overlapping Data and Credit Card Fraud Detection”, Department of Computer Engineering and Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, 1997.

21. ESCOBAR M., “Las aplicaciones del análisis de segmentación: El procedimiento CHAID”. Instituto Juan March de Estudios e Investigaciones, Universidad de Salamanca. Working Paper, Enero 1992.

22. GAO, UNITED STATES GOVERNMENT ACCOUNTABILITY OFFICE. “Lessons Learned from Other Countries on Compliance Risks, Administrative Costs, Compliance Burden and Transition”, Report to Congressional Requesters, Abril 2008.

23. GAO, UNITED STATES GOVERNMENT ACCOUNTABILITY OFFICE. “Data Mining: Agencies have taken key steps to protect privacy in selected efforts, but significant Compliance Issues Remain”, Mayo 2004.

24. GIANNICO F., “Buenas Prácticas para la Segmentación Comportamental de Clientes en la Industria de Telecomunicaciones. III Jornadas de Data Mining, Agosto 2008.

25. GONZÁLEZ F. y DASGUPTA D., “Anomaly Detection Using Real Valued Negative Selection”. Kluwer Academic Publisher. Netherlands, 2004.

26. GRAMATICOV M., “Data Mining Techniques and the Decision Making Process in the Bulgarian Public Administration”, 2003.

27. GROSSER H., BRITOS P., SICRE J., SERVETTO A., GARCÍA MARTÍNEZ R. y PERICHINSKY G., “Detección de Fraude en Telefonía Celular usando Redes Neuronales”, Facultad de Ingeniería, Universidad de Buenos Aires, Facultad de Informática, Universidad Nacional de la Plata, Buenos Aires, Instituto Tecnológico de Buenos Aires, 2003.

28. HENRÍQUEZ R., “Mapas Temporales mediante Redes Neuronales Auto Organizativas”. Tesis para optar al grado de Magíster en Gestión de Operaciones y al título de Ingeniero Civil Electricista, Universidad de Chile, 2008.

29. HARRISON G. y KRELOVE R., “VAT Refunds: A Review of Country Experience”. IMF Working Paper, Noviembre 2005.

30. JANS M., LYBAERT N. y VANHOOF K., “Data Mining as a Methodology for Internal Fraud Risk Reduction”. Journal of Information Systems, 2008.

31. JUNGWON K. y BENTLEY P., “An Evaluation of Negative Selection in an Artificial Immune System for Network Intrusion Detection”. Londres, 2001.

32. KAGAN, R., “On the Visibility of Income Tax Law Violations” en “Taxpayer Compliance”, Vol 2, University of Pennsylvania Press, 1989.

33. FAN W. “Systematic Data Selection to Mine Concept Drifting Data Stream”, Proceedings of SIGKDD04, 2004.

34. FAWCETT T. y PROVOST F. “Adaptive Fraud Detection”. Data Mining and Knowledge Discovery, 1997.

35. FAWCETT T. y PROVOST F. “Activity Monitoring: Noticing Interesting Changes in Behaviour”. Proceedings’ on the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego. CA, 1999.

36. LÓPEZ DE CASTILLA C., “Clasificadores por Redes Bayesianas”. Tesis para optar al título de Maestro en Ciencias de Matemática, Universidad de Puerto Rico, 2005.

37. LUCKEHEIDE S., “Segmentación de los Contribuyentes que declaran IVA utilizando técnicas de Data Mining”. Tesis para optar al título de Ingeniero Civil Industrial, Universidad de Chile, Abril 2007.

38. LUNDIN E., KVARNSTROM H. y JONSSON E., “Synthesizing Test Data for Fraud Detection Systems”, Department of Computer Engineering, Chalmers University of Technology, Sweden. IEEE Computer Society, 2003.

39. MAES S., TUYLS K., VANSCHOENWINKEL B. y MANDERICK, “Credit Card Fraud Detection using Bayesian and Neural Network”. Proc. Of the 1st International NAISO Congress on Neuro Fuzzy Technologies, Enero 2002.

40. MAZHELIS O., PUURONEN S., “Combining One-Class Classifiers for Mobile User Substitution Detection”. International Conference on Enterprise Information Systems, 2004.

41. MUNOZ DELIA J., “Proceso de Reconocimiento de Objetos asistido por computador, aplicando Gases Neuronales y técnicas de Minería de Datos”. Scientia et Technica Año XII, No 30, Mayo de 2006.

42. MUÑOZ GAHETE C. y MARTÍN JIMÉNEZ J., “Evaluación de Clasificadores basados en Redes de Función de Base Radial”. Universidad Carlos III de Madrid, 2007.

43. MYATT GLENN J., “Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining”. Wiley Interscience, 2007.

44. OBERREUTER G., L’HUILLIER G. RIOS S.A Y VELÁSQUEZ J.D., “Approaches for Intrinsic and External Plagiarism Detection: Notebook for PAN at CLEF 2011”.

45. OECD, “Compliance Measurement, Practice Note”. Centre for Tax Policy and Administration, Tax Guidance Series. General Administrative Principles – GAP004 Compliance Measurement, Junio 1999.

46. OECD, “Compliance Risk Management, Use of Random Audit Programs”. Forum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. Septiembre, 2004.

47. OECD, “Compliance Risk Management, Audit Case Selection Systems”. Forum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. Octubre, 2004.

48. PINHEIRO C., EVSUKOFF A. y EBECKEN N, “Revenue Recovering with Insolvency Prevention on a Brazilian Telecom Operator”. ACM SIGKDD Explorations, Volumen 8, Issue 1, 2006.

49. PRIETO RODRÍGUEZ J., SANZO PÉREZ M. J. y SÚAREZ PANDIELLO J, “Análisis Económico de la actitud hacia el Fraude Fiscal en España”. Hacienda Pública Española, Revista de Economía Pública, Instituto de Estudios Fiscales, Abril 2006.

50. SALVATORE S., DAVID F., LEE W., PRODROMIDIS A. y CHAN P.K., “Credit Card Fraud Detection Using Meta Learning Issues and Initial Results”, 2000.

51. SERRA P. y TORO J., “¿Es Eficiente el Sistema Tributario Chileno?”, Cuadernos de Economía, Nº 94, Pontificia Universidad Católica de Chile, 1994.

52. SPARROW M., “Health Care Fraud Control Understanding The Challenge”. Journal of Insurance Medicine. Vol. 28, N° 2, 1996.

53. SPSS INC., “Clementine 12.0 Algorithms Guide” y “Neural Networks 12.0# 2007.

54. STOLFO S., LEE W. y MOK K., “Mining in a Data-Flow Environment: Experience in Network Intrusion Detection”. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 1999.

55. SUPERINTENDENCIA NACIONAL DE ADMINISTRACIÓN TRIBUTARIA (SUNAT), “La Gestión de la Sunat en los últimos cinco años: Principales Avances y Desafíos”, 2006.

56. TANZI V. y SHOME P., “Tax Evasion: Causes, Estimation Methods, and Penalties a Focus on Latin America”, Documento elaborado para el Proyecto Regional de Política Fiscal CEPAL/PNUD, 1993.

57. VELASCO D., “Redes Bayesianas”. Inteligencia Artificial II, 2007.

58. VELÁSQUEZ J. y PALADE V., “Adaptive Web Sites: A Knowledge Extraction from Web Data Approach”. Frontiers in Artificial Intelligence and Applications, Volumen 170, 2008.

59. WEISBERG H. y DERRIG R., “Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims”. Marzo, 1998.

7. ANEXOS

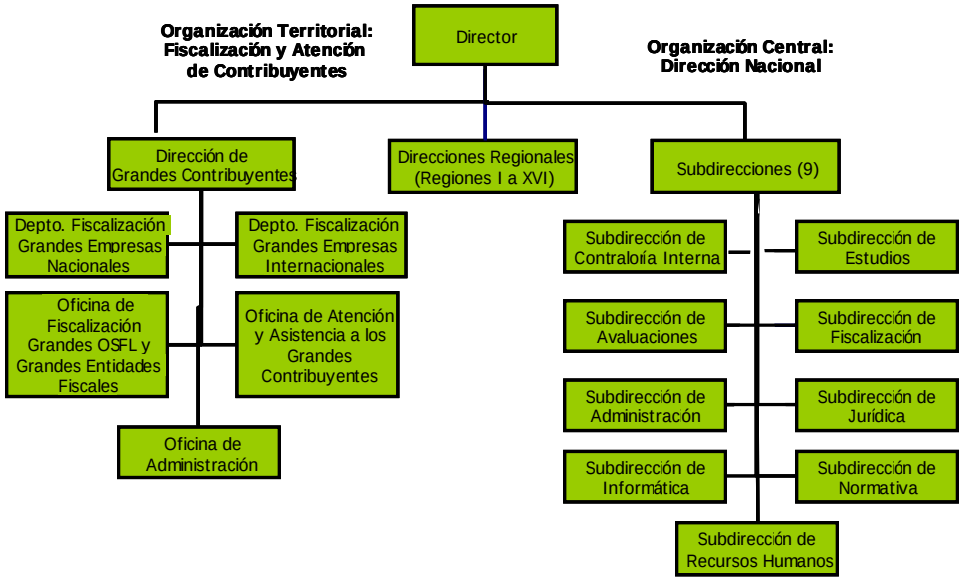
ANEXO A: Estructura del SII para la Detección de Facturas Falsas

La autoridad máxima del SII es el Director, y tiene las atribuciones y deberes inherentes a su calidad de Jefe Superior del Servicio. Su dotación de personal en el año 2010 asciende a 4.183 funcionarios ubicados en las plantas de: Directivos, Directivos de Carrera, Profesionales, Fiscalizadores, Técnicos, Administrativos y Auxiliares.

El Servicio de Impuestos Internos está constituido por la Dirección Nacional, con sede en la capital de la República, y por 16 Direcciones Regionales. Existe una Dirección Regional en cada Región del país y cuatro Direcciones Regionales en la Región Metropolitana.

La Dirección Nacional está a cargo del Director que es el Jefe Superior del Servicio; está compuesta por nueve Subdirecciones que actúan como delegados del Director en la evaluación y desarrollo de los programas de trabajo dentro de sus respectivas áreas y lo asesoran en las materias de su especialidad. De acuerdo a la Ley, deben recomendarle las normas y someter a su aprobación las instrucciones que estimen convenientes impartir al SII, y programar, dirigir, coordinar y supervigilar el funcionamiento de los Departamentos. También se radica en este nivel central la *Dirección de Grandes Contribuyentes*, que es una unidad especialmente creada para la fiscalización de grandes empresas, atendida su importancia económica y tributaria y considerando las características especiales de las operaciones que realizan.

Figura N° 52: Organigrama SII



En particular son 4 las Subdirecciones que intervienen directamente en el proceso de Fiscalización de Facturas Falsas, correspondientes a las Subdirecciones de Fiscalización, Jurídica, Estudios y Normativa:

- La Subdirección de Fiscalización se encarga de estudiar y proponer normas e instrucciones para la fiscalización de los impuestos y procurar que esas funciones alcancen el máximo de eficiencia. Planifica, evalúa y controla el desarrollo y la calidad de las actividades fiscalizadoras, define y entrega criterios operativos para la fiscalización de los distintos sectores económicos, propone normas y procedimientos administrativos y operativos para llevar a cabo su misión, responde consultas al respecto y evalúa el rendimiento de las Direcciones Regionales en materias de su área.

- La Subdirección Jurídica no sólo analiza la jurisprudencia de los Tribunales de Justicia y asesora al Director en materias tributarias, sino que defiende al SII en los recursos que interponen los contribuyentes o se querella contra ellos por los delitos tributarios que le corresponde investigar. Sus Departamentos de Investigación de Delitos Tributarios, Asesoría Jurídica, Defensa Judicial y Oficina Fiscalía Anti Facturas Falsas, se coordinan para desarrollar las labores mencionadas.

- La Subdirección de Estudios está a cargo del control de gestión, la organización y métodos, y los estudios económico-tributarios. Prepara, estudia e investiga las estadísticas de los ingresos tributarios, sus fluctuaciones y su relación con las distintas actividades económicas, para los efectos de interpretar y explicar sus variaciones. Elabora las estadísticas que requiere el Servicio, vela por la simplificación, uniformidad, coordinación y agilización de los métodos y procedimientos, analiza el desarrollo de sus labores, sus costos y productividad para su eficaz funcionamiento.

- La Subdirección Normativa estudia y propone las normas e instrucciones necesarias para la correcta y eficiente aplicación de los impuestos controlados por el SII, y recomienda la interpretación administrativa de las leyes y disposiciones que son de su competencia. Desarrolla estudios relativos a la gestión y modificación de las leyes tributarias y reglamentos, propone respuestas a diversas consultas y asesora al Director en materias de doble tributación internacional.

Por otro lado, la Dirección de Grandes Contribuyentes actúa como unidad generadora de procedimientos y metodologías de fiscalización, las cuales son traspasadas a otras Direcciones Regionales a objeto de mejorar la eficacia de las actuaciones de fiscalización realizadas en ellas y lograr uniformidad en las actuaciones del Servicio frente a los contribuyentes.

Finalmente, las Direcciones Regionales son unidades operativas y corresponden a las autoridades máximas del Servicio dentro de sus territorios jurisdiccionales, dependiendo directamente del Director. Tienen que supervisar el cumplimiento de las leyes tributarias y responder por la buena administración de las unidades a su cargo. En ellas se reciben los problemas y se atiende a los contribuyentes en su primera interacción con el Servicio.

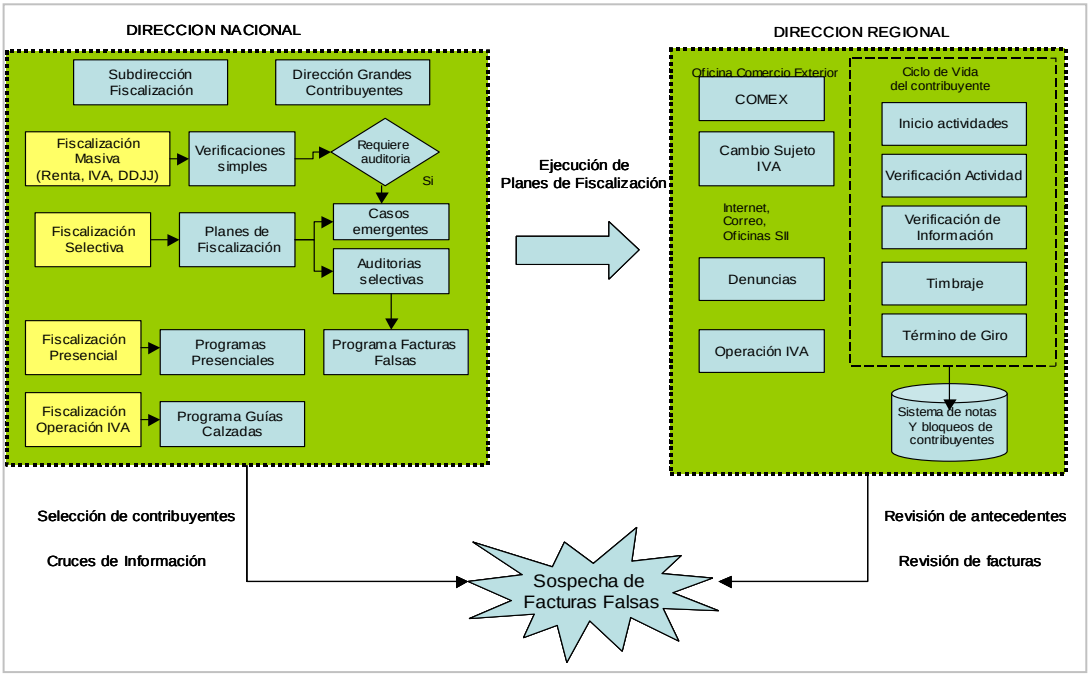
ANEXO B: Procedimiento de Detección de Facturas Falsas

Para detectar una factura falsa, se requiere de la aplicación de un procedimiento de auditoría que permita establecer en forma objetiva que la o las facturas objeto de análisis son falsas material y/o ideológicamente. Para esto será necesario efectuar las pericias contables pertinentes a las facturas, libros de contabilidad, declaraciones y otros antecedentes relacionados con las operaciones que dan cuenta las facturas cuestionadas, como asimismo, reunir las pruebas que avalan los resultados obtenidos.

A) Formas y Fuentes de Detección de Casos

Las sospechas de facturas falsas detectados en el SII pueden tener distintos orígenes, ya sea a través de antecedentes recibidos mediante denuncias, antecedentes generados en el ciclo de vida del contribuyente, antecedentes generados en el área de comercio exterior, en la operación IVA o al aplicar los programas de fiscalización anuales en sus diversas formas, como se muestra en el siguiente esquema.

Figura N° 53: Orígenes de detección de Facturas Falsas



- Denuncias:

Actualmente cualquier ciudadano puede realizar una denuncia de evasión tributaria a través del sitio web del SII, vía correo o directamente en las oficinas, entregando los antecedentes a funcionarios del Servicio o en los buzones dispuestos para ello. Una vez recibidos los antecedentes, las denuncias son ingresadas a un sistema y son derivadas a la Dirección Regional que corresponda, quien asigna un funcionario para que revise los antecedentes y determine si la denuncia es válida, requiere auditoría o faltan antecedentes para concluir.

- Ciclo de Vida del contribuyente:

Durante las distintas etapas del ciclo de vida tributario, se tiene una interacción directa con el contribuyente desde que éste realiza su inicio de actividades y se le autoriza el timbraje de

documentos, hasta que decide finalizar su actividad económica y realizar el término de giro. En cada etapa de este ciclo se obtiene información valiosa para el Servicio que permite detectar situaciones que pueden dar indicios de que el contribuyente está utilizando facturas falsas o está cometiendo un delito tributario.

- Programas de Fiscalización Masivos

Para efecto de velar por el cumplimiento de las obligaciones tributarias de los contribuyentes, el SII elabora planes o programas de fiscalización masivos, dirigidos a fiscalizar el correcto cumplimiento tributario de las obligaciones que afectan a los contribuyentes. En caso de detectarse irregularidades durante estas fiscalizaciones, se puede determinar la realización de una auditoría posterior en la que se examine información adicional del contribuyente. Estos programas son determinados de manera conjunta entre la Subdirección de Fiscalización, la Dirección de Grandes Contribuyentes y las Direcciones Regionales. El principal programa de fiscalización de este tipo lo constituye la Operación Renta, el que consiste en la verificación computacional de más de 1,5 millones de declaraciones de renta anuales, proceso que se realiza en el mes de mayo. También existen planes masivos en la Operación IVA y en la revisión de las Declaraciones Juradas.

- Programas de Fiscalización Selectiva

Los programas de fiscalización selectiva se generan en respuesta a una determinada figura de evasión tributaria, teniendo en consideración la temporalidad de su aplicación y estimación de rendimiento, entre otros factores asociados al programa específico, y consisten en la realización de auditorías destinadas a verificar el correcto cumplimiento de las obligaciones tributarias por parte de los contribuyentes, de tal forma, que sus declaraciones de impuestos correspondan a las operaciones contabilizadas, a la documentación sustentadora de las mismas y que reflejen todas las transacciones u operaciones por ellos efectuadas. En particular existe un Programa de Fiscalización de Facturas Falsas, cuyo objetivo verificar que las facturas de los proveedores que sustentan el crédito fiscal del impuesto, cumplen con los requisitos de la Ley del IVA y son idóneos para acreditar los costos y/o gastos para la determinación de los impuestos de la Ley sobre Impuesto a la Renta.

- Programas de Fiscalización de Presencia en Terreno

La presencia fiscalizadora en terreno corresponde a un procedimiento de control en el lugar que los contribuyentes realizan sus actividades económicas, y responde al objetivo de conocer la actividad comercial que desarrollan, promover el cumplimiento tributario voluntario y verificar que todos los contribuyentes cumplan con sus obligaciones y disposiciones tributarias vigentes, con respecto a la emisión de los documentos asociados a las operaciones comerciales del giro, su registro, declaración y pago de impuestos respectivos.

- Departamento Operación IVA

El Departamento de Operación IVA y Catastro además de controlar el proceso masivo de fiscalización de Operación IVA mensual, en el que se cita al contribuyente a regularización las declaraciones no efectuadas o que presentan irregularidades, realiza también fiscalizaciones selectivas. Entre los planes de fiscalización ejecutados se encuentran: Plan de Renotificación de Contribuyentes No Declarantes, Plan de Fiscalización de Contribuyentes Calzados, Plan de Control de Cotización Adicional, Plan de Créditos Especiales, Plan Importaciones, Plan de Término de Giro.

- Devolución IVA Exportadores

La legislación tributaria chilena contempla la aplicación de la tasa cero a las exportaciones. Esto significa que los exportadores tienen derecho a recuperar el IVA pagado en la adquisición de bienes intermedios y de capital o contratación de servicios, orientados a la actividad de exportación. En la práctica esto se materializa a través de solicitudes de devolución de créditos IVA que mensualmente presentan los exportadores ante el Servicio de Tesorerías, organismo responsable de las funciones de recaudación y cobranza en la administración tributaria chilena, las cuales son revisadas por el SII.

- Cambio de Sujeto del IVA

El cambio de sujeto es una medida de fiscalización instaurada con el propósito de controlar y reducir la evasión del IVA en sectores económicos tradicionalmente evasores, como también para regular y fiscalizar las transacciones afectas a impuestos que se efectúan dentro de estos sectores. Se aplica a mercados caracterizados por un reducido grupo de compradores de un determinado producto y un gran número de proveedores, con escasa significación económica, o con antecedentes tributarios irregulares, en los que se incurre en maniobras para eludir el pago de los tributos, en particular el Impuesto al Valor Agregado.

B) Análisis de los Casos de Facturas Falsas

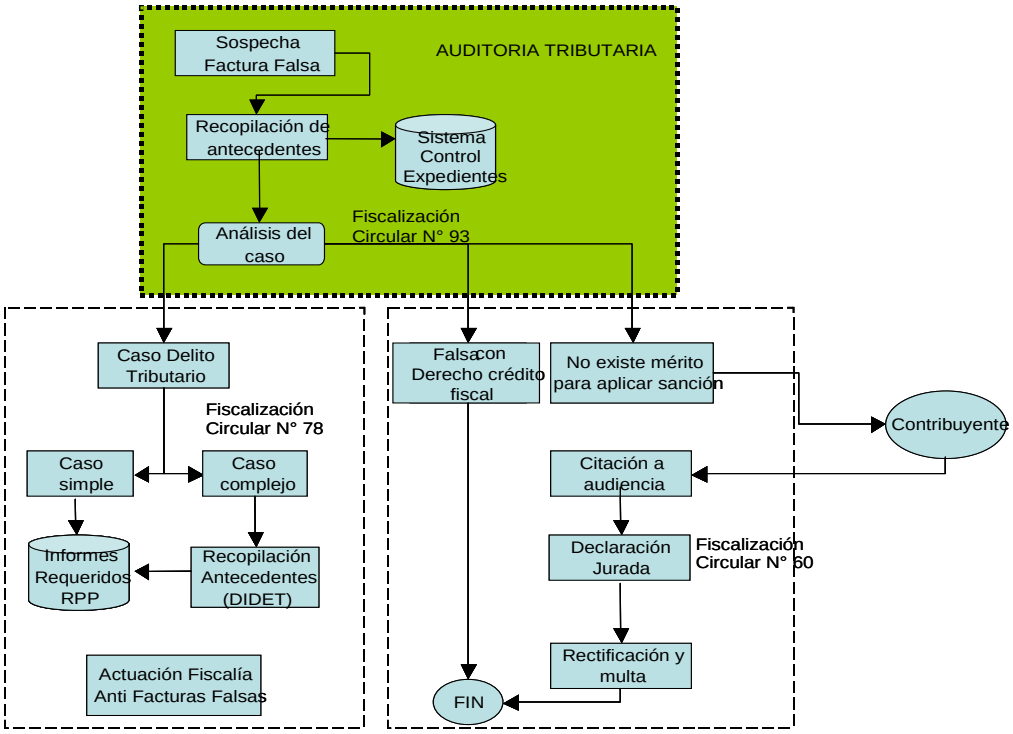
Una vez que se detecta una sospecha de factura falsa o no fidedigna, por alguno de los mecanismos de detección antes mencionado, los fiscalizadores deben seguir el procedimiento establecido en la Circular N° 60/2001 y dar cuenta inmediatamente del ilícito a la jefatura de su área, disponiendo las tareas propias de registro y revisión del caso que correspondan.

La primera labor que desarrolla el área de fiscalización será la incorporación de los antecedentes del caso en el "Sistema de Control de Expedientes", con la finalidad que la información de la existencia de dicho documento falso quede registrada en las "Bases de Datos del Servicio", a objeto que la información que en ellas se contienen esté permanentemente actualizada y a disposición de todos los estamentos de fiscalización.

En este punto debe determinarse si se cuenta con todos los antecedentes e información necesaria que permita iniciar una auditoría, si faltan antecedentes para acreditar la falsedad de las facturas o no se cuenta con los antecedentes necesarios para concluir., para lo cual se deberán ponderar las siguientes circunstancias:

i. Monto del perjuicio al interés fiscal involucrado en el ilícito detectado.
ii. Relación del monto del crédito fiscal amparado en las facturas impugnadas con el monto del crédito total empleado en los períodos revisados.
iii. Reiteración en la utilización de facturas falsas en calidad de tenedor, o participación en casos anteriores como proveedor de las mismas.
iv. Existencia de pruebas que permitan obtener una absoluta convicción respecto del uso malicioso de los documentos.
v. Efecto ejemplarizador o pedagógico que podría alcanzar una eventual persecución penal de los hechos de acuerdo a la realidad regional y al giro o actividad del infractor.

Figura N° 54: Procedimiento en caso de detección de sospecha de Factura Falsa



La circular N° 60, del 4 de septiembre de 2001, distingue las siguientes situaciones

- Caso 1: No obstante la falsedad de los documentos, el contribuyente tiene derecho al uso del crédito fiscal que ellos amparan.
- Caso 2: Los hechos detectados no revisten mérito suficiente para concluir la necesidad de perseguir la aplicación de sanciones por delito tributario.
- Caso 3: Los hechos detectados revisten la gravedad suficiente como para concluir la necesidad de perseguir la aplicación de sanciones por delito tributario.

En este último caso, se seguirá el procedimiento contemplado en la Circular 78, de diciembre de 1997. De esta manera, una vez detectados y verificados los hechos que configuran el ilícito, se pondrá en conocimiento del caso al Jefe de Departamento de Fiscalización a través de un Informe fundado, quien lo remitirá, con los documentos que le sirvan de antecedente al Jefe de Oficina Jurídica, el que elaborará un informe con su opinión legal, para que el Director Regional decida si debe propenderse sólo al cobro civil de los impuestos, o a la aplicación de una sanción pecuniaria y cobro de impuestos, o bien, que el caso reviste mérito suficiente como para perseguir también la aplicación de pena corporal.

Si nos encontramos en presencia de irregularidades susceptibles de ser sancionadas con pena corporal y multa, conforme a lo dispuesto en los artículos 30, 97 y 100 del Código Tributario, en que de acuerdo al mérito de los antecedentes reunidos, remitiendo los antecedentes al Subdirector Jurídico. Si éste estima improcedente la aplicación de sanciones pecuniarias, ya sea porque de acuerdo a lo informado por el Jefe del Departamento de Investigación de Delitos Tributarios estime que no resulta oportuna la aplicación de este tipo de sanciones, y que es necesario recabar mayores antecedentes a objeto de presentar una eventual denuncia o querella por delito tributario, señalará, a través del Departamento de Investigación de Delitos Tributarios, al Director Regional respectivo, dentro del mismo plazo de 72 horas, las acciones a seguir.

Cuando el Director Regional respectivo considere la necesidad que el SII inste por la interposición de una querella ante la justicia ordinaria, deberá distinguir entre dos opciones:

- Caso simple:

Los antecedentes detectados permiten presumir fundadamente la existencia del delito tributario, y no existe necesidad de reunir mayores pruebas, entonces, la Oficina Jurídica Regional redactará un proyecto de querella criminal en contra de las personas que aparezcan como responsables de dicho delito, el que conjuntamente con los antecedentes probatorios recopilados, se remitirán directamente al Departamento de Defensa Judicial. Este Departamento revisará los antecedentes y con su mérito elaborará inmediatamente un breve informe legal con su recomendación para la decisión del Director, sobre la procedencia o no de la interposición de la querella.

- Caso complejo

Si las irregularidades detectadas se consideran de gravedad suficiente como para justificar la interposición de una acción penal, pero por su complejidad se hace necesario reunir más antecedentes y efectuar diligencias adicionales, los antecedentes serán remitidos al Departamento de Investigación de Delitos Tributarios (DIDET), para que prosiga con aquellas. Una vez efectuadas todas las diligencias pertinentes, y en el evento de que se considere procedente la persecución de pena corporal, se elaborará un informe pericial que será remitido al Departamento de Defensa Judicial, el que a su vez, confeccionará un informe legal para la decisión del Director, sobre la interposición de querella.

C) Querellas

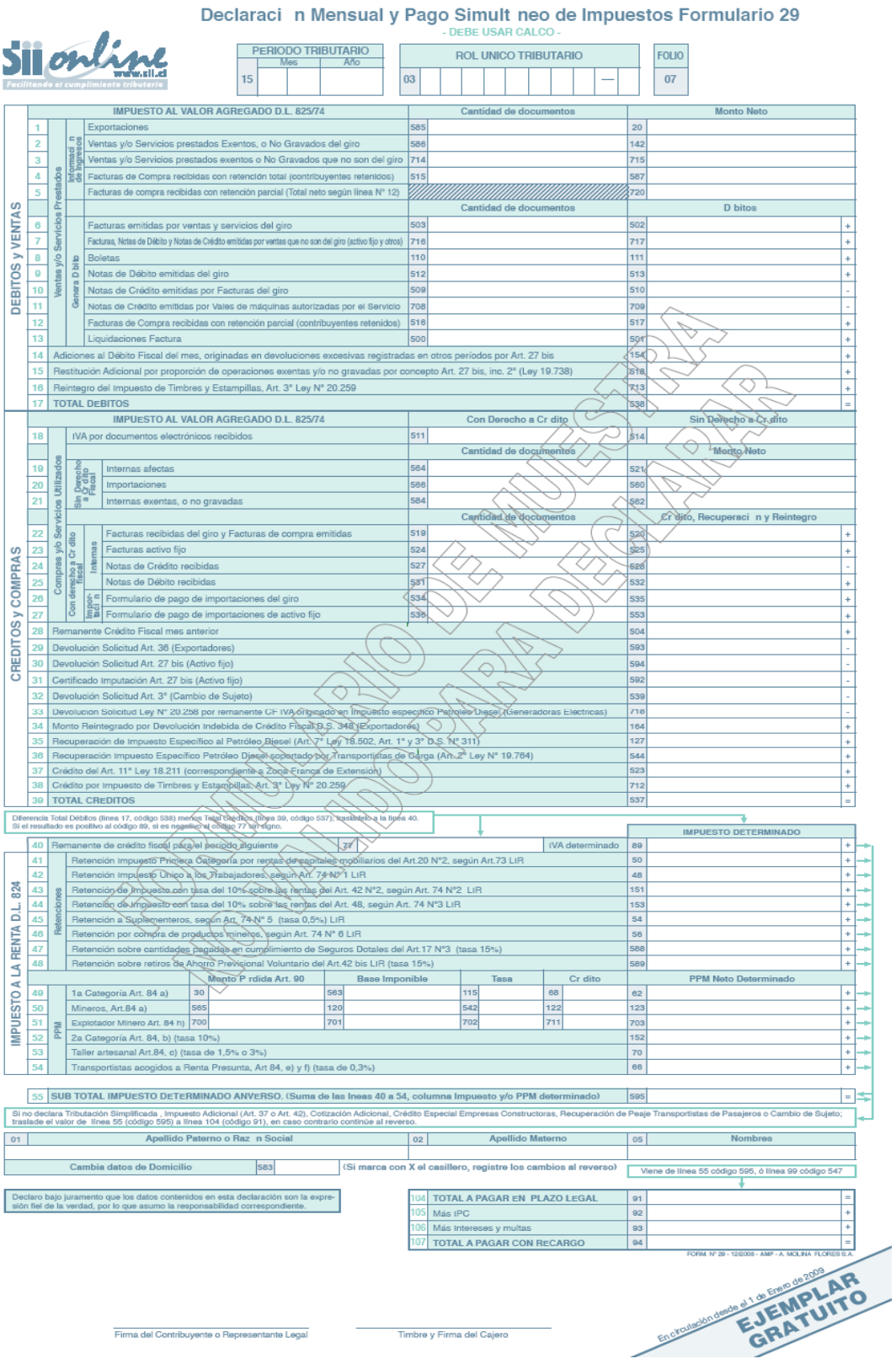
Para que el Director del Servicio pueda tomar la decisión de interponer o no una querella en contra del contribuyente, es necesario que cuente con todos los antecedentes necesarios. El Oficio Circular N° 36 de 2006, señala que este informe debe contener la descripción circunstanciada de la conducta infraccional, de la o las personas que han participado en ella, así como los conclusiones de los peritos de acuerdo a su oficio.

Para ellos su confección debe considerar los siguientes contenidos:

i. Individualización del contribuyente
ii. Introducción, origen de los antecedentes y documentos analizados, enunciación de ellos y procedimientos empleados en la pericia:
iii. Análisis y detalle de las irregularidades detectadas, considerando:
iv. Otros antecedentes:
v. Perjuicio fiscal:
vi. Conclusiones:

Con estos antecedentes, la Oficina Jurídica o Fiscalía, según proceda, deberá definir la interposición de la acción criminal que corresponda en contra de quienes aparezcan como responsables del ilícito. Si del análisis que se efectúe, aparece la necesidad de contar con mayores antecedentes, éstos podrán ser requeridos al área de Fiscalización. De este modo, y detectada una factura falsa en cualquier lugar del país, ya sea que tenga timbre del SII o carezca de él, la Fiscalía evaluará rápidamente los antecedentes y determinará si presenta o no de forma inmediata una querella criminal por estas conductas ilícitas, privilegiando las acciones criminales cuando se trate de grupos organizados o concertados para defraudar al fisco.

ANEXO C: Formulario F29 de Declaración Mensual del IVA



Declaraci n Mensual y Pago Simult neo de Impuestos Formulario 29

- DEBE USAR CALCO -

Sii online www.sii.cl Facilitando el cumplimiento tributario

PERIODO TRIBUTARIO		ROL UNICO TRIBUTARIO	FOLIO
Mes	Año		
15		03	07

DEBITOS y VENTAS

		IMPUESTO AL VALOR AGREGADO D.L. 825/74	Cantidad de documentos		Monto Neto	
1	Informaci n de Ingresos	Exportaciones	585		20	
2		Ventas y/o Servicios prestados Exentos, o No Gravados del giro	586		142	
3		Ventas y/o Servicios prestados exentos o No Gravados que no son del giro	714		715	
4		Facturas de Compra recibidas con retención total (contribuyentes retenidos)	515		587	
5		Facturas de compra recibidas con retención parcial (Total neto según línea N° 12)			720	
			Cantidad de documentos		D bitos	
6	Genera D bito	Facturas emitidas por ventas y servicios del giro	503		502	+
7		Facturas, Notas de Débito y Notas de Crédito emitidas por ventas que no son del giro (activo fijo y otros)	716		717	+
8		Boletas	110		111	+
9		Notas de Débito emitidas del giro	512		513	+
10		Notas de Crédito emitidas por Facturas del giro	509		510	-
11		Notas de Crédito emitidas por Vales de máquinas autorizadas por el Servicio	708		709	-
12		Facturas de Compra recibidas con retención parcial (contribuyentes retenidos)	516		517	+
13		Liquidaciones Factura	500		501	+
14		Adiciones al Débito Fiscal del mes, originadas en devoluciones excesivas registradas en otros períodos por Art. 27 bis			154	+
15		Restitución Adicional por proporción de operaciones exentas y/o no gravadas por concepto Art. 27 bis, inc. 2° (Ley 19.738)			518	+
16		Reintegro del Impuesto de Timbres y Estampillas, Art. 3° Ley N° 20.259			713	+
17		TOTAL DEBITOS			538	=

CREDITOS y COMPRAS

		IMPUESTO AL VALOR AGREGADO D.L. 825/74	Con Derecho a Cr dito		Sin Derecho a Cr dito	
18	Sin Derecho a Cr dito Fiscal	IVA por documentos electrónicos recibidos	511		514	
			Cantidad de documentos		Monto Neto	
19		Internas afectas	564		521	
20		Importaciones	566		560	
21		Internas exentas, o no gravadas	584		562	
			Cantidad de documentos		Cr dito, Recuperaci n y Reintegro	
22	Internas	Facturas recibidas del giro y Facturas de compra emitidas	519		520	+
23		Facturas activo fijo	524		525	+
24		Notas de Crédito recibidas	527		528	-
25		Notas de Débito recibidas	531		532	+
26	Importaci n	Formulario de pago de importaciones del giro	534		535	+
27		Formulario de pago de importaciones de activo fijo	536		553	+
28		Remanente Crédito Fiscal mes anterior			504	+
29		Devolución Solicitud Art. 36 (Exportadores)			593	-
30		Devolución Solicitud Art. 27 bis (Activo fijo)			594	-
31		Certificado Imputación Art. 27 bis (Activo fijo)			592	-
32		Devolución Solicitud Art. 3° (Cambio de Sujeto)			539	-
33		Devolucion Solicitud Ley N° 20.258 por remanente CF IVA originado en Impuesto especifico Petroleo Diesel (Generadoras Electricas)			716	-
34		Monto Reintegrado por Devolución Indebida de Crédito Fiscal D.S. 348 (Exportadores)			164	+
35		Recuperación de Impuesto Específico al Petróleo Diesel (Art. 7° Ley 18.502, Art. 1° y 3° D.S. N° 311)			127	+
36		Recuperación Impuesto Específico Petróleo Diesel soportado por Transportistas de Carga (Art. 2° Ley N° 19.764)			544	+
37		Crédito del Art. 11° Ley 18.211 (correspondiente a Zona Franca de Extension)			523	+
38		Crédito por Impuesto de Timbres y Estampillas, Art. 3° Ley N° 20.259			712	+
39		TOTAL CREDITOS			537	=

Diferencia Total Débitos (línea 17, código 538) menos Total Créditos (línea 39, código 537); trasládelo a la línea 40. Si el resultado es positivo al código 89, si es negativo al código 77 sin signo.

IMPUESTO A LA RENTA D.L. 824

				IMPUESTO DETERMINADO	
40		Remanente de crédito fiscal para el periodo siguiente	77	IVA determinado	89	+
41	Retenciones	Retención Impuesto Primera Categoría por rentas de capitales mobiliarios del Art.20 N°2, según Art.73 LIR			50	+
42		Retención Impuesto Único a los Trabajadores, según Art. 74 N° 1 LIR			48	+
43		Retención de Impuesto con tasa del 10% sobre las rentas del Art. 42 N°2, según Art. 74 N°2 LIR			151	+
44		Retención de Impuesto con tasa del 10% sobre las rentas del Art. 48, según Art. 74 N°3 LIR			153	+
45		Retención a Suplementeros, según Art. 74 N° 5 (tasa 0,5%) LIR			54	+
46		Retención por compra de productos mineros, según Art. 74 N° 6 LIR			56	+
47		Retención sobre cantidades pagadas en cumplimiento de Seguros Dotales del Art.17 N°3 (tasa 15%)			588	+
48		Retención sobre retiros de Ahorro Previsional Voluntario del Art.42 bis LIR (tasa 15%)			589	+

		Monto P rdida Art. 90		Base Imponible		Tasa		Cr dito		PPM Neto Determinado	
49	1a Categoría Art. 84 a)	30		563		115		68		62	+
50	Mineros, Art.84 a)	585		120		542		122		123	+
51	Explotador Minero Art. 84 h)	700		701		702		711		703	+
52	2a Categoría Art. 84, b) (tasa 10%)									152	+
53	Taller artesanal Art.84, c) (tasa de 1,5% o 3%)									70	+
54	Transportistas acogidos a Renta Presunta, Art 84, e) y f) (tasa de 0,3%)									66	+

PPM

55	SUB TOTAL IMPUESTO DETERMINADO ANVERSO. (Suma de las lneas 40 a 54, columna Impuesto y/o PPM determinado)	595	=

Si no declara Tributación Simplificada , Impuesto Adicional (Art. 37 o Art. 42), Cotización Adicional, Crédito Especial Empresas Constructoras, Recuperación de Peaje Transportistas de Pasajeros o Cambio de Sujeto; traslade el valor de línea 55 (código 595) a línea 104 (código 91), en caso contrario continúe al reverso.

01	Apellido Paterno o Raz n Social	02	Apellido Materno	05	Nombres

Cambia datos de Domicilio 583 (Si marca con X el casillero, registre los cambios al reverso)

Viene de línea 55 código 595, ó línea 99 código 547

Declaro bajo juramento que los datos contenidos en esta declaración son la expresión fiel de la verdad, por lo que asumo la responsabilidad correspondiente.

104	TOTAL A PAGAR EN PLAZO LEGAL	91	=
105	Más IPC	92	+
106	Más Intereses y multas	93	+
107	TOTAL A PAGAR CON RECARGO	94	=

FORM. N° 29 - 12/2008 - AMF - A. MOLINA FLORES S.A.

Firma del Contribuyente o Representante Legal

Timbre y Firma del Cajero

ORIGINAL

- DEBE USAR CALCO -

ORIGINAL - SERVICIO DE IMPUESTOS INTERNOS

SIMPLIFICADA

	SISTEMA DE TRIBUTACIÓN SIMPLIFICADA DEL IVA, ART. 29 D.L. 825			IMPUESTO DETERMINADO		
56	Ventas del período	529				
57	Crédito del período	530				
58	IVA determinado por concepto de Tributación Simplificada			409		+

IMPUESTO ADICIONAL ART. 37 D.L. 825

59	Letras e), h), i), l) (tasa 15%)			522	+
60	Letra j) (tasa 50%)			526	+
61	Débito de Impuesto Adicional Ventas Art. 37 letras a), b) y c) y Art. 40 D.L.825 (tasa 15%)	113	+		
62	Crédito de Impuesto Adicional Art.37 letras a), b) y c) D.L. 825	28	-		
63	Monto reintegrado por devolución indebida de crédito por exportadores D.L. 825	548	-		
64	Remanente crédito Art. 37 mes anterior D.L.825	540	-		
65	Devolución Solicitud Art.36 relativa al Impuesto Adicional Art.37 letras a), b) y c) D.L. 825	541	+		

Diferencia Débito menos Crédito Impuesto Art. 37 D.L. 825/74 (operación aritmética de las líneas 61 a la 65), traslade el valor a la línea 66. Si el resultado es positivo al código 550, en caso contrario al código 549 sin signo.

66	Remanente crédito Impuesto Art.37 para período siguiente	549	Impuesto Adicional Art. 37 y Art.40 determinado	550	+

IMPUESTO ADICIONAL ART. 42 D.L. 825

		Débitos		
67	Pisco, Licores, Whisky y Aguardiente (tasa 27%)	577	+	
68	Vinos, Champaña, Chichas (tasa 15%)	32	+	
69	Cervezas (tasa 15%)	150	+	
70	Bebidas analcohólicas (tasa 13%)	146	+	
71	Notas de Débito emitidas	545	+	
72	Notas de Crédito emitidas por Facturas	546	-	
73	Notas de Crédito emitidas por Vales de máquinas autorizadas por el Servicio	710	-	
74	Total Débitos Art. 42 DL 825	602	=	

		Total crédito recargado en facturas recibidas		Crédito imputable del período	
75	Pisco, Licores, Whisky y Aguardiente (tasa 27%)	575	576		+
76	Vinos, Champaña, Chichas (tasa 15%)	574	33		+
77	Cervezas (tasa 15%)	580	149		+
78	Bebidas analcohólicas (tasa 13%)	582	85		+
79	Notas de Débito recibidas		551		+
80	Notas de Crédito recibidas		559		-
81	Remanente crédito Art. 42 mes anterior		508		+
82	Devolución Art. 36 D.L.825 relativas impuesto Art.42		533		-
83	Monto reintegrado devoluciones indebidas de crédito por exportaciones		552		+
84	Total créditos Art. 44 DL 825		603		=

Diferencia Débito menos Crédito Impuesto Art. 42 D.L. 825/74 (código 602 menos el código 603), traslade el valor a la línea 85. Si el resultado es positivo al código 506, en caso contrario al código 507 sin signo.

85	Remanente crédito Imp. Adic. Art. 42 para período siguiente	507	Impuesto Adicional Art. 42 determinado	508	+

CAMBIO DE SUJETO D.L. 825

	ANTICIPO CAMBIO DE SUJETO (CONTRIBUYENTES RETENIDOS)		
86	IVA anticipado del período	556	+
87	Remanente del mes anterior	557	+
88	Devolución del mes anterior	558	-
89	Total de Anticipo	543	=

Registre Total de anticipo (código 543) en el código 598, con tope del valor del código 89 línea 40, el saldo restante se debe registrar en el remanente para el mes siguiente, código 573

90	Remanente Anticipos Cambio Sujeto para período siguiente	573	Anticipo a imputar	598	-

	CAMBIO DE SUJETO (AGENTE RETENEDOR)		
91	IVA total retenido a terceros (tasa Art. 14 DL 825)	39	+
92	IVA parcial retenido a terceros (según tasa)	554	+
93	Retención de margen de comercialización	597	+
94	Retención Anticipo de Cambio de Sujeto	555	+

Registre en el código 596 la suma de las retenciones (código 39, 554, 597 y 555)

	Retención Cambio de Sujeto	596			+

ESPECIALES

95	Imputación del Pago Patente Aguas Ley 20.017	704	Remanente anterior	705	Total a Imputar	706		-
96	Cotización Adicional Ley 18.566	160	Remanente mes anterior	161	Total Crédito mes	570		-
97	Crédito Especial Empresas Constructoras	126	Remanente mes anterior	128	Total Crédito mes	571		-
98	Recup. Peajes Transportistas Pasajeros, Ley 19.764	572	Remanente mes anterior	568	Total Crédito mes	590		-

Realice la operación aritmética de las líneas 55 a 98 (columna Impuesto Determinado). Registre el valor resultante en el código 547 (línea 99), si es negativo anótelo entre paréntesis.

99	TOTAL DETERMINADO	547	=

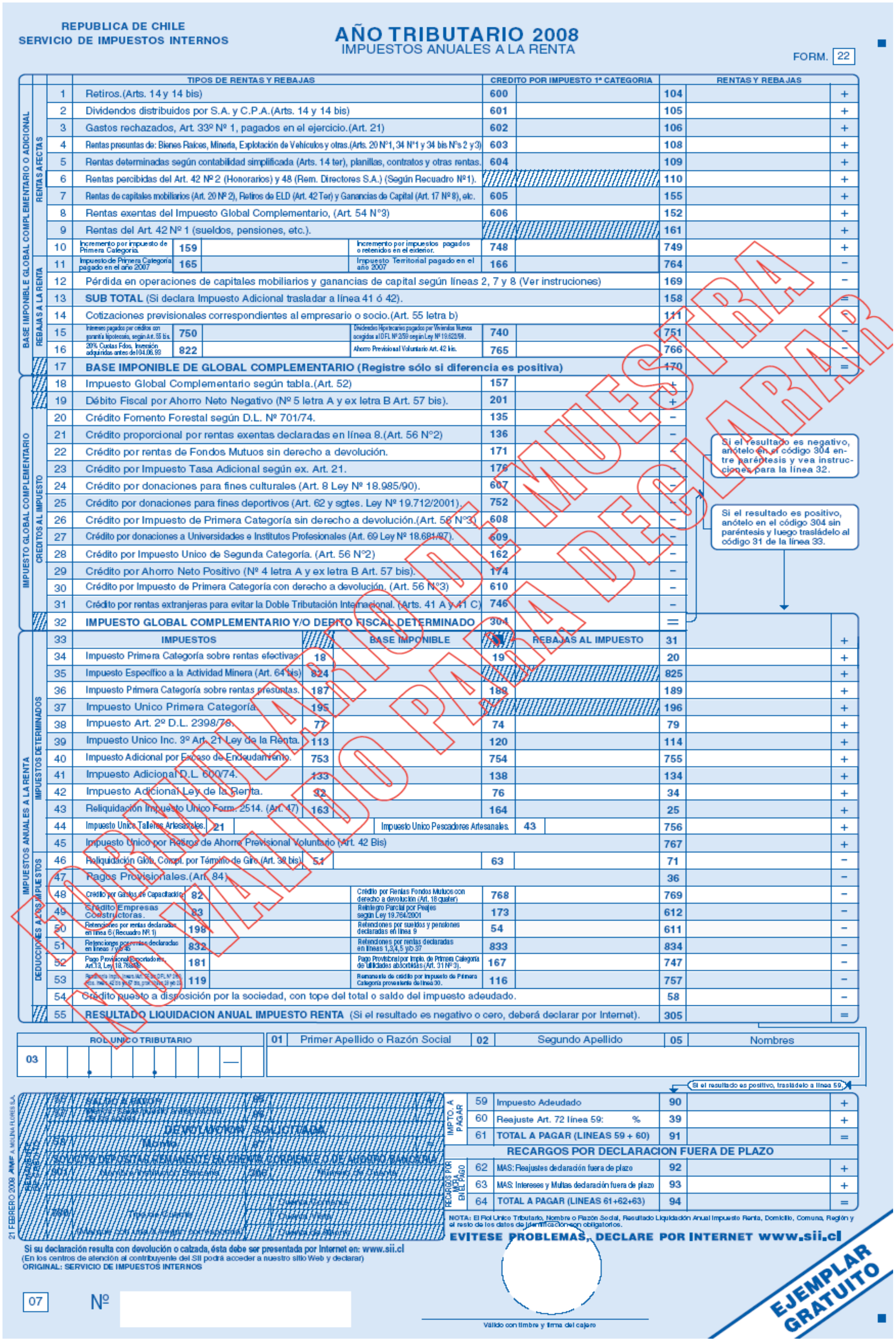
100	Remanente período siguiente Patente Aguas, Ley 20.017	707
101	Remanente de Cotizacion Adicional Ley 18.566	73
102	Remanente Crédito Especial Empresas Constructoras	130
103	Remanente Recup. de Peajes Trans. Pasajeros Ley 19.764	591

Si código 547 es positivo, trasládelo al anverso (código 91, línea 104), en caso contrario regístrelo en los códigos de remanente (línea 100 a 103) teniendo presente las instrucciones.

REGISTRE SI CAMBIA ALGUNO DE LOS SIGUIENTES ANTECEDENTES

06	Calle				610	N	611	Departamento	612	Villa o Poblaci n	
08	Comuna	53	Regi n	613	C d. rea tel fono	09	Tel fono	601	Fax	604	Tel fono celular	
55	Correo Electr nico	44	Domicilio Postal		726	Comuna Postal	313	Rut Contador	314	Rut Representante Legal	

ANEXO D: Formulario F22 de Declaración del Impuesto a la Renta



REPUBLICA DE CHILE
SERVICIO DE IMPUESTOS INTERNOS

AÑO TRIBUTARIO 2008
IMPUESTOS ANUALES A LA RENTA

FORM. 22

N°	TIPOS DE RENTAS Y REBAJAS		CREDITO POR IMPUESTO 1ª CATEGORIA		RENTAS Y REBAJAS		
1	Retiros.(Arts. 14 y 14 bis)		600		104		+
2	Dividendos distribuidos por S.A. y C.P.A.(Arts. 14 y 14 bis)		601		105		+
3	Gastos rechazados, Art. 33º Nº 1, pagados en el ejercicio.(Art. 21)		602		106		+
4	Rentas presuntas de: Bienes Raíces, Minería, Explotación de Vehículos y otras.(Arts. 20 N°1, 34 N°1 y 34 bis N°s 2 y 3)		603		108		+
5	Rentas determinadas según contabilidad simplificada (Arts. 14 ter), planillas, contratos y otras rentas.		604		109		+
6	Rentas percibidas del Art. 42 Nº 2 (Honorarios) y 48 (Rem. Directores S.A.) (Según Recuadro Nº1).				110		+
7	Rentas de capitales mobiliarios (Art. 20 Nº 2), Retiros de ELD (Art. 42 Ter) y Ganancias de Capital (Art. 17 Nº 8), etc.		605		155		+
8	Rentas exentas del Impuesto Global Complementario, (Art. 54 N°3)		606		152		+
9	Rentas del Art. 42 Nº 1 (sueldos, pensiones, etc.).				161		+
10	Incremento por impuesto de Primera Categoría	159	Incremento por impuestos pagados o retenidos en el exterior. 748		749		+
11	Impuesto de Primera Categoría pagado en el año 2007	165	Impuesto Territorial pagado en el año 2007 166		764		–
12	Pérdida en operaciones de capitales mobiliarios y ganancias de capital según líneas 2, 7 y 8 (Ver instrucciones)				169		–
13	SUB TOTAL (Si declara Impuesto Adicional trasladar a línea 41 ó 42).				158		=
14	Cotizaciones previsionales correspondientes al empresario o socio.(Art. 55 letra b)				111		–
15	Intereses pagados por créditos con garantía hipotecaria, según Art. 55 bis	750	Dividendos Hipotecarios pagados por Viviendas Nuevas acogidas al DFL Nº 2/59 según Ley Nº 19.622/99. 740		751		–
16	20% Cuotas Fdos. Inversión adquiridas antes del 04.06.93	822	Ahorro Previsional Voluntario Art. 42 bis. 765		766		–
17	BASE IMPONIBLE DE GLOBAL COMPLEMENTARIO (Registre sólo si diferencia es positiva)				170		=
18	Impuesto Global Complementario según tabla.(Art. 52)		157		+		
19	Débito Fiscal por Ahorro Neto Negativo (Nº 5 letra A y ex letra B Art. 57 bis).		201		+		
20	Crédito Fomento Forestal según D.L. Nº 701/74.		135		–		
21	Crédito proporcional por rentas exentas declaradas en línea 8.(Art. 56 N°2)		136		–		
22	Crédito por rentas de Fondos Mutuos sin derecho a devolución.		171		–		
23	Crédito por Impuesto Tasa Adicional según ex. Art. 21.		176		–		
24	Crédito por donaciones para fines culturales (Art. 8 Ley Nº 18.985/90).		607		–		
25	Crédito por donaciones para fines deportivos (Art. 62 y sgtes. Ley Nº 19.712/2001).		752		–		
26	Crédito por Impuesto de Primera Categoría sin derecho a devolución.(Art. 56 N°3)		608		–		
27	Crédito por donaciones a Universidades e Institutos Profesionales (Art. 69 Ley Nº 18.681/87).		609		–		
28	Crédito por Impuesto Unico de Segunda Categoría. (Art. 56 N°2)		162		–		
29	Crédito por Ahorro Neto Positivo (Nº 4 letra A y ex letra B Art. 57 bis).		174		–		
30	Crédito por Impuesto de Primera Categoría con derecho a devolución. (Art. 56 N°3)		610		–		
31	Crédito por rentas extranjeras para evitar la Doble Tributación Internacional. (Arts. 41 A y 41 C)		746		–		
32	IMPUESTO GLOBAL COMPLEMENTARIO Y/O DEBITO FISCAL DETERMINADO		304		=		
33	IMPUESTOS	BASE IMPONIBLE		REBAJAS AL IMPUESTO	31		+
34	Impuesto Primera Categoría sobre rentas efectivas.	18	19		20		+
35	Impuesto Específico a la Actividad Minera (Art. 64 bis)	824			825		+
36	Impuesto Primera Categoría sobre rentas presuntas.	187	188		189		+
37	Impuesto Unico Primera Categoría	195			196		+
38	Impuesto Art. 2º D.L. 2398/78.	77	74		79		+
39	Impuesto Unico Inc. 3º Art. 21 Ley de la Renta.	113	120		114		+
40	Impuesto Adicional por Exceso de Endeudamiento.	753	754		755		+
41	Impuesto Adicional D.L. 600/74.	133	138		134		+
42	Impuesto Adicional Ley de la Renta.	32	76		34		+
43	Reliquidación Impuesto Unico Form. 2514. (Art. 47)	163	164		25		+
44	Impuesto Unico Talleres Artesanales.	21	Impuesto Unico Pescadores Artesanales. 43		756		+
45	Impuesto Unico por Retiros de Ahorro Previsional Voluntario (Art. 42 Bis)				767		+
46	Reliquidación Glob. Compl. por Término de Giro (Art. 38 bis)	51	63		71		–
47	Pagos Provisionales.(Art. 84)				36		–
48	Crédito por Gastos de Capacitación	82	Crédito por Rentas Fondos Mutuos con derecho a devolución (Art. 18 quáter) 768		769		–
49	Crédito Empresas Constructoras.	83	Reintegro Parcial por Peajes según Ley 19.764/2001 173		612		–
50	Retenciones por rentas declaradas en línea 6 (Recuadro Nº 1)	198	Retenciones por sueldos y pensiones declaradas en línea 9 54		611		–
51	Retenciones por rentas declaradas en líneas 7 y 45	832	Retenciones por rentas declaradas en líneas 1,3,4,5 y/o 37 833		834		–
52	Pago Provisional Exportadores, Art.13, Ley 18.768/88	181	Pago Provisional por Impto. de Primera Categoría de Utilidades absorbidas (Art. 31 Nº 3). 167		747		–
53		119	Remanente de crédito por Impuesto de Primera Categoría proveniente de línea 30. 116		757		–
54	Crédito puesto a disposición por la sociedad, con tope del total o saldo del impuesto adeudado.				58		–
55	RESULTADO LIQUIDACION ANUAL IMPUESTO RENTA (Si el resultado es negativo o cero, deberá declarar por Internet).				305		=

Si el resultado es negativo, anótelo en el código 304 entre paréntesis y vea instrucciones para la línea 32.

Si el resultado es positivo, anótelo en el código 304 sin paréntesis y luego trasládelo al código 31 de la línea 33.

ROL UNICO TRIBUTARIO	01 Primer Apellido o Razón Social	02 Segundo Apellido	05 Nombres
03			

Si el resultado es positivo, trasládelo a línea 59.

SALDO A FAVOR
DEVOLUCION SOLICITADA
Monto
SOLICITO DEPOSITAR DEVOLUCION EN CUENTA CORRIENTE O DE AHORRO BANCARIA
Nombre Institución Bancaria
Número de Cuenta
Tipo de Cuenta
Cuenta Corriente
Cuenta Vista
Cuenta de Ahorro

IMPTO. A PAGAR	59	Impuesto Adeudado	90		+
	60	Reajuste Art. 72 línea 59: %	39		+
	61	TOTAL A PAGAR (LINEAS 59 + 60)	91		=
RECARGOS POR DECLARACION FUERA DE PLAZO					
RECARGOS POR MORA EN EL PAGO	62	MAS: Reajustes declaración fuera de plazo	92		+
	63	MAS: Intereses y Multas declaración fuera de plazo	93		+
	64	TOTAL A PAGAR (LINEAS 61+62+63)	94		=

NOTA: El Rol Unico Tributario, Nombre o Razón Social, Resultado Liquidación Anual Impuesto Renta, Domicilio, Comuna, Región y el resto de los datos de identificación son obligatorios.

EVITESE PROBLEMAS, DECLARE POR INTERNET www.sii.cl

Si su declaración resulta con devolución o calzada, ésta debe ser presentada por Internet en: www.sii.cl
(En los centros de atención al contribuyente del SII podrá acceder a nuestro sitio Web y declarar)
ORIGINAL: SERVICIO DE IMPUESTOS INTERNOS

07 Nº

Válido con timbre y firma del cajero

21 FEBRERO 2008

FORMULARIO DE MUESTRA NO VALIDO PARA DECLARAR

EJEMPLAR GRATUITO

TODOS LOS CONTRIBUYENTES DEBEN COMPLETAR LOS SIGUIENTES DATOS:

06 Calle	Nº	Of. Depto.	Ciudad

08 Comuna	53 Región	13 Actividad, Profesión o Giro del Negocio	14 Cód.Actividad Económica

Domicilio Postal 44	Comuna Postal 726

Teléfono 9	Fax 48	Correo Electrónico 55

Marque con X según corresponda

FINALIDAD TRIBUTARIA

Ley 18.392/19.149	95		E.I.R.L. (Ley 19.857)	787		D.S. 341 (Zona Franca)	73		D.L. 701 (F.F.)	72		Sin Contabilidad	613		Asoc. o Cuentas en Participación	616	
			S.p.A. (art. 424 y sgtes. C.Bo.C.)	846								Contabilidad Completa	614				
Ley 19.709 (Tocopilla)	786		D.L. 600 (E.I.E.)	68		Instit. Art. 40 Nº 2 y 4 LIR	69		P. de Negocios (Art. 41 DLIR)	788					Artículo 14 bis LIR	42	
Contribuyente Acogido al Régimen Tributario Simplificado D.S. Nº 344/2004			OPCION	805		RETIRO	813		Artículo 57 bis LIR	46		Contabilidad Simplificada	615				

DEFINIDOS DEFINITIVAMENTE

RECUADRO Nº 1 : HONORARIOS

Rentas de 2ª Categoría		Renta Actualizada	Impuesto Retenido Actualizado	
Honorarios Anuales Con Retención	461		492	+
Honorarios Anuales Sin Retención	545			+
Total Ingresos Brutos	547			=
Participación en Soc. de Profes. de 2ª Categ.	617			+
Monto Ahorro Previsional Voluntario Art. 42 bis	770			−
Gastos Efectivos (sólo del Total Ingresos Brutos)	465			−
Gastos Presuntos: 30% sobre el código 547, con tope $ 8.139.960	494			−
Total Honorarios	467			=
Total Remuneraciones Directores S.A.	479		491	+
	618		619	=
Total Rentas y Retenciones	(Trasladar línea 6 sólo Personas Naturales)		(Trasladar línea 50 código 198)	

RECUADRO Nº 2: BASE IMPONIBLE DE PRIMERA CATEGORIA

Ingresos del Giro Percibidos o Devengados	628		+
Intereses Percibidos o Devengados	629		+
Otros Ingresos Percibidos o Devengados	651		+
Costo Directo de los Bienes y Servicios	630		−
Remuneraciones	631		−
Depreciación sobre bienes de propiedad del contribuyente	632		−
Intereses Pagados o Adeudados	633		−
Gastos por Donaciones para fines Sociales	792		−
Gastos por Donaciones para fines Políticos	793		−
Gastos por otras Donaciones según Art. Nº 10, Ley 19.885	772		−
Otros Gastos Deducidos de los Ingresos Brutos	635		−
Renta Líquida	636		=
Corrección Monetaria Saldo Deudor (Art. 32)	637		−
Corrección Monetaria Saldo Acreedor (Art. 32)	638		+
Gastos Rechazados (Art. 33 Nº 1)	639		+
Gastos Rechazados por Donaciones para fines Sociales	794		+
Gastos Rechazados por Donaciones para fines Políticos	812		+
Gastos Rechazados por otras Donaciones según Art. Nº10, Ley 19.885	811		+
Impuesto Específico a la Actividad Minera	827		+
Pérdidas de Ejercicios Anteriores (Art. 31 Nº 3)	634		−
Ingresos No Renta (Art. 17)	640		−
Otras Partidas	807		−
Rentas Exentas Impto. 1ª Categoría (Art. 33 Nº 2)	641		−
Dividendos y/o Utilidades Sociales (Art. 33 Nº 2)	642		−
Renta Líquida Imponible (o Pérdida Tributaria)	643		=
Base imponible Renta Presunta	808		
Rentas afectas al Impuesto Unico de Primera Categoría	758		
Rentas por arriendos de Bienes Raíces Agrícolas	809		
Rentas por arriendos de Bienes Raíces No Agrícolas	759		
Otras rentas afectas al Impuesto de Primera Categoría.	760		

RECUADRO Nº3: DATOS CONTABLES BALANCE SOLUNA Y OTROS

Saldo de Caja (sólo dinero en efectivo y documentos al día según arqueo)	101		
Saldo cuenta corriente bancaria según conciliación	794		
Cuentas por Cobrar Empresas Relacionadas	778		
Cuentas por Cobrar Empresas NO Relacionadas	816		
Total préstamos efectuados a los socios en el ejercicio	783		
Existencia Final	129		
Activo Inmovilizado	647		
Depreciación tributaria del ejercicio	785		
Bienes Adquiridos Contrato Leasing	648		
Monto Inversión Ley Arica	815		
Monto Inversión Ley Austral	741		
Total del Activo	122		
Cuentas por Pagar Empresas Relacionadas	779		
Cuentas por Pagar Empresas NO Relacionadas	817		
Total del Pasivo	123		
Capital Efectivo	102		
Capital Propio Tributario Positivo	645		
Capital Propio Tributario Negativo	646		
Patrimonio Financiero	843		
Capital Enterado	844		

RECUADRO Nº 4 DATOS ART. 57 BIS LETRA A Y CONTABILIDAD COMPLETA

Total A.N.P. del Ejercicio	701		+
A.N.P. utilizado en el Ejercicio	702		−
Remanente A.N.P. Ejercicio Siguiente	703		=
Total A.N.N. del Ejercicio	704		
Base Débito Fiscal del Ejercicio	705		

RECUADRO Nº5: ENAJENACION DE ACCIONES

Régimen Tributario de la LIR	Nº Acciones Vendidas	Monto Total Venta Actualizado	Costo de Venta Total Actualizado
Régimen General	796	799	802
Régimen Impto. Único 1ª Categ.	797	800	803
Régimen Art. 18 Ter. (Ver Instr.)	798	801	804

RECUADRO Nº 6: DATOS DEL FUT

Saldo rentas e ingresos al 31.12.83	224		
Remanente FUT ejercicio anterior con crédito	774		+
Remanente FUT ejercicio anterior sin crédito	775		+
Saldo negativo ejercicio anterior	284		−
R.L.I.1ª Categoría del ejercicio	225		+
Pérdida Tributaria 1ª Categoría del ejercicio	229		−
Gastos Rechazados afectos al Art. 21	623		−
Gastos Rechazados no gravados con el Art. 21	624		−
Inversiones recibidas en el ejercicio (Art. 14)	227		+
Diferencia entre depreciación acelerada y normal	776		+
Dividendos y retiros recibidos, participaciones en contabilidad simplificada y otras provenientes de otras empresas	777		+
FUT devengado recibido de sociedades de personas	781		+
FUT devengado traspasado a sociedades de personas	821		−
Reposición Pérdida Tributaria	782		+
Rentas presuntas o participación en rentas presuntas	835		+
Otras Partidas que se agregan	791		+
Partidas que se deducen (Retiros presuntos, etc.)	275		−
Retiros o Distrib. Imputados al FUT en el ejercicio	226		−
Remanente FUT para el Ejerc. Sgte., con crédito	231		=
Remanente FUT para el Ejerc. Sgte., sin crédito	318		=
Saldo negativo para el ejercicio siguiente	232		=
Remanente Crédito Impto. 1ª Categ. ejercicio anterior	625		+
Crédito Impto. 1ª Categ. del Ejercicio	626		+
Crédito Impto. 1ª Categ. Utilizado en el ejercicio	627		−
Remanente Crédito Impuesto 1ª Categ. ejercicio siguiente	838		=
Saldo ajustado por diferencia entre depreciación acelerada y normal (Art. 31 Nº 5 LIR)	845		
Remanente FUNT ejercicio anterior	818		+
Saldo negativo FUNT ejercicio anterior	842		−
FUNT positivo generado en el ejercicio	819		+
FUNT negativo generado en el ejercicio	837		−
Retiros o Distrib. Imputados al FUNT en el ejercicio	820		−
Remanente FUNT para el ejercicio siguiente	228		=
Saldo negativo FUNT para el ejercicio siguiente	840		=
Dividendos afectos no imputados al FUT	836		
Exceso de retiros para el ejercicio siguiente	320		
Crédito IEAM ejercicio	828		+
Crédito IEAM utilizado en el ejercicio	830		-
Remanente crédito IEAM a devolver	829		=

RECUADRO Nº 7: CREDITOS IMPUTABLES AL IMPUESTO PRIMERA CATEGORIA (LINEA 24 y 26)

Crédito por contribuciones de bienes raíces	365	
Crédito por rentas de Fondos Mutuos sin derecho a Devolución	368	
Crédito por donaciones para fines culturales	373	
Crédito por donaciones para fines educacionales	382	
Crédito por donaciones para fines deportivos	761	
Crédito por donaciones para fines sociales	773	
Crédito por inversiones en exterior, según Arts. 41 A letra A y 41 C	841	
Crédito por rentas de zonas francas y otros	392	
Crédito por Impuesto Específico a la Minería (Art. 7º tr. Ley Nº 20.026/2005)	831	
Crédito por bienes físicos del activo inmovilizado del ejercicio	366	
Remanente de Crédito por bienes físicos del activo inmovilizado proveniente de inversiones A.T. 1999-2002	839	
Crédito por donaciones Universidades e Inst. Profesionales	384	
Crédito por Impto. 1ª Categ. Contribuyentes Art. 14 bis	385	
Crédito por inversiones Ley Arica	390	
Crédito por inversiones Ley Austral	742	
Crédito por inversiones en el exterior, según Arts. 41A letra B y C y 41 C	387	

RECUADRO Nº 8 CREDITOS DISPONIBLES

Saldo crédito contribuyentes Art. 14 bis	236	
Saldo crédito ex Art. 21: Tasa 40%	238	
Saldo crédito ex Art. 21: Tasa 30%	239	
Saldo crédito ex Art. 21: Tasa 15%	240	

Fecha de Presentación

650 Rut del Contador	903 Rut del Representante

Firma del Contribuyente o Nombre y Firma del Representante Legal

Declaro bajo juramento que la información contenida en este documento es la expresión fiel de la verdad, por lo que asumo la responsabilidad correspondiente.

ANEXO E: Información utilizada para construir el vector de características

Códigos de IVA utilizados en el estudio

Códigos F29	Descripción
503	Cantidad de facturas emitidas
502	Débito por facturas emitidas
110	Cantidad de boletas emitidas
111	Débito por boletas emitidas
509	Cantidad de notas de crédito emitidas por facturas
510	Débito de notas de crédito emitidas por facturas
538	Total de débitos
519	Cantidad de facturas del giro recibidas y facturas de compra emitidas
520	Crédito por facturas recibidas
524	Cantidad de facturas por activo fijo
525	Crédito por facturas de activo fijo
527	Cantidad de notas de crédito recibidas
528	Crédito por notas de crédito emitidas
504	Remanente de crédito fiscal del mes anterior
593	Solicitud devolución exportadores
537	Total de créditos
77	Remanente de crédito fiscal del mes siguiente
596	Retención por cambio de sujeto
89	Total IVA determinado
91	Total IVA a pagar en el plazo legal

Códigos de Renta utilizados en el estudio

Códigos F22	Descripción
628	Ingresos del giro percibidos o devengados
651	Otro ingresos pagados o devengados
630	Costo directo de bienes y servicios
631	Remuneraciones
635	Otros gastos deducidos de los ingresos brutos
639	Gastos rechazados
634	Pérdida de los ejercicios anteriores
643	Renta líquida imponible o pérdida tributaria
600	Retiros por crédito impuesto primera categoría
170	Base imponible del global complementario
304	Impuesto global complementario
18	Base imponible de primera categoría rentas efectivas
305	Resultado liquidación anual del impuesto a la renta
101	Saldo caja (dinero efectivo y documentos al día según arqueo)
784	Saldo cuenta corriente bancaria según conciliación
778	Cuentas por cobrar a empresas relacionadas
816	Cuentas por cobrar a empresas no relacionadas
783	Total préstamos a socios en ejercicio
129	Existencia Final
647	Activo inmovilizado
785	Depreciación tributaria del ejercicio
648	Bienes adquiridos por contrato leasing
815	Monto inversión Ley Arica
741	Monto inversión Ley Austral
122	Total activos

779	Cuentas por pagar a empresas relacionadas
817	Cuentas por pagar a empresas no relacionadas
123	Total pasivos
102	Capital efectivo
645	Capital propio tributario positivo
646	Capital propio tributario negativo

Ratios Tributarios Generados en el estudio

Ratio	Descripción
Deb/Cred Total	Débito total anual / Crédito total anual
Promedio Deb/Cred	Promedio del ratio débito/crédito mensual en el año
Desv Deb/Cred	Desviación standard del ratio débito/crédito mensual en el año
Deb/Fact	Débito facturas emitidas/ Cantidad de facturas emitidas en el año
Deb/Bol	Débito boletas/ Cantidad de boletas emitidas en el año
Deb/NCred	Débito notas de débito / Cantidad de notas de débito del año
DebFact/DebTot	Débito facturas emitidas/ Débito total
Cred/Fact	Crédito facturas recibidas / Cantidad de facturas recibidas
Cred/ActFijo	Crédito facturas activo fijo /Cantidad de facturas de activo Fijo
Cred/NCred	Crédito notas de crédito /Cantidad de notas de crédito
CredFact/CredTot	Crédito facturas recibidas /Crédito total
Rem/Cred	Remanente promedio/Crédito promedio
IVA/Fact	IVA determinado/ Facturas recibidas del giro
IVA/Ing	IVA determinado/ Total ingresos renta
PagoIVA/Ing	Pago de IVA / Total ingresos renta
IVANeto	IVA determinado + Remanente períodos anteriores – Remanente períodos siguientes
DevExp/Cred	Devolución IVA exportadores /Crédito total
Factem/Factim	Cantidad de facturas emitidas en el año/ Cantidad de facturas timbradas en los últimos tres años
RLI/Ingreso	Renta líquida imponible /Ingresos
Ingreso/Costo	Ingresos renta percibidos /Costo directo de bienes y servicios
Exist/Costo	Existencias/Costo directo de bienes y servicios
Gastos/Ingreso	Costos directo de bienes y servicios + Gastos/ Ingresos renta percibidos
RLI/Activo	Renta líquida imponible / Activos
Ingreso/Activo	Ingresos del giro / Activos
Costo/Activo	Costo directo de bienes y servicios/ Activos
Rem/Activo	Remuneraciones / Activos
OtGastos/Activo	Otros gastos / Activos
GastosRec/Activo	Gastos rechazados/ Activos
PerdAnt/Activo	Perdida ejercicios anteriores/Activos
Capef/Activo	Capital efectivo /Activos
Activo/Pasivo	Activos/Pasivos

Características del contribuyente utilizadas en el estudio

Variable	Descripción
Edad	Edad del contribuyente al año 2006
Antigüedad	Antigüedad de la empresa al año 2006
NroRegiones	N° de regiones en que la empresa tiene sucursales
NroComunas	N° de comunas en que la empresa tiene sucursales
NroSucursales	N° de sucursales activas de la empresa
FactElectronico	La empresa es facturadora electrónica
ContabComput	La empresa registra marca de contabilidad computacional

ActecoAct	N° de actecos activos al año 2006
ActecoTerm	N° de actecos inactivos al año 2006
ActecosDF	N° de actecos de difícil fiscalización activos al año 2006
Cambiosujeto	N° de actividades de marcas de cambio de sujeto vigentes
DeclaInternet	La empresa registra marca declaración por Internet de manera obligada
Espropia	El domicilio principal de la empresa es propio (0=No, 1=Si)
NroSucPropias	N° de sucursales activas propias de la empresa

Características de los relacionados a los contribuyentes utilizados en el estudio

Variable	Descripción
NroSocios	N° de socios activos de la empresa al año 2006
NroRepAct	N° de representantes legales activos de la empresa al año 2006
NroRepInac	N° de representantes legales inactivos de la empresa al año 2006
NroSociedades	N° de sociedades en la que participa el contribuyente como socio
NroRepresentac	N° de empresas en la que el contribuyente es representante legal activo
MandatarioInves	N° de veces que registra anotaciones de mandatario investigado
RelacionadoBloq	N° de veces que registra bloqueo de socios y representantes legales
FamilarSCE	N° de veces que un familiar registra antecedentes en SCE
ArriendaFam	N° de familiares asociados a los domicilios de la empresa
NroDelitoSoc	N° de veces que un socio de la empresa registra antecedentes en SCE
NroDelitoRL	N° de veces que un representante legal de la empresa registra antecedentes en SCE
NroDelitoConta	N° de veces que un contador de la empresa registra antecedentes en SCE
NroDelitoMand	N° de veces que un mandatario de la empresa registra antecedentes en SCE
NroDelitosAsoc	N° de veces que un asociado de la empresa registra antecedentes en SCE

Información relativa al ciclo de vida de los contribuyentes utilizados en el estudio

Variable	Descripción
MovimientoPrev	Registra movimiento antes de realizar inicio de actividades en el SII
VerificaAct	N° de veces que registra verificación de actividades
VIANeg	N° de veces que registra marca de VIA negativa
VATNegSuc	N° de veces que registra marca VAT de sucursal negativa
VIATimbInicial	N° de veces que registra VIA condicionada a los primeros timbrajes
ModifCompleja	N° de veces que registra alguna modificación compleja
FrecTimb06	N° de timbrajes de facturas realizados en el año 2006
FrecProm	N° promedio de timbrajes de facturas realizados en los 3 últimos años
FrecTotal	N° de timbrajes de facturas realizados en los 3 últimos años
Fact06	Cantidad de facturas timbradas en el año 2006
Fact0406	Cantidad de facturas timbradas en los 3 últimos años
Factprom0406	Cantidad promedio de facturas timbradas en los 3 últimos años
Factmax0406	Cantidad máxima de facturas timbradas en los 3 últimos años
MesUltimbraje	Meses transcurridos desde el último timbraje a diciembre 2006
Tgiro	La empresa tiene antecedentes de término de giro previos

Información relacionada con incumplimiento tributario de los contribuyentes del estudio

Variable	Descripción
FF 2006	Registra Facturas Falsas en el año 2006
NroFFPrev	N° de expedientes previos en que registra facturas falsas en SCE
NroVFPrev	N° de expedientes previos en que registra venta de facturas en SCE
NroIRRPrev	N° de expedientes previos en que registra irregularidades en SCE
NroDelitoPrev	N° de expedientes previos en que registra delito en SCE

Investigado	N° de veces que la empresa registra marca investigado
AlertaNomDF	N° de veces que registra marca de alerta de difícil fiscalización
VenceArriendo	Registra marca de vencimiento de arriendo vigente al año
DomInexistente	N° de anotaciones de domicilio inexistente de la empresa histórico
NoUbicadoRcte	N° de anotaciones de no ubicado de la empresa reciente (dos años)
NoubicadoTotal	N° de anotaciones de no ubicado de la empresa total
Deudaregularizada	N° de veces que registra deuda regularizada
DestruccDoc	N° de veces que registra destrucción de documentos
PérdidaRut	N° de veces que registra marca de pérdida de Rut
PerdFact0406	N° de veces que registra pérdida de facturas en los 3 últimos años
PerdFacHist	N° de veces que registra pérdida de facturas históricas
FactObsRcte	N° de veces que registra marca facturas observadas en los 3 últimos años
FactObsHist	N° de veces que registra marca de facturas observadas previas
TimbRestring	N° de veces que registra marca de timbraje restringido
Denuncios06	N° de denuncios cursados al contribuyente durante el año 2006
DenunciosHist	N° de denuncios históricos cursados al contribuyente
Clasuras06	N° de clausuras cursadas al contribuyente durante el año 2006
ClausurasHist	N° de clausuras históricas cursados al contribuyente
InconRta06	N° de inconcurrencias por procesos de renta en el año 2006
InconRtaHist	N° de inconcurrencias por procesos de renta previos al año 2006
InconIVA06	N° de inconcurrencias a procesos de IVA en el año 2006
InconIVAHist	N° de inconcurrencias por procesos de IVA previos al año 2006
InconNotif06	N° de inconcurrencias a notificaciones en el año 2006
InconNotfHist	N° de inconcurrencias a notificaciones previas al año 2006
InconTot06	N° de inconcurrencias totales registradas en el año 2006
InconTotHist	N° de inconcurrencias totales registradas antes del año 2006
GirosTotHist	N° de Giros totales registrados previo al 2006
MarcaPrevent	N° de veces que registra marca preventivas de un fiscalizador
NotifF29Hist	N° de veces que el contribuyente registra notificaciones (fuera de plazo F29 o de recopilación de antecedentes)
NotifFiscHist	N° de veces que el contribuyente registra notificaciones de fiscalización
FiscaSelHist	N° de Fiscalizaciones selectivas realizadas previo al 2006
FiscaSelPos	N° de Fiscalizaciones Selectivas con resultado positivo previo al 2006
FiscaSelNeg	N° de Fiscalizaciones Selectivas con resultado negativo previo al 2006
RendTotal	Montos de fiscalizaciones selectivas realizadas
RendLiquid	Montos de fiscalizaciones selectivas realizadas que no paga
RendRect	Montos de fiscalizaciones selectivas realizadas que paga voluntariamente
RendOculto	Montos de fiscalizaciones selectivas realizadas en forma oculta

ANEXO F: Resultados Análisis Componentes Principales

- Grupo Micro y Pequeñas Empresas

Rotated Component Matrix(a)

	Component														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Fact0406_01	**,952**	,026	,026	,122	,012	,020	,016	,034	,015	,008	-,002	,002	,004	,023	,005
Factmax0406_01	**,936**	,032	,028	,055	,029	,007	,018	,037	,025	,015	,002	-,001	,001	,034	,011
Factprom0406_01	**,932**	,038	,030	,010	,030	-,015	,040	,036	,019	,020	-,009	,000	-,002	,036	,016
Facturas06_01	**,893**	,016	,014	,196	-,001	,004	-,007	,024	,013	-,001	-,009	,001	,014	,014	-,010
FactObsHist_01	,041	**,975**	-,008	,011	,006	,006	,016	,012	,006	,012	,006	,002	-,017	,015	-,003
FactObsRcte_01	,037	**,928**	-,017	,012	,002	,004	,010	,011	,002	,011	,008	-,001	-,037	,019	-,007
DelitoPrev_01	,023	**,766**	,086	,003	,022	,041	,016	,003	,012	,001	,007	,014	,094	-,022	,017
FiscSelecPos_01	,008	,024	**,872**	,012	,044	,067	,049	,023	,024	,130	-,004	,009	,043	,011	,024
RendTotal_01	,008	,010	**,748**	-,020	,068	-,011	-,005	-,002	,020	-,186	-,014	-,017	,052	,002	-,039
FiscSelecPrev_01	,019	,026	**,710**	,009	,067	,070	,061	,023	,041	,638	-,004	,012	,045	-,005	,008
NotifFiscPrev_01	,079	,033	**,407**	,059	-,016	,194	,030	,037	-,012	,173	,040	,017	-,093	,081	,086
FrecTimb06_01	,166	,005	-,004	**,821**	-,014	,059	-,144	,019	-,044	-,009	,000	,013	,036	,031	-,049
FrecProm_01	,251	,015	,038	**,791**	-,011	,282	,005	,069	-,026	-,013	,018	,013	,011	,079	-,026
MesesUltimb_01	-,009	-,008	-,013	**-,679**	-,025	,123	,082	,072	-,038	-,019	-,072	,004	,041	,060	-,034
NRepresentac_01	,024	,015	,041	,004	**,906**	,021	,046	,000	,065	,041	-,001	-,005	-,033	,094	,051
NSociedades_01	,023	,013	,044	,008	**,901**	,012	,067	-,002	,085	,053	-,005	-,002	-,037	,089	,065
RelacionadoBloq_01	,018	,008	,061	-,009	**,368**	,074	-,042	-,014	-,030	-,026	,027	,032	,203	-,184	-,209
InconTotPrev_01	-,003	,010	,090	,003	,076	**,723**	,167	,069	,062	,058	-,009	-,006	,180	,024	-,008
GirosPrev_01	-,014	,015	,156	,203	-,059	**,678**	,081	,074	-,047	-,031	-,002	-,005	-,021	,105	,029
InconTot06_01	,012	-,002	-,015	-,050	,049	**,657**	-,102	-,052	,044	,016	,025	,007	-,009	-,065	,023
NoubicadoRcte_01	,003	,096	,028	-,016	,025	**,231**	-,064	,170	,028	-,020	,219	,004	,062	-,126	-,003
Edad_01	,031	,011	,019	-,149	,014	,055	**,742**	,015	-,019	,027	-,075	,011	,088	,007	-,045
Espropia_01	-,023	,010	,026	,010	-,002	-,118	**,721**	-,038	,014	,007	-,033	,007	-,051	-,099	,043
Antig_01	,083	,022	,050	-,124	,093	,241	**,648**	,124	,116	,034	-,071	,011	,048	,161	-,033
ClausurasHist_01	,040	,000	,012	,012	,001	,003	,017	**,890**	,039	,009	,006	,002	,008	-,025	,025
DenunciosHist_01	,074	,013	,041	-,020	-,014	,094	,047	**,873**	,058	,008	-,007	,002	,087	,029	-,003
Nsucursales_01	,041	,007	,027	,012	,021	,004	,090	,098	**,767**	,035	-,012	,019	-,036	,027	-,003
Ncomunas_01	,034	,004	,001	-,067	,084	,101	-,064	,013	**,722**	,011	,014	-,015	,044	,141	-,087
ActecoDF06_01	-,061	,025	,059	,150	,046	-,109	,310	-,065	**,432**	-,013	,303	,007	,046	-,271	,167
FiscSelecNeg_01	,025	,011	,091	-,007	,056	,014	,035	,005	,036	**,920**	-,001	,007	,015	-,023	-,027
VlANegativa_01	,005	-,008	-,004	-,045	-,016	,052	,028	-,020	-,115	,014	**,785**	,011	-,025	,122	-,038
VerificaAct_01	-,018	-,002	,002	,163	,019	-,023	-,220	,002	,184	-,009	**,695**	,018	,043	-,066	,031
DelitoMand_01	,002	,000	,010	,025	,001	,009	-,022	-,003	,003	,003	,001	**,772**	,025	-,017	,031
DelitoConta_01	-,001	,013	-,004	-,010	,010	-,008	,045	,007	,003	,006	,023	**,767**	-,014	,036	-,020
PerdFactHist_01	,015	-,008	,027	,028	,006	-,070	,039	,033	-,053	-,028	,002	,006	**,678**	,184	-,040
NoubicadoHist_01	,007	,036	-,009	-,043	,004	,218	,004	,080	,037	,028	,050	,006	**,634**	-,093	-,004
FactElect_01	-,009	,005	,044	-,001	-,076	-,081	-,047	-,057	,115	-,085	-,013	,020	,110	**,575**	,016
ContabCompleta_01	,095	,010	,038	,019	,183	,108	,039	,048	-,003	,050	,042	,011	-,083	**,535**	-,012
DelitoRL_01	-,009	,003	-,028	,027	,019	-,001	-,022	-,024	,121	,083	-,082	,040	,052	-,039	**-,530**
CambioSujeto_01	,003	-,010	-,037	,027	,015	-,023	,037	,005	,013	,194	-,013	-,031	,065	,306	**,507**
DelitoFam_01	,022	-,003	,043	-,028	,028	,082	-,141	,003	,044	-,084	-,085	,116	-,051	-,216	**,479**
AlertaNDF06_01	-,024	,056	,025	,030	,019	,009	,085	-,041	,131	,088	-,061	-,006	,348	-,142	**,372**

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 15 iterations.

Communalities

Variable	Initial	Extraction	Variable	Initial	Extraction	Variable	Initial	Extraction	Variable	Initial	Extraction
FactObsHist_01	1,000	,954	ClausurasHist_01	1,000	,798	Edad_01	1,000	,594	InconTot06_01	1,000	,457
FiscSelecPrev_01	1,000	,930	DenunciosHist_01	1,000	,792	DelitoConta_01	1,000	,594	CambioSujeto_01	1,000	,399
Fact0406_01	1,000	,925	FiscSelecPos_01	1,000	,790	Ncomunas_01	1,000	,579	FactElect_01	1,000	,384
Factmax0406_01	1,000	,886	FrecProm_01	1,000	,784	Antig_01	1,000	,577	ContabCompleta_01	1,000	,358
Factprom0406_01	1,000	,878	FrecTimb06_01	1,000	,733	GirosPrev_01	1,000	,556	DelitoFam_01	1,000	,339
FactObsRcte_01	1,000	,866	VlANegativa_01	1,000	,653	Espropia_01	1,000	,552	AlertaNDF06_01	1,000	,323
FiscSelecNeg_01	1,000	,863	Nsucursales_01	1,000	,613	ActecoDF06_01	1,000	,527	DelitoRL_01	1,000	,318
NRepresentac_01	1,000	,844	InconTotPrev_01	1,000	,609	PerdFactHist_01	1,000	,508	NotifFiscPrev_01	1,000	,271
NSociedades_01	1,000	,842	DelitoPrev_01	1,000	,607	MesesUltimb_01	1,000	,503	RelacionadoBloq_01	1,000	,269
Facturas06_01	1,000	,838	RendTotal_01	1,000	,605	NoubicadoHist_01	1,000	,472	NoubicadoRcte_01	1,000	,166
VerificaAct_01	1,000	,601	DelitoMand_01	1,000	,600						

- Grupo Medianas y Grandes Empresas

Rotated Component Matrix(a)

	Component															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ncomunas_01	**,964**	,029	,019	,032	,045	,027	,023	,065	-,009	,058	,015	,020	,004	,012	-,006	-,004
Nsucursales_01	**,916**	,032	,053	,024	,089	-,012	,027	,021	,000	,064	,011	-,009	,034	,000	,005	,000
NRegiones_01	**,899**	,058	,003	,052	-,007	-,003	,027	,086	-,004	,066	,016	,011	,007	,019	-,004	,001
FiscSelecPrev_01	,056	**,937**	,025	,064	,047	-,040	,066	,071	-,003	,035	,169	-,037	,081	,067	,015	,034
FiscSelecNeg_01	,044	**,845**	-,004	,046	,029	-,056	-,123	,122	-,005	-,004	-,161	-,051	,081	,125	,038	,042
FiscSelecPos_01	,039	**,613**	,040	,056	,037	,000	,244	-,021	-,004	,072	,493	-,013	,031	-,044	-,016	,008
NotifFiscPrev_01	,017	**,464**	,014	,007	,026	,007	,221	-,096	,032	,006	,082	,162	-,112	-,104	,014	-,037
ActecoAct_01	,035	,005	**,854**	-,064	,046	,040	-,009	-,045	-,009	-,007	,013	-,011	,047	,007	-,027	-,024
ActecoDF06_01	-,034	,044	**,773**	-,034	-,052	,036	,031	-,123	,027	-,047	-,024	,025	,263	,012	-,041	-,059
DestrucDoc_01	,064	,005	**,586**	,021	,041	,067	,026	,112	,007	,026	-,019	,008	-,228	,013	,074	,112
DeclaInternet_01	,049	,117	-,046	**,843**	,067	-,203	,086	,063	-,024	,084	-,013	-,032	,042	,036	,103	,001
ContaComp_01	,055	-,006	-,046	**,792**	-,047	,156	-,052	,127	-,025	-,044	-,017	,004	-,035	-,065	-,151	,001
Antig_01	,030	,147	-,026	**,429**	,137	-,418	,208	-,048	-,016	,095	,080	-,054	,124	,181	,361	,039
ClausuraHist_01	,007	,034	,021	-,020	**,880**	,022	,019	-,013	,006	-,009	-,012	,015	,040	-,015	-,020	,006
DenunHist_01	,112	,061	,018	,051	**,855**	-,056	,073	-,032	,008	,097	,019	,018	-,002	,028	,026	-,003
ViaNegativa_01	-,005	-,020	-,064	,041	,000	**,803**	,022	-,038	-,018	,002	,037	-,038	,066	,056	,103	,022
VerificaActiv_01	,025	-,032	,314	-,098	-,014	**,686**	-,055	,073	,009	-,026	-,016	,133	-,073	,045	-,068	,016
GirosPrev_01	,055	,066	,052	,121	,027	-,022	**,661**	-,037	,013	,045	,191	,043	-,080	-,043	,003	,041
InconIVA06_01	-,023	-,001	-,042	-,304	,045	,033	**,558**	,012	-,004	-,045	-,193	,033	,081	,005	-,043	,052
InconTotPrev_01	,050	,200	,033	,166	,059	-,107	**,544**	,143	-,004	,025	,103	,042	,135	,221	,137	-,017
NReprinactivo_01	,066	,049	,039	,070	,011	-,012	,087	**,786**	-,008	,018	,049	,006	-,051	,061	,043	,004
NRepractivo_01	,085	,025	-,078	,080	-,056	,038	-,044	**,783**	,006	,051	,003	,002	,012	-,108	-,061	-,014
DelitoMandPrev_01	-,001	,017	,000	-,019	-,007	,000	,012	-,019	**,789**	-,009	-,003	-,003	-,010	,038	,001	,027
DelitoRLPrev_01	-,018	-,008	,011	-,008	,031	-,026	,023	,044	**,651**	,101	-,027	,008	,033	-,013	-,168	-,168
DelitoContaPrev_01	,016	,007	,021	-,030	-,020	,041	-,049	-,053	**,460**	-,089	,049	,004	-,039	-,024	,292	,351
IRRPrev_01	,008	-,024	,007	-,012	,003	,031	,031	-,002	,074	**,695**	,137	-,062	,019	,085	,013	-,008
Factprom0406_01	,203	,084	-,053	,066	,090	-,083	,001	,102	-,034	**,664**	-,102	,003	-,047	-,025	,071	,034
RendTotal_01	,030	,288	,005	,003	,007	,068	,147	,023	-,015	,109	**,663**	-,023	-,006	-,071	-,072	-,037
Mesultimbr_01	,015	-,127	-,140	-,021	-,016	-,124	-,132	,177	,017	-,315	**,408**	-,044	,094	,136	,236	,123
InconNotif06_01	,030	,013	-,013	,006	,012	,012	,114	-,038	-,005	-,078	-,050	**,638**	-,042	,041	-,070	-,033
NoubicadoReciente_01	,000	,012	,024	-,055	,014	,079	,045	,056	-,004	-,073	-,016	**,584**	-,011	,116	,020	-,057
FacObsRcte_01	-,030	-,008	,023	,039	,006	-,087	-,201	-,032	,025	,319	,100	**,570**	,155	-,089	,091	,170
AlertaNominaDF06_01	,037	-,032	,044	-,015	,035	-,012	-,051	-,032	,007	-,065	,136	,059	**,748**	-,007	-,017	,014
Cambiosujeto06_01	,012	,192	,023	,056	,010	,018	,228	,002	-,005	,096	-,319	-,090	**,523**	-,033	,023	,014
NoubicadoHist_01	,025	,004	,013	,064	,032	,069	,028	,022	,015	-,017	,021	,043	,082	**,651**	,072	-,058
tgiro_01	-,003	-,004	,039	-,136	,007	-,064	-,058	-,111	-,052	-,003	-,014	,011	-,190	**,605**	-,085	,055
MarcaPrevent_01	,002	,089	-,025	,071	-,050	,081	,174	,044	,083	,102	-,062	,152	,046	**,323**	-,030	,043
TimbrajeRestr_01	,013	-,021	-,004	,023	,007	-,032	-,027	,005	,081	-,080	,030	,010	,009	,025	**-,769**	,123
DeudaRegul_01	-,015	,003	,013	-,004	,002	,008	,065	,009	-,082	,078	,025	-,047	,060	,056	-,195	**,822**
PerdidaRut_01	,024	,103	,036	,034	,020	,048	,060	-,016	,038	-,130	-,211	,056	-,135	-,209	,215	**,297**

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

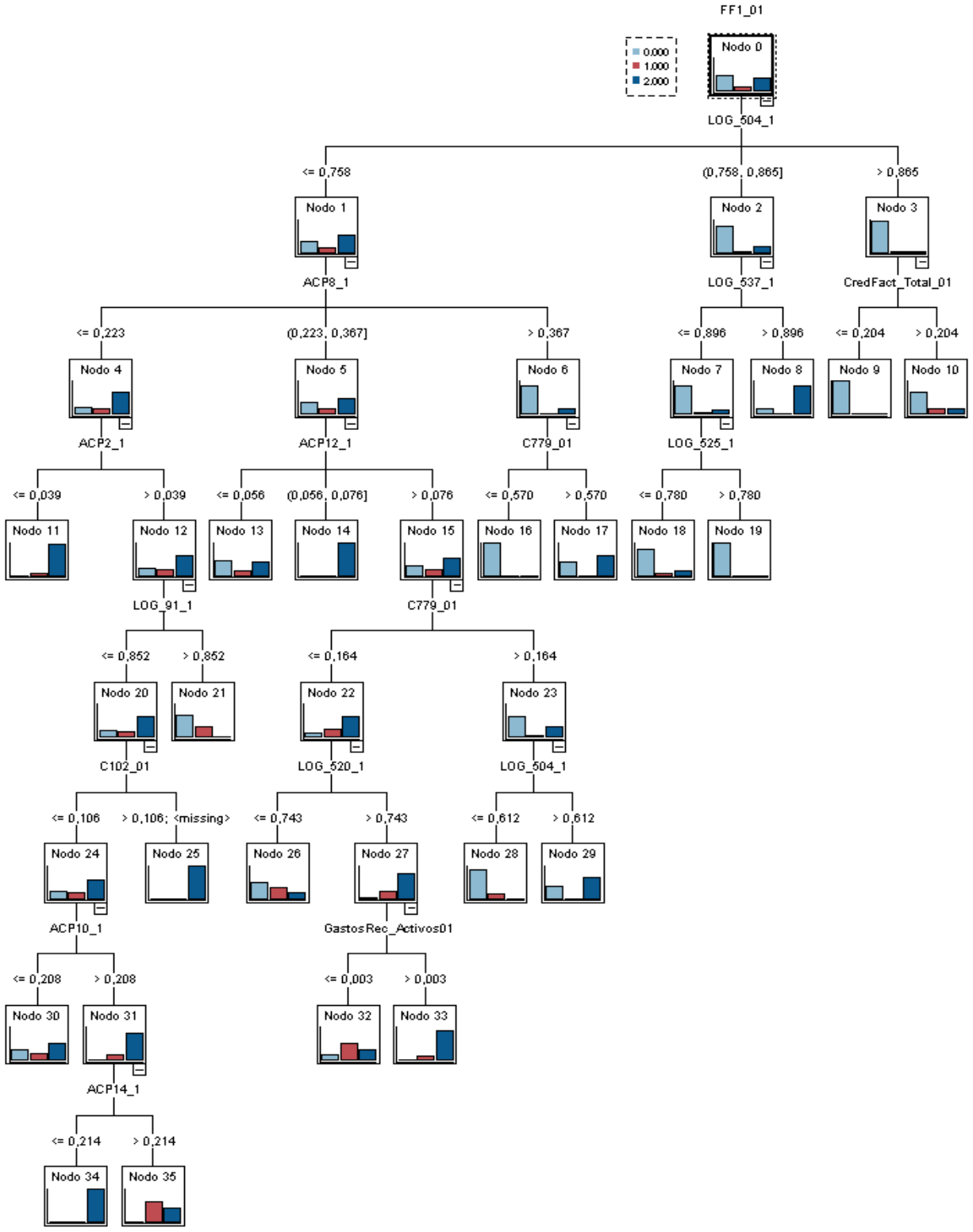
a. Rotation converged in 18 iterations.

Communalities

Variable	Initial	Extraction	Variable	Initial	Extraction	Variable	Initial	Extraction	Variable	Initial	Extraction
Antig_01	1,000	,644	ContaComp_01	1,000	,709	Cambiosujeto06_01	1,000	,487	ActecoAct_01	1,000	,744
NRepractivo_01	1,000	,658	Ncomunas_01	1,000	,943	tgiro_01	1,000	,456	DelitoContaPrev_01	1,000	,441
Nsucursales_01	1,000	,858	NRegiones_01	1,000	,828	PerdidaRut_01	1,000	,283	InconNotif06_01	1,000	,442
NReprinactivo_01	1,000	,650	DelitoRLPrev_01	1,000	,497	DeudaRegul_01	1,000	,742	GirosPrev_01	1,000	,515
FiscSelecPrev_01	1,000	,942	InconTotPrev_01	1,000	,502	NoubicadoHist_01	1,000	,454	MarcaPrevent_01	1,000	,208
FiscSelecNeg_01	1,000	,806	FiscSelecPos_01	1,000	,695	Mesultimbr_01	1,000	,467	Factprom0406_01	1,000	,543
DeclaInternet_01	1,000	,808	IRRPrev_01	1,000	,521	NotifFiscPrev_01	1,000	,334	VerificaActiv_01	1,000	,619
InconIVA06_01	1,000	,461	ActecoDF06_01	1,000	,701	DelitoMandPrev_01	1,000	,626	DestrucDoc_01	1,000	,438
ClausuraHist_01	1,000	,780	DenunHist_01	1,000	,771	AlertaNominaDF06_01	1,000	,595	NoubicadoReciente_01	1,000	,379
RendTotal_01	1,000	,575	TimbrajeRestr_01	1,000	,624	FacObsRcte_01	1,000	,558	ViaNegativa_01	1,000	,674

ANEXO G: Árbol de Decisión - Micro y Pequeñas Empresas

ANEXO H: Árbol de Decisión - Medianas y Grandes Empresas



ANEXO I: Reglas Predictivas del Árbol de Decisión - Micro y Pequeñas Empresas

LOG_CFTOT_1 <= 0,06 [Mode: 0]
    ACP10_1 <= 0,06 [Mode: 0]
        LOG_REMCRED_1 <= 0,33 [Mode: 1] => **Fraude**
        LOG_REMCRED_1 > 0,33 [Mode: 0] => ***Sin Fraude***
    ACP10_1 > 0,06 [Mode: 0] => ***Sin Fraude***
LOG_CFTOT_1 IS MISSING [Mode: 0]
    ACP10_1 <= 0,06 [Mode: 1]
        LOG_SUMA89_1 <= 0,80 [Mode: 1] => **Fraude**
        LOG_SUMA89_1 > 0,80 [Mode: 2] => **Fraude**
    ACP10_1 > 0,06 [Mode: 0] => ***Sin Fraude***
LOG_CFTOT_1 > 0,24 and LOG_CFTOT_1 <= 0,44 [Mode: 2]
    ACP10_1 <= 0,03 [Mode: 2]
        ACP2_1 <= 0,04 [Mode: 2]
            ACP4_1 <= 0,32 [Mode: 1] => **Fraude**
            ACP4_1 > 0,32 and ACP4_1 <= 0,38 [Mode: 1] => **Fraude**
            ACP4_1 > 0,38 and ACP4_1 <= 0,46 [Mode: 2] => **Fraude**
            ACP4_1 > 0,46 [Mode: 2] => **Fraude**
        ACP2_1 > 0,04 [Mode: 2] => **Fraude**
    ACP10_1 > 0,03 and ACP10_1 <= 0,19 [Mode: 1]
        ACP10_1 <= 0,06 [Mode: 1] => **Fraude**
        ACP10_1 > 0,06 [Mode: 2] => **Fraude**
    ACP10_1 > 0,19 [Mode: 0] => ***Sin Fraude***
LOG_CFTOT_1 > 0,44 [Mode: 2]
    ACP10_1 <= 0,03 [Mode: 2]
        LOG_CFACT_1 <= 0,59 [Mode: 2]
            LOG_SUMA503_1 <= 0,15 [Mode: 1] => **Fraude**
            LOG_SUMA503_1 > 0,15 [Mode: 2]
                ACP2_1 <= 0 [Mode: 2]
                    LOG_SUMA91_1 <= 0,84 [Mode: 2] => **Fraude**
                    LOG_SUMA91_1 > 0,84 [Mode: 1] => **Fraude**
                ACP2_1 > 0 [Mode: 2] => **Fraude**
        LOG_CFACT_1 > 0,59 and LOG_CFACT_1 <= 0,63 [Mode: 2]
            LOG_FEMTIM_1 <= 0,06 [Mode: 2] => **Fraude**
            LOG_FEMTIM_1 IS MISSING [Mode: 2]
                ACP2_1 <= 0,04 [Mode: 2] => **Fraude**
                ACP2_1 > 0,04 [Mode: 2] => **Fraude**
            LOG_FEMTIM_1 > 0,11 [Mode: 2] => **Fraude**
        LOG_CFACT_1 > 0,63 and LOG_CFACT_1 <= 0,66 [Mode: 2]
            LOG_SUMA111_1 <= 0 [Mode: 1] => **Fraude**

LOG_SUMA111_1 > 0 [Mode: 2] => **Fraude**
LOG_CFACT_1 > 0,66 and LOG_CFACT_1 <= 0,74 [Mode: 2]
ACP5_1 <= 0 [Mode: 2]
LOG_CTOACT_1 <= 0,21 or LOG_CTOACT_1 MISSING [Mode: 2]
LOG_SUMA519_1 <= 0,29 [Mode: 2] => **Fraude**
LOG_SUMA519_1 > 0,29 [Mode: 2] => **Fraude**
LOG_CTOACT_1 > 0,21 [Mode: 2] => **Fraude**
ACP5_1 > 0 [Mode: 2] => **Fraude**
LOG_CFACT_1 > 0,74 [Mode: 2] => **Fraude**
ACP10_1 > 0,03 and ACP10_1 <= 0,19 [Mode: 2]
ACP2_1 <= 0,04 [Mode: 1]
ACP3_1 <= 0,05 [Mode: 1] => **Fraude**
ACP3_1 > 0,05 and ACP3_1 <= 0,09 [Mode: 1] => **Fraude**
ACP3_1 > 0,09 [Mode: 1] => **Fraude**
ACP2_1 > 0,04 [Mode: 2] => **Fraude**
ACP10_1 > 0,19 [Mode: 0] => ***Sin Fraude***

ANEXO J: Reglas Predictivas del Árbol de Decisión - Medianas y Grandes Empresas

LOG_504_1 <= 0,76 [Mode: 2]
ACP8_1 <= 0,22 [Mode: 2]
ACP2_1 <= 0,04 [Mode: 2] => **Fraude**
ACP2_1 > 0,04 [Mode: 2]
LOG_91_1 <= 0,85 [Mode: 2]
C102_01 <= 0,11 [Mode: 2]
ACP10_1 <= 0,21 [Mode: 2] => **Fraude**
ACP10_1 > 0,21 [Mode: 2]
ACP14_1 <= 0,21 [Mode: 2] => **Fraude**
ACP14_1 > 0,21 [Mode: 1] => **Fraude**
C102_01 > 0,11 or C102_01 MISSING [Mode: 2] => **Fraude**
LOG_91_1 > 0,85 [Mode: 0] => ***Sin Fraude***
ACP8_1 > 0,22 and ACP8_1 <= 0,37 [Mode: 2]
ACP12_1 <= 0,06 [Mode: 0] => ***Sin Fraude***
ACP12_1 > 0,06 and ACP12_1 <= 0,08 [Mode: 2] => **Fraude**
ACP12_1 > 0,08 [Mode: 2]
C779_01 <= 0,16 [Mode: 2]
LOG_520_1 <= 0,74 [Mode: 0] => ***Sin Fraude***
LOG_520_1 > 0,74 [Mode: 2]
GastosRec_Activos01 <= 0,00 [Mode: 1] => **Fraude**
GastosRec_Activos01 > 0,00 [Mode: 2] => **Fraude**
C779_01 > 0,16 [Mode: 0]
LOG_504_1 <= 0,61 [Mode: 0] => ***Sin Fraude***
LOG_504_1 > 0,61 [Mode: 2] => **Fraude**
ACP8_1 > 0,37 [Mode: 0]
C779_01 <= 0,57 [Mode: 0] => ***Sin Fraude***
C779_01 > 0,57 [Mode: 2] => **Fraude**
LOG_504_1 > 0,76 and LOG_504_1 <= 0,87 [Mode: 0]
LOG_537_1 <= 0,90 [Mode: 0]
LOG_525_1 <= 0,78 [Mode: 0] => ***Sin Fraude***
LOG_525_1 > 0,78 [Mode: 0] => ***Sin Fraude***
LOG_537_1 > 0,90 [Mode: 2] => **Fraude**
LOG_504_1 > 0,87 [Mode: 0]
CredFact_Total_01 <= 0,20 [Mode: 0] => ***Sin Fraude***
CredFact_Total_01 > 0,20 [Mode: 0] => ***Sin Fraude***

ANEXO K: Paper publicado en Revista de Ingeniería en Sistemas, Volumen XXV, Septiembre 2011 (Versión español)

CARACTERIZACIÓN DE CONTRIBUYENTES QUE PRESENTAN FACTURAS FALSAS AL SII MEDIANTE TÉCNICAS DE DATA MINING

Pamela Castellón González
Servicio de Impuestos Internos de Chile – pamela.castellon@sii.cl

Juan Velásquez Silva
Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas,
Universidad de Chile – jvelazquez@dii.uchile.cl

Resumen:

En este trabajo se entregan evidencias que es posible caracterizar y pronosticar a aquellos usuarios potenciales de facturas falsas en un año determinado, en función de la información de su pago de impuestos, el comportamiento histórico y sus características particulares, utilizando para ello distintas técnicas de Data Mining. En una primera instancia se aplican técnicas de SOM, Gas Neuronal y Árboles de Decisión para identificar aquellas variables que están relacionadas con un comportamiento de fraude y/o no fraude y detectar patrones de conducta asociada a esta problemática. Posteriormente se utilizan Redes Neuronales y Redes Bayesianas para establecer en qué medida se pueden predecir casos de fraude y no fraude con la información disponible. De esta forma se contribuye a identificar patrones de fraudes y generar conocimiento que pueda ser utilizado en la labor de fiscalización que realiza el Servicio de Impuestos Internos para detectar este tipo de delito tributario.

PALABRAS CLAVES: FACTURAS FALSAS, FRAUDE TRIBUTARIO, DATA MINING, CLUSTERIZACIÓN, PREDICCIÓN.

1. INTRODUCCIÓN

El fraude, en sus diversas manifestaciones, es un fenómeno del que no está libre ninguna sociedad moderna. Todas las instituciones, independiente de si son grandes o pequeñas, públicas o privadas, locales o multinacionales, se ven afectada por esta realidad que atenta gravemente contra los principios de solidaridad y de igualdad de los ciudadanos ante la Ley y pone en riesgo los negocios. De acuerdo a un estudio realizado por Ernst&Young en el año 2006 en el cual se encuestó a 150 empresas chilenas, medianas y grandes, un 41% de ellas declaró haber sido víctima de algún tipo de fraude en los dos últimos años [8]. Esto plantea grandes desafíos en materia de detección y prevención, considerando que el fraude normalmente es mayor que lo declarado por las empresas, debido a que de alguna manera se resiente la imagen de la compañía y en muchos casos, incluso, hay empresas que no están en conocimiento de que han sido víctimas de un fraude.

La Evasión Tributaria y el Fraude Fiscal un tema que ha sido una constante preocupación de todas las administraciones tributarias, en especial de aquellas pertenecientes a países en vías de desarrollo[50]. Si bien es cierto, los impuestos no son la única fuente de financiamiento de un gobierno, es un hecho que éstos marcan una señal muy importante respecto al compromiso y la eficacia con que el Estado puede ejecutar sus funciones, y condicionar el acceso a otras fuentes de ingresos. En el caso de Chile, los ingresos tributarios proporcionan aproximadamente un 75% de los recursos con que año a año el Estado sustenta sus gastos e inversiones, alcanzando durante el año 2010 un monto de $17,7 billones de pesos[51].

La utilización y venta de facturas falsas como mecanismo de evasión, es particularmente relevante, pues no sólo provoca una elusión de los impuestos, sino que en la mayoría de los casos implica un delito tributario. Por otra parte, junto a la generación de una merma en la recaudación, se producen efectos económicos negativos en el resto de las empresas, por el hecho de generar una competencia desleal frente a aquellas empresas que cumplen adecuadamente con sus obligaciones tributarias.

Asimismo, se requiere que los recursos invertidos en fiscalización sean bien enfocados, detectando a aquellos de mayor riesgo de cumplimiento y no importunar ni desperdiciar tiempo y recursos en aquellos que si cumplen con sus obligaciones. Para ello, las técnicas de data mining ofrecen un gran potencial, ya que permiten extraer y generar conocimiento de grandes volúmenes de datos para caracterizar y detectar conductas fraudulentas y de incumplimiento para optimizar el uso de los recursos.

Este artículo se organiza de la siguiente forma: en la sección 2 se describe la problemática e implicancias del uso de facturas falsas sobre la recaudación de los impuestos. La sección 3, describe la manera en que las técnicas de inteligencia artificial han facilitado la detección del fraude fiscal en otras administraciones tributarias. La sección 4 describe el acercamiento propuesto para caracterizar y detectar fraude en la emisión de facturas a través de las técnicas de data mining. La sección 5 presenta las principales conclusiones y las líneas de investigación futuras.

2. NECESIDAD DE DETECTAR FRAUDE EN UNA INSTITUCIÓN RECAUDADORA DE IMPUESTOS

El Servicio de Impuestos Internos (SII) es la Institución responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario y propiciar la reducción de los costos de cumplimiento, en pos del desarrollo económico de Chile y de su gente. Para ello cuenta con 4.183 funcionarios, de los cuales el 31% corresponde a fiscalizadores, quienes deben velar por el cumplimiento de 3.4 millones de contribuyentes, considerando los declarantes del Impuesto al Valor Agregado (IVA) y el Impuesto a la Renta.

Particularmente el IVA se ha convertido en un componente clave de la recaudación fiscal, representando durante el año 2010, el 47% del total de los ingresos tributarios recaudados, por un monto de $8,3 billones de pesos [19]. Actualmente existen 708 mil contribuyentes que declaran IVA, de los cuales 28.000 están autorizados para emitir facturas electrónicas, lo cual ha ido aumentando progresivamente desde el año 2003, como parte de la política adoptada por el SII para modernizar su gestión y asegurar la autenticidad de los emisores de documentos tributarios.

[50] Habitualmente se habla de “elusión fiscal” cuando se hace referencias a conductas que, dentro de la Ley, evitan o reducen el pago de impuestos, mientras que la “evasión o fraude fiscal” supone un quebrantamiento de la legalidad para obtener para obtener esos mismos resultados.

51 Información publicada en la Cuenta Pública SII 2010 de Marzo 2011, considerando los Ingresos Tributarios del Gobierno Central (sin incluir a Codelco, las Municipalidades y la Seguridad Social).

Del total de facturas emitidas, un 60% se emite en formato papel y un 40% en formato electrónico, generándose cerca de 400 millones de facturas al año.

El fenómeno de las facturas falsas respecto del IVA se explica por la mecánica de determinación del impuesto. Cuando una empresa recibe una factura falsa, aparenta con ello una compra que nunca existió, con lo que aumenta fraudulentamente su crédito fiscal y disminuye su pago de IVA. Asimismo se produce una disminución del pago en el Impuesto a la Renta, debido al aumento de los costos y gastos declarados.

La falsedad del documento puede ser "material", si en él se han adulterado los elementos físicos que conforman la factura o "ideológica", cuando la materialidad del documento no está alterada, pero las operaciones que en ella se consignan son adulteradas o inexistentes. Ésta última es más difícil y compleja de detectar, ya que implica transacciones ficticias, en las cuales se requiere una auditoría para revisar los libros de compra y las rectificaciones o la realización de cruces de información con proveedores. Por otra parte, estos casos son más costosos para el Servicio, ya que requieren una mayor cantidad de tiempo destinado a la recopilación de antecedentes y pruebas, las cuales son más difíciles de encontrar.

Los casos más conocidos de falsedad material son la adulteración física del documento, la utilización de facturas colgadas en la que se falsifica una factura para suplantar a un contribuyente de buen comportamiento tributario, y el uso de doble juego de facturas, en la que se tiene dos facturas de igual numeración pero una de ellas ficticia y por un monto mayor. En el caso de la falsedad ideológica se encuentran las facturas utilizadas para registrar una operación inexistente o que adulteran el contenido de una operación existente. Adicionalmente existen otros delitos comúnmente relacionados, como la falsificación del inicio de actividades a través de palos blancos, con la única finalidad de adquirir facturas timbradas que posteriormente son vendidas a otros contribuyentes.

De acuerdo a un método de estimación de la evasión del IVA por concepto de facturas falsas y otros abultamientos de créditos, aplicado en el período 1990-2004 por el SII, la evasión por facturas falsas ha representado entre un 15% y un 25% de la evasión total del IVA, aumentando considerablemente en años de crisis económicas. Es así como en el año 1992, el porcentaje de participación aumentó a un 30% y en la crisis del año 1998-1999 alcanza su punto máximo con un 38% de participación, año en que alcanza una cifra cercana a los $317.000 millones de pesos. Esto adquiere relevancia producto que recientemente se produjo una crisis económica mundial que afectó a Chile a fines del 2008 y mediados del 2009, provocando un aumento de la tasa de evasión del IVA a un 18%, por un monto evadido de $1,5 billones de pesos.

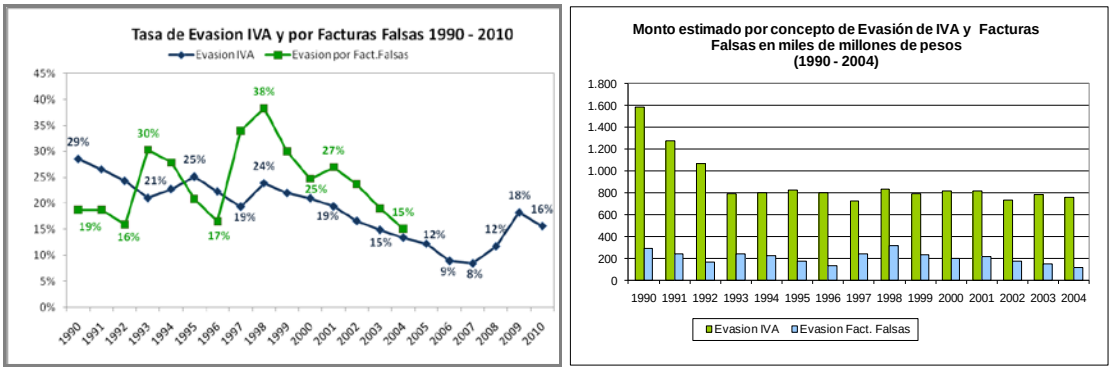


Figura N° 1: Tasa y Monto de Evasión en el IVA y por Facturas Falsas, Período 1990-2010,
Fuente: Subdirección de Estudios, SII

Asimismo, la detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera un importante costo administrativo para las áreas de fiscalización y jurídica. Durante el año 2010, el costo de recaudación de $100 fue de $0,91, es decir, aproximadamente un 1% del valor recaudado. En el período 2001-2007 se han presentado más de 2.300 querellas por facturas falsas y otros delitos de defensa judicial, las cuales involucraron a más de 4.000 querellados, por un monto de perjuicio fiscal cercano a los $274.130 millones de pesos.

ESTADISTICAS SCE	2001	2002	2003	2004	2005	2006	2007	Acumulado
Cantidad de Querellas	171	394	358	407	451	306	243	2.330
Cantidad de Querellados	371	835	667	839	801	537	386	4.436
Monto Perjuicio Fiscal (MM$)	29.370	36.407	49.751	58.812	47.856	21.620	30.314	274.130
Casos SCE[52]	830	2.081	1.794	1.609	1.553	1.052	870	9.789

Cuadro N° 1: Estadísticas de acciones legales relacionadas con facturas falsas 2001-2007
Fuente: Cuenta Pública SII, 2005, 2006, 2007

El SII utiliza diversos métodos para seleccionar contribuyentes a ser controlados. En el caso de las fiscalizaciones masivas, los contribuyentes se determinan como resultado de un proceso de cruce de información de las declaraciones recibidas y otras fuentes de información en la cual se detectan inconsistencias y diferencias tributarias. Las fiscalizaciones selectivas, en cambio, se generan en respuesta a determinadas figuras de evasión, ya sea a nivel nacional o local, utilizando para ello distintos ratios tributarios y condiciones, los cuales utilizan información parcial del contribuyente. Para ello resulta fundamental, aprovechar la gran cantidad de información disponible en los sistemas respecto del comportamiento de cada contribuyente en el tiempo.

3. TRABAJOS RELACIONADOS

La mayor parte de las administraciones tributarias planifican su lucha contra el fraude fiscal. No obstante, existen importantes diferencias en los mecanismos, alcances, enfoque, contenido y énfasis puestos en dicha labor. Para detectar el fraude fiscal, las instituciones comenzaron aplicando auditorías de selección aleatoria o enfocándose en aquellos casos que no tuvieran fiscalizaciones en períodos anteriores recientes y seleccionando casos de acuerdo a la experiencia y conocimiento de los auditores [18]. Posteriormente, se desarrollan metodologías basadas en análisis estadísticos y en la construcción de ratios tributarios o financieros, lo cual evolucionó a la creación de sistemas basados en reglas y modelos de riesgo, que transforman la información tributaria en indicadores que permitan rankear a los contribuyentes por riesgo de cumplimiento. Durante los últimos años, las técnicas de Data Mining e Inteligencia Artificial, han sido incorporadas en las actividades de planificación de auditorías, principalmente para detectar patrones de fraude o de evasión, las cuales han sido utilizadas por las instituciones tributarias con fines específicos.

La Internal Revenue Service, institución a cargo de administrar los impuestos en Estados Unidos, ha utilizado técnicas de Data Mining con distintos fines, entre los que se encuentran la medición del riesgo de cumplimiento de los contribuyentes, la detección de la evasión tributaria y actividades financieras delictivas, la detección de fraude electrónico, la detección de abusos en impuesto de las viviendas, la detección de fraude en contribuyentes que reciben ingresos

[52] El Sistema de Control de Expedientes (SCE) contiene los antecedentes de los casos asociados a delitos tributarios. Contiene más casos que el número de querellados, pues contiene la información de todos los contribuyentes involucrados en la investigación del caso, incluyendo a mandatarios, proveedores, socios, contadores, entre otros.

obtenidos por crédito fiscal y lavado de dinero [10]. Para ello ha utilizado modelos de regresión logística, árboles de decisión, redes neuronales, algoritmos de clustering y técnicas de visualización como Link Analysis, entre otros.

En la Administración Tributaria de Australia, el "Compliance Program" se basa en un modelo de riesgos, que utiliza estadísticas y Data Mining con el objetivo de realizar comparaciones, encontrar asociaciones y patrones mediante modelos de regresión logística, árboles de decisión y SVM [18]. Un caso de interés ha sido el enfoque utilizado por Denny, Williams y Christen [6] de descubrimiento de pequeños clusters o subpoblaciones inusuales, denominadas "Hot Spots", utilizando técnicas como el Self Organizing Map (SOM) para explorar sus características, algoritmos de agrupación como k-means y representaciones visuales, que son fáciles de entender para usuarios no técnicos.

En el caso de Nueva Zelanda, el modelo existente asocia el grado de cumplimiento con la atención del control, el cual coincide con el utilizado por la administración australiana [18]. El Plan incluye un análisis del entorno económico, internacional, poblacional, de diversidad étnica y de estructura familiar. Por su parte, Canadá utiliza redes neuronales y árboles de decisión para distinguir las características de los contribuyentes que evaden o cometen fraude, en base a los resultados de auditorías pasadas, para detectar los patrones de incumplimiento o evasión [18].

A nivel latinoamericano, Perú fue uno de los primeros en aplicar estas técnicas para detectar evasión tributaria, incorporando al sistema de selección en la Aduana Marítima del Callao una herramienta de inteligencia artificial basada en redes neuronales [3]. Durante el año 2004, este modelo fue mejorado a través de la aplicación de reglas difusas y de asociación para el pre-procesamiento de las variables y árboles de clasificación y regresión (CART) para seleccionar las variables más relevantes.

Por su parte, Brasil desarrolló el proyecto HARPIA (Análise de Risco Aduaneiro e Inteligência Artificial Aplicada) de manera conjunta entre la Brazilian Federal Revenue y las universidades de ese país [7]. Este proyecto consiste en desarrollar un sistema de detección de puntos atípicos que ayude a los fiscalizadores a identificar operaciones sospechosas basado en la visualización gráfica de información de importaciones y exportaciones históricas, y un sistema de información de exportación de productos, apoyado en cadenas de markov, para ayudar a los importadores en el registro y clasificación de sus productos, evitar duplicidades y calcular para la probabilidad de que una cadena es válida en un determinado dominio.

En el caso de Chile, la primera experiencia fue desarrollada en el año 2007, utilizando SOM y K-means para segmentar contribuyentes de IVA de acuerdo a sus declaraciones de F29 y características particulares [13]. Posteriormente, siguiendo la tendencia internacional, en el año 2009 se construyen modelos de riesgos en distintas etapas del ciclo de vida del contribuyente, en los que se aplican técnicas de redes neuronales, árboles de decisión y regresión logística. Adicionalmente se desarrolla la primera experiencia para detectar potenciales usuarios de facturas falsas a través de redes neuronales artificiales y árboles de decisión, utilizando principalmente información de su declaración de IVA y Renta en micro y pequeñas empresas.

4. APLICACIÓN DE DATA MINING PARA LA DETECCIÓN DE FRAUDE EN LA EMISIÓN DE FACTURAS

A diferencia del estudio anterior desarrollado en el año 2009 relacionado con esta problemática, este trabajo busca complementar el uso de información de impuestos con variables adicionales relacionadas a su comportamiento histórico y su comportamiento en el año de análisis, así como incluir aspectos concernientes a sus relacionados directos, tales como mandatarios, socios y representantes legales. Por otra parte, se desarrolla un modelo para

medianas y grandes empresas, en los que existe menor conocimiento de forma de operar respecto del uso de facturas falsas, debido a que tienen procedimientos más complejos de evasión.

4.1. DATOS UTILIZADOS

Para efectos de la caracterización se escoge el año 2006 como año de estudio. Si bien el peak de contribuyentes usuarios de facturas falsas detectados ocurre en el año 2002, se determina utilizar información más reciente, debido a que las dinámicas de evasión se van modificando en el tiempo, al igual que lo hicieron los formularios de pago de impuestos en ese período. Por otra parte, las auditorías se realizan hasta un período de 3 años atrás, lo que dificulta utilizar información más reciente, pues durante el año 2010 aún se estaban generando casos que podrían haber utilizado facturas falsas desde el año 2007 hacia adelante.

De esta forma, el universo queda compuesto por todos aquellos contribuyentes que hayan presentado al menos una declaración de IVA entre el año 2005 y 2007, correspondiente a 582.161 empresas. Para caracterizar a los casos de fraude/no fraude se utiliza información de aquellas auditorías en las que existe certeza que se le revisaron sus facturas del año 2006, independiente del momento en el que fue realizada, generando un total de 1.692 empresas.

Contribuyentes del análisis	MI y PE	ME y GR	Total
Empresas activas en el período 2005-2007	558.319 (96%)	23.842 (4%)	582.161 (100%)
Empresas auditadas por facturas en el 2006 con resultado de fraude o no fraude conocido	1.280 (76%)	412 (24%)	1.692 (100%)

Cuadro Nº 2: Número de contribuyentes utilizados en el análisis

Uno de los mayores inconvenientes para obtener la información de casos con fraude y no fraude se produce por la forma en la que se registra la información, pues se conoce la fecha de inicio y término de la auditoría, así como los períodos tributarios revisados y el resultado obtenido, pero la información de los períodos en los que ocurren las diferencias no está automatizada. Por lo tanto, para saber si la factura falsa detectada correspondía al año 2006 específicamente, hubo que revisar las anotaciones y comentarios efectuados por el auditor y las rectificatorias efectuadas en códigos relacionados con facturas de ese año.

Los casos de fraude y no fraude se categorizaron en tres tipos: “0” indica que el contribuyente fue auditado y no se encontraron facturas falsas en ninguno de los períodos revisados, “1” que indica que el contribuyente no utilizó facturas falsas en el año de análisis pero sí en otros períodos revisados (normalmente el año anterior o siguiente) y “2” que indica que el contribuyente utilizó facturas falsas en el año de estudio.

Concepto	Tipo de Información
Pago de Impuestos	Declaraciones de IVA (F29), Declaración de Renta (F22), Ratios Tributarios de IVA/Renta.
Características Propias	Edad, Antigüedad Empresa, Cobertura, Facturador electrónico, Contabilidad computacional, Actividades económicas, Cambio sujeto, Declara por internet, Tiene domicilio y sucursales propias
Comportamiento Histórico y en el año	Fiscalizaciones selectivas, Delitos Previos, Problemas con el domicilio, Inconcurrencias, Denuncias y Clausuras, Pérdidas de Rut, Destrucción de documentos, Deuda regularizada, Pérdida de Facturas, Facturas observadas y/o bloqueos, Marcas Preventivas.
Ciclo de Vida	Inicio de actividades, Verificación de actividades, Timbraje de documentos, Modificaciones de información, Términos de giro previos

Relacionados	Mandatarios, Representantes Legales, Socios, Familiares, Proveedores, Contadores, Sociedades y Representaciones (activos, antecedentes de delito, investigados, bloqueados)

Cuadro Nº 3: Tipo de Información utilizada para construir el vector de características

Para la construcción del vector de características se seleccionaron 20 códigos del Formulario de Pago Mensual de IVA (F29), 31 códigos del Formulario del Impuesto Anual de la Renta (F22) asociados a la generación de la base imponible de primera categoría y datos contables de la empresa, y 31 ratios tributarios que relacionan la información de IVA y Renta y la rentabilidad de la empresa con su liquidez, entre otros. Adicionalmente se generan 92 indicadores que pueden dar indicios de un buen o mal comportamiento en el tiempo, relacionados con su comportamiento histórico, el comportamiento de sus relacionados, sus características particulares e información generada en las distintas etapas del ciclo de vida, como se muestra en el Cuadro Nº 3.

4.2. TECNICAS DE DATA MINING IMPLEMENTADAS

Para efecto de la caracterización e identificación de patrones, se aplican tres técnicas de data mining: el Self- Organizing Maps (SOM), el Gas Neuronal (NG) y Árboles de Decisión. Posteriormente para la predicción, se utiliza Redes Neuronales con Backpropagation y Redes Bayesianas, las que se describen a continuación:

- ***Self-Organizing Maps (SOM)*:** es uno de los modelos de redes neuronales artificiales más utilizado para el análisis y visualización de datos de alta dimensión, basado en aprendizaje competitivo no supervisado. La red consiste en un conjunto de neuronas dispuestas en una grilla de dimensión a, normalmente rectangular, cilíndrica o toroidal, que genera un espacio de salida de dimensión d, con $a \leq d$, sobre el cual se construyen relaciones de vecindad. Durante el entrenamiento de la red, las neuronas generan cierta actividad ante el estímulo de los datos de entrada, lo que permite determinar qué neuronas han aprendido a representar los patrones de la entrada, los cuales pueden ser agrupados dentro de una misma categoría o cluster, basándose en una medida de distancia, normalmente Euclideana. Esta herramienta usualmente es aplicada para clusterización y segmentación, generando grupos con objetos de comportamiento similar entre sí, pero diferentes a los objetos de otro grupo.
- ***Gas Neuronal (NG:Neural Gas):*** es un algoritmo relativamente nuevo de redes neuronales no supervisada, orientada a la cuantización vectorial de estructuras arbitrarias. La mayor diferencia con el SOM es que este método no define una grilla que impone relaciones topológicas entre unidades de la red y cada neurona puede moverse libremente a través del espacio de datos. Esta libertad permite al algoritmo una mejor capacidad para aproximar la distribución de los datos en el espacio de entrada, ya que las neuronas no están obligadas a tener que mantener ciertas relaciones de vecindad, sin embargo, requiere tener algunos antecedentes respecto del número de grupos que se espera obtener.
- ***Árboles de Clasificación:*** es uno de los métodos más utilizado para realizar clasificaciones, y se destaca por su sencillez y su aplicabilidad a diversas áreas e intereses. Básicamente el algoritmo consiste en formar todos los pares posibles y combinaciones de categorías, agrupando aquellas que se comportan homogéneamente con respecto a la variable respuesta en un grupo, manteniendo separadas las categorías que se comportan de forma heterogénea. Para cada posible par, se calcula el estadístico correspondiente a su cruce con la variable dependiente (estadístico chi-cuadrado en caso de campos de destino categóricos o estadístico F para salidas continuas). Para las categorías fusionadas se procede a realizar nuevas

fusiones de los valores del pronosticador, pero esta vez con una categoría menos, El proceso se acaba cuando ya no pueden realizarse más fusiones porque los estadísticos entregan resultados significativos.

- ***Red Neuronal de Perceptrón Multicapa (MLP):*** es un modelo de red neuronal artificial de varias capas utilizado para la clasificación y agrupación, basado en la funcionalidad del cerebro humano a través de un conjunto de vértices interconectados. La red debe encontrar la relación existente entre los atributos de entrada y la salida deseada para cada caso. Esto lo realiza a través de un método de aprendizaje llamado "Backpropagation" o "Retropropagación del error", que minimiza el error de predicción mediante un ajuste a los pesos de la red. Este método posee dos etapas: en la primera se calculan las salidas basado en las entradas y los pesos asignados a la red inicial, para la cual se calcula el error de la predicción y en la segunda fase, se calcula el error hacia atrás a través de la red, desde las unidades de salida hacia las unidades de entrada. De esta forma se actualizan los pesos a través de un método de descenso por gradiente. Este proceso es iterativo, por lo que tras realizar varias veces el algoritmo, la red va convergiendo hacia un estado que permita clasificar todos los patrones que minimizan el error[53].
- ***Redes Bayesianas:*** son un grafo dirigido acíclico, utilizado para predecir la probabilidad de ocurrencia de diferentes resultados, sobre la base de un conjunto de hechos. La red consta de un conjunto de nodos que representan las variables del problema y de un conjunto de arcos dirigidos que conectan los nodos e indican una relación de dependencia existente entre los atributos de los datos observados. Las redes bayesianas describen la distribución de probabilidad que gobierna un conjunto de variables, especificando suposiciones de independencia condicional junto con probabilidades condicionales. Típicamente, este problema se divide en dos partes: un aprendizaje estructural, que consiste en obtener la estructura de la red, y un aprendizaje paramétrico, en el que conocida la estructura del grafo, se obtienen las probabilidades correspondientes a cada nodo. Su principal ventaja es que permite obtener la probabilidad de ocurrencia de un determinado suceso en función de un conjunto de acciones, entregando una vista clara de las relaciones mediante un gráfico de red.

4.3. PRE PROCESAMIENTO DE LOS DATOS

La preparación de los datos es una parte fundamental del proceso KDD, ya que la información puede provenir de muchas fuentes, tener errores, ambigüedades o ser redundante, consumiendo gran parte del tiempo del proyecto. Por otra parte, los datos deben ser transformados de manera apropiada para realizar el análisis.

4.3.1. LIMPIEZA

La calidad de los datos tiene una incidencia directa en los resultados, ya que si los datos no son de calidad, los resultados tampoco lo serán. Para lo anterior, se eliminan los puntos atípicos o outliers, utilizando como regla aquellos casos que superan la media más cinco veces la desviación estándar, considerando únicamente los casos con valor positivo de cada código. En la mayoría de las variables la distribución era decreciente, debido a que un gran porcentaje de contribuyentes paga montos bajos de impuestos, y sólo un pequeño grupo paga montos altos, por lo que la eliminación de datos se hizo de manera cuidadosa, considerando el juicio experto de los involucrados en el negocio, de manera de no eliminar casos que estuvieran correctos pero

[53] Normalmente se calcula el error cuadrático medio

alejados del promedio. Lo mismo sucede con las variables de comportamiento, ya que constituyen conductas irregulares que sólo tiene un grupo pequeño de contribuyentes. Por lo tanto, al eliminar los casos con valores más altos, se elimina a aquellos contribuyentes que en general tienen un peor comportamiento, los cuales son el grupo de interés de este trabajo.

Las variables de comportamiento, no tenían grandes inconsistencias debido a que fueron construidas en forma manual, sin embargo, se presentaban algunos problemas en los códigos del F29. Por ejemplo, se declaraban ventas con facturas pero no se indica una cantidad de facturas emitidas o viceversa. Dado que estos casos no eran muchos, se determina eliminarlos de la base. El mismo criterio se utilizó para el resto de los códigos de débitos y créditos.

Luego de quitar los outliers y los casos inconsistentes, el conjunto de datos final queda compuesto por 532.755 contribuyentes que son micro y pequeñas empresas, y 22.609 medianas y grandes empresas, eliminando un 4.6% del primer grupo y un 3.4% del segundo.

4.3.2. TRANSFORMACIÓN Y NORMALIZACIÓN

Debido a que la declaración del pago de IVA se realiza mensualmente y la declaración de impuesto a la renta se realiza en forma anual, la primera transformación fue considerar el total anual, sumando los montos mensuales de cada código del F29 en el año para hacerlo comparable con la información de renta. Respecto de la completitud de datos nulos, la información de IVA es más completa que la de renta, debido a que los códigos del reverso del F22, sólo deben ser presentados por contribuyentes que llevan contabilidad completa. Por lo tanto, se utiliza información de débitos y créditos de IVA para completar datos de ingresos y costos del período, debido a la relación directa existente entre ambos. Para el resto de los campos de renta, se utiliza la mediana del código para contribuyentes del mismo tramo de ventas. Finalmente, producto de la distribución decreciente de las variables de impuesto, se aplica una transformación logarítmica para disminuir el efecto de los datos extremos como se muestra en la Figura N° 2.

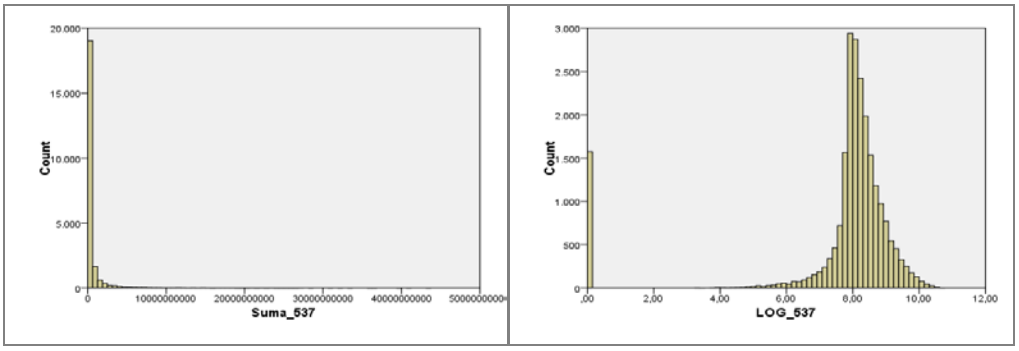


Figura N° 2: Ejemplo de distribución original y transformada de códigos de impuestos.
(a) Variable original x (b) Variable transformada Log(x+1)

Para evitar que las variables con un mayor rango de valores le quiten importancia a otras con un rango menor, se procede a normalizar las variables de manera que sean comparables la una con la otra, utilizando la normalización “Min-Max” en el rango [0,1]. Adicionalmente, previo a la selección de las variables de utilizar en los modelos, se procede a reducir las variables de comportamiento a través del Análisis de Componentes Principales (ACP). Como resultado se generan 15 componentes principales para el grupo de las micro y pequeñas empresas, que explican un 61,3% de la varianza de los datos. Del mismo modo, se generan 16 componentes principales para las medianas y grandes empresas, que explican un 59,9% de la varianza de los datos, las que se presentan en el Cuadro N° 4.

Micro y Pequeñas Empresas	%	Medianas y Grandes Empresas	%
(1) Nivel de facturas timbradas en los últimos años	9,7	(1) Cobertura de la empresa	9,2
(2) Delitos e irregularidades de facturas previos	7,0	(2) Fiscalizaciones previas	6,2
(3) Fiscalizaciones previas con resultado positivo	5,6	(3) Nº Actividades económicas	5,5
(4) Frecuencia de Timbraje	5,1	(4) Nivel de formalidad de la empresa y antigüedad	4,2
(5) Participación en otras empresas	4,5	(5) Clausuras y denuncios históricos	3,8
(6) Problemas de localización	4,2	(6) Verificaciones de actividad	3,4
(7) Antigüedad	3,5	(7) Giros e inconcurrencias	3,2
(8) Clausuras y denuncios históricos	3.4	(8) Representantes legales	3,2
(9) Cobertura de la empresa	3,0	(9) Delitos de los relacionados	2,9
(10) Fiscalizaciones previas con resultado negativo	2,9	(10) Irregularidades de facturas y nivel de timbraje	2,8
(11) Verificaciones de actividad	2,6	(11) Rendimiento de fiscalizaciones previas	2,8
(12) Delitos de relacionados indirectos	2,6	(12) Irregularidades recientes	2,7
(13) Irregularidades previas (pérdida facturas)	2,5	(13) Cambio de sujeto	2,6
(14) Nivel de formalidad de la empresa	2,4	(14) Antecedentes de término de giro y no ubicado	2,6
(15) Delitos de relacionados directos	2,4	(15) Antecedentes de timbraje restringido	2,5
		(16) Regularización de deudas y pérdidas de rut.	2,5

Cuadro N° 4: Conceptos asociados a cada Componente Principal y el porcentaje de la varianza explicada

Dado que nuestro interés era generar variables de comportamiento relacionadas al uso y venta de facturas falsas y no a otros comportamientos, se seleccionan sólo aquellas variables que tienen una correlación mediana-alta con la variable de uso de facturas falsas en el año 2006, eliminando aquellas que tienen más de un 10% de probabilidad que el coeficiente de pearson sea cero, exceptuando algunos códigos de interés como el total de débitos, total de créditos y pago de IVA, entre otros.

Igualmente, se descartan aquellas variables que tienen un gran porcentaje de valores nulos. De esta forma se seleccionan 42 variables en el segmento micro y pequeñas y 36 variables medianas y grandes para el análisis. En el primer grupo, un 35% de las variables corresponde a códigos de la declaración de IVA, un 35% a códigos relacionados con renta y un 30% a variables relacionadas al comportamiento. En el segundo grupo en cambio estos porcentajes varían a un 31%, 38% y 31% respectivamente, con mayor preponderancia de variables asociadas a la renta.

4.4. MODELAMIENTO

Para efectos de caracterización e identificación de patrones, en una primera instancia se aplican las técnicas de data mining al universo de empresas, con el objetivo de identificar relaciones entre su pago de impuestos (IVA y Renta) y variables de comportamiento asociadas a la utilización de facturas falsas. Posteriormente se aplican técnicas de clasificación para aquellos casos en los que la condición de fraude y no fraude es conocido, de manera de identificar patrones específicos de este conjunto de contribuyentes. Finalmente se aplican herramientas de clasificación para detectar casos de fraude y no fraude con la información generada.

4.4.1. CARACTERIZANDO AL UNIVERSO DE EMPRESAS

Inicialmente se aplica el método SOM al universo de contribuyentes, para identificar clusters o grupos de empresas de comportamiento similar. La hipótesis de trabajo suponía que al considerar sólo las variables de comportamiento relacionadas al uso de facturas falsas combinadas con variables de impuestos, era posible detectar grupos de contribuyentes que tienen un buen o mal comportamiento tributario y conocer cómo realizaban su pago de impuesto. Para ello se utiliza el paquete “som” de R, considerando una topología de red rectangular, con 3

neuronas de entrada y 24x24 neuronas de salida en el grupo de las micro y pequeñas empresas y 36x36 neuronas de salida en el grupo de las medianas y grandes empresas, con un número máximo de 100 iteraciones. En el primer grupo se considera una muestra de 100.000 empresas, debido a restricciones computacionales.

En el caso de las micro y pequeñas empresas se generan 5 clusters, mientras que en las medianas y grandes se identifican 6 clusters, como se muestra en la Figura N° 3.

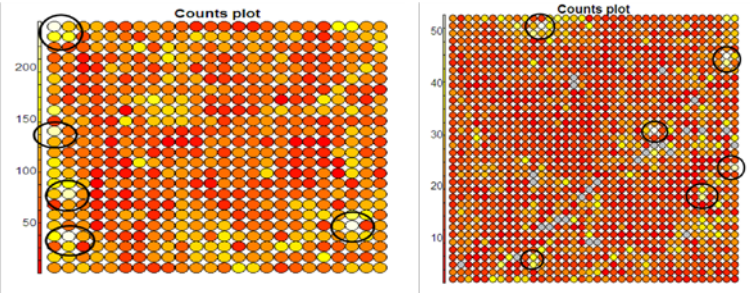


Figura N° 3: Mapa resultante aplicación SOM en MI y PE (izquierda) y ME y GR (derecha)

Los clusters obtenidos en el primer grupo se diferencian principalmente por la utilización de boletas y/o facturas, el nivel de pago de IVA, el nivel de costos declarados, el nivel de formalidad de la empresa y participación en otras empresas y algunos problemas de localización. Mientras que en las medianas y grandes, se diferencian por la utilización de boletas y/o facturas, niveles de uso de remanentes, notas de crédito y facturas de activo fijo, pasivos y activos, así como los resultados de fiscalizaciones previas y el nivel de formalidad de la empresa, como se indica en los Cuadros N° 5 y N° 6.

Cluster 1	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel alto de pago de IVA y costos altos. Con algunos problemas de localización, mayor nivel de participación en otras empresas y formalización de la contabilidad.
Cluster 2	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel intermedio-alto de pago de IVA y costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad y participación en otras empresas.
Cluster 3	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad.
Cluster 4	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara niveles altos de costos y problemas de localización.
Cluster 5	Tiene niveles altos de débitos con boletas, nivel intermedio de uso de facturas y pago de IVA, y costos altos. Relativamente joven con algunos problemas de localización y nivel intermedio de formalización.

Cuadro N° 5: Clusters resultantes aplicación SOM en MI y PE

Cluster 1	No utiliza boletas. Tiene nivel intermedio de remanentes y costos bajos. Presenta monto alto de créditos por factura de activo. Con un nivel alto de formalidad.
Cluster 2	No utiliza boletas. Tiene nivel intermedio de remanentes y pocas fiscalizaciones previas. Nivel intermedio de formalidad.
Cluster 3	No utiliza boletas. Tiene nivel alto de remanentes, pasivos y activos. Tiene bajo porcentaje de crédito asociado a facturas. Nivel alto de formalidad.
Cluster 4	Nivel alto de uso de boletas. Tiene nivel intermedio de remanentes y de notas de crédito. Nivel alto de formalidad.
Cluster 5	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel bajo de formalidad. Pocas fiscalizaciones previas.
Cluster 6	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel alto de formalidad. Tiene nivel intermedio de uso de notas de crédito.

Cuadro N° 6: Clusters resultantes aplicación SOM en ME y GR

Si bien se encontraron algunos patrones de comportamiento con éste método, estos no estaban relacionados específicamente a la utilización de facturas falsas, ya que los casos conocidos de fraude y no fraude se encontraban distribuidos en todo el mapa sin un patrón definido.

Posteriormente se aplica el Gas Neuronal, considerando el mismo número de clusters que el Mapa de Kohonen, utilizando el paquete "Clust" de R, el cual genera una matriz con las características de los centroides de cada variable y un vector de clasificación que señala el grupo al que pertenece cada contribuyente. En este caso, los grupos generados también se encuentran influenciados por el pago de impuestos, aunque con mayores diferencias en términos de comportamiento. Esto, permite diferenciar cuáles grupos tienen mejor y peor comportamiento, y relacionarlo con su pago de impuesto, aunque no necesariamente los casos de facturas falsas se encontraban en un mismo grupo.

De acuerdo a esto, se identificaron los siguientes patrones asociados a un mal y buen comportamiento, considerando los puntos comunes obtenidos en ambos métodos.

Buen Comportamiento MI y PE	Declaran montos más altos de débitos (emite más boletas) y pagan más IVA. Declaran bajos niveles de créditos y de remanentes, mayor relación ingresos/costos y costos/activos. Tienen mayor cantidad de facturas timbradas y frecuencia de timbraje, menor cantidad de delitos e irregularidades previas y delitos de los relacionados indirectos. Registran pocas verificaciones de actividad.
Buen Comportamiento ME y GR	Declaran mayor nivel de costos y gastos y mayor nivel de activos y pasivos. Tienen montos más altos de créditos y remanentes. Registran un mayor nivel de formalización de su contabilidad y mayor cobertura, mayor número de representantes legales y cantidad de fiscalizaciones previas.
Mal Comportamiento MI y PE	Declaran niveles bajos de pago de IVA y una relación débito/crédito baja. Registran una mayor cantidad de créditos y acumulación de remanentes. Tienen un nivel más bajo del ratio ingresos/activo, mayor cantidad de fiscalizaciones previas con resultado positivo y un menor nivel de facturas timbradas. Registran varias verificaciones de actividad.
Mal Comportamiento ME y GR	Declaran mayores costos y remuneraciones respecto de sus activos, menor nivel de pasivos y mayor cantidad de porcentaje de débitos con boleta, aunque con un número menor de boletas. Registran mayor cantidad de anotaciones de timbraje restringido, términos de giro previos y antecedentes de no ubicado. Tienen mayor cantidad de denuncias y clausuras históricas, menor cantidad de fiscalizaciones previas y cobertura, así como un menor nivel de formalización de su contabilidad y antigüedad.

Cuadro N° 7: Caracterización de grupos con buen y mal comportamiento

4.4.2. CARACTERIZANDO A LOS CASOS CON FRAUDE Y NO FRAUDE

Si bien las dos técnicas anteriores implementadas permiten caracterizar al universo de contribuyentes e identificar algunos patrones diferenciadores, considerando aquellas variables más relacionadas con el uso de facturas falsas. Éstas tienden a darle mayor importancia al pago de impuestos que a las variables de comportamiento, creando grupos que se diferencian en el tipo de operación (ventas con facturas y/o boletas), el nivel de actividad (alto-bajo nivel de ventas, costos) y pago de impuestos (alto-bajo), debido a la mayor variabilidad de estas variables en comparación a las de comportamiento.

Por otra parte, al analizar la distribución de cada variable, se observa que los casos con fraude normalmente se encuentran en los casos extremos de cada una de ellas. Por este motivo se determina aplicar árboles de decisión al conjunto de datos con resultado de auditoría conocido, ya que permite identificar el punto de corte de cada variable frente al cual se produce un cambio de

comportamiento, considerar casos extremos y generar reglas que pueden ser validadas e implementadas.

El tipo de árbol utilizado es el CHAID (Chi-square automatic interaction detection), el cual permite clasificaciones no binarias y generar un número distinto de ramas a partir de un nodo considerando tanto variables continuas como categóricas. Un punto a considerar de éste método es que se requiere disponer de tamaños de muestra significativos, ya que al dividirse en múltiplos grupos, cabe el riesgo de encontrar grupos vacios o poco representativos si no se dispone de suficientes casos en cada combinación de categorías. Adicionalmente se evalúa el método del CHAID exhaustivo, el cual es una modificación del algoritmo tradicional, que busca hacer frente algunas debilidades del CHAID tradicional.

Se realizan varios experimentos que consideran distinto número de variables y tipos de salidas (categóricas y numéricas) para identificar si se producen diferencias entre una formato de salida y otro.

Exp. N°	Segmento	Método	N° variables	Tipo de salida	N° Niveles	N° Nodos finales
1	Micro y Peq.	Árbol CHAID	30	Categórica	6	33
2	Micro y Peq.	Árbol CHAID	30	Numérica	5	36
3	Med. y Grandes	Árbol CHAID	38	Numérica	4	22
4	Med. y Grandes	Árbol CHAID	24	Numérica	6	24

Cuadro N° 8: Caracterización de grupos con buen y mal comportamiento, según el gas neuronal

Finalmente esta técnica resultó ser altamente efectiva para encontrar patrones diferenciadores entre fraude y no fraude, ya que los nodos finales estaban compuestos mayoritariamente por casos de un solo tipo, o en su defecto combinado con casos con valor de salida “1”, los cuales se aproximan más al comportamiento de los casos con fraude “2”.

Como se indica en el Cuadro N° 8 el número de nodos finales fue similar en ambos experimentos realizados en cada grupo, obteniéndose 33 y 36 nodos en el segmento de las micro y pequeñas empresas y 22 y 24 nodos en el segmento de las medianas y grandes.

A modo de ejemplo se presenta un extracto del resultado de la aplicación del experimento N° 1, en el cual se identifican patrones bastante claros asociados a fraude y no fraude, debido a la preponderancia de nodos finales con casos de fraude y no fraude. Como se indica en la Figura N° 4, los factores que tienen mayor incidencia fueron el resultado de las fiscalizaciones previas (ACP10) y el porcentaje de las compras sustentado en facturas (CFTOT). Esto indica que aquellos que han sido más veces fiscalizados en el pasado y no se les ha encontrado nada y sus compras no se basan principalmente en facturas, tienen menos probabilidad de utilizar facturas falsas, que aquellos que mayoritariamente registran compras con facturas y tienen fiscalizaciones productivas en el pasado. De hecho, estas dos variables por sí solas, determinan varios nodos finales con preponderancia de casos sin fraude.

Adicionalmente, la variable que indica una mayor preponderancia de delitos e irregularidades asociadas a facturas históricas combinado con la frecuencia de timbraje, genera nodos finales con preponderancia de casos con facturas falsas. Particularmente el nodo 12 que contiene casi la mitad de los casos (46%) se descompone en varias ramas en función del valor que toma el crédito promedio por factura emitida (mientras mayor sea este indicador, mayor posibilidad hay de que cometa fraude). De igual manera, la preponderancia de casos con fraude en cada rama depende del número de facturas emitidas, el IVA pagado, el total de débitos por boletas, la relación entre costos y activos y el nivel de participación en otras empresas.

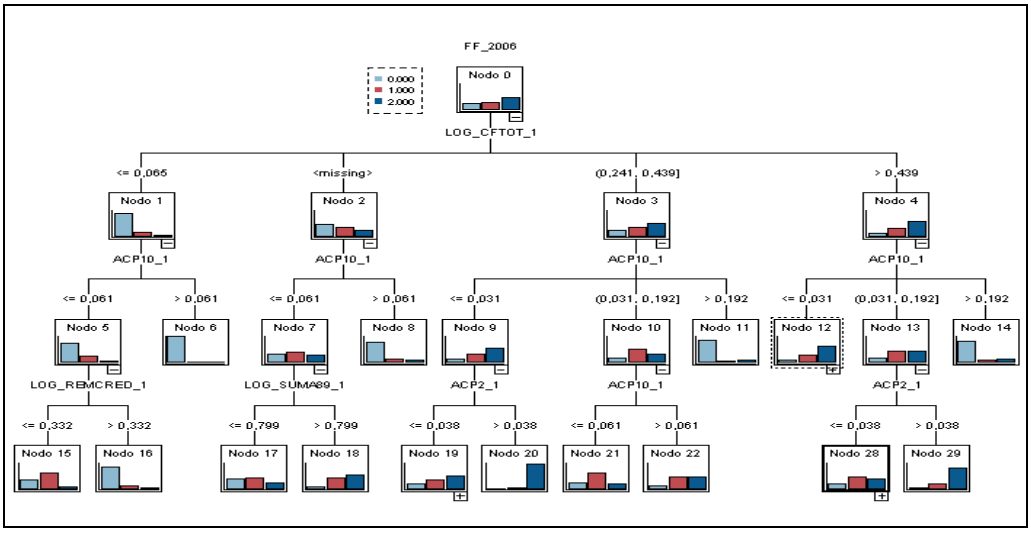


Figura N° 4: Clasificación resultante de la aplicación del árbol CHAID – Experimento N° 1

Cómo se señala en la Figura N° 5, las variables más relevantes para distinguir casos de fraude en las micro y pequeñas empresas fueron el resultado de las fiscalizaciones previas, el Total de IVA determinado, el porcentaje de crédito sustentado en facturas, la relación entre remanentes y créditos, el total de débitos por boletas y la relación entre facturas timbradas y emitidas. Mientras que en las medianas y grandes las variables corresponden a total de remanente, porcentaje de crédito respaldado en facturas, el número de representantes legales, nivel de formalización de la contabilidad, la relación entre remuneraciones y activos, entre otros.

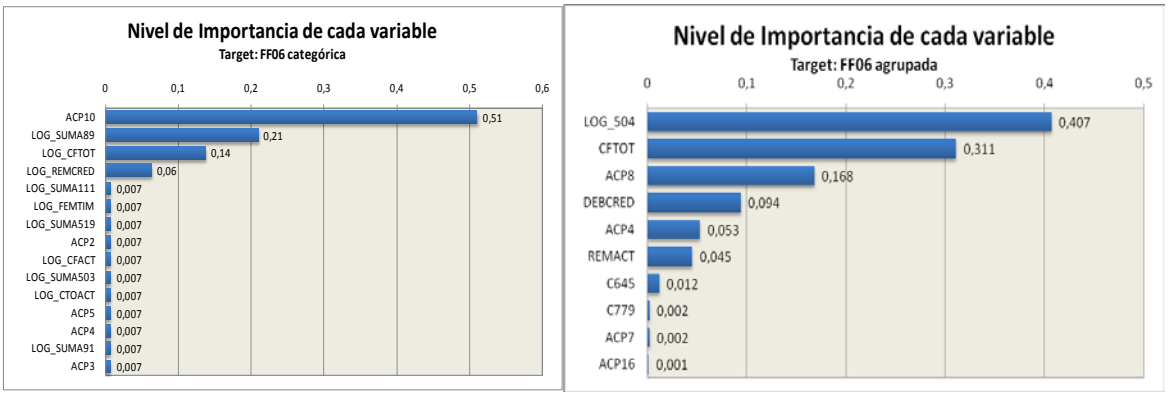


Figura Nº 5: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

Considerando los patrones y reglas que se repiten en cada rama del árbol para diferenciar entre casos de fraude y no fraude, en el Cuadro N° 9 se presenta un extracto de los comportamientos asociados a cada uno de ellos en cada segmento, que resume las variables principales consideradas y las relaciones que generan nodos con y sin utilización de facturas falsas en el año de estudio.

Comportamiento asociado a Fraude MI y PE	Registran menor porcentaje de créditos asociados a facturas y más fiscalizaciones previas con resultado negativo. Emiten menor cantidad de facturas emitidas y un valor más bajo del indicador facturas emitidas/facturas timbradas. Registran un mayor monto del indicador remanentes/crédito promedio.
Comportamiento asociado a Fraude ME y GR	Registran menor porcentaje de crédito asociado a facturas. Declaran más remanente acumulado del período anterior. Tienen valores bajos del indicador costos/activos. Registran menor cantidad de irregularidades previas asociadas a facturas y de timbraje.

Comportamiento asociado a No Fraude MI y PE	Tienen mayor porcentaje de créditos asociados a facturas y débitos con boletas. Tienen valor alto del indicador costos/activos. Emiten mayor cantidad de facturas y tienen valor alto del indicador facturas emitidas/facturas timbradas. Tienen montos altos de IVA determinado. Registran menos fiscalizaciones previas con resultado negativo y más fiscalizaciones con resultado positivo. Tienen más antecedentes de delitos e irregularidades previas asociadas a facturas y mayor frecuencia de timbraje en los últimos dos años.
Comportamiento asociado a No fraude ME y GR	Tienen mayor porcentaje de créditos asociados a facturas. Declaran monto menor de remanente acumulado en el mes anterior y tienen valores altos del indicador costos/activos. Tienen mayor nivel de informalidad en su contabilidad y son de menor antigüedad. Registran mayor número de actividades económicas activas e irregularidades previas asociadas a facturas y timbraje. Tienen mayor cantidad de giros e inconcurrencias a notificaciones.

Cuadro N° 9: Caracterización de casos con y sin fraude según árbol CHAID

4.4.3. PREDICCIÓN DEL FRAUDE

Para la predicción, se aplicaron redes neuronales artificiales y redes bayesianas. En ambos procesos para evitar el sobreajuste de la red, los datos se dividen en dos conjuntos: uno de entrenamiento y uno de testeo, utilizando la regla 70/30. Por otra parte, ambos métodos fueron implementados utilizando la herramienta tecnológica clementine del SPSS.

Uno de las complejidades de las redes neuronales, es determinar el número de capas y nodos ocultos, así como la cantidad de épocas o iteraciones. Para determinar tales parámetros se consideraron distintos números de ciclos y nodos en las capas ocultas, de manera de establecer a través de ensayo y error los valores más adecuados. Para las iteraciones se utilizaron los valores: 1.000, 5.000, 10.000 y 20.000. En el caso de los nodos se utiliza el número que el software calcula por defecto en función de los datos del modelo y otra correspondiente a la mitad del número de nodos de entrada, es decir, 3 y 20 nodos respectivamente.

En el caso de las redes bayesianas se evalúan dos métodos para construir la red: el algoritmo TAN y el algoritmo de estimación de Markov-Blanket disponibles en el software clementine del SPSS. Adicionalmente se utiliza un preprocesamiento previo de las variables para identificar cuáles son las variables más relevantes y mejorar el tiempo de procesamiento y rendimiento del algoritmo. De igual forma se utiliza un test de independencia de máxima verosimilitud y chi-cuadrado para el aprendizaje paramétrico

Los resultados de los experimentos se presentan en el Cuadro Nº 10, el que contiene los siguientes indicadores obtenidos en el grupo de testeo: (1) Sensibilidad: Indica la proporción de casos con fraude clasificados en forma correcta, (2) Especificidad: Indica la proporción de casos sin fraude en los que la clasificación fue correcta, (3) Concordancia: Indica la proporción de casos con y sin fraude en los que la clasificación fue correcta y (4) Tasa de error: Indica la proporción de casos con y sin fraude que fueron asignados a una clase incorrecta.

Exp. N°	Segmento	Método	Sensitividad (1)	Especificidad (2)	Concordancia (3)	Tasa de error (4)
1	Micro y Peq.	Red Neuronal	92.6%	72.9%	87.2%	12.8%
2	Micro y Peq.	Red Bayesiana	82.3%	64.1%	77.9%	22.1%
3	Med. y Grandes	Red Neuronal	88.8%	59.1%	72.5%	27.5%
4	Med. y Grandes	Red Bayesiana	73.3%	66.7%	70.3%	29.7%

Cuadro N° 10: Experimentos realizados para detectar los casos con fraude por facturas falsas

En ambos segmentos, los mejores resultados de predicción de casos con facturas falsas se obtuvieron con la técnica de red neuronal. En el grupo de las micro y pequeñas empresas, el

experimento 1 arrojó que en un 92,6% los casos con fraude fueron asignados a la clase correcta, mientras que en el grupo de las medianas y grandes empresas la proporción de casos con fraude correctamente asignada fue de 88,8%. Por otra parte, el poder de generalización del modelo fue bastante bueno, ya que los resultados del testeo fueron similares a los obtenidos en el entrenamiento de la red, cuya predicción fue casos con y sin fraude fue de 93.7% y 89.6% respectivamente.

La red neuronal generada para las micro y pequeñas empresas, indica una preponderancia de variables asociadas al pago de IVA y al comportamiento, y en menor medida, a variables relacionas a la renta. Las más relevantes corresponden a los antecedentes obtenidos de la verificación de actividades, la relación entre remanentes y créditos, el total de débitos por facturas emitidas, la relación entre ingresos del giro y los activos y la relación entre el IVA pagado y el Ingreso declarado. En el caso de las medianas y grandes empresas, las variables más relevantes corresponden a la relación entre remanentes y créditos, las cuentas por pagar a empresas relacionadas, el total de pasivos, la proporción de créditos asociado a facturas y el IVA determinado en el período.

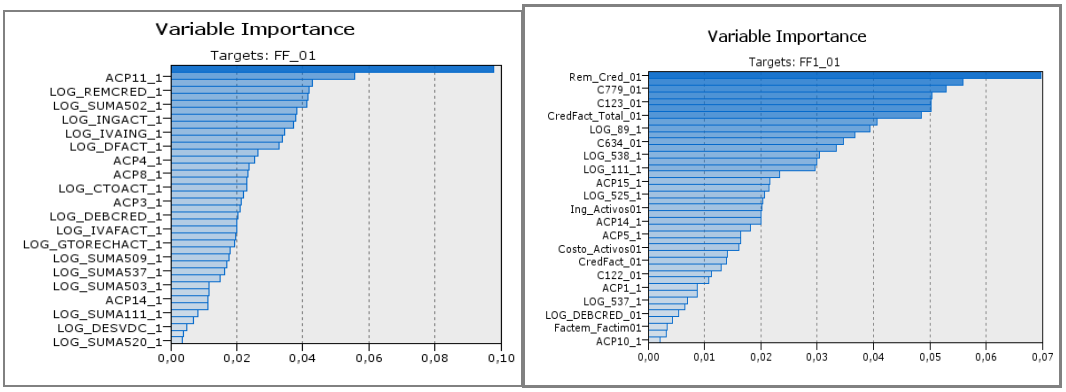


Figura Nº 6: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

5. CONCLUSIÓN Y TRABAJO FUTURO

La utilización y venta de facturas falsas tiene un impacto significativo en la recaudación que percibe el Estado para financiar sus proyectos. La detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera además un importante costo administrativo para el SII, lo que da cuenta de la relevancia que tiene focalizar los esfuerzos en la detección de casos de evasión y fraude fiscal.

Los métodos de clusterización y clasificación utilizados para caracterizar a los contribuyentes que tienen buen o mal comportamiento tributario asociado a la utilización de facturas falsas, demuestran que es posible identificar algunas características diferenciadoras entre un grupo y otro, las cuales hacen sentido con lo que sucede en la realidad. Particularmente el método de gas neuronal arrojó que era posible determinar algunas variables relevantes para diferenciar entre un buen o mal comportamiento, los que no necesariamente se asocian a la utilización y venta de facturas falsas. El método de kohonen, en cambio, no permitió obtener patrones de comportamiento relacionados con la utilización de facturas falsas, sino más bien, se detectaron clusters en relación al pago de impuestos, en la que las variables con mayor cantidad de ceros y varianza resultaron ser las que más impacto tuvieron en la conformación de los grupos. Los árboles de decisión aplicados a los casos en el que el resultado de fraude y no fraude era conocido resultó ser una buena técnica para detectar variables que permiten distinguir entre casos de fraude y no fraude. Esto debido que al analizar la distribución de las variables en cada grupo,

se observa que los casos con fraude tendían a tomar valores más extremos de las variables, por lo que era posible distinguir rangos a partir de los cuales, existe una probabilidad de tener o no tener fraude. Por otro lado, los resultados obtenidos fueron coherentes con lo observado en la realidad, de acuerdo a la vista experta.

Es así como en el caso de las micro y pequeñas empresas las variables que permitían distinguir entre fraude y no fraude se relacionaban principalmente con el porcentaje de créditos generado por facturas respecto del crédito total y las fiscalizaciones previas con resultado negativo. En la medida que el contribuyente fue fiscalizado más veces en el pasado y no se encontró nada, es más probable que no tenga fraude en el futuro. Por otro lado, mientras su crédito esté más asociado a otros ítems distintos a las facturas (activo fijo u otros), es menos probable que utilice facturas para respaldar sus créditos. Otras variables relevantes fueron la cantidad de facturas emitidas en el año y su relación con las facturas timbradas en los últimos dos años, el monto de IVA total declarado, la relación entre remanentes y créditos promedio, las fiscalizaciones previas con resultado positivo y los delitos e irregularidades históricos asociadas a facturas. Mientras que en las medianas y grandes empresas, las variables más relevantes fueron la cantidad de remanente acumulado en los períodos anteriores, el porcentaje de crédito asociado a facturas, la relación entre costos y activos, el nivel de informalidad en su contabilidad y la antigüedad, así como la cantidad de irregularidades previas asociadas a facturas y la cantidad de giros e inconcurrencias históricas.

En relación a los modelos predictivos, los que tuvieron mejor desempeño fueron los modelos de red neuronal de perceptron multicapa, que para efectos del estudio contaban con una capa de entrada que contenía las variables explicativas, una capa intermedia de procesamiento y una capa de salida. En el caso de las micro y pequeñas empresas el porcentaje de casos con fraude asignado correctamente fue un 92%, mientras que en las medianas y grandes empresas, este porcentaje fue de 89%.

Considerando que en la práctica sólo es posible fiscalizar a un grupo más bien reducido de empresas en un año, se recomienda realizar una combinación de los resultados obtenidos con las redes neuronales y las redes bayesianas, de manera de seleccionar para fiscalización a aquellos que aparecen catalogados como fraude en la red neuronal y que tienen las probabilidades más altas de cometer fraude según la red bayesiana.

En términos de recaudación, la predicción de un caso de fraude en una micro y pequeña empresa aporta un beneficio neto de $ 86.282, mientras que para una mediana y gran empresa, esta cifra aumenta a un $3.424.083, lo que permitiría reducir la evasión por concepto de IVA de manera significativa, si consideramos el total de casos auditados en un año.

De acuerdo a estudios que ha realizado el SII, se estima que aproximadamente un 20% de los contribuyentes utilizan facturas para evadir impuesto. No existe información desagregada por tipo de contribuyente, pero suponiendo que este porcentaje se repite en cada segmento y considerando los porcentajes de clasificación de casos con fraude y no fraude de los modelos de red neuronal, se tiene que el universo de potenciales usuarios de facturas es de 116.000 micro y pequeñas empresas y 4.768 medianas y grandes empresas, que generan un ingreso por fiscalización de $21.344 millones de pesos y $80.102 millones de pesos respectivamente, generando un potencial de recaudación de $101.446 millones de pesos.

Finalmente, para probar la capacidad predictiva real del modelo desarrollado y siendo concordante con el punto anterior, resulta vital su aplicación en actividades que permitan determinar en terreno el nivel de acierto en la clasificación de los contribuyentes seleccionados en la muestra, para lo cual se recomienda la implementación de un programa piloto que estará dirigido a los dos segmentos económicos estudiados, que será concluyente en términos de la efectividad real del modelo.

6. REFERENCIAS

1. Arnaiz, T., García, J. A. y López, J.M. (2006). Los Planes Integrales para la Prevención y Corrección del Fraude Fiscal. Banco Interamericano de Desarrollo (BID).
2. Bolton, R. y Hand, D. (2002). Statistical Fraud Detection: A Review. Statistical Science, Vol. 17- N° 3.
3. Centro Interamericano de Administraciones Tributarias (2004). Métodos de Selección de Declaraciones sujetas al Control Concurrente ocupando Herramientas de Minería de Datos. Programa Regional (TC-00-05-00-8-RG). Superintendencia Nacional de Administración Tributaria, Perú.
4. Clifton, P. y Chun, W. (2003). Investigative Data Mining in Fraud Detection. School of Business Systems, Monash University.
5. Davia, H.R., Coggins, J.W. y Kastantin, J. (2000). Accountant's Guide to Fraud Detection and Control (2da edición).
6. Denny, Williams, G., Christe, P. (2007). Exploratory Multilevel Hot Spot Analysis: Australian Taxation Office Case Study. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70.
7. Digimpietri, L., Trevisan, N., Meira, L., Jambeiro, J., Ferreira, C. y Kondo, A. (2008). Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System. Proceedings of the 9th Annual International Digital Government Research Conference.
8. Ernst&Young (2006). 9th Global Fraud Survey 2006: Fraud Risk in emerging markets. Junio.
9. Fayyad, U., Piatestky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. American association for artificial intelligence 0738-4602, 37-54.
10. Government Accountability Office (GAO), United States (2004). Data Mining: Agencies have taken key steps to protect privacy in selected efforts, but significant Compliance Issues Remain. Mayo.
11. Government Accountability Office (GAO), United States (2008). Lessons Learned from Other Countries on Compliance Risks, Administrative Costs, Compliance Burden and Transition. Report to Congressional Requesters, Abril.
12. Harrison, G. y Krelove, R. (2005). VAT Refunds: A Review of Country Experience. Noviembre. International Monetary Fund (IMF) Working Paper.
13. Luckeheide, S. (2007). Segmentación de los Contribuyentes que declaran IVA aplicando herramientas de clustering. Revista de Ingeniería en Sistemas. Volumen XXI.
14. Munoz, D.J. (2006). Proceso de Reconocimiento de Objetos asistido por computador, aplicando Gases Neuronales y técnicas de Minería de Datos". Scientia et Technica Año XII, No 30, Mayo.
15. Myatt Glenn, J. (2007). Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining. Wiley Interscience.
16. OECD (1999). Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Serie. General Administrative Principles - GAP004 Compliance Measurement, Junio.

17. OECD (2004). Compliance Risk Management, Use of Random Audit Programs. Forum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. Septiembre.
18. OECD (2004). Compliance Risk Management, Audit Case Selection Systems. Forum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. Octubre.
19. Servicio de Impuestos Internos (2011). Información de Cuenta Pública 2010. http://www.sii.cl/cuenta_publica/
20. Superintendencia Nacional de Administración Tributaria (2006). La Gestión de la Sunat en los últimos cinco años: Principales Avances y Desafíos.
21. Tanzi, V. y Shome, P. (1993). Tax Evasion: Causes, Estimation Methods, and Penalties a Focus on Latin America, Documento elaborado para el Proyecto Regional de Política Fiscal CEPAL/PNUD.
22. Velasco, D. (2007). Redes Bayesianas. Inteligencia Artificial II.
23. Velásquez, J. y Palade, V. (2008). "Adaptive Web Sites: A Knowledge Extraction from Web Data Approach". Frontiers in Artificial Intelligence and Applications, Volumen 170.

ANEXO L: Paper Presentado en Revista Internacional (Versión inglés)

Characterization and Detection of Taxpayers with False Invoices using Data Mining Techniques

Pamela Castellón González
Servicio de Impuestos Internos de Chile – pamela.castellon@sii.cl

Juan D. Velásquez [54]
Department of Industrial Engineering, School of Engineering and Science, University of Chile
jvelasqu@dii.uchile.cl

Abstract

In this paper we give evidence that it is possible to characterize and detect those potential users of false invoices in a given year, depending on the information in their tax payment, their historical performance and characteristics, using different types of data mining techniques. First, clustering algorithms like SOM and Neural Gas are used to identify groups of similar behaviour in the universe of taxpayers. Then decision trees, neural networks and Bayesian networks are used to identify those variables that are related to conduct of fraud and/or no fraud, detect patterns of associated behaviour and establishing to what extent cases of fraud and /or no fraud can be detected with the available information.This will help identify patterns of fraud and generate knowledge that can be used in the audit work performed by the Tax Administration of Chile (in Spanish Servicio de Impuestos Internos) to detect this type of tax crime.

Keywords: False Invoices, Fraud Detection, Data Mining, Clustering, Prediction.

1. INTRODUCTION AND MOTIVATION

Tax evasion and tax fraud[55] have been a constant concern for tax administrations, especially when pertaining to developing countries [1]. While it is true that taxes are not the only source of government funding, the fact is that they send a very important signal about the commitment and effectiveness with which the State can carry out its functions and restrict access to other sources of income.

In particular, the Value Added Tax (VAT), implemented in over 130 countries at different stages of economic development has become a key component of tax revenues [2]. For OECD countries[56], VAT accounts provide about 25% of total tax revenue collected. In the case of Chile, taxes provide about 75% of the resources from which the State each year pays its expenses and investments, collecting during 2011 a total of USD$41.6 billion dollars[57]. VAT represents 45%, amounting to USD$18.7 billion dollars and generating over 400 million invoices a year, of which 56% is issued in paper format and 44% in electronic format.

[54] Contact for correspondence
[55] Usually refers to "tax avoidance" when referring to behaviors that, within the law, prevent or reduce taxes, while "evasion or tax fraud" involves a violation of the law to obtain the same results.
[56] The OECD member countries represent countries that have reached a relatively high level of development and share a commitment to a market economy and pluralist democracy. Its members represent 60% of gross national product in the world, three quarters of world trade and 14% of the population.
[57] Considering only Central Government tax revenue (excluding Codelco, municipalities and social security).

The phenomenon of false invoices in respect of VAT is explained by the mechanics of determining the tax payable. When a company receives a false invoice, it simulates a purchase that never existed, thus increasing its tax credit fraudulently and decreasing VAT payment. Also, there is a decrease of payment in the income tax due to increased costs and expenditures declared.

The falsity of the document may be *"material"* if the physical elements that make up the invoice have been adulterated, or *"ideological"* when the materiality of the document is not altered, but the operations recorded in it are adulterated or nonexistent. The latter is more complex and difficult to detect because it involves fictitious transactions in which an audit is required to examine the sales books and corrections, or cross referencing the information with suppliers. Moreover, these cases are more expensive for SII, as they require a greater amount of time dedicated to collecting and testing evidence, which is harder to find.

The best known cases of material falsification are the physical adulteration of the document, the use of *"hanging"* invoices in which an invoice is counterfeited to impersonate a taxpayer of good behavior, and the use of a double set of tax invoices, which has two same-numbered invoices, but one of which is fictional and for a higher amount. In ideological falsification, invoices are used to register a nonexistent operation or adulterate the contents of an existing operation.

According to a method used by the SII to estimate VAT evasion [3] resulting from false invoices and other credit enlargements applied in the period 1996-2004, evasion by false invoices has historically represented between 15% and 25% of total VAT evasion, increasing signicantly in years of economic crisis. This is why in the crisis of 1998-1999 the participation rate increased to 38%, reaching an amount close to USD$1 billion dollars. This becomes relevant since recently there was a global economic crisis that hit Chile in late 2008 and the middle of 2009, causing an increase in the rate of VAT evasion to 23%, in the amount of evasion of USD$4 billion dollars.

It also requires that resources be invested in well-focused monitoring, detecting those taxpayers who have greater compliance risk and not bother or waste time and resources on those who do comply [4]. For this, data mining techniques offer great potential, because they allow the extraction and generation of knowledge from large volumes of data to detect and characterize fraudulent behavior and failure to pay tax, in the end improving the use of resources [5].

This paper is organized as follows: Section 2 describes how artificial intelligence techniques have facilitated the detection of tax evasion in tax administrations. Section 3 describes the data mining techniques applied. Section 4 describes the type of information used and the main results obtained in the characterization and detection of fraud in the issuance of invoices, and Section 5 presents the main conclusions and future lines of research.

2. RELATED WORK

Fraud in its various manifestations is a phenomenon that no modern society is free of. All governments, regardless of whether they are large or small, public or private, local or multinational, are affected by this reality, which seriously undermines the principles of solidarity and equality of citizens before the law and threatens business.

There are many fields and industries affected by this phenomenon. A study conducted by [6] in 2006, surveyed 150 medium and large Chilean companies to consult on this issue. The results show that 41% of them were victims of fraud in the past two years. This poses great challenges in prevention and opportunities for detection [7], given that fraud is usually higher than reported by companies, because somehow disturbs the image of the company towards customers and suppliers. In many cases there are even companies that are not known to have been victims of fraud.

Many fraud detection problems involve a large amount of information [8]. Processing these data in search of fraudulent transactions requires a statistical analysis which needs fast and efficient algorithms, among which data mining provides relevant techniques, facilitating data interpretation and helping to improve understanding of the processes behind the data [9]. These techniques have facilitated the detection of tax evasion and irregular behavior in other areas such as banking, telecommunications, insurance, IT, money laundering, and in the medical and scientific fields, among others [10].

To detect tax fraud, tax institutions began using random selection audits or focusing on those taxpayers who had no previous audits in recent periods and selecting cases according to the experience and knowledge of the auditors. Later methodologies were developed based on statistical analysis and construction of financial or tax ratios which evolved into the creation of rule-based systems and risk models [11]. These transform tax information into indicators which permit ranking of taxpayers by compliance risk. In recent years, the techniques of data mining and artificial intelligence have been incorporated into the audit planning activities [12,13], mainly to detect patterns of fraud or evasion, which are used by tax authorities for specific purposes.

The Internal Revenue Service, the institution responsible for administering taxes in the United States, has used data mining techniques for various purposes [12], among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities [14], electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering [13,15,16]. Among techniques used have been logistic regression models, decision trees, neural networks, clustering algorithms and visualization techniques such as link analysis.

In the Australian Tax Office, the "Compliance Program" is based on a risk model which uses statistical techniques and data mining in order to make comparisons, to find associations and patterns by logistic regression, decision trees and SVM [12,17]. A case of interest has been the approach used by [18], of discovering small clusters or unusual subpopulations, called "Hot Spots", using techniques such as the Self Organizing Map (SOM) to explore its features, clustering algorithms like k-means and visuals that are easy to understand for non-technical users.

Technique Applied	USA	Canada	Australia	UK	Bulgaria	Brazil	Peru	Chile
Neural Networks	✓	✓		✓	✓		✓	✓
Decision Tree	✓	✓	✓				✓	✓
Logistic Regression	✓		✓	✓	✓			
SOM			✓					✓
K-means			✓					✓
Support Vector Machines	✓		✓					
Visualization Techniques	✓					✓		
Bayesian Networks			✓					
K-Nearest Neighbour			✓					
Association Rules							✓	
Fuzzy Rules							✓	
Markov Chains						✓		
Time Series		✓						
Regression				✓				
Simulations	✓							

Figure No. 1: Data Mining techniques used by tax administrations to detect tax fraud

In New Zealand, the existing model associates the degree of compliance with attention to auditing, which coincides with that used by the Australian counterpart [15]. The plan includes an analysis of the economic, international, population, ethnic diversity and family structure. For its part, Canada uses neural networks and decision trees to distinguish the characteristics of

taxpayers who evade or commit fraud, based on the results of past audits, to detect patterns of noncompliance or evasion [13].

In Latin America, Peru was one of the first to apply these techniques to detecting tax evasion [19,20], adding to the selection system of the Maritime Customs of Callao an artificial intelligence tool based on neural networks. During 2004, this model was improved through the application of fuzzy rules and association for pre-processing variables and classification and regression trees (CART) to select the most relevant variables.

Brazil has developed project HARPIA (Risk Analysis and Applied Artificial Intelligence) jointly with the Brazilian Federal Revenue and universities in the country [21]. This project consists of developing a detection system of atypical points to help the regulators to identify suspicious transactions based on a graphic display of information on historical imports and exports and a system of export product information based on Markov chains, to help importers in the registration and classification of their products, avoid duplication and to calculate the probability that a string is valid in a given domain.

In the case of Chile the first trial was developed in 2007 [22], using the SOM and k-means to segment VAT taxpayers according to their F29 statements and characteristics. Later, in 2009, following the international trend, risk models were built of different stages of the life cycle of the taxpayer, in which neural networks, decision trees and logistic regression techniques are applied. The first trial was further developed to identify potential users of false invoices through artificial neural networks and decision trees, mainly using information from tax and income declarations in micro and small enterprises.

3. DATA MINING TECHNIQUES APPLIED

For purposes of characterization and identification of patterns three data mining techniques are applied: Self Organizing Maps (SOM), neural gas (NG) and decision trees. Backpropagation neural networks and Bayesian networks are subsequently used for detection, and are described below:

3.1. SELF-ORGANIZING MAPS

The Self-Organizing Map (SOM) [23] is one of the models most widely used in artificial neural networks for analysis and visualization of high dimensional data, based on unsupervised competitive learning. Specifically, the network consists of a set of neurons arranged in a grid dimension, usually rectangular, cylindrical or toroidal, which generates an output space of dimension d, with a <= d, on which neighborhood relations are defined, and whose aim is to discover the underlying structure of the data entered into it. By construction all the same neurons receive input at any given time.

During training, neurons in the network generate some activity from the stimulation of the input data, allowing a more specific identification of which areas or which neurons have learned to represent certain input patterns. Activity patterns generated in the same area have similar characteristics and can be grouped into a single category or cluster, based on a distance measure, usually Euclidean. The winning neuron output layer or "Best Matching Unit" (BMU) is one whose weight vector is most similar to input information.

This tool is usually applied to clustering and segmentation, creating groups of objects with behavior similar to each other but different from the objects of another group.

3.2. NEURAL GAS (NG)

Neural gas (NG) [24] is a relatively new algorithm for unsupervised neural networks, focused on vector quantization of arbitrary structures. The major difference with the SOM is that this method does not define a grid that imposes topological relationships between units of the network and each neuron can move freely through the data space. This freedom allows the algorithm a better ability to approximate the distribution of data in the input space, as the neurons are not required to have to maintain specific neighborly relations. However, having some background on the number of groups is expected to be required.

During the network training, the neurons change their position and adapt themselves to the data cloud. In this algorithm, each input pattern generates an excitation in each unit of the network. In each iteration a random data vector is presented to all neurons. For each data vector the nearest neuron is found, according to the Euclidean distance. This neuron is called “winning”. In the next step the neighborhood (diameter) of the winning neuron is established, which decreases exponentially with the number of iterations.

3.3. CLASSIFICATION TREES

Classification trees [25] are one of the non-parametric supervised learning methods most commonly used, being notable for their simplicity and applicability to different areas and interests. In general, the tree construction algorithms differ in the strategies used to partition nodes and prune the tree. In our case, we use trees based on CHAID methodology, which generate a different number of branches from a node considering both continuous and categorical variables. Basically the algorithm consists in forming all possible pairs and combinations of categories, grouping the categories that behave homogeneously with respect to the response variable in a group and maintaining separate those categories that behave differently.

For each possible pair, we calculate the statistics for their cross with the dependent variable (chi-square statistic for categorical target fields or F statistic for continuous outputs). The pair with the lowest value of this indicator will be a new category of two merged values, provided it is not statistically significant. For merged categories further consolidation of the values of the predictor is done, but this time with one category less, the process ending when no more mergers can be effected because statistically significant results occur.

3.4. MULTILAYER PERCEPTRON NEURAL NETWORK (MLP)

The multilayer perceptron model (MLP) [26] is an artificial neural network model of layers used for classification and grouping, based on human brain function through an interconnected set of vertices. The network must find the relationship between input attributes and the desired output for each case. This is done through a learning method called *"backpropagation"* or *"retropropagation"* which minimizes the prediction error by adjusting the weights of the network. This method has two stages. The first departures are calculated based on the inputs and the weights assigned to the initial network, for which the prediction error is calculated. In the second phase, the error is calculated backward through the network from the output units to the input units, getting an error in each unit. In this way the weights are updated through a gradient descent method. This process is iterative, so that after repeating the algorithm several times, the network will converge to a state that allows the classification of all training patterns, which minimizes the error[58].

[58] Usually calculates the mean square error

3.5. BAYESIAN NETWORKS

Bayesian networks [27,28] are directed acyclic graphs, used to predict the likelihood of different outcomes, based on a set of facts. The network consists of a set of nodes representing the variables of the problem and a set of directed arcs connecting the nodes and indicating a relationship of dependency between the attributes of the observed data. Bayesian networks describe the probability distribution that governs a set of variables, specifying assumptions of conditional independence with conditional probabilities. Typically, this problem is divided into two parts: structural learning, which is to obtain the network structure, and parametric learning, in which through known graph structure, we obtain the probabilities for each node. Their main advantage is that the probability of occurrence of a given event based on a set of actions can be obtained, giving a clear view of the relationship through a web graph.

4. DATA, ANALYSIS AND RESULTS

Unlike the previous study developed by SII related to this problem, this paper aims to complement the use of tax information with additional variables related to its historical performance and its performance in the year of analysis, and include aspects concerning direct associates, such as agents, partners and legal representatives and their characteristics, such as level of coverage, age, and whether electronic invoices or full accounting are used, among others. Moreover, a model for medium and large companies is developed, where there is less knowledge of how to operate regarding the use of false invoices, since they have more complex evasion procedures. This will build models differentiated by the size of the taxpayer, grouping on one hand the micro and small enterprises and on the other medium and large enterprises.

4.1. DATA AND ATTRIBUTE SELECTION

The year 2006 is chosen for characterization and detection, because the audits are performed up to a period of three years previous, which makes it difficult to use the latest information, as in 2010 cases were still being generating that could have used false invoices from 2007 onwards. Thus, for the characterization of contributors the universe of taxpayers is considered to be all those taxpayers who had filed at least one VAT return between 2005 and 2007, which corresponds to 582,161 enterprises. In the case of detection, information is used from those audits where there is certainty that the invoices were checked in 2006, independent of when that was done, considering a total of 1,692 companies. Table 1 shows a taxonomy with the taxpayers consider in our analysis.

Taxpayers	Micro and Small	Medium and Large	Total Enterprises
Enterprises active in period 2005-2007	558.319 (96%)	23.842 (4%)	582.161 (100%)
Companies audited by invoices in 2006 resulting in fraud or no fraud	1.280 (76%)	412 (24%)	1.692 (100%)

Table No.1: Number of taxpayers used in the analysis

One of the biggest drawbacks to defining cases with and without fraud occurring is related to the way in which information is recorded, for example the date of the start and completion of the audit, the revised tax periods and the result are known, but the information on the periods in which differences occur is not automated. Therefore, to see if the false invoice detected corresponded specifically to 2006, the notes and comments made by the auditor would have to be

reviewed and corrections done in the codes related to the invoices of that year. Cases with and without fraud were categorized into three types: "0" indicates that the taxpayer was audited and no false invoices found in any of the periods reviewed, "1" indicating that the taxpayer did not use false invoices in the year of analysis but did in other periods reviewed (usually the previous year or the next) and "2" indicating that the taxpayer used false invoices in the year of study.

To construct the feature vector, 20 codes were selected from the Monthly VAT Tax Payment Form (F29) related to the operative payment of VAT, 31 codes from the Annual Income Tax Form (F22) associated with the generation of taxable income class and business financial data and 31 tax ratios relating the VAT and Income Tax information with profitability and the company's liquidity, among others. Regarding the behavior and features of the company, this generates 92 indicators that can signal good or bad behavior over time related to its historical performance, its particular characteristics and information generated at different stages of the life cycle as show in Table N° 2.

Concept	Type of Information
Payment of Taxes	VAT Tax Declarations (F29), Declarations of Income Tax (F22), Tax Rates and Income Taxes
Personal Characteristics	Age of Taxpayer, Age of Company, Level of Coverage, Electronic Biller, Computer Accounting, Economic Activities, Change of Subject, Declares Online, Whether Domiciled and Owns Branches
Historical behavior and within year of study	Selective Audits, Previous Offenses, Address Problems, Failures to Attend, Accusations and Closures, Losses of RUT, Destruction of documents, Debt Regularization, Loss of Invoices, Invoices Investigated and/or Closures, Warnings
Life Cycle	Start-ups, Verification of Activities, Stamping of Documents, Changes of Information, Expiration of Prior Suspension of Activities
Relationships	Agents, Legal Representatives, Partners, Family, Suppliers, Accountants, Associations and Representations (assets, a history of offenses, investigations, closures)

Table No.2: Type of information used to construct the feature vector

In the pre-processing of data, using a rule to carry out data cleansing, those cases that exceed the mean plus five times the standard deviation are considered as outliers, leaving only those cases with a positive value of each code. In most cases, the distribution of variables with which they worked was declining, where a large percentage of taxpayers pay low amounts of taxes, and only a small group pays high amounts. The elimination of data is done carefully, while considering the expert opinion of those involved in the business, so as not to eliminate cases that were correct but far from the average. The same applies to the behavioral variables, because they constitute the misconduct of only a small group of taxpayers. Therefore, eliminating cases with higher values removes those taxpayers who generally have worse behavior, which are the focus group of the study.

Since the declaration of payment of VAT is done monthly and the declaration of income tax is done on an annual basis, the first transformation is to consider the annual total sum of the monthly amounts for each F29 code during the year to make them comparable with income tax information. Regarding the completeness of null data, the VAT information is more complete than the income tax information, because these codes should only be filed by taxpayers who are full accounting. Therefore, VAT debit and credit information is used to complete the revenue and cost data for the period, due to the direct relationship between them. For the rest of the income fields, the median is used for taxpayers in the same sales code section. Finally, due to the decreasing distribution of the tax variables, logarithmic transformation is applied to reduce the

impact of extreme data. To avoid variables with a greater range of values downplaying others with a smaller range, it is necessary to normalize the variables in ways that are comparable with each other, using the min-max standard deviation in the range [0.1].

Additionally, prior to selecting the variables used in behavioral models it is necessary to reduce them through principal component analysis (PCA). As a result 15 principal components in the micro and small enterprises are generated, explaining 61.3% of the variance in the data. Similarly, 16 principal components for medium and large businesses are generated that explain 59.9% of the variance in the data.

Since our interest was to generate behavioral variables related to the use and sale of false invoices and not other behaviors, those variables were selected that have a medium-high correlation with the variable use of false invoices. Those variables were discarded that have more than a 10% chance that the Pearson correlation coefficient is zero, except for some codes of interest such as the total debits, total tax credits and VAT payments.

Similarly, we discarded those variables that have a large percentage of null values. In this way 42 variables are selected in the micro and small segment and 36 variables in the medium and large segment for analysis. In the first group. 35% of the code variables correspond to the VAT, 35% of code variables are related to income tax and 30% to variables related to behavior. In the second group on the other hand these percentages vary by 31%, 38% and 31% respectively, with a higher prevalence of variables related to income tax.

After removing the outliers and inconsistent cases, the final data set is composed of 532,755 taxpayers who are micro and small enterprises and 22,609 medium and large enterprises, eliminating 4.6% of the first group and 3.4% of the second.

4.2. MODELING

In order to effect the characterization and identification of patterns, in a first stage, data mining techniques are applied to the universe of companies, in order to identify relationships between their payment of taxes (VAT and income) and behavioral variables associated with the use of false invoices. Then classification techniques are applied in those cases where the condition of fraud and no fraud is known, in order to identify specific patterns of this group of taxpayers. Finally, classification tools are applied to detect cases with and without fraud with the information generated.

4.2.1. CHARACTERIZING THE UNIVERSE OF COMPANIES

Initially, the SOM method is applied to the universe of taxpayers, to identify clusters or groups of taxpayers who have similar behavior. The working hypothesis assumed that when considering only the behavioral variables related to the use of false invoices combined with tax variables, it was possible to detect groups of taxpayers who had good or bad fiscal behavior, and know how they made their tax payment.

For the generation of experiments the "R-SOM" package is used, based on a rectangular network topology, with three input neurons and 24x24 output neurons in the case of micro and small enterprises, and 36x36 output neurons in the case of medium and large enterprises, with a maximum of 100 iterations. In the first group a random sample of 100,000 businesses is considered due to computational constraints. As a result 5 clusters are generated in the segment of micro and small enterprises and 6 clusters in the medium and large enterprises, as shown in Figure N° 2.

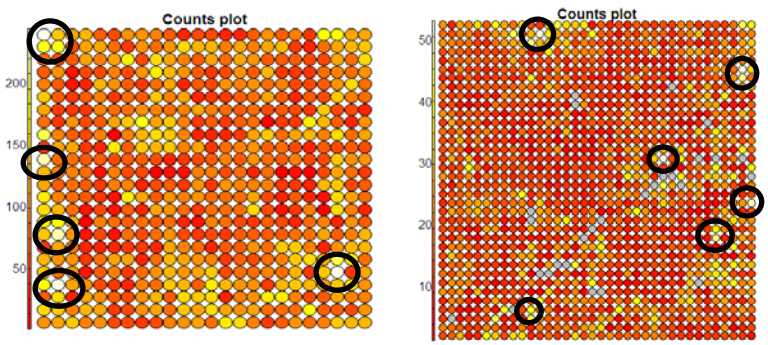


Figure No.2: Map resulting from SOM application in MI and PE (left) and ME, and GR (right)

The clusters obtained in the first group are mainly differentiated by the use of sales slips and/or invoices, the VAT payment level, the level of reported costs, the level of formality of the company, participation in other companies and some tracking issues. The medium and large group is differentiated by the use of sales slips and / or invoices, level of use of tax credit balances, credit notes and invoices of fixed assets, liabilities and assets as well as the results of previous audits and the level of formality of the company. While some patterns of behavior were found, these were not specifically related to the use of false invoices. Moreover, behavioral variables associated with historical characteristics and irregularities do not vary much from one group to another. While some patterns of behavior were found with this method, these were not specifically related to the use of false invoices, since the known cases with and without fraud were distributed across the map without a definite pattern. Moreover, behavioral variables associated with historical characteristics and irregularities do not vary much from one group to another.

Neural gas is then applied, considering the same number of clusters as the Kohonen map, using the "R- cclust" package, which generates an array with the characteristics of the centroids of each variable and a vector classification, marking the group each taxpayer belongs to. In this case, the groups generated are also influenced by the payment of taxes, but with major differences in terms of behavior. This allows the differentiation of which groups have better and worse behavior, and relates it to their tax payment, although the cases of false invoices were not necessarily found in the same group.

While these techniques can characterize the universe of taxpayers and identify some distinguishing patterns, considering those variables most correlated with the use of false invoices tends to give more prominence to tax than behavior variables, creating groups that differ in the type of transaction (sales with invoices and / or sales slips), the level of activity (high-low level of sales, costs) and tax payment (high-low), due to greater variability in these variables compared to those of behavior. Accordingly, the following patterns were identified associated with bad and good performance, considering the common points obtained by both methods, as shown in Figure 3.

MICRO AND SMALL				
Variable	Period	Concept	Good	Bad
Sales Slip Debits	t	VAT	↑	
Payment of VAT			↑	↓
Credits			↓	↑
Tax credits balances			↓	↑
Ratio debts/credits	t	Ratio Income VAT		↓
Ratio income/assets				↓
Stamped invoices	t-2	Stamping	↑	↓
Stamping frecuency			↑	
Activity checks	< t	Lyfe cycle	↓	↑
Crimes and irregularities	< t	Historical Behavior	↓	
Crimes indirectly related			↓	
Positive previous audit				↑

MEDIUM AND LARGE				
Variable	Period	Concept	Good	Bad
Costs and expenditures	t	Income	↑	
Assets			↑	
Liabilities			↑	↓
Credits	t	VAT	↑	
Tax credit balances			↑	
Number of sales slips				↓
Ratio costs/assets	t	Ratio Income VAT		↑
Ratio earnings/assets				↑
Ratio Invoice debits/total debits				↑
Formalization of accounting	t	Characteristics	↑	↓
Coverage			↑	↓
Legal representatives			↑	
Previous closures	< t	Historical Behavior		↑
Stamping restricted				↑
Accusations and closures				↑
Failures to attend				↑
Audits			↑	↓

Figure No. 3: Variables associated with good and bad behavior in the universe of taxpayers

4.2.2. **CHARACTERIZING CASES WITH FRAUD OR WITHOUT FRAUD**

When analyzing the distribution of each variable, it is noted that fraud cases are usually found among the extreme cases of each. For this reason, it is determined to apply decision trees to all audit data with known results, since it permits the identification of the cutoff point of each variable against which there is a change of behavior, the consideration of extreme cases and the generation of rules that can be validated and implemented.

The type of tree used is the CHAID (Chi-square automatic interaction detection), which allows non-binary classifications and the generation of branches from a node considering both continuous and categorical variables. This method requires access to significant sample sizes, since when divided into multiple groups there is a risk of finding empty or unrepresentative groups if there are not sufficient cases in each combination of categories. In addition, the exhaustive method is evaluated, a modification of the traditional algorithm which seeks to address some weaknesses of the traditional CHAID.

As shown in Figure 4, the factors that have the greatest impact were the result of previous audits and the percentage of purchases supported by invoices. This indicates that those who have been audited more times in the past and nothing was found, and whose purchases are not based primarily on invoices, are less likely to use false invoices than those whose purchases were mainly recorded by sales slips and had productive audits in the past. In fact, these two variables alone identify a number of end nodes with a preponderance of cases without fraud.

Additionally, the variable indicating a greater preponderance of crimes and irregularities associated with historical invoices combined with the frequency of stamping generates end nodes with a preponderance of cases with false invoices. In particular node 12, which contains nearly half the cases (46%), is decomposed into several branches according to the value taken by the average credit by invoice issued (the higher this indicator, the greater potential there is to commit fraud). Similarly, the preponderance of cases of fraud in each branch depends on the number of invoices issued, VAT paid, the total debits per invoice/sales slips, the relationship between costs and assets and the level of participation in other companies.

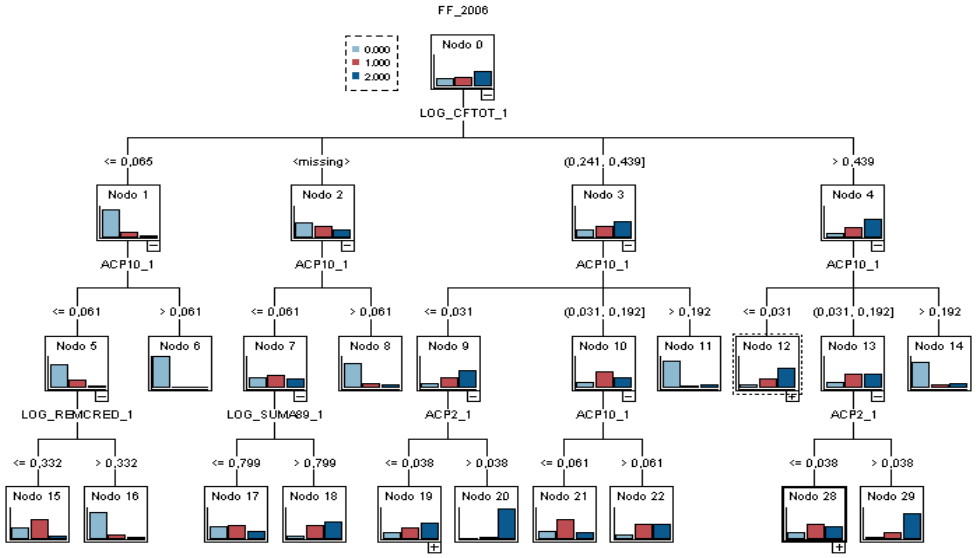


Figure No. 4: Main branches of decision tree in micro and small companies

This technique was highly effective in identifying patterns associated with fraud and without fraud, since the end nodes consisted mainly of cases of a single type, or were otherwise combined with cases with output value "1", which more closely approximate the behaviour of fraud cases "2". Considering the patterns and rules that are repeated in each branch of the tree to differentiate between cases with fraud and without fraud, Figure 5 shows the behaviours associated with each of them in each segment, which summarizes the main variables considered and the relationships that generate nodes with and without the use of false invoices.

The most important variables to distinguish cases of fraud in the micro and small enterprises were the result of previous audits, the total VAT determined, the percentage of credit supported by invoices, the relationship between tax credit balances and credits, total debits by invoice/sales slips and the relationship between stamped and issued invoices. The medium and large variables correspond to the total of tax credit balances, percentage of credit supported by invoices, the number of legal representatives, level of formalization of accounting and the relationship between earnings and assets, among others.

MICRO AND SMALL				
Variable	Period	Concept	No Fraud	Fraud
Invoice Debits	t	VAT		↑
Issued invoices			↓	
VAT				↑
Ratio credits invoices/total credits	t	Ratio Income VAT	↓	↑
Ratio tax credit balances/ credit mean			↑	
Ratio costs/assets				↑
Stamping frecuency	t-2	Stamping		↑
Ratio issued invoices/stamping invoices			↓	↑
Crimes and irregularities	< t	Historical Behavior		↑
Negative previous audits			↑	↓
Positive previous audits				↑
MEDIUM AND LARGE				
Variable	Period	Concept	No Fraud	Fraud
Tax credit balances	t	VAT	↑	↓
Ratio credit invoices/total credits	t	Ratio Income VAT	↓	↑
Ratio costs/assets			↓	↑
Age of company	t	Characteristic		↓
Formalization of accounting				↓
Economic activities				↑
Amount of orders to pay	< t	Historical Behaviour		↑
Failures to answer notifications				↑
Irregularities with invoices			↓	↑

Figure No.5: Variables associated with fraudulent and non- fraudulent behavior by false invoices

4.2.3. **FRAUD DETECTION**

For detection, artificial neural networks, decision trees and Bayesian networks were applied. To avoid over-adjustment of the network, the data are divided into two sets, a training set and a testing set, using the 70/30 rule. Moreover, both methods were implemented using the "SPSS Clementine" technological tool.

One of the complexities of neural networks is to determine the number of layers and hidden nodes and the number of epochs or iterations. To determine these parameters different numbers of cycles and nodes in the hidden layers were considered, in order to establish the appropriate values through trial and error. For the iterations the values used are 1,000, 5,000, 10,000 and 20,000. In the case of the nodes, using the number the software calculates by default based on the model and other data corresponding to half the number of input nodes, gives 3 and 20 nodes respectively.

In the case of Bayesian networks two methods are evaluated for constructing the network, the TAN algorithm and the Markov Blanket estimation algorithm available in the "SPSS Clementine" software. Additionally, a previous pre-processing of the variables is used to identify which are the most relevant variables and improve the processing time and performance of the algorithm. Likewise, an independent test of maximum likelihood and a chi-square test for parametric learning are used.

The experimental results are presented in Table N° 5, which contains the following indicators obtained in group testing: (1) Sensitivity- indicates the proportion of cases with fraud classified correctly, (2) Specificity- indicates the proportion of cases without fraud where the classification was correct, (3) Consistency- indicates the proportion of cases with and without fraud in which the classification was correct and (4) Error Rate- indicates the proportion of cases with and without fraud which were assigned to an incorrect class.

Exp. N°	Segment	Method	Sensitivity	Specifity	Consistency	Error Rate
1	Micro-Small	Neural Network	92.6%	72.9%	87.2%	12.8%
2	Micro-Small	Bayesian Network	82.3%	64.1%	77.9%	22.1%
3	Micro-Small	Decision Tree	89.0%	79.0%	87.0%	13.0%
4	Medium-Large	Neural Network	88.8%	59.1%	72.5%	27.5%
5	Medium-Large	Bayesian Network	73.3%	66.7%	70.3%	29.7%
6	Medium-Large	Decision Tree	79.0%	85.0%	82.0%	18.0%

Table No.5: Experiments in detection fraud by false invoices

In both segments, the best detection results of cases with false invoices were obtained with the neural network method. In the group of micro and small enterprises, experiment N° 1 showed that 92.6% of fraud cases were assigned to the correct class, while in the group of medium and large enterprises the proportion of fraud cases correctly allocated was 88.8%. Moreover, the power of generalization of the model was quite good, as test results were similar to those obtained in the network training, where detection of cases without fraud was 93.7% and 87.4 % respectively.

The neural network generated for the micro and small enterprises indicates a preponderance of variables associated with the payment of VAT and behavior, and to a lesser extent to income-related variables. The most relevant correspond to background information obtained from the activity checks, the relationship between tax credit balances and average credits, the total debits by invoices issued, the relationship between money income and assets and the relationship between VAT paid and the income declared as shown in Figure 6.

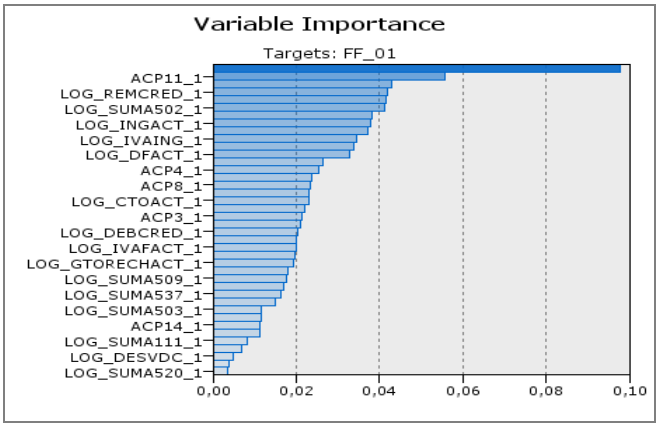


Figure No.6: Level of importance of the variables in micro and small companies according to neural network

In the case of medium and large companies, the most important variables correspond to the relationship between tax credit balances and average credits, accounts payable to related companies, total liabilities, the proportion of tax credits associated with invoices and the VAT determined in the period as shown in Figure 7.

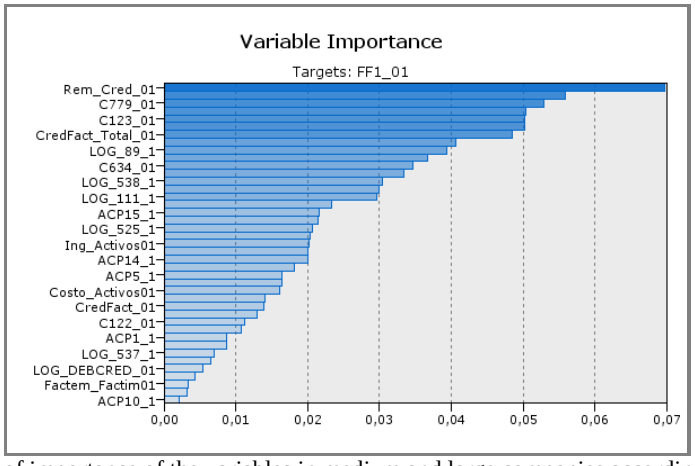


Figure No.7: Level of importance of the variables in medium and large companies according to neural network

5. CONCLUSIONS AND FUTURE WORK

The clustering and classification methods used to characterize the taxpayers who have good or bad fiscal behavior associated with the use of false invoices show that it is possible to identify some distinguishing characteristics between one group and another, which accord with what happens in reality. Particularly the neural gas method found that it was possible to identify some relevant variables to differentiate between good or bad behavior, not necessarily associated with the use and sale of false invoices. Kohonen's method however, did not provide any behavioral patterns associated with the use of false invoices, but rather clusters were detected in relation to taxation, in which the variables with the largest number of zeros and variance proved to have more impact in shaping the groups.

The decision tree method applied to cases in which the result of fraud and no fraud was known was a good technique to detect variables that could distinguish between fraud and no fraud. This is because when analyzing the distribution of variables in each group, it is noted that fraud cases tend to take more extreme values of the variables, so it was possible to distinguish

ranges in which there is a chance of having or not having fraud. On the other hand, the results were consistent with those observed in reality, according to the expert view.

Thus, in the case of micro and small enterprises the variables that allowed distinguishing between fraud and no fraud were mainly related to the percentage of tax credits generated by invoices with respect to total credit and previous audits with negative results. To the extent that the taxpayer was audited several times in the past and nothing was found, they are more likely to have no fraud in the future. On the other hand, where their credit is more associated with other items than invoices (fixed or other assets) they are less likely to use invoices to support their claims. Other important variables were the number of invoices issued during the year and its relation to the invoices stamped in the past two years, the total amount of VAT declared during the year, the ratio of average tax credit balances and positive prior audits and historical crimes and irregularities associated with invoices.

In the medium and large companies, the most important variables were the amount of surplus credit accumulated in prior periods, the percentage of credit associated with invoices, the relationship between costs and assets, the level of informality in their accounting and the age of the company, as well as the number of irregularities associated with previous invoices and the amount of orders to pay and historical failures to answer notifications.

In relation to the detection models, those which performed better were the multilayer perceptron neural network models, which for purposes of the study had an input layer containing the explanatory variables, an intermediate layer of processing and an output layer. In the case of micro and small businesses the percentage of correctly detected fraud cases was 92%, while in the case of medium and large enterprises, this percentage was 89%.

Given this result, and considering that in practice only a rather small group of companies in a year can be monitored, we recommend a combination of the results obtained with neural networks, decision tree and bayesian networks, in order to select for audit those that appear labeled as fraud in the neural network and have the highest odds of committing fraud under the Bayesian network and decision tree.

According to studies made by the SII, about 20% of taxpayers use false invoices to evade taxes. No information disaggregated by type of taxpayer exists but considering the percentage of classification of cases with and without fraud by neural network models, it is estimated that the universe of potential users of false invoices is 116,000 micro and small enterprises and 4,768 medium and large enterprises, generating a potential collection of USD$210 million dollars.

Finally, to test the actual detection model developed, and being consistent with the previous point, its implementation in activities in the field is vital to determine the level of accuracy in the classification of taxpayers selected in the sample. The implementation of a pilot program that will target the two economic sectors studied is recommended, which shall be conclusive in terms of the real effectiveness of the model.

For future work, we recommend generating new historical behavioral variables related to specific audits and level of coverage of these, considering other methods for preprocessing and selection of variables as well as cross-validation techniques to explore and implement other data mining techniques to improve the detection of cases with and without fraud.

Acknowledge

We are very grateful to the Chilean Millennium Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16), which partially supported this paper.

6. REFERENCES

[1] H. R. Davia, P. Coggins, J. Wideman, J. Kastantin, Accountant's Guide to Fraud Detection and Control, 2nd edition, Wiley, 2000.

[2] G. Harrison, R. Krelove, VAT refunds: A review of country experience, International Monetary Fund (IMF), 2005.

[3] F. Schneider, D. Enste, Shadow economies: size, causes and consequences, Journal of Economic Literature XXXVIII (2000) pp. 77-114.

[4] J. Slemrod, S. Yitzhaki, Tax avoidance, evasion, and administration, Handbook of Public Economics 3 (2002) pp. 1423-1470.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, American Association for Artificial Intelligence (1996) pp. 37-54.

[6] H. Chen, S. Huang, C. Kuo, Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets, Expert Systems with Applications 36 (2009) pp. 1478-1484.

[7] F. Bonchi, F. Giannotti, G. Mainetto, D. Pedreschi, A classification based methodology for planning audit strategies in fraud detection, in: Proceedings of the Fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99, ACM, 1999, pp.175-184.

[8] E. L. Barse, H. Kvarnström, E. Jonsson, Synthesizing test data for fraud detection systems, in: Computer Security Applications Conference, 2003. Proceedings. 19th Annual, CSAC Press, pp. 384-394.

[9] G. J. Myatt, Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, Wiley Interscience, 2007.

[10] M. Cecchini, H. Aytug, G. Koehler, P. Pathak, Detecting management fraud in public companies, Management Science 56 (2010) pp.1146-1160.

[11] OECD, Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Serie. General Administrative Principles- GAP004 Compliance Measurement, OECD Press, 1999.

[12] U. S. Government Accountability Office, Data Mining: Agencies have taken key steps to protect privacy in selected efforts, but significant Compliance Issues Remain, GAO Press, 2004.

[13] OECD, Compliance risk management, Audit case selection systems. Forum on tax administration compliance subgroup. Centre for Tax Policy and Administration, OECD Press, 2004.

[14] J. Dubin, Criminal investigation enforcement activities and taxpayer noncompliance, Public Finance Review 35 (2007) pp.500-529.

[15] OECD, Compliance risk management: Managing and improving tax compliance. Forum on tax administration compliance subgroup. Centre for Tax Policy and Administration, OECD Press, 2004.

[16] R. C. Watkins, K. M. Reynolds, R. Demara, M. Georgiopoulos, A. Gonzalez, R. Eaglin, Tracking dirty proceeds: Exploring data mining technologies as tools to investigate money laundering, Police Practice and Research: An International Journal 4 (2003) pp.163-178.

[17] U.S. Government Accountability Office, Lessons learned from other countries on compliance risks, administrative costs, compliance burden and transition. Report to Congressional Requesters, GAO Press, 2008.

[18] G. Williams, P. Christen, Exploratory multilevel hot spot analysis: Australian taxation office case study, in: Conferences in Research and Practice in Information Technology, volume 70, CRPIT Press, pp. 73-80.

[19] B. Torgler, Tax morale in Latin America, Public Choice 122 (2005) pp.133-157.

[20] V. García, J. Valderrama, Toward a more efficient tax policy, in: M. Giugale, V. Fretes-Cibils, N. J.L. (Eds.), An Opportunity for a Different PERU Prosperous, Equitable, and Governable, The World Bank, Washington, DC, USA, 2007, pp. 103-134.

[21] L. A. Digiampietri, N. T. Roman, L. A. A. Meira, J. J. Filho, C. D. Ferreira, A. A. Kondo, E. R. Constantino, R. C. Rezende, B. C. Brandao, H. S. Ribeiro, P. K. Carolino, A. Lanna, J. Wainer, S. Goldenstein, Uses of artificial intelligence in the Brazilian customs fraud detection system, in: Proceedings of the 2008 international conference on Digital government research, Digital Government Society of North America, 2008, pp. 181-187.

[22] S. Luckeheide, J. D. Velásquez, L. Cerda, Segmentación de los contribuyentes que declaran IVA aplicando herramientas de clustering, Revista de Ingeniería de Sistemas 21 (2007) pp. 87-110.

[23] J. Vesanto, Clustering of the self-organizing map, IEEE Transactions on Neural Networks 11 (2000) pp.586-600.

[24] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognition 41 (2008) pp.176- 190.

[25] S. Murthy, Automatic construction of decision trees from data: A multidisciplinary survey, Data Mining and Knowledge Discovery 2 (1998) pp.345-389.

[26] A. Parlos, K. Chong, A. Atiya, Application of the recurrent multilayer perceptron in modeling complex process dynamics, IEEE Transactions on Neural Networks 5 (1994) pp.255-266.

[27] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) pp. 131-163.

[28] D. Heckerman, D. Geiger, D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) pp. 197-243.