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

MODELO DE PREDICCIÓN DE DEFAULT TRIBUTARIO DE CONTRIBUYENTES DEL SEGMENTO DE MICRO Y PEQUEÑA EMPRESA DEL SERVICIO DE IMPUESTOS INTERNOS DE CHILE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

TRINIDAD RETTIG INFANTE

PROFESOR GUÍA:
CRISTIÁN BRAVO ROMÁN.

MIEMBROS DE LA COMISIÓN:
RICHARD WEBER HAAS.
JOSÉ MIGUEL CRUZ GONZÁLES.

SANTIAGO DE CHILE
2013

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: TRINIDAD RETTIG INFANTE
FECHA: 04/08/2013
PROFESOR GUÍA: CRISTIÁN BRAVO

MODELO DE PREDICCIÓN DE DEFAULT TRIBUTARIO DE CONTRIBUYENTES DEL SEGMENTO DE MICRO Y PEQUEÑA EMPRESA DEL SERVICIO DE IMPUESTOS INTERNOS DE CHILE

En Chile, durante los últimos años, ha habido un aumento en la tasa de evasión de impuestos, observándose que, en particular para el IVA esta cifra ha crecido un 8,5% con respecto el año 2007. Cada punto porcentual se traduce en una pérdida de recaudación de 350 millones de dólares, por lo que se hace necesario el diseño de un plan que revierta este efecto.

La presente memoria consiste en el desarrollo de un modelo estadístico que permita predecir el default en la declaración y pago de IVA para los contribuyentes de Micro y Pequeña empresa. Teniendo como resultado la detección de las variables que más influyen en esta conducta, la probabilidad de default de cada contribuyente para el período tributario julio 2013 y la identificación del porcentaje de default según zona geográfica.

La metodología de trabajo contempla las etapas de entendimiento del negocio, extracción de la información, preparación de datos, modelamiento y finalmente, la interpretación de los resultados. Dada la inexistencia de trabajos publicados en materia de default tributario, el sustento teórico de esta investigación se basa en el *credit scoring,* técnica utilizada en la industria bancaria.

Se toma como fuente de información el Data Warehouse del SII, con el que se construyen 55 variables que reflejan características demográficas, índices de liquidez, movimientos de caja, comportamiento y tamaño del negocio. Se prueban tres algoritmos de clasificación: árboles de decisión, regresión logística y redes neuronales. Se elige como modelo definitivo el de regresión logística, dada su clara interpretabilidad y buena capacidad de predicción, que alcanza una precisión global de 68,81%, especificidad de 67,29% y sensibilidad de 68,88%, cumpliendo con el objetivo inicial de obtener el mejor modelo predictivo balanceado posible.

Respecto de las variables, los resultados indican que las de mayor relevancia se relacionan con el historial de cumplimiento del F29, dónde se presenta una clara segmentación de los contribuyentes en tres tipos de conducta. Respecto de la identificación por zona geográfica, se aprecia el mayor porcentaje de default para la zona del norte grande del país, y el menor para la Región de Magallanes y la zona Centro y Oriente de la Región Metropolitana.

Se concluye que la presencia de errores en la información proporcionada por los contribuyentes, así como la cantidad de campos nulos encontrados, hace que aún no se cuente con una base de datos óptima para aplicar la técnica de *credit scoring*. El SII se encuentra trabajando en esta línea, lo que permitirá en un futuro obtener mejores resultados. Como recomendación final se propone utilizar las predicciones obtenidas para diseñar un plan de medidas preventivas, así como también evaluar el desarrollo de un modelo de alta precisión y baja especificidad que se enfoque en la detección de los defaulters más críticos.

AGRADECIMIENTOS

Quiero agradecerles a todos los que han estado conmigo en este largo ciclo que ya termina.

Agradezco a mi familia por confiar en mí y acompañarme. En especial a mi abuela Yolanda, abuelo Julio y tío Mario que creyeron en mí en todo momento y que han estado presentes en cada uno de mis logros. Por alegrarse por cada etapa que se va cumpliendo en mi vida y por darme palabras de aliento siempre que lo necesité.

Le agradezco a mi madre, Verónica, todo el cariño que me ha dado y las cosas que me ha enseñado. Por el esfuerzo que hace día a día para darnos todo a mí y a mis hermanos. Por la confianza y preocupación que siempre ha mostrado, y por ser el mejor ejemplo de perseverancia en mi vida. A ella le debo todo lo que soy y todo lo que he logrado.

A mis hermanos, Isidora, Rodrigo y Bárbara, por ser como son. Por sus muestras de cariño y por todos los recuerdos llenos de alegría. Por la paciencia que han tenido y por contenerme en los momentos difíciles. Simplemente los amo.

A mí querida Mari por su dedicación y paciencia incondicional en los momentos más críticos. Por regalonearme y entregarme el calor de hogar que me dio la fuerza para luchar día a día.

A mi profesor guía, Cristián Bravo, por su apoyo y correcciones a lo largo del semestre. Sin su ayuda esta memoria no hubiera salido adelante.

A la Universidad de Chile, institución que me formó durante estos siete maravillosos años.

Al Servicio de Impuestos Internos por facilitarme el uso de sus instalaciones y recursos.

A mis compañeros de trabajo, Brandon, Jorge y Constanza, por acogerme con calidez y gentileza, y darme la confianza necesaria para finalizar de la mejor manera este proceso.

Finalmente agradecer a mis amigas, Sofía, Consuelo, Fabiola, Nicole, Natalia, Carolina y Elisa, por hacerme reír y soportarme. Gracias por las conversaciones, anécdotas, viajes y jornadas de estudio. Por estar siempre que las he necesitado y por no dejar que el tiempo nos separe.

Cierro esta etapa orgullosa, tranquila y agradecida.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

1.1 EXPOSICIÓN DEL TEMA

De acuerdo a la Ley orgánica el Servicio de Impuestos Internos (SII) es el organismo estatal al que le corresponde la aplicación y fiscalización de todos los impuestos internos de Chile de carácter fiscal y de otro carácter, en el que tenga interés el Fisco. Depende directamente del Ministerio de Hacienda y, en términos generales, posee facultades judiciales y administrativas que permiten la recolección de los fondos estatales [14].

En relación a las facultades judiciales, el SII es el encargado de ejercer la acción penal por delitos tributarios y resolver las denuncias por infracción. En relación a las administrativas el SII se encarga de representar al fisco en la aplicación de impuestos, asesorar al Ministerio de Hacienda en materia tributaria e interpretar la ley tributaria y sus normas complementarias [14].

Dentro de las facultades administrativas, la principal actividad que se debe realizar es la de fiscalizar el cumplimiento de la normativa tributaria por parte de los contribuyentes, siendo la contracara del cumplimiento la evasión y elusión fiscal. Estas prácticas generan que la brecha tributaria aumente y, por consiguiente, se reduzca la recaudación.

El proyecto a realizar consiste en el desarrollo de un modelo estadístico para predecir una de las tantas formas de evasión tributaria. Lo que se busca es modelar el default en la declaración mensual de impuestos a las ventas y servicios (Formulario 29) para los contribuyentes de Micro y Pequeña empresa.

En la investigación se trabaja directamente con el Servicio de Impuestos Internos de Chile, quienes facilitan las bases de datos que tienen a su disposición, en particular la del Formulario 29, que constituye la fuente de información primaria para la construcción de un modelo predictivo que permita orientar la fiscalización a los contribuyentes más riesgosos.

1.2 JUSTIFICACIÓN

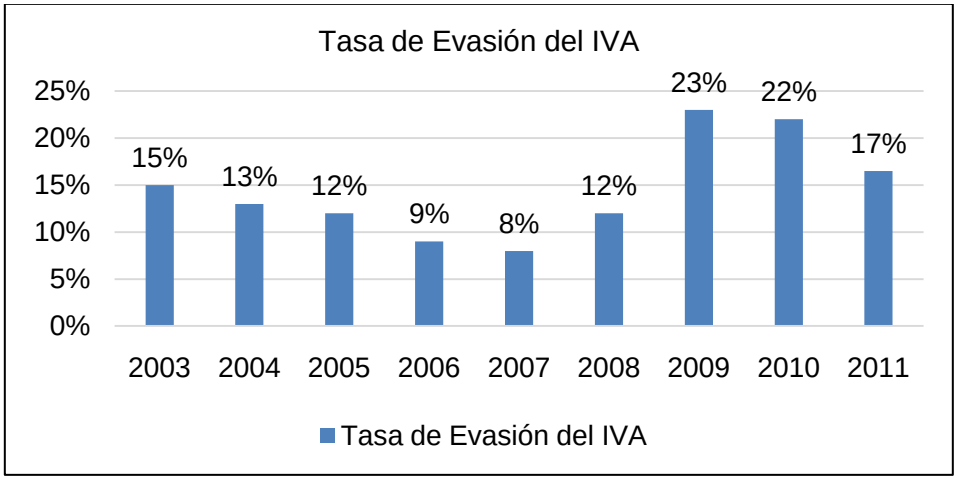
La relevancia de este proyecto radica en las tasas de evasión de impuesto que se ha presentado en el país durante los últimos años. Teniendo en cuenta los efectos de las crisis económicas, no deja de sorprender la marcada alza de la tasa de evasión del IVA desde un 8% en el 2007 a un 16,5% en el 2011 como se observa en la figura 1.

Este aumento es aún más preocupante si se considera el doble efecto que tiene la evasión del IVA. Una subdeclaración en este impuesto influye directamente en las utilidades de la empresa y por ende en la declaración de renta de un contribuyente. Jorrat y Serra [8] estimaron que en Chile cerca del 75% de la evasión del impuesto a la renta de las empresas se explica por la evasión del IVA.

Según el exdirector del SII, señor Julio Pereira, cada punto de evasión en el IVA significa una pérdida de recaudación de unos US$300 a US$350 millones, lo que equivale a 0,1% del PIB. Reducir la tasa actual de evasión en IVA desde 16,5%, a su nivel histórico más bajo del 8%, permitiría una recaudación adicional de US$3.600, lo

que representa un 0,6% del PIB [4]. Esto sin considerar su impacto adicional en los impuestos a la renta.

Figura 1: Tasa de evasión del IVA.



Fuente: Elaboración propia con datos de [8] y [7].

El exdirector del SII reconoció interés en reducir tasa de evasión del IVA hasta menos del 10% [4], para lograr llegar a esos niveles y seguir avanzando en la reducción del incumplimiento, se vislumbra la necesidad de elaborar un plan contra la evasión tributaria que considere, entre otras materias, orientar la fiscalización hacia el análisis de riesgo y la inteligencia fiscal.

Avanzar en el desarrollo de funciones que permitan diferenciar y hacer una selección más precisa de los casos a fiscalizar es fundamental para el SII. Si se logra una mayor focalización en los procesos de selección de contribuyentes a fiscalizar, los costos de Administración Tributaria debieran disminuir.

Disponer de información respecto de quienes tienen una mayor probabilidad de no declarar el F29 es crucial, pues, permitirá orientar mejor la fiscalización y facilitará la creación de planes de contención para ayudar a aquellos contribuyentes cuya propensión a cometer default sea mayor.

1.3 Objetivos

1.3.1 Objetivos Generales

Construir un modelo predictivo a un horizonte de tres meses[1] que estime la probabilidad de que un contribuyente que pertenece al segmento de micro y pequeña empresa deje de cumplir sus obligaciones de declaración y pago de IVA.

1.3.2 Objetivos Específicos

- Identificar las variables de mayor incidencia en la probabilidad de default[2] de un contribuyente.

[1] La razón para tomar como horizonte temporal tres meses radica en el tiempo que toma recolectar la información de las declaraciones e ingresarlas al data warehouse.

- Obtener un modelo o un conjunto de reglas que permitan estimar la probabilidad de que un contribuyente no presente su Declaración Mensual y Pago Simultáneo Formulario 29, en un período tributario determinado.
- Determinar la probabilidad de default de cada contribuyente para el período tributario de Julio 2013, documentado a través de un listado innominado de contribuyentes.
- Segmentar a los contribuyentes por zona geográfica, para luego caracterizar los segmentos según su probabilidad de default.
- Documentar el proceso para asegurar replicabilidad posterior.

1.4 METODOLOGÍA

La cantidad de antecedentes almacenados en las bases de datos del SII está por sobre la capacidad humana para analizar la información sin el uso de técnicas automatizadas. Si se considera que el volumen y variedad de registros se incrementa cada año, se hace necesario encontrar una técnica que permita trabajar con esta fuente de datos de manera eficiente.

El Descubrimiento de Conocimiento en Grandes Bases de Datos (*Knowledge Discovery in Databases*, KDD) es un método completo que incluye la extracción de información, la preparación de los datos, el modelamiento y la interpretación de los resultados [5].

Basando el desarrollo de la memoria en las etapas del KDD, se muestra a continuación el detalle de la metodología a seguir.

Instruirse respecto a la problemática de la investigación

Esta etapa incluye el entendimiento del negocio, comprensión de la problemática, estudio de posibles soluciones y definición de requisitos y alcances. Se indaga en las soluciones actuales que existen en el SII para detectar contribuyentes con situaciones irregulares desde la perspectiva tributaria y se comprende el funcionamiento de la institución tanto organizacionalmente como técnicamente (comprensión detallada del Formulario 29 y 22).

Estudiar la técnica de *credit scoring*

Investigar el estado del arte de la técnica de credit scoring en el sector bancario y averiguar sobre estudios en el área de default tributario para crear un marco conceptual.

Recopilar y comprender los datos existentes

Identificar las fuentes de información primaria y secundaria, comprendiendo la arquitectura, estructura y terminologías del *Data Warehouse* del SII. Realizar una inducción del uso del software *IBM SPSS Modeler*; y finalmente, recopilar las tablas del DW con los datos que se van a utilizar.

[2] Se define como default la no presentación del F29.

Construir variables

Diseñar las variables a construir, definiendo la variable objetivo y variables independientes de acuerdo a los requerimientos del problema y a la nómina de variables utilizadas en estudios previos. Realizar también el pre-procesamiento de datos, construcción, limpieza y transformación de variables de acuerdo a las necesidades de cada modelo.

Construir modelos

En esta etapa se construyen modelos de acuerdo a tres algoritmos que se utilizan en *credit scoring* en la actualidad. Estos algoritmos corresponderán a regresión logística, árboles de decisión y redes neuronales. También se procede a construir una muestra válida de contribuyentes para desarrollar cada modelo, y seleccionar los dos que presenten un mejor desempeño para mejorar su capacidad predictiva.

Interpretar y analizar resultados

Siendo esta etapa una de las más importantes, se busca testear la capacidad predictiva de los modelos seleccionados a través de ratios de desempeño, se realiza un análisis de sensibilidad y se evalúa la robustez de cada uno.

Aplicar Modelos

Se aplica el modelo elegido para predecir la probabilidad de default del conjunto de contribuyentes para el período Julio 2013 y se segmenta a los contribuyentes de acuerdo a características geográficas y de comportamiento.

Documentar nuevo conocimiento

Generar un documento explicativo del proceso de construcción del modelo. En el que también se expliquen resultados encontrados y se entregue una propuesta de mejora.

1.5 ALCANCES

Este trabajo considera la construcción de un modelo predictivo para el impuesto al valor agregado (IVA). Se trabaja con la información contenida en la declaración del Formulario 29.

El historial de datos a utilizar es de 6 años, contando desde el año 2007 hasta el año 2012. Esto debido tanto a la disponibilidad de la información, como a la metodología que propone Basilea II para la construcción de modelos de comportamiento.

El modelo solo será útil para determinar el comportamiento de los contribuyentes de la micro y pequeña empresa, de acuerdo a la clasificación creada por el SII desde el año 2009.

Los resultados que se entreguen (en términos de la probabilidad de default) servirán únicamente para el período tributario de julio 2013. Si se quiere predecir la

probabilidad para otros meses, se deberá aplicar el modelo con ciertas modificaciones (quedará documentado en el manual de aplicación).

Por último, es necesario recalcar que el modelo posiblemente tenga un sesgo asociado a la subdeclaración voluntaria de los contribuyentes, pues esa conducta no se considera en la construcción.

1.6 RESULTADOS ESPERADOS

Los resultados esperados del proyecto, de acuerdo a los objetivos planteados, pueden resumirse en los siguientes hitos y/o entregables:

- Una lista de las 10 variables de mayor incidencia en la probabilidad de default.
- Modelo predictivo para el default.
- Lista innominada de cada contribuyente para el período julio 2013 con las probabilidades de default.
- Segmentos de contribuyentes caracterizados por zona geográfica de acuerdo a su probabilidad de default.
- Documento técnico del proceso.

1.7 PLANIFICACIÓN

Tarea	Duración	Inicio	Fin
Actividades			
A. Análisis y entendimiento del negocio	**60**	**01-10-2012**	**21-12-2012**
Reunión Inicial y Presentación del Problema	**20**	**01-10-2012**	**26-10-2012**
Reunión de presentación	1	01-10-2012	01-10-2012
Benchmarking del mercado de *credit scoring*	1	08-10-2012	08-10-2012
Comprensión de la dinámica del SII y el IVA	1	17-10-2012	17-10-2012
Identificación y definición del problema	1	26-10-2012	26-10-2012
Definir requisitos y alcances	**5**	**29-10-2012**	**02-11-2012**
Estudiar soluciones	**35**	**05-11-2012**	**21-12-2012**
Investigación de metodologías	11	05-11-2012	19-11-2012
Reunión de alineamiento de objetivos, definiciones, carta Gantt y factores de riesgo	1	23-11-2012	23-11-2012
Estudio de documentos entregados por el SII	15	02-12-2012	21-12-2012
B. Recopilación y entendimiento de datos existentes	**23**	**26-12-2012**	**25-01-2013**
Determinar las fuentes de información	**8**	**26-12-2012**	**04-01-2013**
Presentación del Data Warehouse	3	26-12-2012	28-12-2012
Entendimiento de las variables existentes (qué formularios se usarán)	3	02-01-2013	04-01-2013
Comprensión de datos	**10**	**07-01-2013**	**18-01-2013**
Documentación de variables existentes (estructuras, modelo entidad relación)	5	07-01-2013	11-01-2013
Identificación de datos relevantes	4	14-01-2013	17-01-2013
Reunión de consenso de variables descartables con el equipo del SII	1	18-01-2013	18-01-2013

Inducción al uso del software	**3**	**21-01-2013**	**23-01-2013**
Clase explicativa del uso de *IBM SPSS Modeler*	2	21-01-2013	22-01-2013
Habilitación del Software para el proyecto y del lugar de trabajo	1	23-01-2013	23-01-2013
Recopilar los datos	**5**	**21-01-2013**	**25-01-2013**
Reunión de levantamiento de de información de los datos existentes	5	21-01-2013	25-01-2013
Hito 1: Fin de Análisis Inicial y Recopilación de Datos.	0	25-01-2013	25-01-2013
C. Preparación, Transformación y Selección de datos	**45**	**11-03-2013**	**10-05-2013**
Preparación de Datos	**20**	**11-03-2013**	**05-04-2013**
Integración de las bases de dato	5	11-03-2013	15-03-2013
Limpieza (anómalos, *missing values* y faltantes)	**15**	18-03-2013	05-04-2013
Análisis de datos	**10**	**08-04-2013**	**19-04-2013**
Proyecciones y estadísticas de datos recopilados	10	08-04-2013	19-04-2013
Reunión con el equipo del SII para definir la variable objetivo	1	17-04-2013	17-04-2013
Registro e importación de datos	**15**	**22-04-2013**	**10-05-2013**
Primera Selección de variables	**2**	22-04-2013	23-04-2013
Primera etapa de transformación de variables	**4**	24-04-2013	29-04-2013
Diseño del módulo de importación de datos	**3**	30-04-2013	02-05-2013
Creación de consultas de importación	**5**	03-05-2013	09-05-2013
Entregable 1: manual y módulo de registro e importación de datos	0	10-05-2013	10-05-2013
Hito 2: Fin de construcción del ETL y poblamiento del Datamart	0	10-05-2013	10-05-2013
Selección del método de minería a utilizar	**0**	**03-05-2013**	**09-05-2013**
D. Modelo	**35**	**13-05-2013**	**28-06-2013**
Construcción de Modelos (4 modelos)	**30**	**13-05-2013**	**21-06-2013**
Elección de Algoritmos	2	13-05-2013	14-05-2013
Análisis de variables	2	15-05-2013	16-05-2013
Selección de variables	5	17-05-2013	23-05-2013
Reunión de presentación de Análisis y modelos	1	24-05-2013	24-05-2013
Construcción de Algoritmos	5	27-05-2013	31-05-2013
Identificar el conjunto de entrenamiento	2	03-06-2013	04-06-2013
Calibración de los modelos	7	05-06-2013	13-06-2013
Evaluación de Modelos	**5**	**24-06-2013**	**28-06-2013**
Análisis de error	1	24-06-2013	24-06-2013
Calcular estadísticas de incertidumbre	1	25-06-2013	25-06-2013
Reunión de presentación de modelos y Retroalimentación. Iteración hasta llegar a nivel de aceptación.	1	26-06-2013	26-06-2013
Pruebas de validación	1	27-06-2013	27-06-2013
Entregable 2: Variables con mayor incidencia en la probabilidad de default y selección de modelo predictivo para	0	28-06-2013	28-06-2013

default			
Hito 3: Fin de construcción del modelo definitivo para predecir default	0	28-06-2013	28-06-2013
E. Implementación e Interpretación de resultados	**15**	**01-07-2013**	**19-07-2013**
Interpretar los resultados del modelo	3	01-07-2013	03-07-2013
Análisis de sensibilidad	5	04-07-2013	10-07-2013
Script con reglas de cada modelo: puntos de corte, cotas, implicancias para el SII	7	11-07-2013	19-07-2013
Entregable 3: Planilla con probabilidad de default para cada contribuyente y segmentación de contribuyentes.	0	01-07-2013	19-07-2013
F. Difusión del nuevo conocimiento	**8**	**22-07-2013**	**31-07-2013**
Generar informe con el proceso y los resultados	8	22-07-2013	31-07-2013
Entregable 4: manual de aplicación del modelo y descripción del proceso	0	31-07-2013	31-07-2013
Hito 4: Fin del proyecto	0	31-07-2013	31-07-2013

2. ADMINISTRACIÓN TRIBUTARIA EN CHILE

La Administración Tributaria en Chile está representada por tres instituciones dependientes del Ministerio de Hacienda: el Servicio de Impuestos Internos, el Servicio Nacional de Aduanas y el Servicio de Tesorería General de la República [14].

El Servicio de Impuestos Internos tiene por función la aplicación y fiscalización de todos los impuestos internos actualmente establecidos, fiscales o de otro carácter en que tenga interés el Fisco. El Servicio Nacional de Aduanas tiene por función la aplicación y fiscalización de todos los impuestos asociados al comercio con el exterior. Por otra parte el Servicio de Tesorería tiene como función la recaudación y cobranza de todos los impuestos y el manejo de la cuenta única tributaria de los contribuyentes.

2.1 SERVICIO DE IMPUESTOS INTERNOS (SII)

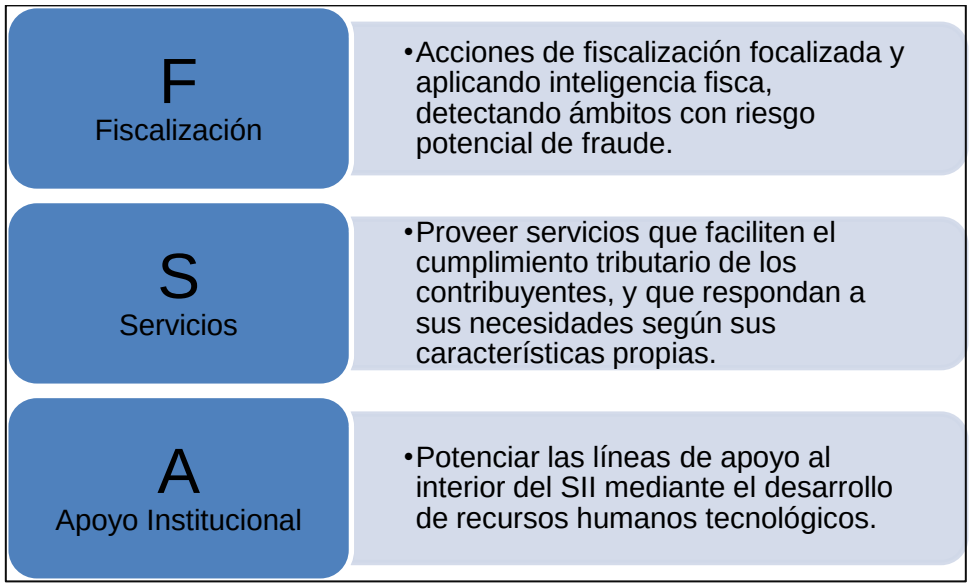
La misión del SII como institución del Estado es: “Fiscalizar y proveer servicios, orientados a la correcta aplicación de los impuestos internos; de manera eficiente, equitativa y transparente, a fin de disminuir la evasión y proveer a los contribuyentes servicios de excelencia, para maximizar y facilitar el cumplimiento tributario voluntario”[15].

En lo que se refiere a sus funciones, de acuerdo al Código Tributario y la Ley Orgánica del Servicio, al SII le corresponde:

- Interpretar administrativamente las disposiciones tributarias, fijar normas, impartir instrucciones y dictar órdenes a fin de asegurar su aplicación y fiscalización.
- Supervigilar el cumplimiento de las leyes tributarias que le han sido encomendadas; conocer y fallar como tribunal de primera instancia los reclamos que presenten los contribuyentes y asumir la defensa del Fisco ante los Tribunales de Justicia en los juicios sobre aplicación e interpretación de leyes tributarias.

- Crear conciencia tributaria, informar sobre el destino de los impuestos y las sanciones a que se exponen por el no cumplimiento de sus deberes.

Figura 2: Modelo de gestión de servicios del SII.



Fuente: Elaboración propia con información de [15].

En materia de gestión, se ha intentado generar una mirada global e integrada enfatizando el logro de resultados [15]. Esta mirada se hace cargo del rol que el SII tiene asignado por ley, como así también de la individualidad del contribuyente y de la demanda de servicios de atención e información tanto para contribuyentes como para usuarios externos.

Se mantiene como foco principal al contribuyente, entregando productos y servicios de calidad, adaptados a cada tipo de contribuyente, que faciliten el cumplimiento voluntario de las obligaciones tributarias, promoviendo el uso de la oficina virtual como canal de comunicación.

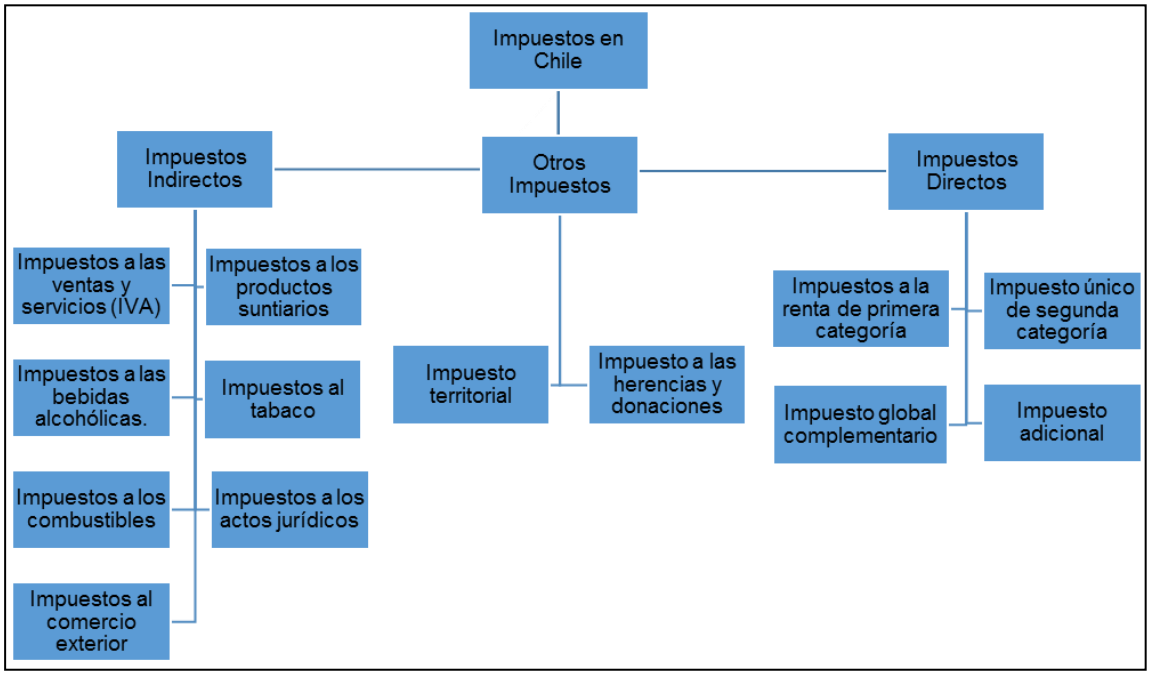
En la figura 2 se detalla el modelo de gestión del Servicio, basada en 3 focos que dan respuesta a los desafíos estratégicos; reducir la tasa de evasión, aumentar la calidad de servicio e incrementar la excelencia institucional.

2.2 IMPUESTOS

La información de esta sección fue extraída de [11].

Para el SII los impuestos son: “pagos obligatorios de dinero que exige el Estado a los individuos y empresas que no están sujetos a una contraprestación directa, con el fin de financiar los gastos propios de la administración del Estado y la provisión de bienes y servicios de carácter público”.

Figura 3: Clasificación de impuestos en Chile.



Fuente: Elaboración propia con información de la Subdirección de Estudios del SII.

Con el objetivo de clasificar los distintos tipos de gravámenes que actualmente posee el sistema tributario en Chile se clasificará en dos grandes grupos:

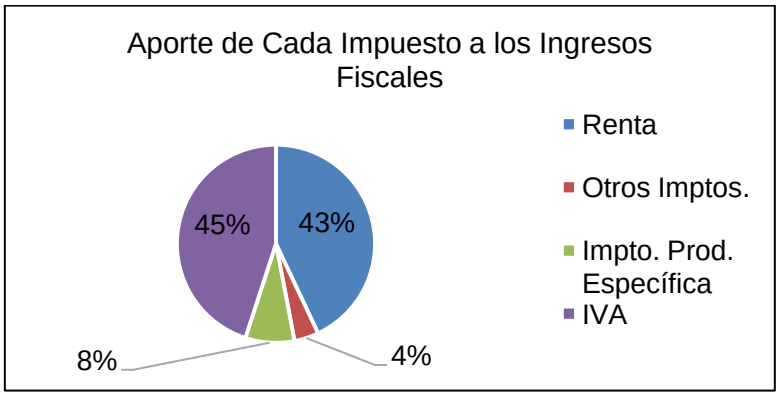
- **Impuestos Indirectos,** Impuesto que se aplica por el uso de la riqueza sobre las personas y, por lo tanto, indirectamente. Los impuestos son indirectos sobre las ventas, la propiedad, el alcohol, las importaciones, la gasolina, entre otros.
- **Impuestos Directos,** Impuestos que se aplican directamente al titular de la renta o riquezas que los paga, de manera que se puede reconocer quién lo pagó y su monto. Dentro de los impuestos directos están aquellos contemplados en la Ley de la Renta, como los impuestos a las utilidades de las empresas o los impuestos personales.

En la figura 3 se puede ver en detalle cada impuesto y su clasificación dentro de estos grupos.

Respecto al impacto de cada tipo de impuesto a los ingresos tributarios, se tiene que el año 2011 de un total de recaudación de $21,5 billones de dólares, el mayor ingreso ($18,9 billones de dólares) provino del impuesto a la renta y del impuesto al valor agregado [10]. Siendo estos dos, los de mayor importancia para las arcas tributarias como se puede observar en la figura 4.

En relación al objeto de investigación, se explicará brevemente el Impuesto a la Renta y el Impuesto al Valor Agregado (IVA) que competen directamente al desarrollo del estudio.

Figura 4: Distribución de ingresos tributarios por tipo de impuesto.



Fuente: Elaboración Propia con datos de [10].

2.2.1 IMPUESTO A LA RENTA

Es uno de los principales impuestos de Chile. Busca gravar las rentas definidas como "los ingresos que constituyan utilidades o beneficios que rinda una cosa o actividad, utilidades e incrementos de patrimonio que se perciban o devenguen, cualquiera sea su naturaleza".

Se compone de los impuestos de primera categoría, segunda categoría, impuesto adicional e impuesto global complementario. Cada uno de ellos está definido en el diccionario tributario del SII como:

- Los impuestos de primera categoría son un 20% y gravan las rentas que provienen del capital (puede provenir de una empresa o persona natural o jurídica).
- Los impuestos de segunda categoría se aplican a los ingresos que se perciben por trabajo independiente, y varía de acuerdo al tramo en que se encuentre.
- El impuesto adicional tiene una tasa de 35% y se aplica sobre rentas chilenas que son enviadas al extranjero.
- El impuesto global complementario se aplica a personas naturales residentes en Chile. Puede llegar a una tasa de un 40%.

2.2.2 IVA

El Impuesto al Valor Agregado es un impuesto interno que grava las ventas de bienes corporales muebles e inmuebles y también la prestación de servicios que se efectúen o utilicen en el país.

El Impuesto al Valor Agregado 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. El crédito fiscal equivale al impuesto recargado en las facturas de compra y de utilización de servicios, y en el caso de importaciones el tributo pagado por la importación de especies.

Si del período resulta un remanente, éste se acumulará al período tributario siguiente y así sucesivamente hasta su extinción, ello debido a la existencia de un sistema de reajustabilidad que se aplica hasta la época de su pago efectivo. Asimismo, existe un mecanismo especial para la recuperación del remanente del crédito fiscal acumulado durante seis o más meses consecutivos[3], originado en la adquisición de activo fijo.

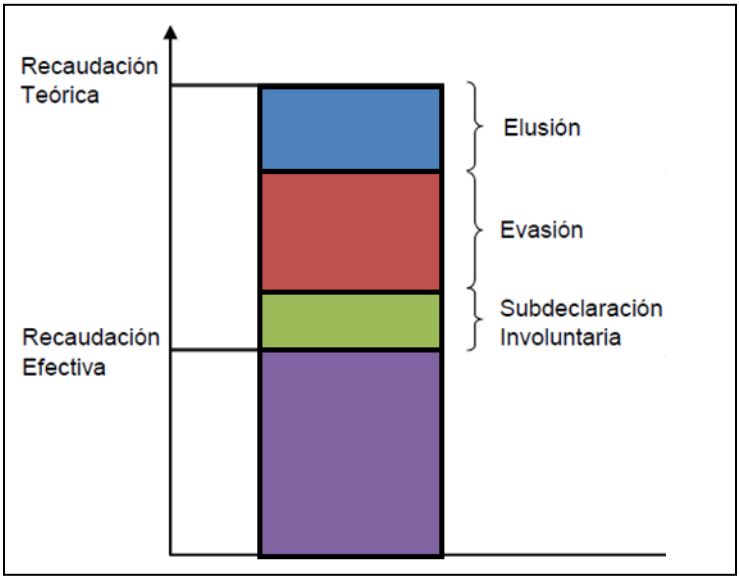
A partir del 1 de Octubre de 2003 el Impuesto al Valor Agregado (IVA) se aplica con una tasa del 19% sobre la base imponible de ventas y servicios que establece la ley, y la forma de declaración es a través del Formulario 29[4] por vía presencial o a través de internet.

2.3 BRECHA TRIBUTARIA

La recaudación potencial de un sistema tributario es aquella que se obtendría con un cien por ciento de cumplimiento tributario. La brecha que separa a la recaudación potencial de la recaudación efectiva es llamada brecha tributaria, y se define como la "diferencia entre lo que el gobierno debería recaudar de acuerdo a la legislación tributaria, y lo que realmente recauda" [8].

De acuerdo al estudio de Jorrat y Serra hay diversas causas que explican que los contribuyentes paguen menos impuestos de los que deberían. Dichas razones se pueden agrupar en tres categorías: subdeclaración involuntaria, elusión tributaria y evasión tributaria [8] como se observa en la figura 5.

Figura 5: Descomposición de la brecha tributaria.



Fuente: Elaboración propia con información extraída de [8].

[3] Artículo 27 Bis.
[4] Ver anexo 1: Formulario 29.

La subdeclaración tributaria es consecuencia de los errores involuntarios que puede cometer un contribuyente al momento de preparar su declaración de impuestos Estos errores se atribuyen al desconocimiento de la normativa.

La elusión tributaria es un concepto que hace referencia al uso abusivo de la legislación, es decir, a no respetar el espíritu de la ley, con el propósito de reducir el pago de impuestos.

Por último, la evasión tributaria corresponde a la subdeclaración ilegal o a la no presentación de declaración de los impuestos. En este caso hay un acto deliberado por parte del contribuyente de reducir sus obligaciones tributarias.

Para términos de este trabajo es preciso entrar en detalle sobre lo que se considerará como evasión tributaria, puesto que incluirá la definición de subdeclaración tributaria y de evasión tributaria expuestas en el estudio de Jorrat y Serra [8].

2.3.1 EVASIÓN

Se entenderá por evasión, la acción que se produce cuando un contribuyente deja de cumplir con su declaración y pago de un impuesto según lo que señala la ley. Esta acción puede ser involuntaria (debido a ignorancia, error o distinta interpretación de la buena fe de la ley) o culposa (ánimo preconcebido de burlar la norma legal, utilizando cualquier medio que la ley prohíbe y sanciona.)

En el IVA, que es el principal impuesto del sistema tributario chileno, los mecanismos más utilizados para evadir su pago pasan por una subdeclaración de los débitos, o bien, por un abultamiento de los créditos. En términos simples, lo anterior significa que el evasor registra menos ventas y por tanto menos débitos de IVA (o bien, más compras y más créditos de IVA) de los que en realidad realiza.

También se observa frecuentemente la no presentación del formulario, estando obligado a hacerlo, y es este mecanismo de evasión el que se tomará en cuenta para términos del estudio, definiéndose como default la no presentación del F29 en el tiempo correspondiente.

2.4 SEGMENTACIÓN DE CONTRIBUYENTES

La segmentación resulta útil para el SII pues establece procedimientos de fiscalización focalizada en el control del incumplimiento, aplicando procedimientos de auditoría ajustados a las características de los contribuyentes, privilegiando las acciones preventivas para evitar las infracciones tributarias. Se distinguen 2 líneas de negocio orientadas al contribuyente, en 2 áreas separadas y especializadas [15]:

Servicios para el Cumplimiento: Atención y Asistencia.

- Facilitación del cumplimiento de obligaciones tributarias.
- Controles sistémicos y formalizados adecuados para el registro de entrada de contribuyentes.
- Facilitación en el emprendimiento y operación de los negocios.

Fiscalización: Focalización del Control según tipo y riesgo.

- Actuaciones planificadas para segmentos: Personas, Micro y Pequeñas Empresas, Medianas Empresas y Grandes Empresas.
- Evaluación de riesgo y focalización de la fiscalización.

La segmentación de contribuyentes en el Servicio de Impuestos Internos se realiza de acuerdo a un criterio de monto de ingreso, que se definen por el nivel de ventas que tienen. Dada esta definición, existen 5 grupos de contribuyentes [12]:

Gran Empresa: Contribuyentes que tributen en primera o segunda categoría con ingresos mayores o iguales a 60.000 UTM o capital propio tributario mayor o igual a 300.000 UTM o compras mayores o iguales 60.000 UTM, en alguno de los dos últimos años.

Mediana Empresa: Contribuyentes que tributen en primera o segunda categoría con ingresos mayores o iguales a 15.000 UTM y menores a 60.000 UTM o capital propio tributario mayor o igual a 75.000 UTM y menor a 300.000 UTM o compras mayores o iguales 15.000 UTM y menores a 60.000 UTM, en alguno de los dos últimos años.

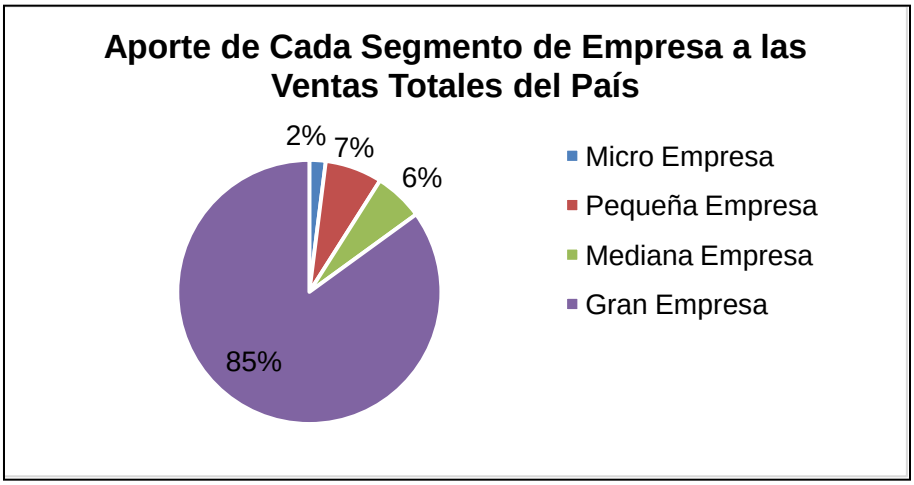
Pequeña Empresa: Contribuyentes que tributen en primera o segunda categoría con ingresos mayores o iguales a 1.400 UTM y menores a 15.000 UTM, en alguno de los dos últimos años.

Micro Empresa: Contribuyentes que tributen en primera o segunda categoría con ingresos menores a 1.400 UTM.

Persona: Contribuyentes personas naturales que tributen en Segunda Categoría o que no posean actividades económicas registradas en el Servicio de Impuestos Internos.

El aporte que cada grupo tiene a las ventas totales del país se muestra en la figura 6, donde se observa la relevancia de la grande y mediana empresa para el PIB del país y por ende para la recaudación fiscal.

Figura 6: Aporte de cada segmento empresarial a las ventas totales del país.



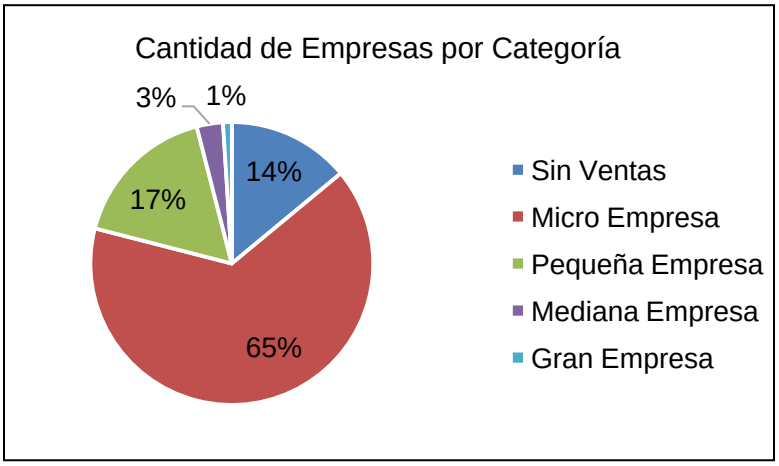
Fuente: Elaboración propia con datos de [13].

2.4.1 MICRO Y PEQUEÑA EMPRESA

Si bien este segmento no muestra un aporte significativo para el monto, tiene una gran importancia en función de su volumen, pues, la pequeña y micro empresa representa un 82% del total de empresas, como se puede apreciar en la figura 7.

Para efectos de este estudio, en el segmento de micro empresa se considerará 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. Para el segmento de pequeña empresa se incluirá a todas aquellas cuyas ventas sean superiores a las 2 400 UF, pero inferiores a las 25 000 UF al año[5].

Figura 7: Porcentaje de empresas según clasificación acuñada por el SII.



Fuente: Elaboración propia con datos de [12].

Las actividades económicas más frecuentes de ambos segmentos se resumen a continuación de acuerdo al rubro [13]:

- Agricultores, ganaderos, silvicultores
- Talleres artesanales, mineros, pirquineros, pescadores
- Cooperativas
- Comerciantes (mayoristas, minoristas, suplementeros)
- Subcontratistas de empresas constructoras
- Elaboradores de productos (Industria Manufacturera)
- Prestadores de servicios
- Profesionales
- Transportistas

3. CREDIT SCORING

En el contexto de detección de default tributario no existen investigaciones académicas públicas. Esto puede deberse tanto a que no se han realizado estudios al respecto o bien a que las que se han desarrollado, no son accesibles debido a la ley de

[5] Para hacer la analogía con UF se aproximó el monto de acuerdo al valor de la UTM y de la UF de Mayo 2013.

secreto tributario que prohíbe exhibir cualquier información asociada. En el ámbito internacional tampoco se ha encontrado estudios relacionas, debiéndose primordialmente a la segunda razón.

Para efectos de esta memoria, se muestra a continuación el estado del arte del *credit scoring*, técnica estadística utilizada para predecir el default bancario, y que se tomará como base para el desarrollo de una teoría de default aplicada al pago de impuestos.

El *credit score* es una expresión numérica basada en un análisis estadístico sobre el historial de crédito de una persona, que representan la solvencia de la misma [16]. La aplicación del *credit scoring* nació en el rubro bancario pero no se limita a este. Un gran número de organizaciones como compañías de telefonía móvil, de seguros y los departamentos gubernamentales emplean las mismas técnicas.

El *credit scoring* permite evaluar el riesgo potencial de los préstamos a los consumidores y para mitigar las pérdidas por deudas incobrables. Este índice determina quién califica para un préstamo, a qué tasa de interés y cuáles son los límites de monto.

La teoría tras el *credit scoring* se basa en el desarrollo de modelos estadísticos que permitan categorizar a los solicitantes de créditos bancarios en dos grupos [16]: "buenos" y "malos" dependiendo de su propensión a caer en default, es decir, a presentar un retraso en el pago de la cuota correspondiente. Se requiere como primera instancia proporcionar una estimación cuantitativa de la probabilidad de que un cliente vaya a caer en default para luego generar reglas de corte en la clasificación.

Con el fin de estimar esa probabilidad, es necesario construir un modelo que resuma adecuadamente una gran base de datos de información, que puede incluir tanto el historial de comportamiento del individuo como las características intrínsecas del mismo [9]. Para la construcción de este modelo estadístico se debe conocer la metodología KDD en la que se basa la técnica de *credit scoring* y que se resume a continuación.

3.1 METODOLOGÍA KDD

El *Knowledge Discovery in Databases* (KDD) se ha definido como un proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y comprensibles en los datos, siendo su principal función la de interpretar grandes cantidades de información para encontrar relaciones que permitan posteriormente tomar decisiones [5]. En la figura 8 se puede apreciar el diagrama del proceso KDD que será utilizado en la presente memoria. A continuación se describe en detalle cada etapa de acuerdo a la información contenida en [18]:

Análisis y Entendimiento del Negocio: Consiste en comprender el dominio de la aplicación, del conocimiento relevante y de los objetivos del proyecto.

Selección: Consiste en la selección del conjunto de datos, del subconjunto de variables y de la muestra de datos sobre la cual se va a realizar el descubrimiento.

Preprocesamiento: Involucra la limpieza y pre-procesamiento de los datos. Se deciden las estrategias sobre la forma en que se van a manejar los campos de datos no

disponibles o vacíos, se realizan cambios fundamentales a la base de datos y se analiza la información.

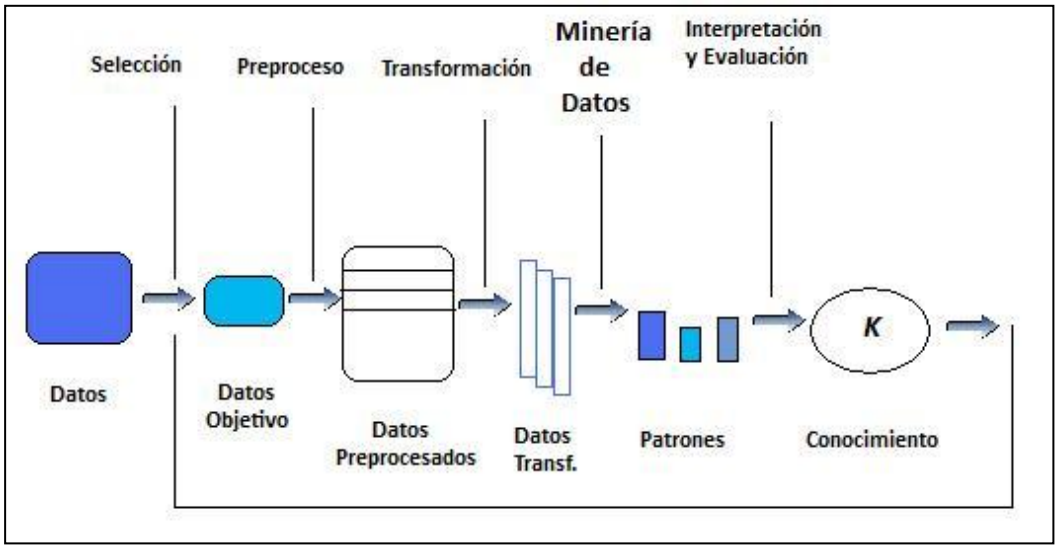
Transformación: Consiste en encontrar otras representaciones de los datos y reducir el número de variables a ser consideradas, encontrando las características más significativas.

Data Mining: Se debe decidir el objetivo del modelamiento (regresión, clasificación o agrupamiento), esto con la finalidad de definir el algoritmo que se aplicará. Incluye tanto la selección de métodos a ser utilizados para la búsqueda de patrones en los datos, como la decisión sobre que parámetros son los más apropiados para cada modelo.

Interpretación de Resultados: Interpretación de los patrones encontrados, determinando su impacto para el negocio y como se puede aplicar el nuevo conocimiento. Incluye encontrar puntos de corte, cotas y reglas de decisión.

Consolidación: Consiste en documentar el nuevo conocimiento descubierto, incorporándolo al funcionamiento del sistema e informando a las partes interesadas.

Figura 8: Diagrama de etapas del KDD



Fuente: Elaboración propia con información de [5].

El proceso descrito puede involucrar varias iteraciones y puede contener ciclos entre dos de cualquiera de los pasos, haciendo que la secuencia no siga un orden estricto. La mayor parte de los trabajos realizados sobre KDD se centran en la etapa de minería de datos. Sin embargo, los otros pasos se deben considerar igual o más importantes para el éxito de la metodología. Por eso, gran parte del esfuerzo debe recaer sobre la fase de preparación de los datos, siendo crucial para el éxito en la veracidad de los resultados que se obtendrán.

3.2 DEFINICIONES PRELIMINARES

Esta etapa es la que toma más tiempo y requiere de mayor profundidad investigativa. Se debe definir en detalle los parámetros y conceptos del proyecto para asegurar la correcta aplicación del método. Dichas definiciones incluyen la ventana de la muestra, ventana de comportamiento y la definición de "buenos" y "malos" (variable objetivo).

3.2.1 CASOS "BUENOS" Y "MALOS"

Hacer una correcta clasificación de los pagadores "buenos" y "malos" depende intrínsecamente de la definición que se asigne a cada categoría. Por lo mismo, no se debe pasar por alto las siguientes consideraciones:

- Definir los casos "buenos" y "malos" de acuerdo a un planteamiento estratégico que considere los objetivos de la organización. Se recomienda definir como "malos" aquellos casos que sean nocivos o no rentables para la empresa.
- Construir una definición lo suficientemente precisa para identificar correctamente el comportamiento de los contribuyentes [16]. Una definición muy restringida de los casos "malos" puede llevar a muestras muy pequeñas de este segmento, mientras que si la definición es muy amplia no genera diferenciación y produce un *score* que no es útil.
- La definición debe ser fácil de entender, interpretar y monitorear [16]. Estas características hacen que la construcción de la variable sea viable, y simplifica la toma de decisiones a partir de los resultados que entregue el modelo.
- Considerar los requerimientos y reglas del gobierno o los entes reguladores, independientemente de los requerimientos operacionales de la organización [16]. Si existe una pre-definición de lo que se considera como "default" esto se debe evaluar. De acuerdo a Basilea II, para el mercado bancario el default se considera como 90 días de atraso en el pago de una cuota.

Por último es necesario aclarar que deben existir aproximadamente un mínimo de 2000 casos clasificados como "malos" y 2000 "buenos". Esto para tener suficiente información para entrenar el modelo [16].

3.2.2 VENTANA DE MUESTRA Y DE COMPORTAMIENTO

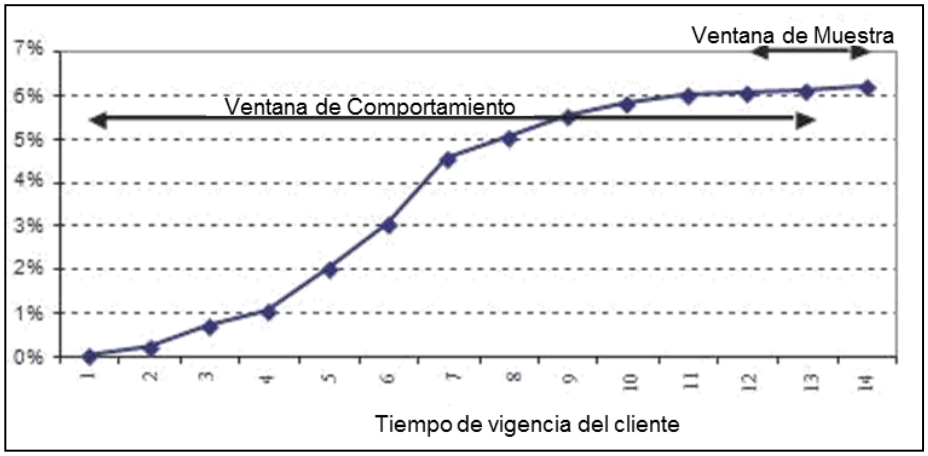
El *credit scoring* se basa en el supuesto de que "el comportamiento futuro de las personas refleja el comportamiento pasado" [16]. Por esta razón se hace necesario recopilar el historial de los clientes durante un período de tiempo definido, de manera de determinar si se clasifican como "buenos" o "malos", y así poder predecir casos futuros.

En el *credit scoring* existen dos conceptos asociados a definir los tiempos que se deben considerar en el análisis: la "ventana de comportamiento" y la "ventana de muestra". La "ventana de comportamiento" (*performance window*) es un parámetro que hace alusión al intervalo de tiempo en el que se observa al individuo para poder clasificarlo dentro de una de las dos categorías de la variable objetivo. La "ventana de muestra" (*sample window*) hace referencia al intervalo de tiempo del que se van a

seleccionar los casos de “buenos” y “malos” para construir la muestra con la que se trabajará [16].

En algunos casos es relevante realizar un análisis en el que se determine cuál es el marco de tiempo ideal a utilizar. Una forma sencilla de definir ambos conceptos es estudiar los pagos o cumplimientos de la cartera de clientes, en conjunto con la evolución de los casos "malos" definidos en el tiempo.

Figura 9: Evolución de casos "malos" en el tiempo.



Fuente: Imagen obtenida de [16].

La “ventana de muestra” es elegida luego de un período de tiempo en que se considera que la tasa de casos “malos” pasa a ser estable. Por ejemplo, en la Figura 9, de los meses 12 al 14 es una buena estimación para la “ventana de muestra”, determinando así una “ventana de comportamiento” de 12 meses.

La selección de casos que se hace al aplicar el análisis de madurez, permite minimizar las posibilidades de error en la clasificación del comportamiento, pues, se ha considerado un tiempo suficiente para que puedan expresar sus características. Esto también evita la subestimación de casos “malos” en la población que se quiere estudiar.

3.3 DESARROLLO DEL MODELO DE *CREDIT SCORING*

A continuación se explica las etapas en el desarrollo del *credit scoring* y las principales herramientas que se aplican en cada una.

3.3.1 ELECCIÓN DE LA MUESTRA

El muestreo es una técnica que busca seleccionar un subconjunto de casos de una población con la finalidad de ahorrar recursos de procesamiento, pero obteniendo resultados similares a los que se alcanzarían trabajando con la población completa.

Al elegir la muestra se espera que esta sea representativa de las características de la población. Para que esta selección sea válida existen diversas formas de llevar a cabo el muestreo, siendo la más utilizada la técnica de intervalos de confianza.

El intervalo de confianza es una medida en la que se define a priori el porcentaje de confianza que se quiere tener en los resultados. Bajo el supuesto de que se quiere un 95% de confianza, el tamaño de la muestra (n) se define como [16]:

$$n = \left(\frac{z * \sigma}{d}\right)^2 \qquad z = 1{,}96$$

Teniendo el tamaño de la muestra se puede hacer una selección aleatoria estratificada, en donde se seleccione una cantidad de casos de cada categoría de la variable objetivo guardando equivalencia con las proporciones en que se da esa variable en la población. Para comprobar que el muestreo sea el correcto basta con ver que la distribución de la muestra sea similar a la distribución de los casos de la población.

En el caso del *credit scoring* (y como se mencionó en el apartado anterior) existe ocasiones que la población tiene un muy bajo porcentaje de casos "malos". Lo que normalmente se hace entonces es generar una muestra equilibrada con el método de *over –sampling* [16] que propone aumentar la cantidad de casos "malos" seleccionados, asegurando una cantidad mínima de 2000 casos "malos" y 2000 casos "buenos" para obtener predicciones más realistas.

Finalmente, la muestra se debe dividir en dos subconjuntos: conjunto de entrenamiento, y de validación. El primer grupo se utilizará en el entrenamiento del modelo y el segundo se utilizará para testear el desempeño del modelo. Normalmente se divide la muestra en la siguiente proporción: un 70% u 80% para el conjunto de entrenamiento, el resto (20-30%) para validar el modelo [16].

3.3.2 SELECCIÓN DE LOS DATOS

Se inicia con la identificación de los datos relevantes que caractericen el fenómeno en estudio. Para lograr este primer paso se debe tener algún acercamiento con los datos que se requieren, en cuanto a su importancia para la problemática específica, su disponibilidad y costo; se determinan las fuentes de información primarias y secundarias.

El objetivo de esta etapa es crear un conjunto de datos para la investigación, que consiste en la selección del subconjunto de variables, o atributos (variables independientes) que representan de mejor forma el problema que se quiere resolver (variable dependiente).

La etapa de selección es recurrente en el desarrollo del *credit scoring*. Existen a modo general dos etapas donde se hace selección. La primera es previa al tratamiento de datos, y permite identificar los campos del Data Warehouse que son útiles para la construcción de variables. La segunda, que se realiza después del tratamiento de datos permite obtener las variables relevantes y potencialmente significativas para el modelo.

Se han desarrollado variados algoritmos para la segunda etapa de selección, que dependen tanto de los criterios de evaluación como de las estrategias de búsqueda.

De acuerdo al criterio de evaluación del algoritmo se tienen dos métodos [9]:

- Métodos de selección que utilizan una medida para evaluar individualmente cuán explicativo es cada atributo (medidas de ganancia de información, medidas de distancia y las medidas de dependencia).
- Métodos de selección que evalúan cuán explicativo es un subconjunto de atributos. El objetivo es tratar de encontrar grupos de características buenas o malas. La búsqueda de subconjuntos requiere un tiempo computacional extenso, por lo tanto, algunas estrategias de búsqueda se aplica para disminuir el número de subconjuntos a ser evaluada. A modo de ejemplo, se tiene la selección *forward y backward*.

Cabe destacar que el proceso de selección de variables se incorpora a veces en los algoritmos de clasificación asociados al modelo final. De acuerdo a esto, los métodos de selección de atributos se pueden dividir en "algoritmos de filtro" y los "algoritmos de envoltura" (*wrapper*) [9].

- Los algoritmos wrapper evalúan los atributos de acuerdo a la precisión de la clasificación proporcionada, utilizando un algoritmo de clasificación.
- Los algoritmos de filtro, son independientes de cualquier algoritmo de aprendizaje y utilizando una medida concreta que refleje las características del conjunto de datos.

3.3.3 PREPROCESAMIENTO

Los datos que no van a ser obtenidos del DW pueden ser extraídos de varias fuentes de datos, por lo que deben ser transformados a un formato común manteniendo consistencia e integridad referencial.

Para realizar esta tarea, se crea una etapa de pre-procesamiento de datos. Si no se tratan problemas de consistencia, los datos pueden contener potencialmente valores erróneos, faltantes o inconsistentes, situación que puede causar resultados alterados en la etapa siguiente.

En esta etapa es vital la detección de datos fuera de rango (*outliers*), valores perdidos (*missing values*) o datos erróneos. Los datos fuera de rango se definen como “valores u observaciones en una muestra que se alejan tanto del promedio que sugieren ser resultado de un error en la medición” [2]. Los datos perdidos son los casos en donde para determinado campo no se tiene registro, ya sea porque se perdió o bien porque el cliente no llenó ese campo (valores nulos).

El reconocimiento de estos datos es difícil de automatizar, y la limpieza que requiere una corrección de errores automática puede resultar incluso más ardua. Cada vez que una operación manual es realizada, se debe tratar de identificar el camino tomado para luego eliminar el proceso manual y realizar procesos automáticos que lo reemplacen. Las técnicas más utilizadas para el tratamiento de estos datos son:

Fuera de rango

Una primera aproximación es eliminar los valores extremos de la base de datos. Sin embargo es necesario realizar un test previo, pues se puede estar reduciendo la varianza de la población. Si bien el test puede indicar donde hay datos anómalos, no

determinan con certeza que estos sean erróneos. Es por lo mismo que se hace necesaria la identificación y análisis manual de cada caso antes de eliminarlo.

Una herramienta que se utiliza para esto es el Test de normalidad, que identifica los datos que estén fuera de un intervalo que incluye tres desviaciones estándar sobre la media. El supuesto que se toma acá es la normalidad en la distribución de los datos por lo que se debe tener en consideración al momento de aplicarlo [2].

Se debe tener en cuenta que si más de un 20% de la data está clasificada como *outlier*, entonces alguna de las consideraciones es errónea y se debe cambiar el tipo de herramienta utilizada para su detección. En este caso, resulta muy útil el conocimiento experto en la materia que puedan orientar en los criterios de eliminación.

Valores perdidos

- Eliminación por casos: excluir todos los casos (o clientes) que tienen valores nulos o perdidos en una o más de los campos seleccionados. Esta técnica no es recomendada en bases de dato compleja donde la distribución es aleatoria y podría terminar borrándose la mayoría de los casos [16].
- Eliminación de campos: excluye las variables completas donde hayan muchos casos faltantes. Esto es recomendable en variable que mantengan una alta correlación con otras donde no hayan tantos datos faltantes [16].
- Sustitución de medias (o mediana): sustituye los casos faltantes por la media de la variable. Una complicación de esta técnica es que subestima la dispersión de la variable y modifica el valor de algunos estadísticos descriptivos [16].
- Asignar valores especiales: en los valores faltantes se puede ingresar números o categorías que estén fuera del rango de la variable, con la finalidad de incluirlos en el análisis [16].

Si bien la opción de eliminar es la más conveniente, se debe considerar la última alternativa, pues, en el caso del *credit scoring,* los datos faltantes pueden contener información relevante sobre el comportamiento de los clientes que se relacionen con otras características no observadas.

3.3.4 TRANSFORMACIÓN DE DATOS

Desde una visión operacional es necesario encontrar las variables más significativas para representar los datos. Dependiendo de los objetivos del proceso la transformación de variables puede incluir la agrupación de campos o creación de variables completamente nuevas (categorizar, ajustar escala y cambiar espacio dimensional). De acuerdo a la naturaleza de su construcción se describen a continuación las técnicas más utilizadas.

Técnicas de Agrupación

Para reducir esta dimensionalidad, se debe estudiar la correlación y la dinámica entre los atributos y el comportamiento de default. Teniendo en cuenta las características del negocio, se deben generar *clusters* de atributos para luego elegir una o más variables representativas y eliminar el resto. Las técnicas para el agrupamiento de variables que se utilizarán en esta memoria son:

- ACP (Análisis de componentes principales): es una técnica de reducción de dimensionalidad en donde se crean nuevas dimensiones a partir de proyecciones de las existentes, para representar mejor los datos [16].
- *Clusters*: agrupación por conglomerado de las variables similares. Su resultado varía dependiendo de la medida que se utilice [16].
- Análisis de correlación: variables que estén altamente correlacionadas se evaluarán para eliminar una de ellas dos [16].

Técnicas de Transformación

Con el objeto de encontrar otras representaciones de los datos que pudieran ser más significativas en la búsqueda del conocimiento, se pueden utilizar distintos métodos de transformación. Se presentan clasificados de acuerdo a su objetivo primordial:

- **Discretizar:**
  Categorización: Transformar variables de valores continuos, nominales o discretos en variables categóricas.

- **Ajustar escala**
  Logaritmo: permite expandir la escala de medida y visualizar diferencias en variables que presentan datos muy condensados (por ejemplo: ingresos).
  Inversa de la variable.

- **Estandarizar**
  Z-score: Lleva a la variable a un espacio dimensional con media 0 y desviación estándar 1. La ecuación de conversión es la siguiente:

$$x' = \frac{(x - \bar{x})}{\sigma}$$

- **Normalizar.**
  Min-Max: Lleva a la variable a un espacio dimensional entre 0 y 1. La ecuación de conversión es la siguiente:

$$x' = \frac{(x - min_x)}{(max_x - min_x)} \cdot (max_{x'} - min_{x'}) + min_{x'}$$

- **WOE (*Weight of Evidence*)** [16]
  Transforma la variable en términos de la cantidad de información que aporta. Su rango varía entre 0 y 1.

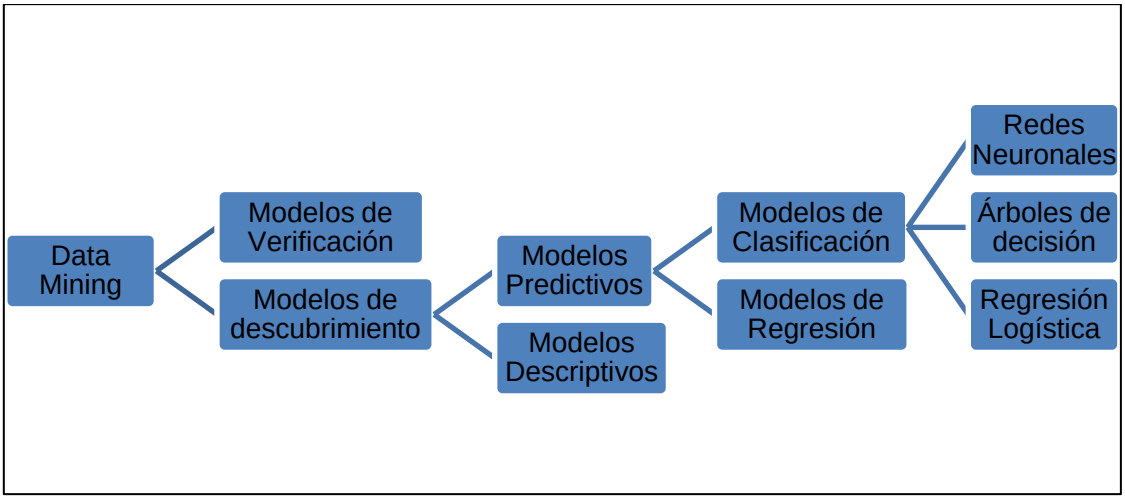
3.3.5 Creación de Modelos (Técnicas de Data Mining)

La información para esta sección se obtuvo de [18]. Los métodos de minería de datos se utilizan para diferentes propósitos. En la figura 10 se muestra un diagrama resumen con los distintos tipos de métodos. Para ayudar en la comprensión de su variedad se debe hacer distinción en lo siguiente:

- Métodos orientados a la verificación (el sistema verifica la hipótesis del usuario)
- Métodos orientados al descubrimiento (el sistema encuentra nuevas reglas y patrones de forma autónoma).

Los métodos de descubrimiento son los que "identifican automáticamente los patrones en los datos". En esta rama se tiene dos lineamientos principales: métodos de predicción y métodos de descripción.

Figura 10: Diagrama de modelos de data mining.



Fuente: Elaboración propia con información de [18].

Los métodos de descripción están "orientados a la interpretación de datos, centrándose en la comprensión de la forma en que las variables se relacionan entre ellas". Los métodos de predicción tienen como objetivo construir un modelo de comportamiento, que entregue información nueva y nunca antes vista y que sea capaz de predecir los valores de una o más variables relacionadas con la muestra.

La mayoría de las técnicas de minería de datos orientada al descubrimiento se basan en el aprendizaje inductivo, en donde se construye un modelo a partir de un número suficiente de ejemplos de entrenamiento. El supuesto básico del enfoque inductivo es que el modelo entrenado es aplicable a futuros ejemplos aún no conocidos.

Los métodos de verificación, por otro lado, se enfocan en la evaluación de una hipótesis propuesta por una fuente externa (como un experto). En ellos se utilizan las técnicas más comunes de estadísticas, como test de bondad de ajuste, test de hipótesis y el análisis de la varianza (ANOVA). Estos métodos están menos asociados a la minería de datos, porque la mayoría de los problemas tienen que ver con comprobar una hipótesis ya conocida en lugar de descubrir una hipótesis nueva.

Otra clasificación común utilizada en relación a los métodos de predicción son los que hacen alusión al tipo de aprendizaje: supervisado y no supervisado. El aprendizaje no supervisado se refiere a modelar la distribución de los casos sin un atributo especificado de antemano.

Los métodos supervisados intentan descubrir la relación entre los atributos de entrada (variables independientes) y un atributo de destino (variable dependiente u objetivo). La relación descubierta se representa en un modelo. Normalmente el modelo describe y explica los fenómenos que están ocultos en el conjunto de datos, y pueden ser utilizado para predecir el valor de la variable objetivo si se conocen los valores de los atributos de entrada.

Es útil hacer la distinción entre dos modelos usados para métodos supervisados: los modelos de clasificación y modelos de regresión. La principal diferencia entre ambos es que los primeros asignan cada caso a de una clase que ha sido predefinida. Estos modelos son los que se utilizan con mayor frecuencia para la predicción de casos “buenos” y “malos” en el caso del default.

Existen muchas alternativas para realizar clasificaciones. Ejemplos típicos incluyen árboles de decisión, regresiones logísticas y redes neuronales.

Árboles de Decisión

Un árbol de decisión es una herramienta de análisis estadístico que se aplica a la minería de datos. Los árboles de decisión son ideales para realizar clasificación y predicción en bases de dato donde la cantidad de registros es de gran envergadura, pues los tiempos computacionales que utiliza son menores que los de otras técnicas.

Un árbol de decisión es un modelo de predicción que permite la división de los datos en conjuntos más pequeños a partir de reglas de decisión. Se construyen diagramas en base a los atributos entregados que sirven para representar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema.

Cada nodo representa una de estas reglas o condiciones sobre el valor de un atributo, las ramas representan los resultados de la evaluación del atributo, y las hojas (o nodos finales) son las clases de la variable dependiente, es decir, 1 o 0 dependiendo de si comete o no default.

Para la construcción de un árbol de decisión se utilizan técnicas de partición en donde se busca aquellos atributos con mejor capacidad de diferenciar las clases. Dentro de las técnicas de partición, las más utilizadas son:

- Gini Index: mide la diversidad de la población descrita con el atributo.
- Entropía: mide la ganancias de información que entrega el atributo
- Chi- cuadrado: mide la diferencia del atributo entre las clases.

La utilización de cada una de estas técnicas dependerá del tipo de atributo que se tiene, si es nominal, categórico o continuo.

La principal ventaja de esta metodología es que no está sujeta a supuestos estadísticos referentes a distribuciones o formas funcionales. Presentan relaciones visuales entre los atributos, las clases de la variable dependiente y el riesgo, lo que permite generar fácilmente una interpretación de los resultados. Los algoritmos más comunes para construir los árboles de decisión son el ID3, C5 y CART (*Classification and Regression)* [18].

Regresión Logística

La regresión logística es un técnica estadística que intenta predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes [16].

Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. Las probabilidades se modelan como una función de los atributos o variables independientes, utilizando una función logística. Puede usarse para correlacionar la probabilidad de una variable cualitativa binaria con una o más variables representadas por un vector x. La probabilidad del suceso se aproximará mediante una función del tipo:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}+1}$$

Donde x es el vector de atributos, β_0 es el punto de intersección y β_1 es el vector de pesos que ajustan el modelo.

Este modelo presenta la ventaja de medir la probabilidad de incumplimiento al mantener la variable explicada siempre dentro de un rango de variación entre cero y uno. La principal ventaja del modelo de regresión logística radica en que no tiene requerimientos sobre la distribución de las variables continuas de entrada. Su principal desventaja radica en que no se puede incluir variables categóricas ni ordinales [16]. Se ha demostrado que la precisión mejora cuando las variables de entrada de tipo continua se encuentran en el intervalo [0,1].

Redes Neuronales

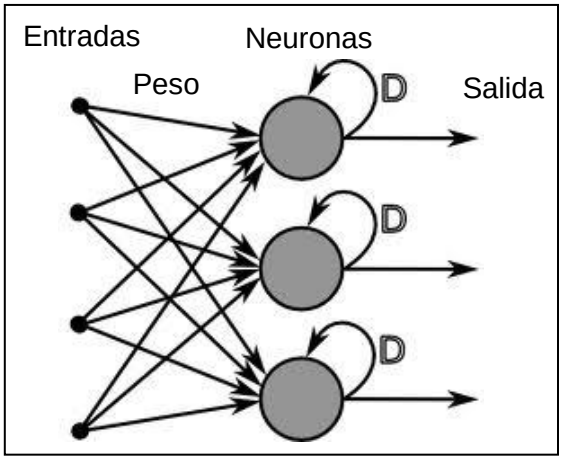
Las redes neuronales artificiales son estructuras computacionales que tratan de imitar al sistema nervioso, de modo que construyen sistemas con cierto grado de inteligencia [1]. Están formadas por una serie de procesadores simples, denominados nodos o neuronas, que se encuentran interconectados entre sí. Como nodos de entrada se considera las variables independientes, como nodo de salida la variable objetivo (en este caso default). Los datos de entrada deben ser numéricos y deben estar normalizados para evitar que algunos atributos tengan más importancia que otros.

Las conexiones entre los nodos juegan un rol fundamental ya que determina el funcionamiento del modelo. Se debe determinar la cantidad de capas intermedias y el número de neuronas a utilizar. Además, a cada conexión entre las capas y los atributos se le asigna un peso como se puede ver en la figura 11. Todo lo anterior influye en la capacidad de predicción del modelo.

Cada neurona en determinada capa toma un conjunto de pesos multiplicado por el atributo correspondiente. Inicialmente los pesos son asignados de manera aleatoria, pero junto con la optimización del algoritmo los pesos son reajustados de acuerdo a un criterio apropiado como puede ser el error cuadrático medio.

La tasa de aprendizaje es un factor crucial para esta técnica, pues una tasa muy pequeña hará que la convergencia sea muy lenta, mientras que si la tasa es muy grande la convergencia probablemente no se dará.

Figura 11: Diagrama de redes neuronales.



Fuente: Elaboración Propia.

Resumiendo, la dificultad de este algoritmo radica en que el proceso interno de aprendizaje funciona como una “caja negra”, donde la comprensión de lo que ocurre dentro requiere de conocimientos especializados, y muchas veces puede producir un sobreajuste. Sus fortalezas se centran en ser un algoritmo útil para la modelación no lineal ya que la existencia de capas puede reproducir dicho comportamiento [1].

3.3.6 COMPARACIÓN DE RESULTADOS

Existe una gran cantidad de técnicas dedicadas a medir la performance de los modelos para poder compararlos. Todas parten de la base de maximizar la cantidad de aciertos (verdaderos malos y verdaderos buenos) y minimizar el error de clasificación, que se manifiesta como clasificar un caso “bueno” siendo “malo” (falsos positivos) o clasificar un caso “malo” siendo “bueno” (falsos negativos) [16]. Las tres técnicas más usadas para medir performance son:

Matriz de confusión

Muestra la cantidad de verdaderos malos, verdaderos buenos, falsos malos y falsos buenos, es decir, los aciertos y desaciertos. A partir de esta matriz se pueden obtener 4 medidas de performance [16]:

- Precisión = Verdaderos negativos y verdaderos positivos / Total de casos
- Tasa de error = Falsos negativos y falsos positivos / Total de casos
    - Tasa de error tipo II = Falsos negativos/ (Falsos negativos + verdaderos negativos)
    - Tasa de error tipo I = Falsos positivos/ (Falsos positivos + verdaderos positivos)
- Sensibilidad= Verdaderos positivos / Total de casos positivos.
- Especificidad= Verdaderos negativos/ Total de casos negativos.

Curva ROC

Método de visualización de performance. Es el gráfico de la frecuencia acumulada de casos “malos” y casos “buenos”. Mientras más diferencia haya entre las

curvas, mejor es el poder predictivo del modelo.

Kolmogorov-Smirnov (KS)

Expresa la máxima distancia entre la curva de frecuencia acumulada de los casos "buenos" y la curva de frecuencia acumulada de los casos "malos". La deficiencia es que sólo mide un punto de corte y no todo el espectro.

Estadístico C o AUC

Medida que representa el promedio de distancias entre las curvas de frecuencia acumulada de casos "buenos" y "malos". Se calcula como el área bajo la curva ROC (o integral). Es la medida más efectiva ya que mide el espectro completo del scoring.

3.3.7 CONSIDERACIONES

La técnica de *credit scoring* puede presentar dificultades no previstas que deben ser consideradas al momento de obtener conclusiones y generar recomendaciones.

- Datos dinámicos: En la mayoría de las bases de datos, los datos son modificados de forma continua. Cuando el valor de los datos almacenados es función del tiempo, el conocimiento inducido variará según el instante en que se obtenga. Por ello es deseable un sistema que funcione de forma continua, que permita tener actualizado el conocimiento extraído.
- Datos incompletos: El manejo de datos incompletos en una base de datos puede deberse a la pérdida de valores de algún atributo o a la ausencia del mismo. En ambos casos la incidencia en el resultado dependerá de que el dato incompleto sea relevante o no para el objetivo del sistema de aprendizaje.
- Ruido en los datos e incertidumbre: Debe tenerse en cuenta no sólo a la presencia de ruido en la base de datos, sino también a la indeterminación existente en muchos aspectos de la realidad y que repercuten en el modelamiento.
- Tamaño de las bases de datos: El tamaño de las bases de datos suele ser muy superior al de los conjuntos de entrenamiento de los sistemas de aprendizaje. Como es inabordable un análisis de todos los datos, deben emplearse técnicas específicas que aceleren el aprendizaje sobre las mismas.
  1. Elegir el algoritmo de aprendizaje con menor complejidad, para que los requerimientos de tiempos computacionales no crezcan más que de forma proporcional al aumentar el tamaño de la base de datos.
  2. Realizar muestreo dentro de la base de datos, para trabajar con una proporción representativa del total.

4. DESCRIPCIÓN DE LOS DATOS

Se considera la construcción de un modelo predictivo para el impuesto al valor agregado (IVA), para lo cual se trabaja con información extraída del *Data Warehouse* del Servicio de Impuestos Internos, de su base de Sistema de Control de Expedientes (SCE). Para el tratamiento de los datos y la construcción de los modelos se utiliza el software *IBM SPSS Modeler*, cuya interfaz se puede ver en el anexo 2.

El historial de datos a utilizar es de seis años, contando desde el año 2007 hasta el año 2012. Esto debido tanto a la disponibilidad de la información, como a lo que propone la teoría para la construcción de modelos de comportamiento [16].

Dado que se trabaja con información confidencial proporcionada por la Subdirección de Fiscalización, las bases de datos se encuentran innominadas. Por solicitud del SII los resultados obtenidos se presentarán de forma agregada.

4.1 DATOS INICIALES

Al inicio del proceso se hace revisión del *Data Warehouse* completo, observándose que la arquitectura de dicho repositorio es extensa y compleja, contando con un total de 330 tablas. Se estudia el contenido de cada tabla, evaluando si corresponden a formularios que los contribuyentes de micro y pequeñas empresas deben declarar. También se buscan las tablas que describan características demográficas de los contribuyentes. Finalmente la decisión es revisada por expertos del SII, llegando a determinarse la extracción de diez tablas que se pueden ver en la Tabla 1.

Tabla 1: Selección de tablas del DW del SII

Tablas	Descripción
DW_TRN_F29_VW	Declaración y pago mensual del IVA, Formulario 29.
DW_TRN_ALERTAS_VW	Alertas sobre negocios o contribuyentes.
DW_TRN_TIMBRAJES_VW	Detalle de documentos legalizados.
DW_TRN_CONTRIBUYENTES_VW	Detalle de personas naturales y jurídicas con o sin inicio de actividades.
DW_HEC_CONT_COMPORTAMIENTO_VW	Marcas de comportamiento asociadas al negocio de un contribuyente.
DW_TRN_F22_VW	Declaración de renta, Formulario 22.
DW_TRN_FISCALIZACIÓN_SELECTIVA_VW	Auditorías y procesos de fiscalización.
DW_TRN_ACTIVIDAD_ECONOMICA_DW	Contribuyentes y sus actividades económicas (ACTECO).
DW_TRN_NEGOCIOS_VW	Detalle del tipo de negocio del contribuyente.
DW_TRN_REPRESENTANTE_VW	Representantes legales de los contribuyentes que son personas jurídicas.

Fuente: Elaboración propia.

En una segunda etapa de extracción, se analizan los campos contenidos en las tablas. La dificultad de esta etapa radica en la comprensión del contenido de cada campo, pues, la terminología utilizada es técnica y específica. Para sortear esta dificultad se trabaja con el diccionario de datos del DW[6] en el que se detalla la información contenida en cada campo. Así mismo, se realizan diversas reuniones con expertos del área de riesgo para comprender a cabalidad la terminología tributaria.

[6] Archivo Excel proporcionado por el área informática del SII.

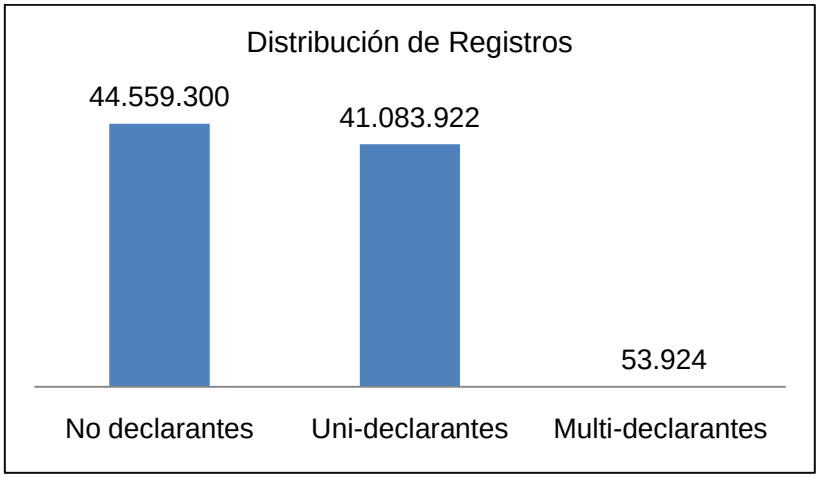
Se cuenta con una cantidad inicial de 15 373 campos de los cuales se seleccionan los 163 contenidos en las tablas previamente mencionadas. Para ver el detalle de estos campos revisar el anexo 3.

Una vez seleccionados los campos y tablas necesarias para la construcción de las variables independientes, se procede a construir la variable objetivo. Se define como "registro" a los pares ordenados de RUT y período tributario. El RUT se considera como el identificador único de cada contribuyente y el período tributario como el mes calendario correspondiente.

Se obtiene una base de datos inicial de 85 697 146 registros, que se clasifican en 3 tipos:

- El contribuyente no presenta declaración para el período tributario.
- El contribuyente presenta solo una declaración válida para el período tributario (atrasada o no).
- El contribuyente presenta más de una declaración válida para el período tributario (atrasadas o no).

Figura 12: Tipos de registro.



Fuente: Elaboración propia.

La frecuencia de registros para cada clasificación se observa en la figura 12, donde se concluye que la cantidad de contribuyentes multi-declarantes es poco relevante en relación a los uni-declarantes. Llama la atención la alta cantidad de contribuyentes que no declaran. Esto puede deberse a que muchos de ellos están actualmente inactivos y no han formalizado su situación.

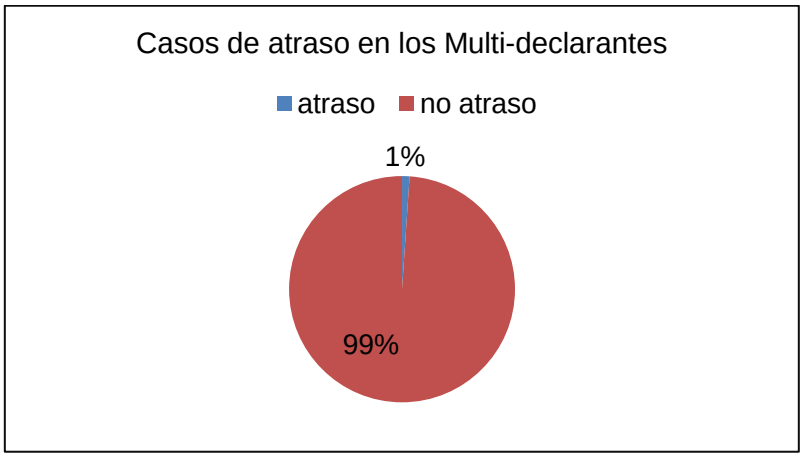
4.2 Definición de Default

Para la construcción de la variable objetivo, se crea primero la variable "Atraso". El atraso se define como cualquiera de las siguientes 3 conductas:

- No presentación de F29.
- Presentación de F29 con posterioridad al mes calendario en que debió declarar.

- Presentación “sin movimiento” del F29 dentro del plazo en el que debe declarar, y luego, fuera de plazo, presenta una rectificatoria con movimiento.

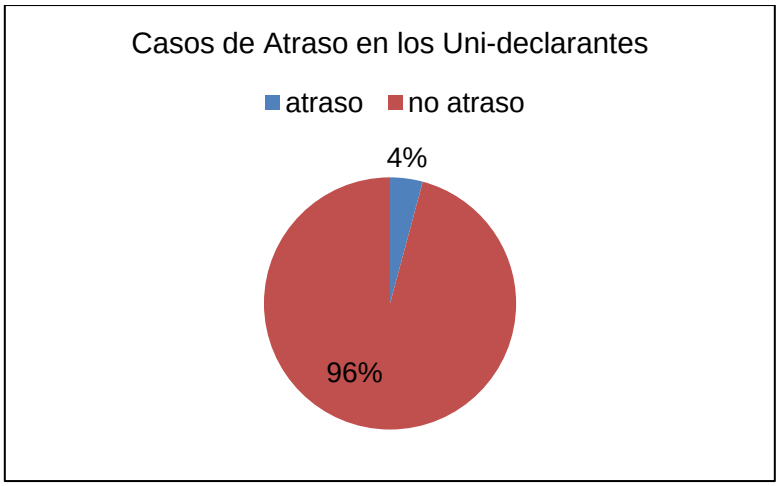
Figura 13: Porcentaje de atrasos para contribuyentes multi-declarantes.



Fuente: Elaboración propia.

En cada tipo de registro se da una proporción de atraso diferente. Para los no declarantes el atraso es de 100%, en el caso de los uni-declarantes el atraso es de un 4% y para los multi-declarantes este se reduce a un 1% como se observa en las figuras 13 y 14.

Figura 14: Porcentaje de atrasos para contribuyentes uni-declarantes.

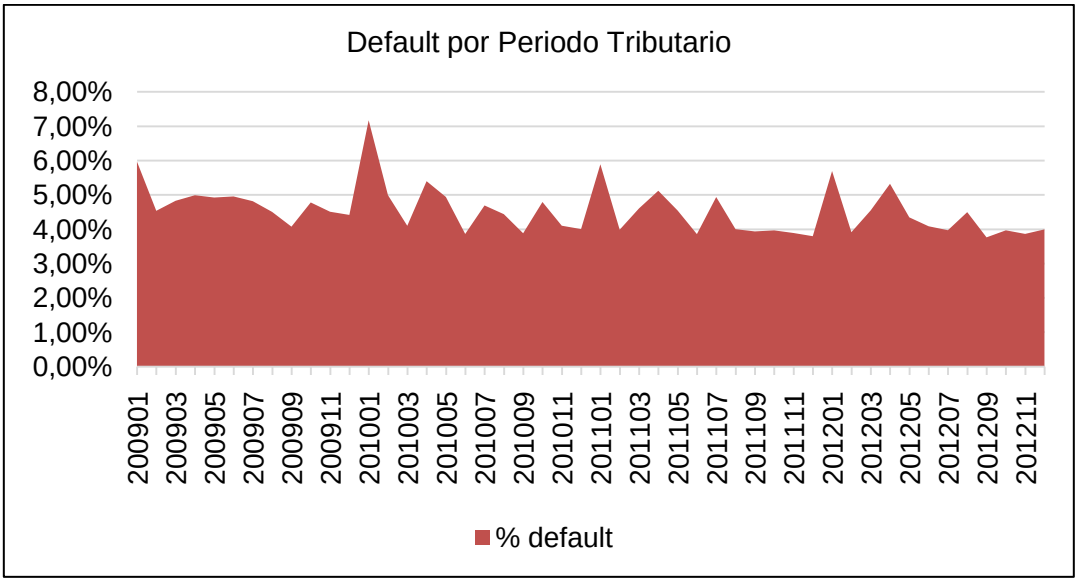


Fuente: Elaboración propia.

Se procede posteriormente a definir el default como el caso en que un contribuyente presente atraso en el período tributario actual, no habiendo presentado atraso el período tributario anterior. Con esta definición la cantidad de registros se reduce a 20 409 665 que se distribuyen de manera uniforme a lo largo de todos los períodos tributarios.

Los casos que presentan default también se distribuyen de manera uniforme en los períodos tributarios. Exceptuando el período de enero 2010 donde aumenta de forma notable como se observa en la figura 15[7]. El default promedio entre los períodos tributarios es de un 4,6%. Llegando a un máximo de 7,1% en enero 2010.

Figura 15: Porcentaje de default en cada período tributario.



Fuente: Elaboración propia.

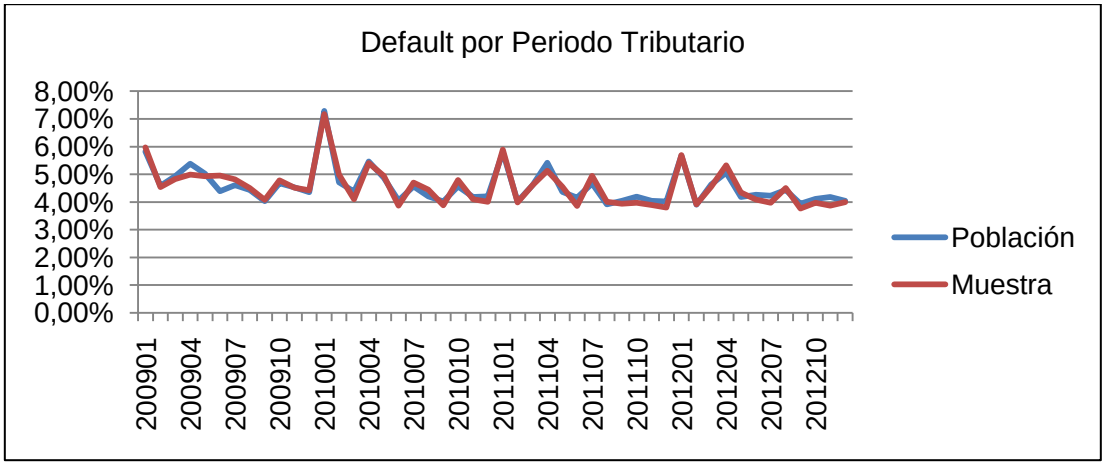
4.3 Elección de la Muestra

Para seleccionar la muestra se utiliza la distribución uniforme de los RUT, tomándose los dos últimos dígitos para segregar a los contribuyentes en grupos. A cada período tributario (72 en total) se le asigna un conjunto de combinaciones de estos dos dígitos, seleccionando los casos de forma aleatoria y generando una muestra de 515.581 registros.

Se comprueba, al 95% de confianza, que dicha muestra se comporta de manera similar a la población. En la figura 16 se constata, mediante un análisis visual, que las distribuciones son prácticamente iguales para cada período tributario. El porcentaje de default de la muestra es de un 4,67% en promedio, alcanzando su máximo en enero 2010 con un 7,34%.

Figura 16: Porcentaje de default en cada período tributario.

[7] El alza del período Enero 2010 se condice con la tendencia de aumento de default en el mes de enero cada año. Sin embargo no hay razones internas ni macroeconómicas para suponer que el alza usual supere en tal nivel el promedio.



Fuente: Elaboración propia.

4.4 DEFINICIÓN DE LAS VARIABLES INDEPENDIENTES

Se diseña un total de 55 variables con la intención de reflejar las características de los contribuyentes de micro y pequeñas empresas. En la Tabla 2 se muestran las ocho categorías en las que se agrupan las variables. Principalmente se construyen atributos asociados a indicadores de liquidez y tamaño de la empresa, pues se cree que los contribuyentes serán más propensos a cometer default cuando no tengan dinero suficiente para sacar de la empresa.

Tabla 2: Categoría de las variables independientes

Categoría	Cantidad de Variables
1. Características del Contribuyente	3
2. Antecedentes Tributarios	11
3. Indicadores de Liquidez	12
4. Sector Económico	3
5. Tamaño y Crecimiento	16
6. Cumplimiento	1
7. Detalle de Incumplimiento	4
8. Comportamiento	5

Fuente: Elaboración propia.

Se cree que las características del contribuyente son relevantes. Sin embargo no es posible construir esas variables debido a que se necesita el historial del contribuyente y en el DW solo se tiene la característica actual.

Los atributos de incumplimiento, cumplimiento y su detalle involucran el historial de pagos, declaraciones, frecuencia con la que reinciden, situaciones de posible comporamiento tributario irregular, entre otros. Mientras que la categoría de

antecedentes tributarios incluye características del proceso de declaración de los formularios.

A cada variable se le asigna un número con el que será referenciada en etapas posteriores. Para ver el detalle revisar el anexo 4.

La creación de estos atributos no es sencilla debido a que se solicita que el módulo de importación de datos y construcción de variables quede automatizado para futuras aplicaciones. Esta etapa consume la mayor cantidad de tiempo de trabajo de la memoria. Se revisa uno a uno los nodos, corroborándose que en cada uno no haya filtros incorrectos.

Posteriormente se calculan las estadísticas descriptivas con la finalidad de comprender la distribución de los datos y el comportamiento de estos atributos para los contribuyentes de micro y pequeñas empresas. Si el tipo era categórico o nominal se calcula la moda y mediana, mientras que para las continuas y discretas se calcula la media y desviación estándar. Los resultados se pueden ver en el anexo 5.

Cabe destacar que dentro de los estadísticos las medias corresponden a lo esperable para cada variable, sin embargo, los máximos y mínimos muchas veces se escapan de los rangos lógicos, por lo que a priori se determina la existencia de *outliers* en la muestra.

5. TRATAMIENTO DE DATOS

5.1 TRATAMIENTO DE DATOS NULOS

Para el tratamiento de datos nulos se opta por la cuarta técnica expuesta en el marco teórico, pues, como se explica, los datos nulos pueden contener información escondida.

Se detalla a continuación el tratamiento que se aplica a cada variable. El detalle de la cantidad de registros nulos que se tienen se encuentra en el anexo 5.

- Para todos los campos que se extraen del *Data Warehouse* y que tienen que ver con montos de dinero o recuento de eventos, los nulos se rellenan con “0”. Esto debido a que la nulidad del campo puede significar que el monto efectivamente es “0” o bien que no existe registro para ese período tributario.

  Es importante destacar que esto no genera un problema, pues, para identificar los casos nulos de los que realmente valen “0”, existe una variable auxiliar que identifica con “1” o “0” si el contribuyente tiene registro en ese período tributario (por ejemplo: para el promedio de ventas del último año, también existe la variable auxiliar de cantidad de veces que presenta el formulario 29 el último año).

  Para atributos correspondientes a ratios o porcentajes, tanto el numerador como denominador puede ser “0”. En tal caso se propone dejar el caso con valor “0” y agregar a la base de datos dos nuevas variables: el numerador y el denominador de manera aislada.

- Para variables *dummy*, se considera como “1”que exista el registro, y “0” que el registro sea nulo (pudiendo significar *missing value* o bien la ausencia de la característica que la variable busca plasmar).

- Para variables que tienen que ver con fechas o recuentos de períodos los valores nulos se reemplazan con un valor no presente entre los casos. La idea es categorizar posteriormente la variable y asignar los valores nulos como una nueva categoría.

5.2 TRATAMIENTO DE OUTLIERS

Para la eliminación de *outliers* primero se generan estadísticas descriptivas de cada variable y se grafican los histogramas para facilitar un análisis visual. Se determina si la variable sigue una distribución normal para poder aplicar un test de desviación estándar. Se observa una baja ocurrencia de normalidad debido a que la cantidad de datos nulos distorsionan los histogramas.

Considerando lo anterior, se aplica un test de datos anómalos. Se considera como datos atípicos aquellos que sean mayores a tres desviaciones estándar sobre la media, y extremos aquellos que sean mayores a cinco desviaciones estándar sobre la media. La suma total de casos atípicos y extremos se puede ver en la tabla 3, y corresponden a un 18% y 6% del total respectivamente. Para el detalle por variable revisar el anexo 7.

Tabla 3: Resumen de tratamiento de *outliers*.

	Atípicos	Extremos
Casos totales	99.373	37.304

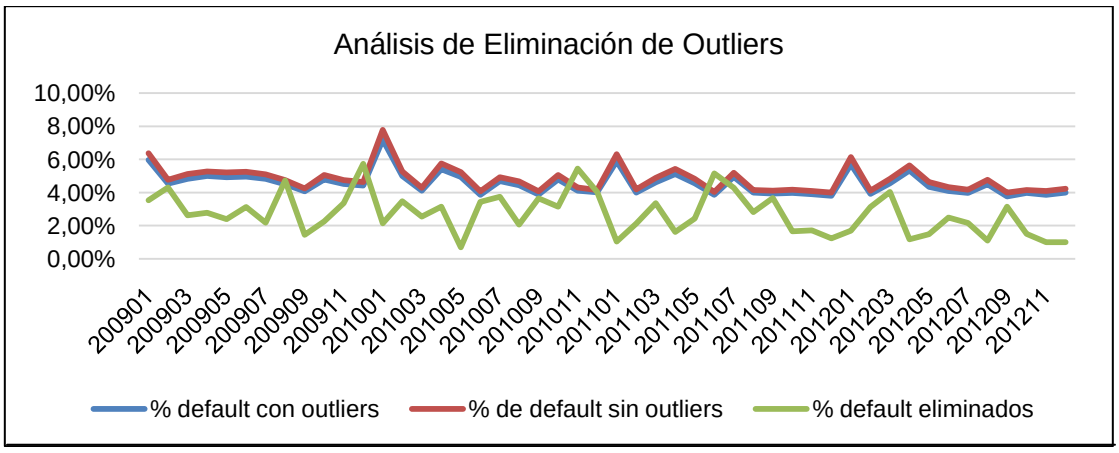
Fuente: Elaboración propia.

Como no se cumplen los supuestos de distribución normal, no es posible aplicar con seguridad este criterio. Se decide eliminar solo los registros de las variables en que la cantidad de datos extremos fuera despreciable.

Posteriormente, se comienza una segunda etapa de detección de *outliers*. En esta se busca eliminar (a juicio experto) aquellos datos en que el mínimo y máximo de la variable estén fuera de rangos aceptables. Se realiza el procedimiento sobre 29 variables, y se elimina una cantidad igual a 7 604 registros que equivale a un 1,4% de la muestra.

Se analiza el comportamiento de los registros eliminados, llegándose a la conclusión de que efectivamente corresponden a *outliers* dada la diferencia que presentan con la muestra. El gráfico del comportamiento de default en cada grupo se puede ver en la figura 17.

Figura 17: Análisis de default para la eliminación de *outliers.*



Fuente: Elaboración propia.

5.3 SELECCIÓN DE VARIABLES

Para tener una primera idea de las variables que resultan relevantes para el modelamiento de default tributario, se aplica la herramienta "selección de características" del software *IBM SPSS Modeler*, que permite hacer un ranking de los atributos en función de una escala de porcentajes definida como 1-P, donde P representa la probabilidad de obtener un resultado igual o más extremo que un caso elegido al azar[8].

Siete variables fueron eliminadas de la base de datos de acuerdo al criterio explicado, pues la importancia del predictor se encuentra por debajo del 80%. Específicamente se extrae la variable 66, 29, 29.b, 7, 47, 25 y 26. El detalle se puede ver en el anexo 8.

En el siguiente paso se analiza la correlación con la finalidad de eliminar aquellas que entreguen información redundante. Son 15 variables las que presentan un índice mayor a 0,7. Para ver el detalle ir a anexo 9.

Tabla 4: Decisión de eliminación de variables de acuerdo a la correlación.

Componentes	Variables de la Componente	Variables Eliminadas
Tamaño	13, 42, 52, 54, 29.a y 37.a	42 y 52
Nivel de Renta	70, 31 y 17	17 y 70
Crecimiento	20 y 21	20
Actividad Económica	7, 40 y 55	40
Comportamiento	23 y 69	ninguna

Fuente: Elaboración propia.

Se decide utilizar el método de análisis de componentes principales (ACP) para tomar la decisión final. El test indica que un 60% de la varianza se explica por las primeras 18 componentes (ver detalles en el anexo 10). No se realiza esta transformación ya que la varianza total explicada por cada componente es marginal.

[8] P calculado con el estadístico de Fischer.

Sin embargo, se observa que las primeras cinco componentes contienen las variables de mayor correlación. Dentro de estos cinco subconjuntos, se eligen los atributos con mayor proyección en cada componente, los atributos restantes son seleccionados a juicio experto de acuerdo a la experiencia del trato con contribuyentes. El resultado final de esta etapa se puede ver en la tabla 4.

5.4 TRANSFORMACIÓN DE VARIABLES

La última etapa consiste en transformar los datos de acuerdo a las necesidades de cada modelo. Se utiliza la herramienta de transformación de *IBM SPSS Modeler* que permite visualizar los histogramas de las variables al aplicarle distintas funciones de re-escalamiento (logaritmo en base 10, logaritmo natural, función inversa, función cuadrática). Se decide aplicar la función de logaritmo natural a 11 variables, que en su mayoría se relacionan con montos de dinero, y se aplicó la transformación inversa a una variable.

El siguiente paso es la categorización de las variables que por el tratamiento que se le da a los datos nulos deben ser transformadas. El criterio para la formación de categorías es maximizar la diferencia de comportamiento (porcentaje de default) entre los grupos. En total se aplicó este tratamiento a cinco variables, y los resultados se pueden ver en el anexo 11.

Los modelos de árboles de regresión pueden utilizarse con variables categóricas, *dummys* y continuas, por lo que aquí culmina la preparación de la base de datos para el modelo 1.

En el caso del modelo de redes neuronales y regresión logística se requiere además transformar las variables categóricas en *dummys* y normalizar las variables continuas a las que se le aplica la transformación Min-Max. En el anexo 10 se puede ver el detalle de las transformaciones aplicadas a cada variable.

No se testea la interacción de variables dado que para el negocio no es relevante. Tampoco se utiliza el método de WOE debido a que al transformarlos se pierde la interpretabilidad de las variables, tema sumamente relevante para la futura obtención de reglas de decisión.

6. APLICACIÓN DE MODELOS

Para la construcción de los modelos se realiza una partición de la muestra en tres subconjuntos: Entrenamiento, Comprobación y Validación en una razón de 50:20:30. En la tabla 5 se puede ver la frecuencia de clases y la cantidad de registros para cada subconjunto de la partición. El porcentaje de default es de 5% e igual para todos los subconjuntos en la partición.

Con la finalidad de que el modelo sea capaz de predecir correctamente ambos casos, se equilibra la muestra. Se crea la variable “peso” como la cantidad de registros que presentan default dividido por la cantidad de registros que no presentan default. Este número se utiliza como multiplicador de los casos sin default en la muestra de entrenamiento, logrando el objetivo expuesto.

Tabla 5: Partición de la muestra.

	Entrenamiento	Comprobación	Validación	Total
No Default	253.085	101.053	152.153	506.291
Default	12.436	4.917	7.523	24.876
Total	265.521	105.970	159.676	531.167

Fuente: Elaboración propia.

Para la regresión logística se aplica un método similar. La variable peso se utiliza de forma multiplicativa, pero esta vez dentro de la función de verosimilitud. La maximización de la función contemplará entonces una disminución del aporte de los casos buenos de acuerdo a este número.

6.1 TÉCNICAS

La elección de las técnicas se realiza en base a los siguientes criterios:

- Interpretabilidad: Facilidad con que el modelo explica el fenómeno y genera reglas de decisión que permiten evaluar la incidencia de cada variable.
- Simplicidad: Complejidad del algoritmo de aprendizaje. Se tiende a buscar el de menor complejidad para que los tiempos computacionales no crezcan más que de forma proporcional al aumentar el tamaño de la base de datos.
- Precisión: Capacidad de clasificar correctamente tanto los casos buenos como los malos. Privilegiándose aquellos modelos que tengan una sensibilidad y especificidad de similar porcentaje, pues, es necesario que identifique correctamente ambos grupos. La tasa de error también será estudiada, dándole especial atención al error de tipo II que indica la clasificación de falsos negativos. Se trabajará con cinco indicadores que son: AUC, KS, precisión general, sensibilidad, especificidad y tasa de error tipo II.

Se debe destacar que la especificidad indica la capacidad de identificar correctamente los casos malos, mientras que la sensibilidad indica la capacidad de predecir correctamente los casos buenos. Ambos indicadores se relacionan de manera inversa, por lo que al aumentar uno disminuye el otro.

Al ser la sensibilidad el indicador de la clase con mayor presencia en la base de datos, influye en mayor proporción para la predicción global. Por consecuencia, se considera que los modelos tienen una buena precisión cuando la predicción global es alta, y los dos indicadores nombrados tienen un porcentaje parejo entre sí. A la vez se debe intentar tener la tasa de error tipo II más baja posible.

Se desea obtener un modelo base que sea el de mejor precisión y un modelo alternativo, que si bien no es el de mejor precisión, sea el más práctico en términos de interpretación y tiempos computacionales.

En la tabla 6 se ve la puntuación que se le da a cada técnica según los criterios nombrados [9]. Se observa que la regresión logística tiene un buen índice de interpretabilidad, precisión y simplicidad lo que la lleva a ser la más utilizada por la

[9] Esta puntuación se da de acuerdo a las características que tiene cada técnica, y que se nombraron en el capítulo 3 de esta memoria.

industria bancaria hoy en día. Sin embargo, las redes neuronales superan a la regresión logística en precisión, logrando una mejora en la clasificación a costa de un empeoramiento en la interpretabilidad y simplicidad. En este último aspecto destacan los árboles de decisión, que generan modelos auto explicativos y que fácilmente se pueden transformar en un set de reglas de decisión.

Tabla 6: Comparación de diferentes modelos de riesgo de crédito.

Método	Precisión	Interpretabilidad	Simplicidad
Regresión Logística	●●	●●●	●●●
Redes Neuronales	●●●	●	●
Árboles de decisión	●●	●●●	●●

Fuente: Elaboración propia.

Para aumentar la performance de cada modelo y obtener el mejor ajuste posible se pueden modificar parámetros en las opciones de configuración del software que permitirá que los algoritmos varíen. A continuación se detallan dichos parámetros para cada una de las técnicas.

6.1.1 Árboles de Decisión

Se hacen pruebas con los algoritmos CHAID, QUEST, C&R y C.5 para evaluar su rendimiento. Para ello se utilizan variables sin transformar y la configuración de parámetros que viene por defecto en el software *IBM SPSS Modeler*. Como se observa en la tabla 7, la tasa de error tipo II alcanza su nivel más bajo con el algoritmo C.5. Sin embargo, pronostica prácticamente todos los casos dentro de una sola clase, lo que indica una capacidad predictiva casi nula. El segundo mejor nivel de error lo tiene el algoritmo QUEST, pero su especificidad empeora notoriamente.

Se decide utilizar el algoritmo CHAID pues es el que obtiene mejores resultados. Los tres indicadores de precisión son similares y alcanzan un porcentaje alto. La precisión global es superada por el QUEST y C.5, pero ambos árboles tienen una especificidad demasiado baja, lo que hace que se rechacen.

Tabla 7: Comparación de algoritmos de árboles de decisión.

Algoritmo	Precisión Global	Especificidad	Sensibilidad	Error II	Elección
QUEST	73,30%	59,36%	74,42%	89,82%	
C&R	63,40%	75,05%	62,77%	91,09%	
C.5	95,36%	1,92%	99,97%	35,45%	
CHAID	68,28%	67,80%	68,30%	90,41%	X

Fuente: Elaboración propia.

El software *IBM SPSS Modeler* permite la modificación de los siguientes parámetros relevantes para la técnica de CHAID:

- Algoritmo de crecimiento del árbol.
- Máxima profundidad del árbol.

- Criterio de parada: porcentaje mínimo de registros en rama principal y en rama secundaria.
- Nivel de significancia para la división.
- Nivel de significancia para la combinación.
- Chi cuadrado para objetivos categóricos.

6.1.2 Redes Neuronales

Para esta técnica se trabaja con el software *IBM SPSS Modeler* que permite la modificación de los siguientes parámetros:

- Modelo de red neuronal.
- Número de capas ocultas.
- Número de neuronas en cada capa oculta.
- Tiempo máximo de entrenamiento de cada componente.

6.1.3 Regresión Logística

Se utiliza el software *SPSS Statistics*, que permite hacer modificación en los siguientes parámetros para mejorar el ajuste:

- Método de selección de variables.
- Máximo de iteraciones para la convergencia.
- Corte de significancia para la entrada de variables.
- Corte de significancia para la salida de variables.

6.2 Prueba Ex-Ante

Utilizando las 55 variables independientes en su estado natural, se procede a realizar una prueba para cada uno de los métodos elegidos. En esta etapa no se modifica los parámetros de los algoritmos y tampoco se aplican los procesos de transformación a las variables.

La idea es obtener una primera aproximación a los resultados que se tendrán con cada método, para evaluar la capacidad predictiva “bruta” de la información contenida en las variables. Se tomará esta información como cota inferior para mejorar la predicción.

6.2.1 Árboles de Decisión

El modelo entrega un árbol de decisión de profundidad cinco que se demora un total de 13 segundos en ser construido y cuya precisión global es de un 68,28%.

Como se observa en la tabla 8, tiene una capacidad similar para predecir tanto los casos buenos como los casos malos, pues la especificidad y sensibilidad muestran un porcentaje que se mantiene en el mismo rango. El error tipo II es alto siendo de 90,41%.

6.2.2 Redes Neuronales

El modelo entrega una red neuronal de una capa oculta y siete neuronas, que se demora un tiempo total de 2 minutos y 23 segundos en ser procesada. La predictividad es menor que la de los árboles de decisión, llegando a tener una precisión global de un 67,46% para el punto de corte de 0,48.

Como se observa en la tabla 8 existe una diferencia notable entre la especificidad y la sensibilidad, lo que indica que el modelo está siendo capaz de predecir bastante bien los casos malos, no así los casos buenos. Esto se refleja también en el error tipo II que es de 91,27%.

6.2.3 Regresión Logística

El tiempo de procesamiento del modelo es de 40 segundos. Como se observa en la tabla 8 la precisión global es superior a todas las otras técnicas, siendo de 70,54% para un punto de corte de 0,45. El comportamiento de los indicadores de especificidad y sensibilidad es totalmente opuesto a lo obtenido con la red neuronal. La sensibilidad es cercana al 71%, indicando la buena capacidad para predecir los casos buenos, pero la especificidad es de un 63,93% bastante más baja que el de cualquiera de las otras técnicas. Como es de esperar, el KS y AUC también son los más bajos de esta prueba, sin dejar de ser buenos.

Tabla 8: Comparación de resultados prueba ex-ante.

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II
Árbol de Decisión	68,28%	67,80%	68,30%	-	-	90,41%
Red Neuronal	67,46%	73,26%	61,64%	0,346	0,734	91,27%
Regresión Logística	70,54%	63,93%	70,87%	0,348	0,737	91,36%

Fuente: Elaboración propia.

En síntesis, el modelo con mejor desempeño al utilizar las variables en bruto es el árbol de decisión. Como se observa en la tabla 8, la precisión global es alta y la capacidad de predecir tanto los buenos como los malos casos es similar, siendo la técnica que muestra la tasa de error tipo II más baja. Si bien la regresión logística logra una precisión global mayor, su especificidad es bastante baja, lo que indica una deficiencia para identificar los casos malos.

El resultado se condice con el hecho de que las variables no han sido transformadas de acuerdo a los requerimientos de los otros dos modelos. Como consecuencia existen variables que por su rango de magnitud puedan estar opacando a otras, cuya capacidad predictiva sea más importante para detectar el default, lo que empeora el desempeño de los modelos de redes neuronales y regresión logística.

A pesar de que unos métodos se comportan mejor que otros, la capacidad explicativa bruta de las variables independientes construidas es buena. Se observa que las variables son capaces de diferenciar de manera global los malos de los buenos, pues, en los dos últimos casos el AUC es cercano a 0,73 y la capacidad predictiva bastante mayor a un 50%. Sin embargo, en la predicción más detallista se observa deficiencias dado que el error tipo II es muy alto.

6.3 PRUEBA GENERAL

En esta etapa se utilizan las variables transformadas, a las que previamente se le aplicaron funciones de escalamiento, categorización y normalización según los requerimientos de las redes neuronales y regresión logística. Esto lleva a un total de 75 variables de entrada debido a la categorización de cinco de ellas.

Se desea observar el desempeño de los modelos con la finalidad de determinar la técnica que mejor funcione (sin considerar la incidencia de la selección de atributos). Para esto se prueba cada una de las técnicas con la configuración de parámetros que trae por defecto el software. A continuación se detalla el proceso.

6.3.1 ÁRBOLES DE DECISIÓN

Los parámetros que inciden en el algoritmo se definen de la siguiente manera:

- Algoritmo de crecimiento del árbol: CHAID
- Máxima profundidad del árbol: 5
- Criterio de parada (porcentaje mínimo de registros): rama principal= 2%
  rama secundaria= 1%
- Nivel de significancia para la división: 0,05
- Nivel de significancia para la combinación: 0,05
- Chi cuadrado para objetivos categóricos: Pearson

El modelo que surge de esta etapa entrega un árbol de profundidad cinco que se demora un total de 13 segundos y tiene una precisión global de 68,39%. Como se aprecia en la tabla 9 los tres indicadores de precisión se mantienen en un rango similar, por lo que el modelo es adecuado para predecir tanto los casos buenos como malos.

Con respecto a la prueba ex-ante se observa una leve mejora en la precisión global, especificidad, sensibilidad y el error tipo II, lo que indica que la transformación de las variables mejora la capacidad predictiva.

6.3.2 REDES NEURONALES

Se define la configuración de los parámetros para la prueba general de redes neuronales de la siguiente manera:

- Modelo de red neuronal: Perceptrón Multicapa.
- Número de neuronas en cada capa oculta: cálculo automático.
- Número de capas ocultas: cálculo automático.
- Tiempo máximo de entrenamiento por cada componente: 15 minutos.

El software demora un total de 2 minutos 10 segundos en ejecutar el modelo. El resultado arroja una red neuronal de una capa oculta que contiene cinco neuronas, y cuya precisión global es de 71,89% para el punto de corte 0,45, lo que indica una mejora con respecto a la prueba ex-ante.

Si se observa la tabla 9, la sensibilidad y especificidad se invierten con respecto a los resultados de la prueba ex-ante, mejorando sustancialmente la primera y

empeorando la segunda. Sin embargo, se obtiene una disminución de la brecha que los separa, pues la diferencia previa era de más de diez puntos porcentuales y ahora se reduce a tan solo nueve.

Se sigue encontrando un peor resultado que para la técnica de árboles de decisión, pues la diferencia de niveles entre los indicadores es muy alta, a pesar de que el error tipo II es un poco menor. Es necesario ajustar los parámetros de configuración para modificar la propensión a predecir pobremente los casos malos.

6.3.3 Regresión Logística

Se define la configuración de los parámetros para la prueba general de regresión logística de la siguiente manera:

- Método de selección de variables: Enter
- Máximo de iteraciones: 10
- Significancia para la entrada: 0,05
- Significancia para la salida: 0,1

El modelo de regresión logística demora un tiempo de 38 segundos en ejecutarse. La precisión global mejora levemente llegando a ser un 70,59%. Si bien no es la mejor de las tres, su AUC y KS indican que es el modelo de mejor desempeño para cualquiera de los puntos de corte.

Como se ve en la tabla 9 la especificidad y sensibilidad son disímiles pero en un rango menor que el de la red neuronal, por lo que se considera un buen modelo, que podría mejorar al modificar los parámetros de ajuste. El error tipo II es el segundo mejor.

Tabla 9: Comparación de resultados modelo general.

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II
Árbol de Decisión	68,39%	67,88%	68,41%	-	-	90,18%
Red Neuronal	71,89%	63,32%	72,31%	0,368	0,741	89,88%
Regresión Logística	70,59%	65,47%	70,84%	0,368	0,746	90,06%

Fuente: Elaboración propia.

Como es esperable, existe una mejora en los modelos de redes neuronales y regresión logística al transformar las variables. Estas técnicas aumentan la precisión general y disminuyen la brecha entre la especificidad y sensibilidad. Así mismo, aumentan su KS, AUC y mejoran su tasa de error, indicando una ganancia global en el desempeño.

Sigue liderando la técnica de árboles de decisión dado que los índices de precisión global, especificidad y sensibilidad son altos y parejos, y su tasa de error no difiere sustancialmente de las otras dos. Cabe destacar que las redes neuronales y regresión logística tienden a predecir mejor los casos buenos, debido a la incidencia que tiene la sensibilidad en la precisión global. Esto puede corregirse modificando los parámetros de ajuste en la siguiente etapa.

6.4 Iteraciones

En la etapa de iteración se intenta mejorar la performance de cada técnica cambiando los parámetros de los algoritmos. Se pretende personalizar los resultados de cada modelo para ajustarlos de mejor forma a las variables que se tienen.

Se toma como caso base los modelos generales y se modifican los parámetros de manera aislada evaluando si la modificación produce mejoras en la precisión. Se adopta el mejor valor testeado para el parámetro.

6.4.1 ÁRBOLES DE DECISIÓN

En una etapa inicial se prueba de manera aislada la inclusión de una matriz de costos para los errores de clasificación. El indicador de error tipo II mejora llegando a un mínimo de 82% a costa de un empeoramiento de los índices de precisión global, especificidad y sensibilidad. No se incluye este análisis en la iteración del modelo, pues se considera que afecta de forma negativa el desempeño. Los detalles se pueden ver en el anexo 13.

Posteriormente se comienza la etapa de iteración donde se modifican cinco parámetros: algoritmo de crecimiento, máxima profundidad del árbol, técnica para el chi-cuadrado, nivel de significancia para la división y nivel de significancia para la agrupación. En la tabla 10 se detalla el rango de modificación de los distintos parámetros y en el anexo 14 se pueden ver los resultados para cada iteración.

Tabla 10 : Rango de los parámetros de modificación.

Parámetro	Rango
Algoritmo de crecimiento	Chaid o Exhaustive Chaid
Máxima profundidad	5 a 9
Test de chi-cuadrado	Pearson o Verosimilitud
Significancia para la división	0,02 a 0,1
Significancia para la combinación	0,02 a 0,2
Proporción entre costos de error	1 a 3

Fuente: Elaboración propia.

El cambio en el algoritmo de crecimiento muestra una mejora en la sensibilidad que repercute en la precisión global pero empeora la diferenciación entre clases, pues, aumenta la brecha entre la especificidad y sensibilidad. Se decide utilizar el algoritmo CHAID.

Se prueban distintas profundidades máximas para el árbol, obteniéndose que con el aumento de ramas mejora la precisión general pero no la capacidad de diferenciar entre clases. Cada vez se distancian más la sensibilidad y especificidad, por lo que se decide optar por las cinco ramas de profundidad máxima.

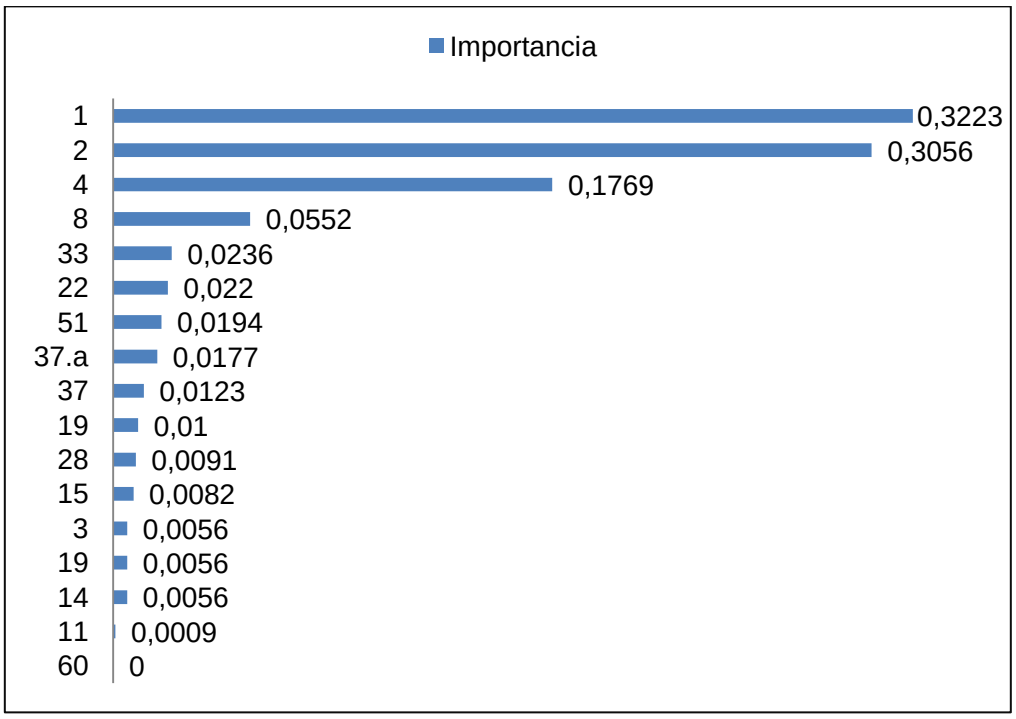
Para crear el modelo, el CHAID aplica un test de independencia chi-cuadrado sobre las tablas de contingencia, intentando descubrir la influencia que tiene cada atributo en la variable objetivo mediante la comparación celda a celda de la frecuencia estimada y observada.

El software *IBM SPSS Modeler* entrega la posibilidad de usar dos tipos de test: Pearson o Verosimilitud. La diferencia radica en que el segundo test es más robusto que el primero, teniendo un mejor resultado en muestras pequeñas. Sin embargo, al modificar este parámetro no se observan cambios en la capacidad predictiva.

Tampoco se observan mejorar al alterar los valores de significancia de entrada y salida, por lo que tras las iteraciones se decide no modificar los parámetros, utilizándose la configuración que trae por defecto el software *IBM SPSS Modeler*.

De este modelo final se obtiene un ranking de importancia de las variables de acuerdo a su capacidad predictiva[10]. En la figura 18 se puede ver un gráfico que muestra la importancia relativa de cada una, indicando la capacidad del atributo para realizar un pronóstico.

Figura 18: Importancia relativa de variables para árboles de decisión.



Fuente: Elaboración propia.

Se obtiene que la variable 1, 2 y 4 son las más relevantes, teniendo una importancia casi diez veces mayor que las otras. La variable 1 da cuenta de la cantidad de atrasos en la declaración del F29 en el último año, la variable 2 explica el número de anotaciones 'no declarante' y la variable 4 el número de boletas emitidas. Después de la variable 11 todas las otras muestran una menor importancia (cercana a 0), es decir, una capacidad casi nula para pronosticar. Por esta razón no aparecen en la figura.

6.4.2 Redes Neuronales

[10] La suma de importancias de las variables debe sumar uno. La importancia de cada variable es un indicador que deriva del p-valor obtenido con la prueba chi-cuadrado. A menor p-valor, mayor es la importancia relativa de la variable.

Para esta técnica se pueden modificar cuatro parámetros de la configuración: modelo, cantidad de nodos en la capa oculta, tiempo máximo por cada iteración y porcentaje del conjunto de prevención. En la tabla 11 se muestran los rangos de modificación de los parámetros, y en el anexo 15 se puede ver en detalle el resultado de las iteraciones.

Tabla 11 : Rangos de los parámetros de modificación.

Parámetro	Rango
Modelo	PMC o FBR
Tiempo de iteración	10 a 20 minutos
Conjunto de prevención de sobreajuste	25 a 35%
Cantidad de neuronas	8 a 30

Fuente: Elaboración propia.

Las mejoras más relevantes se obtienen al modificar el modelo utilizado. Se opta por perceptrón multicapa pues muestra un desempeño mucho mejor en los cinco indicadores.

Si se modifican los tiempos de iteración aumenta la precisión global, pero todos los otros índices empeoran. Al cambiar el conjunto de sobreajuste tampoco se observan mejoras. Se opta por dejar la cantidad de nodos de la capa oculta en modo automático, pues, de acuerdo a la teoría, el rango de prueba dependerá de la cantidad de variables a considerar[11] y por lo mismo se realizará luego de reducir los atributos.

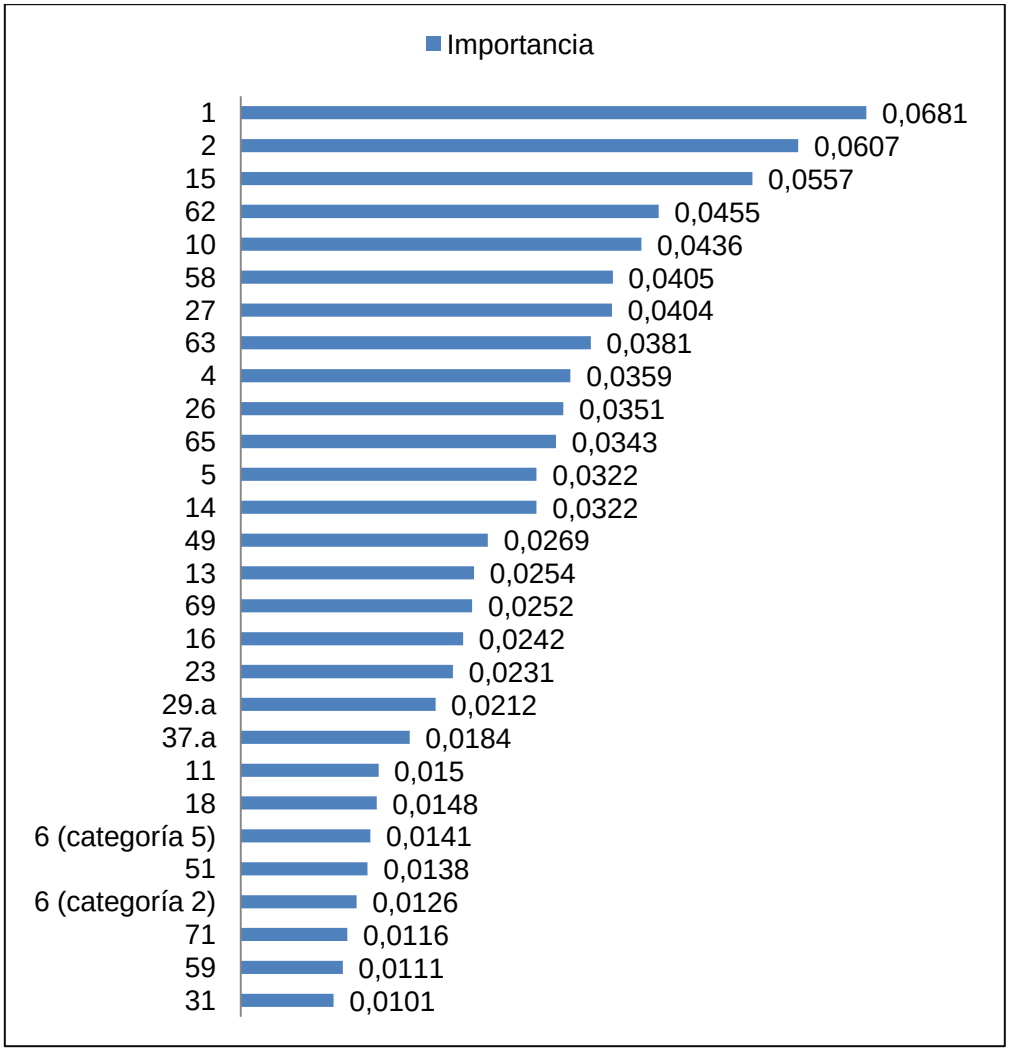
La configuración final de los parámetros contempla la modificación del modelo utilizado al de perceptrón multicapa. Su estructura se mantiene un una capa oculta, pero aumenta la cantidad máxima de neuronas a 11. El tiempo de iteración es de 15 minutos y el conjunto de prevención de sobreajuste no se altera.

Se comienza con un AUC de 0,741 que aumenta a 0,748. El indicador KS también crece desde 0,368 a 0,378 indicando una mejora en la capacidad de diferenciar los casos buenos de los malos. La precisión global pasa de 71,89% a 68,76%, pues disminuye la sensibilidad del modelo pero aumenta la especificidad, y el error tipo II disminuye a 89,87%.

En la figura 19 se muestra el gráfico de importancia de las variables para el modelo final obtenido de la etapa de iteración, exhibiendo solo las 28 variables que tienen una importancia relativa mayor al 1%. Se observa una distribución más equitativa que en el modelo de árboles de decisión. En efecto, las variables 1 y 2 siguen tendiendo la mejor capacidad de predicción, pero esta vez el nivel de la importancia relativa es cercano al 6%, lo que le asigna un mayor peso a otros atributos como el 15, 62 y 10.

Figura 19: Importancia relativa de las variables para el modelo de redes neuronales.

[11] Rango debe fluctuar entre v/2 y 2v, siendo v la cantidad de variables ingresadas.



Fuente: Elaboración propia.

6.4.3 Regresión Logística

Para la regresión logística los parámetros de configuración que se modifican son: tipo de algoritmo *Stepwise*, significancia para el ingreso de variables, significancia para la eliminación de variables, corte de probabilidad y cantidad de iteraciones máxima. En la tabla 12 se muestra el rango de modificación de los parámetros en cada iteración.

La primera prueba se realiza con la alternativa *Enter* que es la configuración que viene por defecto con el software *SPSS Statistics.* Esta opción no contempla la entrada ni salida de variables. Como el objetivo de esta etapa es seleccionar el modelo final, esta opción no facilita la decisión y se descarta.

Tabla 12 : Rango de modificación de los parámetros.

Parámetro	Rango
Técnica stepwise	*Forward* o *Backward*
Estadístico para stepwise	Wald, RL o Condicional
Significancia de entrada	0,1 a 0,3

Significancia de salida	0,1 a 0,3
N° máximo de iteraciones	2 a 20

Fuente: Elaboración propia.

Las dos alternativas restantes corresponden a la técnica *Backward* y *Forward*. La primera arroja modelos de más de 30 variables aún cuando se modifiquen los parámetros de entrada y salida[12]. Se elige la segunda técnica que permite obtener modelos de entre 1 y 25 variables, siendo fácilmente interpretable y manejable.

Los métodos de *Stepwise* permiten trabajar con tres estadísticos: Wald, Condicional y RL. Luego de hacer las 6 combinaciones posibles se determina que no influye en la precisión del modelo el estadístico que se elija. Como resultado de lo anterior, las iteraciones se realizan con la técnica *Forward* Wald, debido a que es el que muestra menores tiempos de ejecución. El detalle se puede ver en la tabla Forward del anexo 16.

La etapa de iteración comienza con un modelo cuyo AUC de 0,746 y cuya precisión global es de 70,59%. Al final de las iteraciones el primer estadístico desciende a 0,745, la precisión global baja a 68,42% y el error disminuye a 90,04%

Lo anterior se debe a un aumento en la especificidad del modelo y por ende una disminución de la sensibilidad. Sin embargo la proporción en que aumenta la capacidad de predecir los malos, es mayor que la disminución en la capacidad de predecir los buenos casos. Esto se refleja en el aumento del KS desde un 0,368 a un 0,369.

No se aprecian mejoras al cambiar la significancia de salida ni de entrada por lo que se asume que las variables elegidas son siempre relevantes al 95% de confianza o más. Tampoco se ven mejoras al modificar la cantidad máxima de iteraciones, por lo que se mantiene en 20 iteraciones. Los detalles de las iteraciones se pueden revisar en la tabla resumen del anexo 16.

La configuración final queda como un modelo de regresión logística que utiliza *Backward* de Wald, realiza un máximo de 20 iteraciones por etapa y tiene una significancia de entrada de 0,05 y de salida de variables de 0,1.

En este método no se evalúa la importancia individual de las variables, porque el algoritmo *Stepwise* permite elegir el conjunto de variables que genera un mejor desempeño del modelo.

Tabla 13: Resumen de los resultados de la etapa iteración.

		Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II
Árbol de Decisión	Inicial	68,39%	67,88%	68,41%	-	-	90,18%
	Final	68,39%	67,88%	68,41%	-	-	90,18%
Red Neuronal	Inicial	71,89%	63,32%	72,31%	0,368	0,741	89,88%
	Final	68,76%	68,33%	68,78%	0,378	0,748	89,87%
Regresión Logística	Inicial	70,59%	65,47%	70,84%	0,368	0,746	90,06%

[12] Para ver detalles ir a la tabla *Backward* del al anexo 16.

	Final	68,43%	68,01%	68,45%	0,369	0,745	90,04%

Fuente: Elaboración propia

De acuerdo a los resultados expuestos en la tabla 13, tras las iteraciones se observan mejoras en el modelo de redes neuronales y de regresión logística. En particular ambos mejoran su KS, AUC y tasa de error de tipo II, debido a que aumenta la capacidad de diferenciar los casos buenos y malos. Esto se refleja en una disminución de la precisión global que pasa de ser sobre un 70% a un 68% en ambos casos.

Los resultados en esta etapa no son comparables debido a que las iteraciones para la regresión logística implicaron la eliminación de una gran cantidad de variables. Sin embargo, se puede ver que en concordancia con la teoría, las redes neuronales tienen el mejor desempeño, seguidos por el modelo de regresión logística y terminando con el algoritmo más simple que es el de árboles de decisión.

6.5 MODELO FINAL

Los modelos finales de cada técnica se construyen con la configuración de parámetros adoptada en la etapa de iteración. Se contempla la selección de máximo 20 atributos para simplificar la interpretación y posterior aplicación del modelo. Las variables se eligen de acuerdo al indicador de importancia obtenido con cada software.

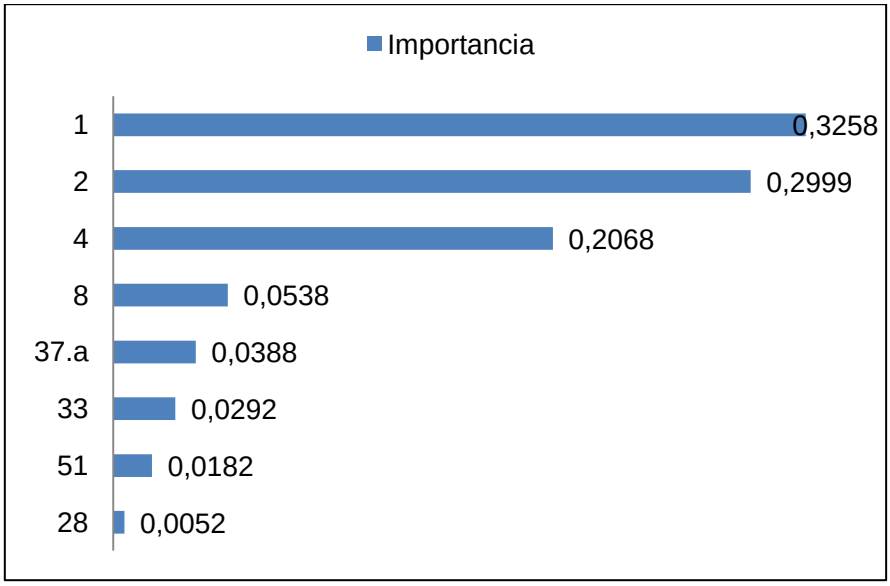
6.5.1 ÁRBOLES DE DECISIÓN

La configuración establecida para el modelo de árboles de decisión consiste en un algoritmo de tipo CHAID, con una profundidad máxima de cinco ramas, cuyo criterio de parada es obtener un porcentaje mínimo de registros de 2% en la rama principal y 1% en las ramas secundarias. El nivel de significancia tanto para la división como para la combinación de ramas es de 0,05 y el estadístico utilizado para el test chi-cuadrado es el de Pearson.

Se eligen las 8 variables más relevantes de acuerdo al modelo final de la etapa de iteración, excluyéndose la variable 37 por representar una característica semejante a la de la 37.a, y la variable 22 y 19 pues al ser agregadas no aportaron en la clasificación.

Como se observa en la figura 20, un 83,25% de la importancia relativa es explicada por los atributos 1, 2 y 4 que guardan relación con el comportamiento previo del contribuyente. Esto indica que la capacidad predictiva del modelo se basa casi en su totalidad en el pronóstico que estos entregan. Más adelante se detallará que refleja cada variable.

Figura 20: Importancia relativa de las variables para el modelo de árboles de decisión.



Fuente: Elaboración propia.

Se realiza una prueba que incluye solo los atributos 1, 2 y 4, pero no se logran buenos resultados debido a que el modelo muestra una muy baja capacidad de detectar los casos negativos.

Los resultados de esta prueba entregan un modelo que toma seis segundos en ser construido, tiene precisión general de 68,37% y un error tipo II de 90,29%. Su profundidad máxima es de 5 ramas e incluye 10 variables en total.

Si bien la precisión global desciende con respecto al último modelo de la etapa de iteración, esto se debe a que se toman menos variables de ingreso, lo que disminuye la cantidad de información disponible para el entrenamiento. Sin embargo, como los árboles de decisión son sensibles al ruido y a las variables irrelevantes, este cambio genera una mejora en la capacidad de predecir los malos casos, que se expresa en un aumento de 1,24%.

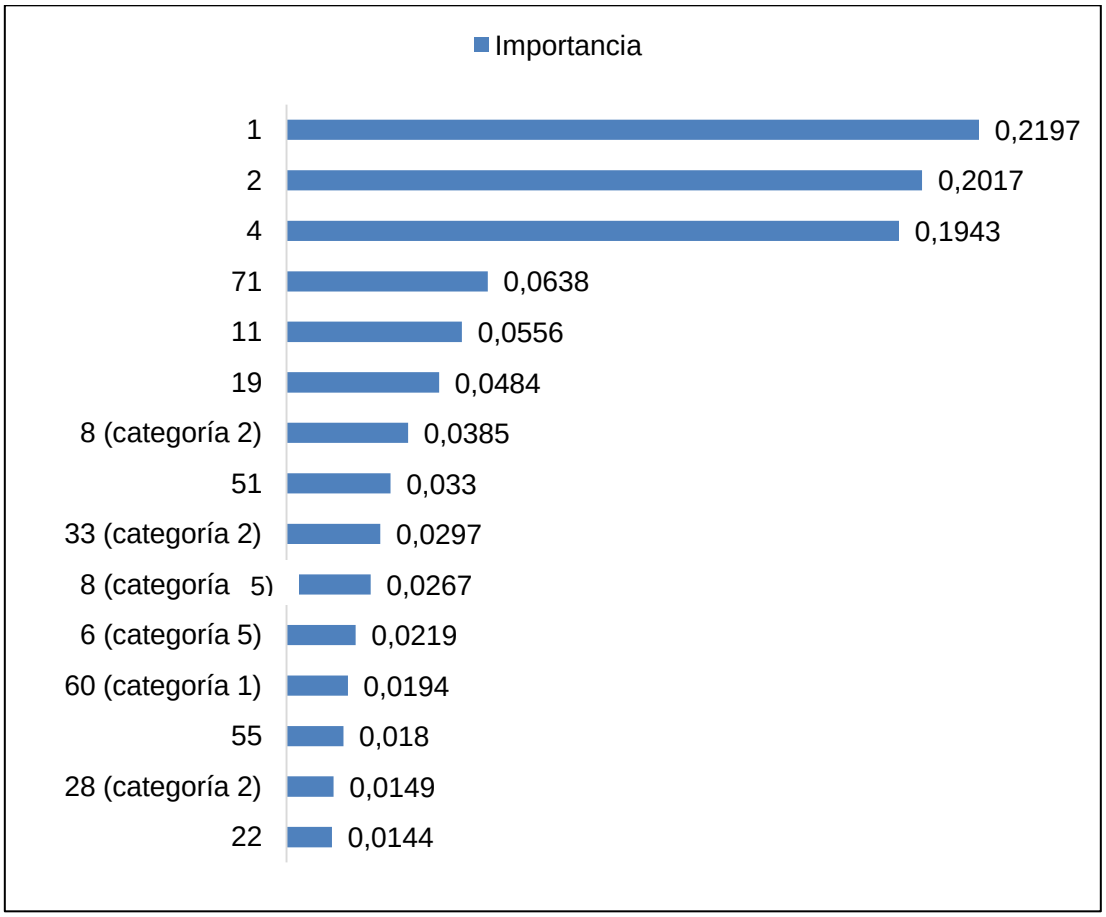
Las reglas de decisión obtenidas del modelo final se pueden ver en el anexo 16. Para interpretar los cortes generados en cada variable, se aplica la función inversa de las transformaciones realizadas.

6.5.2 Redes Neuronales

Para el modelo final de redes neuronales se configuran los parámetros con el algoritmo de perceptrón multicapa, una sola capa oculta y un conjunto de prevención de sobreajuste de 30%.

Se testean subconjuntos de máximo 20 variables. La primera prueba se realiza con las diez variables de mayor relevancia según la figura 19. Luego de probar con 36 subconjuntos, se opta por una fusión entre las variables consideradas inicialmente y los atributos del modelo final de regresión logística y árboles de decisión. En la figura 21 se muestra la selección final y la importancia relativa de cada una dentro del modelo.

Figura 21: Importancia relativa de las variables para el modelo final de redes neuronales.



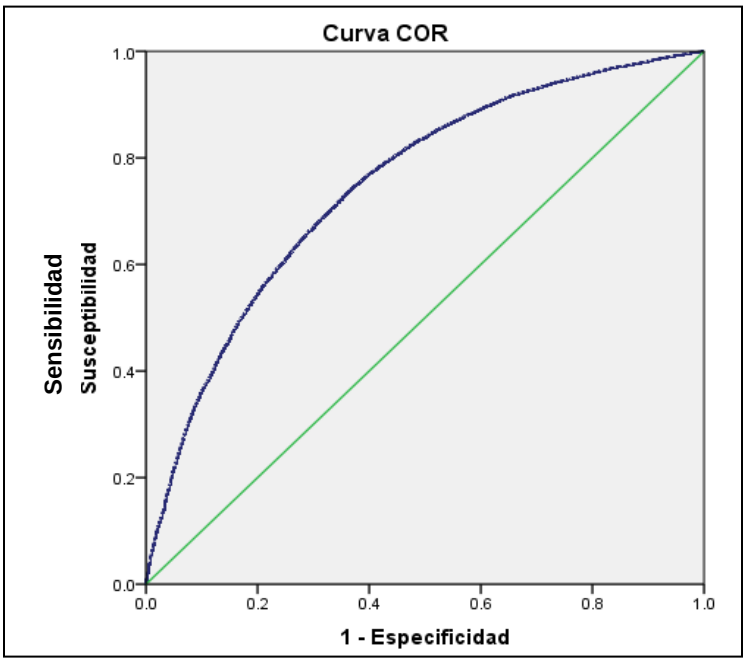
Fuente: Elaboración propia.

Se modifica el número máximo de neuronas que contempla el modelo, estudiándose un rango va desde 8 hasta 30. Se alcanza la mejor precisión con un total de 25 neuronas y un corte de clasificación de 0,5. Los resultados de las pruebas se pueden ver en el anexo 17.

De esta etapa se obtiene un modelo de redes neuronales que tiene un tiempo de ejecución de 50 segundos y una precisión general de 69,03%, una especificidad de 68,16%, una sensibilidad de 69,07% y un error tipo II de 90,20%. Los indicadores mejoran con respecto a lo obtenido en la etapa anterior debido a la modificación de la cantidad de neuronas de forma manual.

En la figura 22 se puede ver que la curva ROC mantiene un buen nivel. La exclusión de más de 30 variables no influyó significativamente en la capacidad predictiva.

Figura 22: Curva ROC para el modelo final de redes neuronales.

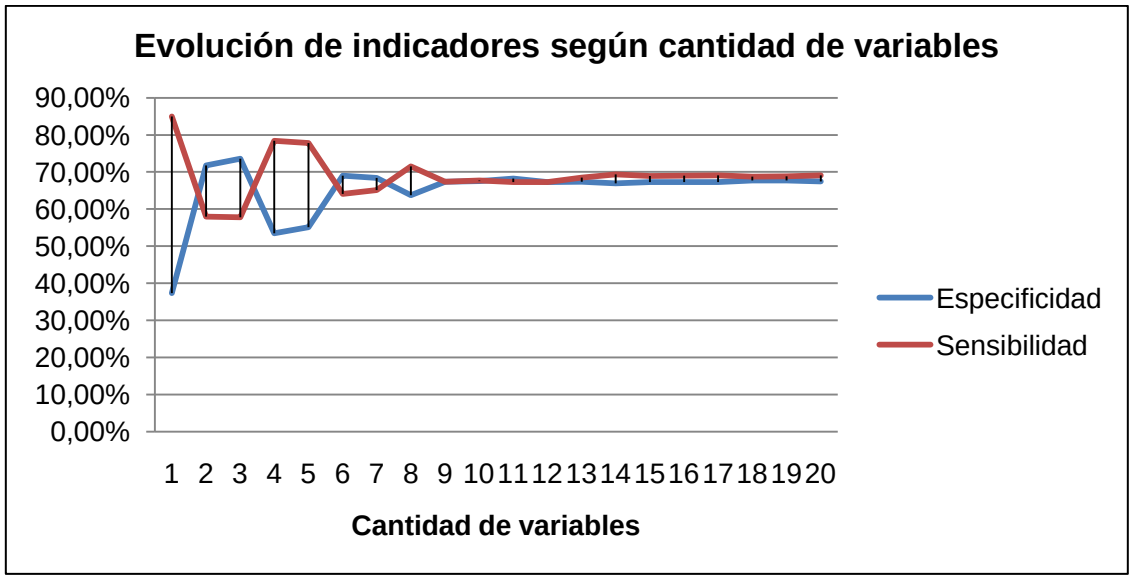


Fuente: Elaboración propia.

6.5.3 Regresión Logística

Para la técnica de regresión logística la elección del modelo final se realiza en la etapa de iteración con la herramienta Stepwise que permite la selección del mejor conjunto de atributos. El método elegido en la etapa de iteración es el Backward con el estadístico de Wald.

Figura 23: Evolución de indicadores según cantidad de variables.



Fuente: Elaboración propia.

La decisión se toma en base a la precisión general, especificidad y sensibilidad que logra el modelo. En la figura 23 se muestra le evolución de los indicadores de

acuerdo a la cantidad de variables que se incluyen. Se observa que de diez atributos en adelante los indicadores se estabilizan, llegando a una meseta en la precisión global cuando se utilizan 14 variables.

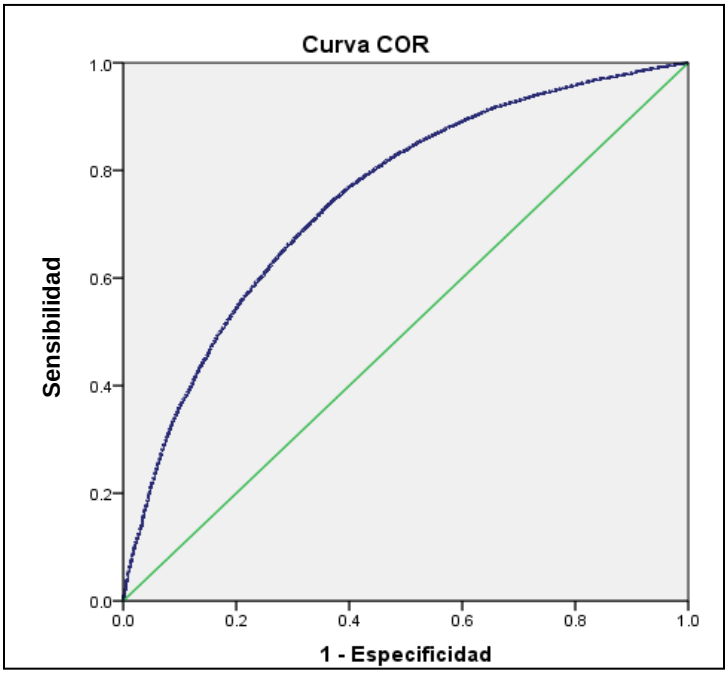
La elección final es un subconjunto de 15 variables, pues, si bien en este la precisión global se mantiene, el AUC sube de 0,472 a 0,473.

La configuración final queda como un modelo de regresión logística con un total de 15 variables, realiza un máximo de 20 iteraciones por etapa y tiene una significancia de entrada de variables de 0,05 y una significancia de salida de variables de 0,1.

Los atributos elegidos son: 1, 2, 4, 6 (categoría 5), 8 (categoría 2), 8 (categoría 5), 11, 19, 22, 28 (categoría 2), 33 (categoría 2), 51, 55, 60 (categoría 1) y 71. El detalle del modelo se puede ver en el anexo 19.

Como se muestra en la tabla 11 se logra un aumento de 0,4% en la precisión global del modelo final, para un punto de corte de 0,488 donde se logra el mayor KS[13]. La eliminación de ciertas variables provoca una disminución en el AUC que se explica por el descenso en la especificidad. Sin embargo, se observa una mejor capacidad predictiva de los buenos casos. La curva ROC que se aprecia en la figura 24 da cuenta de la calidad del modelo para todos los puntos de corte.

Figura 24: Curva ROC para el modelo final de regresión logística.



Fuente: Elaboración propia.

7. RESULTADOS

[13] Ver el análisis para la obtención del KS en el anexo 20.

Los resultados que se muestran en la tabla 14 indican que los tres modelos tienen una capacidad predictiva alta. Sus indicadores de especificidad y sensibilidad demuestran un nivel adecuado de precisión. Sin embargo, se observa una falencia en la tasa de error tipo II.

Tabla 14: Resultados de los modelos finales.

	Precisión General	Especificidad	Sensibilidad	KS	AUC	Error II
Árbol de Decisión	68,37%	69,12%	68,33%	-	-	90,29%
Red Neuronal	69,03%	68,16%	69,07%	0,372	0,748	90,20%
Regresión Logística	68,81%	67,29%	68,88%	0,364	0,742	90,34%

Fuente: Elaboración propia.

La precisión general es similar en los tres casos, siendo el modelo de redes neuronales el que tiene el porcentaje más alto con un 69,03% y el de árboles de decisión el más bajo con un 68,37%, lo que se condice con la teoría que indica que las redes neuronales debiese alcanzar la cota máxima en predicción.

Se observa que la sensibilidad sigue el mismo patrón que la precisión general. Sin embargo, la especificidad es mejor en la técnica de árboles de decisión. En cuanto a los indicadores KS y AUC se observa que los niveles más altos son alcanzados por las redes neuronales, siendo la misma técnica la que alcanza la menor tasa en el error tipo II.

Es importante analizar en conjunto los cinco indicadores, ya que la precisión general no refleja correctamente la capacidad predictiva, pues la influencia de la especificidad se ve opacada por la sensibilidad debido a la reducida proporción con que se observan los casos de default.

Se concluye que el mejor modelo es el de redes neuronales, tanto por la precisión (que incluye precisión general, sensibilidad y especificidad) como por el AUC y KS. Sin embargo, se decide optar por el modelo de regresión logística como modelo base, ya que esta técnica permite entender la fuerza de influencia de cada atributo y explica en mayor profundidad las características que gatillan el comportamiento de default.

Como modelo alternativo se elige el árbol de decisión por la simpleza de interpretación y la facilidad para generar reglas de decisión. Además, este modelo muestra mejor predicción de casos malos, posibilitando la obtención de una perspectiva diferente de las variables que influyen en el comportamiento de los contribuyentes.

7.1 Análisis Validación y Robustez.

Para validar los modelos elegidos se define una partición de la muestra en tres subconjuntos: entrenamiento, comprobación y validación, evaluando cada uno de los indicadores de precisión en las distintas particiones. Esto permitirá definir el nivel de sobreajuste y la confiabilidad de los resultados al aplicarlo a la totalidad de contribuyentes.

También se hace un análisis de robustez, en el que se calculan los indicadores de precisión para los cuatro años que contemplan los modelos. Se estudia la evolución de

estos indicadores en búsqueda de un comportamiento estable que permita inferir que el modelo es estable a lo largo del tiempo.

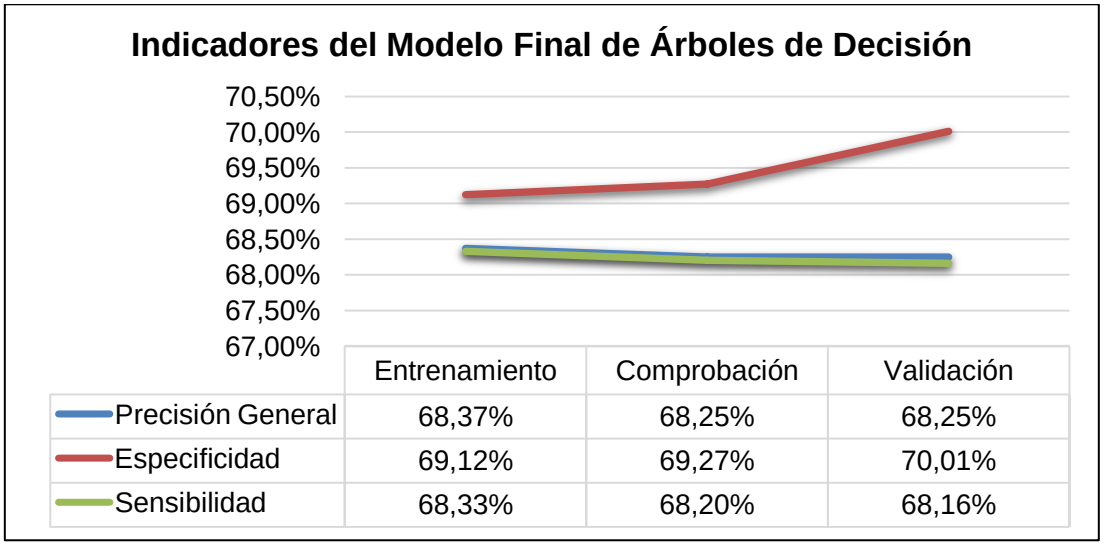
A continuación se muestra el proceso para el modelo definitivo (regresión logística) y el modelo desafiante (árboles de decisión).

7.1.1 Árboles de Decisión

Se calculan los indicadores de precisión de manera independiente para cada una de las particiones de la muestra. En la figura 25 se grafica la precisión general, especificidad y sensibilidad, donde se observa estabilidad entre los diferentes conjuntos de la partición.

Las diferencias más importantes se ven en la especificidad, que aumenta cerca de un 1% para la muestra de validación. Como esta variación es pequeña, y además genera cambios positivos en los indicadores, se concluye que el modelo es válido para el universo de contribuyentes y no muestra un sobreajuste en el entrenamiento.

Figura 25: Indicadores del modelo de árboles de decisión para las particiones.



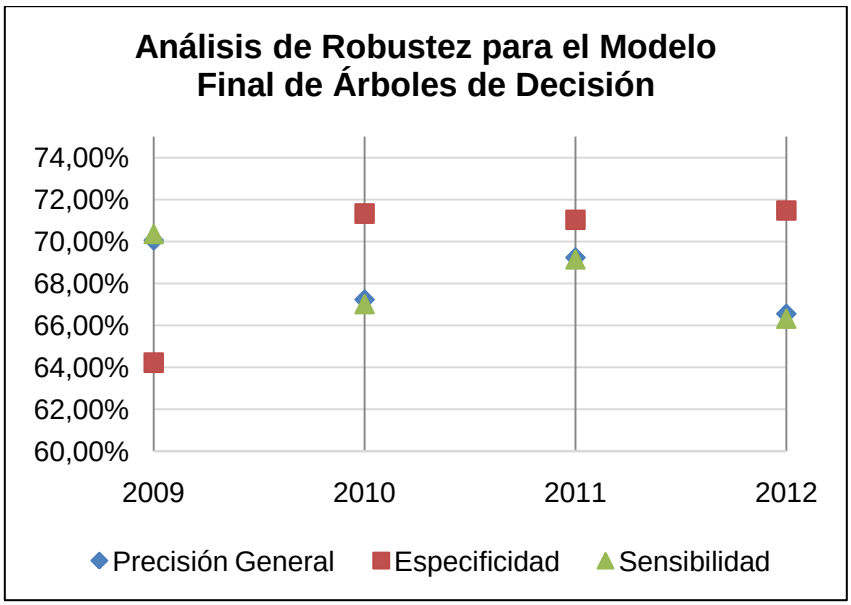
	Entrenamiento	Comprobación	Validación
Precisión General	68,37%	68,25%	68,25%
Especificidad	69,12%	69,27%	70,01%
Sensibilidad	68,33%	68,20%	68,16%

Fuente: Elaboración propia.

Posteriormente se analiza la robustez del modelo, calculando los indicadores de precisión de manera independiente para cada año. Como se muestra en la figura 26, el modelo de árboles de decisión tiene un comportamiento errático en la sensibilidad, mientras que en la especificidad se observa que es más estable.

Se destaca la diferencia entre el período 2009 y los tres restantes. Este fenómeno podría estar indicando que cada cierto tiempo los contribuyentes cambian su comportamiento, haciendo necesaria la reestimación de los parámetros del modelo al cabo de unos años. La razón del cambio de comportamiento es desconocida, pudiendo deberse a variables relacionadas con ciclos económicos, características intrínsecas de los grupos, cambios demográficos de la población, entre otras.

Figura 26: Análisis de robustez para el modelo de árboles de decisión.



Fuente: Elaboración propia.

7.1.2 Regresión Logística

Se calculan los indicadores de precisión de manera independiente para cada una de las particiones de la muestra. Para el modelo final de regresión logística se observa una menor estabilidad. Se utiliza el mismo punto de corte de clasificación para las tres particiones, correspondiente a 0,488 [14]. Se observa que la precisión general y sensibilidad aumentan cerca de un 1% en el conjunto de comprobación, y la sensibilidad disminuye un 1% en el mismo.

Es importante destacar que en la muestra de validación los tres indicadores tienden a un mismo porcentaje (cercano al 69%), lo que es un buen resultado. También se destaca una mejora consistente en los niveles del AUC y KS para la comprobación y validación.

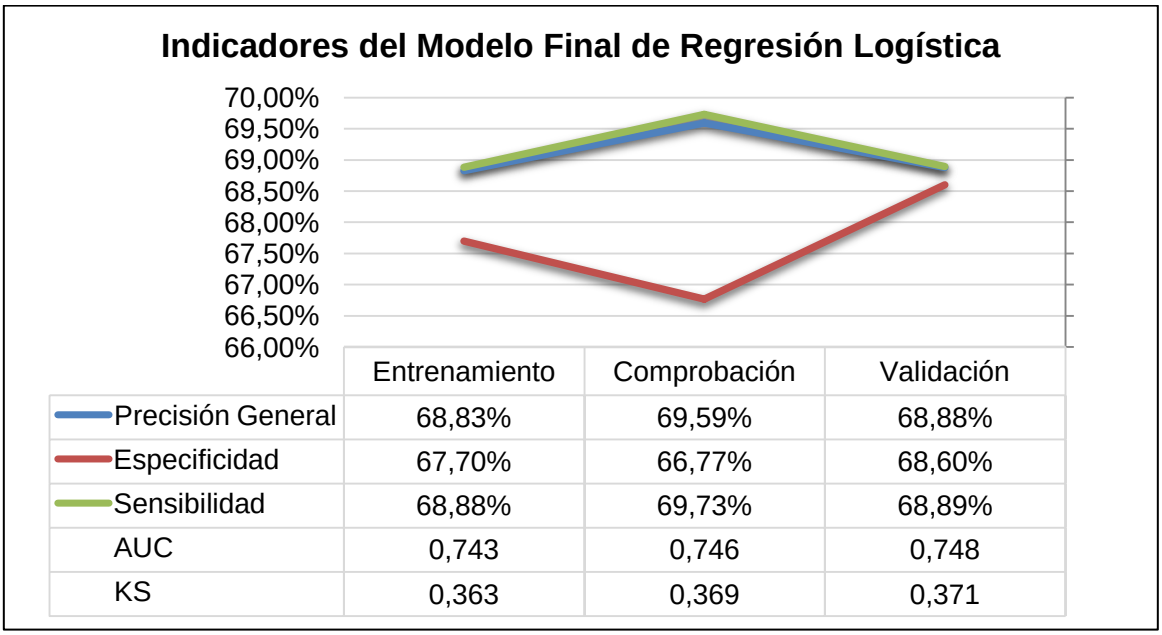
A pesar de que existen fluctuaciones entre las particiones, la proporción en que estas varían es muy pequeña. La mejora en el nivel de los cinco estadísticos en la muestra de validación permite concluir que este modelo es válido para las distintas y no está sobre-ajustado a la muestra de entrenamiento.

Posteriormente se analiza la robustez del modelo, calculando los indicadores de precisión de manera independiente para cada año. En el modelo de regresión logística ocurre lo contrario que en el de árboles de decisión. En este caso los primeros tres años son estables y es en el último que se desestabilizan los indicadores.

En primer lugar se identifica la probabilidad de corte óptima para la muestra de cada año. Se obtiene un corte similar para los cuatro años, siendo de 0,488 en los primeros dos, 0,489 en el tercero y 0,499 en el último año. Este punto óptimo se obtiene de acuerdo a la mejor precisión lograda para cada una de las muestras.

[14] Punto óptimo para el corte de probabilidad de acuerdo al indicador KS para la partición de entrenamiento. Este punto se encuentra la mayor distancia entre las distribuciones de las dos clases de contribuyentes.

Figura 27: Indicadores del modelo de regresión logística para las particiones.

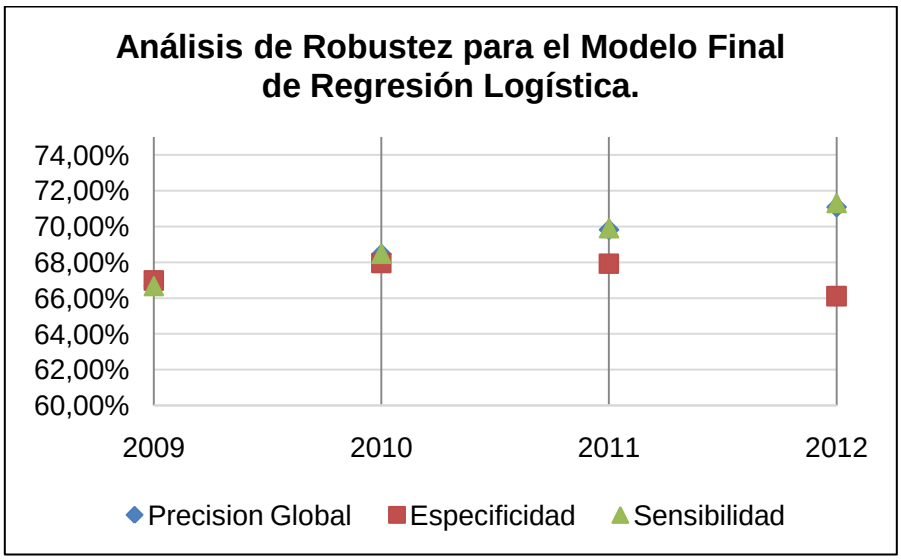


	Entrenamiento	Comprobación	Validación
Precisión General	68,83%	69,59%	68,88%
Especificidad	67,70%	66,77%	68,60%
Sensibilidad	68,88%	69,73%	68,89%
AUC	0,743	0,746	0,748
KS	0,363	0,369	0,371

Fuente: Elaboración propia.

Como se observa en la figura 28, la predicción global y sensibilidad aumenta con los años, mientras que la especificidad disminuye, pero muy levemente, por lo que se considera que permanece constante. Este fenómeno puede deberse a que el modelo de regresión logística logra incluir más variables que el de árboles de decisión, por lo que la incidencia del cambio de comportamiento en una variable no afectará tanto al resultado final como en el caso de los árboles de decisión.

Figura 28: Análisis de robustez para el modelo de regresión logística.

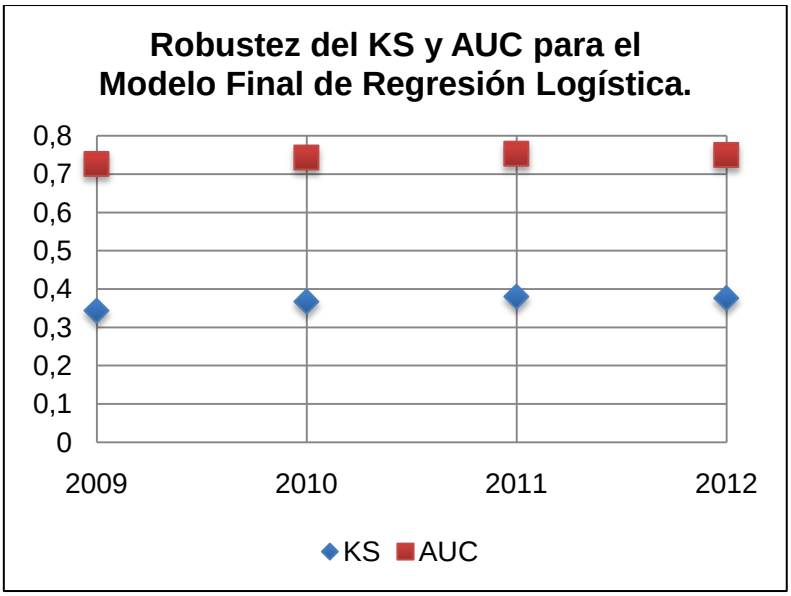


Fuente: Elaboración Propia.

Si se analiza la figura 29 se corrobora el hecho de que este modelo es robusto. Este gráfico indica estabilidad en los indicadores de KS y AUC, que muestran la

capacidad predictiva para todos los niveles de corte. La baja variabilidad que se observa en el tiempo en la especificidad, precisión general, y sensibilidad, sumado a la estabilidad que presentan el AUC y el KS indican que el modelo de regresión logística es robusto en el paso del tiempo. Esto no implica que no se deba realizar una reevaluación al cabo de algunos años, pues la incidencia de las variables podría verse modificada si el comportamiento de los contribuyentes cambia.

Figura 29: Robustez del KS y AUC para el modelo final de regresión logística



Fuente: Elaboración Propia.

Para finalizar se presenta en la tabla 15 el error de tipo II para distintos cortes de probabilidad de la predicción. Se observa que al aumentar el nivel de corte disminuye el error de falsos negativos. En particular para un corte 0,9 se tiene que el error mejora en 12 puntos porcentuales. Mientras más estricta sea la clasificación menor error de tipo II se encontrará. Sin embargo, la cantidad de defaulters identificada también disminuirá.

Tabla 15 : Tasa de error para distintos cortes de probabilidad.

Corte	Error II
0,9	22,8%
0,8	20,6%
0,7	16,7%
0,6	13,2%
0,5	9,9%

Fuente: Elaboración Propia.

7.2 INTERPRETACIÓN

A continuación se analizan en detalle los modelos elegidos, dándole una interpretación a los atributos relevantes y evaluando su influencia en el comportamiento de los contribuyentes que caen default.

7.3.1 ÁRBOLES DE DECISIÓN

El modelo final del árbol de decisión incluye 8 variables, tiene una profundidad de cinco ramas y 41 nodos finales. Se puede ver en el anexo 18 el diagrama detallado de las reglas de decisión.

De las reglas de decisión se desprende que si un contribuyente presenta anotaciones del tipo “no declarante” es muy probable que continúe teniendo esta conducta, conformando así un primer segmento de contribuyentes que no tienen actividad en el SII.

Se identifica un segundo segmento de contribuyentes que corresponde a los que probablemente nunca caigan en default. Este grupo se define como aquellos casos en que no se tiene anotaciones del tipo “no declarante” y tampoco se observa atrasos en la declaración del F29.

El tercer segmento de contribuyentes es aquel que no tiene anotaciones del tipo “no declarantes”, no se tiene registro de atrasos en la declaración del F29 y tampoco se tiene facturas de compra emitidas.

Es en este último grupo donde el resto de los atributos seleccionados logra incidir en la predicción. Se observa que la propensión a caer en default depende del período del año en que se encuentre. Mientras más lejano sea el mes en que se declara el último F22, mayor es probabilidad de que el modelo asigne un 1 a la variable objetivo. Esto se agrava si en los últimos períodos el contribuyente no ha tenido remanente de crédito fiscal, y si sus ventas mensuales son menores a 81.000 pesos.

El análisis individual de cada variable se muestra a continuación.

- **V2: N° de anotaciones “no declarante” teniendo que declarar el último año.** A mayor cantidad de anotaciones “no declarante”, existe una mayor probabilidad de que se cometa default. En particular si la cantidad es mayor o igual a dos, el pronóstico siempre indicará que el contribuyente cae en default. Cuando la cantidad es menor que uno, dependerá del comportamiento que se observa en V1.
  El efecto se explica porque los contribuyentes que cuentan con un historial no declarante, con gran certeza repetirán esta conducta. Tener una anotación “no declarante” implica una infracción mayor que un atraso en la declaración, pues, indica que el contribuyente (hasta la fecha) no ha presentado el F29 dicho mes.
- **V1: N° de veces que declara atrasado el F29 el último año**. Al aumentar el número de veces que el contribuyente paga atrasado el F29, mayor es la propensión a cometer default. En particular el modelo pronostica que si V1 y V2 son iguales a cero, el contribuyente muy pocas veces cometerá default.
  El fenómeno se explica dado que quienes jamás se han atrasado, tienen un alto sentido de responsabilidad. Esta variable refleja la importancia que el contribuyente le atribuye al cumplimiento de sus obligaciones tributarias.

- **V4: N° de facturas de compra emitidas el último año.** A menor cantidad de facturas de compra emitidas, más probabilidad hay de que se cometa default. Esta variable es decidora cuando el contribuyente presenta V1 y V2 con valor cero. En este caso si V4 es mayor que cero, el pronóstico del modelo es que el contribuyente no comete default.
  Las facturas de compra corresponden a documentos mercantiles que reflejan la información de una operación de compra. Para utilizar este recurso se debe cumplir con una serie de requisitos impuestos por el SII, que demuestran la preocupación del contribuyente por cumplir con las normas y la voluntad que tienen por mantener un orden contable y estar al día con sus obligaciones tributarias. Estas son razones que justifican el efecto observado.
- **V8: Meses de antigüedad desde el inicio de actividades.** Si no se tiene anotaciones del tipo "no declarante" y se cuenta con un solo atraso en la declaración del F29, entra en juego este atributo. Cuando el contribuyente inicia actividades hace menos de cuatro años y medio, el modelo pronostica la existencia de default.
  Aquellas empresas que han iniciado actividades recientemente muestran una mayor inestabilidad de flujos y una falta de experiencia en materia de contabilidad y mantención del negocio, por lo que se considera razonable la influencia de la variable observada.
- **V37.a: Promedio de ventas en los últimos seis meses.** Para contribuyentes cuyas ventas mensuales promedio sean mayores a 81.000 pesos y no hayan pagado atrasados el F29 en el último año, el pronóstico será que no caen en default. Si sus ventas son menores a 81.000 pesos serán pronosticados como que caen en default.
  El promedio de ventas refleja el tamaño, la capacidad administrativa y el nivel de organización de la empresa. Cuando se tienen montos de ventas muy pequeños, se puede estar reflejando empresas emergentes, empresas que no han logrado crecer o empresas que estén pasando por dificultades económicas. En cualquiera de los casos existen razones para atrasarse en la declaración del F29, ya sea por inexperiencias, falta de liquidez o desórdenes tributarios que lleven al incumplimiento
- **V28: N° de períodos desde el último período con remanente de crédito fiscal.** Cuando ha pasado un solo período o bien entre cuatro y ocho, se pronostica que el contribuyente no cae en default. Mientras que si ha pasado entre tres y cuatro períodos, o más de ocho, o el registro es nulo, se infiere que caerá en default.
  El remanente de crédito fiscal se genera por una diferencia positiva entre las compras y las ventas realizadas. Si un contribuyente tiene una mayor cantidad de compras que ventas, quedará con un abono de IVA (remanente) para el siguiente mes. Mientras menor sea el tiempo pasado desde ese evento, se puede suponer que el contribuyente cuenta con un "colchón" crediticio, lo que disminuye la probabilidad de no contar con dinero suficiente para declarar el F29.
- **V33: Períodos tributarios desde la última declaración del F22.** Cuando no se tiene anotaciones del tipo "no declarante", no se ha atrasado en presentar el F29 y no se tienen facturas de compra emitidas, se analizan los períodos pasados desde el último período con remanente fiscal. Si esta variable es mayor a un año o bien nunca se ha declarado el F22 entonces se pronostica que habrá default.

En caso de que haya pasado entre medio y un año, se analiza el valor de V51. Si este tiempo es menor a seis meses, se pronostica que no habrá default.
Durante el mes en que se declara el F22 los contribuyentes le otorgan mayor relevancia a la preparación y declaración de este formulario, restándole tiempo y dedicación al F29. Esto puede llevar a un descuido de las obligaciones de declaración de IVA, muchas veces provocado por el olvido o la consiente presentación tardía del formulario. Como se observó en el análisis de la muestra, todos los años en el mes de mayo la tasa de default se eleva, por lo que resulta razonable la influencia de esta variable.

- **V51: N° de períodos con remanente de crédito fiscal en los últimos dos años.** Cuando no se tiene anotaciones de tipo "no declarante", no se ha atrasado en presentar el F29, no se tienen facturas de compra emitidas y el tiempo transcurrido entre el último período con remanente de crédito fiscal comprende entre ocho meses y un año, se analiza esta variable. Si es menor a nueve indica que se cae en default.
  Similar a lo que ocurre con la variable 28, los contribuyentes que han tenido más veces remanente de crédito fiscal, tienen mayor disponibilidad de dinero a su favor y en general tienen un mayor cumplimiento de su obligación de declaración de IVA, por lo que es razonable que la propensión disminuya.

7.3.2 Regresión Logística

El detalle de la estimación de parámetros y del modelo de regresión logística obtenido se puede ver en el anexo 19. El modelo final incluye 15 variables, de las cuales destacan la variable 1, 2, 4, 71 y la categoría 5 de la variable 6, todas relacionadas con el historial de cumplimiento y con características de flujos y contabilidad. Estos resultados se condicen con lo obtenido en el modelo de árboles de decisión, pero esta vez se destaca la incidencia de la variable 71.

A continuación se analiza en detalle la influencia de cada atributo en el comportamiento de los contribuyentes.

- **V4: N° de facturas de compra emitidas el último año.** La proporción entre las chances es de 0,202 para V4. Este ratio no es interpretable dado que la variable fue previamente transformada a un intervalo 0-1. Sin embargo, se puede inferir que un contribuyente que presente más facturas de compras emitidas tiene una menor probabilidad de caer en default que aquel que presenta menos.
  Tal como se explica para el modelo de árboles de decisiones, la utilización de este recurso implica el cumplimiento de una serie de requisitos impuestos por el SII, que demuestran la voluntad del contribuyente por mantener un orden contable y estar al día con sus obligaciones tributarias.
- **V11: Boletas emitidas en el último año.** La proporción entre las chances es de 0,674. En este caso tampoco es posible interpretar el ratio por la transformación hecha a la variable. Se puede inferir que si esta proporción aumenta, menor es la probabilidad de caer en default.
  La emisión de boletas es un mecanismo para llevar registro de las ventas, e indica tanto el tamaño del negocio como la propensión del contribuyente a mantenerse dentro de las reglas establecidas por el SII. Por lo tanto, la

probabilidad de caer en default es mayor para aquellos contribuyentes que no emiten boletas.

- **V19: ¿Tiene contador el último año?** Un contribuyente que realiza las declaraciones con un contador tiene un riesgo relativo de 0,724 con respecto a uno que no tiene, indicando que la probabilidad de caer en default disminuye si hay contador. La cantidad de contribuyentes que declara tener contador corresponde al 41% del total de micro y pequeñas empresas, lo que es esperable dado que los niveles de ingresos de este segmento muchas veces no alcanza para pagar los servicios de un contador.
  Este fenómeno tiene sentido, pues el hecho de tener contador refleja el interés del contribuyente de cumplir con sus obligaciones y mantener ordenado su estado de resultados.
- **V71: ¿Declara renta presunta[15] el último año?** Un contribuyente que declara renta presunta en el F22 del último año, tiene un riesgo relativo de 0,561 con respecto a uno que no. Esto indica que el riesgo disminuye para los contribuyentes que declaran renta presunta, siendo este grupo un 5% de los contribuyentes totales.
  Cabe destacar que en esta variable puede existir un efecto oculto del tipo de actividad económica, ya que solo algunas actividades económicas son las que pueden optar a la declaración con renta presunta. Este tipo de declaraciones puede simplificar el trabajo que significa presentar el F29. Lo anterior pudiera considerarse como un incentivo, o una barrera menos, en el cumplimiento de la obligación de los contribuyentes, produciendo como efecto una menor cantidad de default.
- **V51: N° de períodos con remanente de crédito fiscal en los últimos dos años.** La proporción entre las chances es de 0,774. En este caso no es posible interpretar el ratio por la transformación hecha a la variable. Sin embargo se infiere que un contribuyente que cuenta con más períodos de remanente fiscal tiene una probabilidad menor de caer en default que uno que tenga menos períodos.
  Se espera este efecto, pues los contribuyentes que han tenido más veces remanente de crédito fiscal, tienen mayor disponibilidad de dinero a su favor y en general tienen un mayor cumplimiento de su obligación de declaración de IVA, por lo que es razonable que la propensión disminuya.
- **V22: ¿Tiene contabilidad completa [16] ?** Un contribuyente que declara contabilidad completa tiene un riesgo relativo de 0,823 con respecto al que declara con contabilidad simplificada. Esto indica que la probabilidad de default en el grupo de contabilidad completa es menor que en la contabilidad simplificada, lo que se condice con el efecto de V19.
  Para la contabilidad completa es necesario tener un mayor orden en las cuentas de la empresa, y en la mayoría de los casos, contratar un contador. Dicho

[15] Las empresas o personas que declaran renta presunta pagan sus impuestos según lo que la ley determina para esa actividad y no de acuerdo con los resultados reales obtenidos. Algunas actividades que se consideran en esta clasificación son: el transporte de pasajeros, la minería y explotación de bienes raíces agrícolas y no agrícolas.

[16] Es aquella que comprende los libros Caja, Diario, Mayor e Inventarios y Balances, independiente de los libros auxiliares que exija la ley, tales como Libro de Ventas Diarias, de Remuneraciones, de Impuestos Retenidos, etc.

requerimiento hace que este atributo disminuya la probabilidad de cometer default, pues se refleja la disposición a cumplir con la normativa tributaria de manera detallada. El porcentaje de contribuyentes de micro y pequeñas empresas que tiene contabilidad completa es cercano a un 20%. Esto explica el porcentaje de campos nulos (en materia de contabilidad) que se observa en la base de datos.

- **V55: ¿Tiene ACTECO tipo "saco" [17] ?** Un contribuyente cuya actividad económica (ACTECO) está clasificada dentro del tipo "saco" tiene un riesgo relativo de 1,098 con respecto a un contribuyente que no. Esto indica que la probabilidad de default aumenta si se tiene ACTECO de esta categoría. La razón de este fenómeno es que los ACTECO tipo saco son de difícil fiscalización y por ende los contribuyentes tienden a no ser tan rigurosos en cuanto a sus obligaciones.
- **V1: N° de veces que declara atrasado el F29 el último año.** La incidencia de la variable no se puede interpretar de forma directa dado que fue transformada a un intervalo 0-1. La proporción entre las chances es de 55,49 por lo que se infiere que los contribuyentes que tienen mayor cantidad de atrasos en la declaración del F29 muestran un aumento en la probabilidad de caer en default versus aquellos que presentan menor cantidad de atrasos.
  El fenómeno se explica pues esta variable refleja la importancia que el contribuyente le atribuye al cumplimiento de sus obligaciones tributarias.
- **V2: N° de anotaciones "no declarante" teniendo que declarar el último año.** La proporción entre las chances es de 4,494 lo que permite inferir que aquellos contribuyentes con mayor cantidad de anotaciones del tipo "no declarante" tienen una mayor probabilidad de caer en default.
  Tener una anotación "no declarante" implica una infracción mayor que un atraso en la declaración, pues, indica que el contribuyente (hasta la fecha) no ha presentado el F29 dicho mes. Si ya se ha cometido esta falta una vez, muy probablemente se volverá a caer en default.
- **V6 (categoría 5): No tiene o ha transcurrido más de 10 meses desde el último comportamiento tributario irregular[18] MIPE.** Si la última vez fue hace más de diez meses o nunca ha tenido, su riesgo relativo de 0,572 con respecto a un contribuyente que presenta comportamiento tributario irregular en un período más cercano, es decir, la probabilidad de caer en default disminuye si no se presenta esta sitación.
  Esta variable cumple con el rol de hoja de antecedentes. Los contribuyentes que posean comportamiento irregular, tendrán una facilidad mayor a caer en default, pues, han mostrado faltas en más de un ámbito.
- **V8 (categoría 5): Antigüedad desde el inicio de actividades es mayor a 16 años.** Aquellos contribuyentes cuya antigüedad de inicio de actividades es mayor a 16 años, tienen un riesgo relativo de 0,766 con respecto a los menos antiguos. Al igual que en V8, a mayor antigüedad, menor es la propensión a cometer default. En particular, para este punto de corte, se está identificando a las

[17] Se le llama 'Actividad económica tipo saco' a todas aquellas actividades no registrada y deban clasificarse como 'otras actividades'. Esto se debe principalmente a que son muy específicas y se presentan poco entre los contribuyentes.

[18] Existe muchos eventos en esta clasificación de anotaciones, siendo algunos ejemplos: No ubicado, domicilio inexistente, no concurre a notificación, entre otros.

empresas que cuentan con una vasta trayectoria. El permanecer activo por un periodo mayor a 16 años indica un alto nivel de orden en materia contable, una cartera de clientes segura y un conocimiento de la norma tributaria. Estas empresas son robustas y cuentan con una rutina establecida, lo que hace disminuir su probabilidad de default.

- **V28 (categoría 2): El mes anterior tienen remanente de crédito fiscal[19].** Aquellos contribuyentes el mes anterior tienen remanente de crédito fiscal cuentan con un riesgo relativo de 0,835 con respecto a los que no tienen. Esto indica que los contribuyentes con remanente fiscal presentan una menor probabilidad de caer en default. Si se tiene remanente el mes anterior y las ventas del mes siguiente son iguales o inferiores al remanente, estos no pagarán IVA, razón por la cual es posible inferir que habrá menores dificultades en la presentación del F29.
- **V33 (categoría 2): Han transcurrido menos de siete meses desde la última declaración del F22.** Aquellos contribuyentes que declararon el F22 hace menos de siete meses tienen un riesgo relativo de 0,778 con respecto a los que no lo han declarado nunca o lo declararon hace más tiempo. Esto indica que en los primeros períodos tras la declaración del F22 es menos probable caer en default. La explicación de este fenómeno se debe a que luego de la declaración del F22, las empresas debiesen tener una contabilidad más ordenada y sin la acumulación de deudas pre-existentes, lo que facilitaría la declaración de IVA. Además se observa que el mes de mayo de cada año (fecha en que se declara el F22) el porcentaje de default sube, justificando el efecto de esta variable.
- **V60 (categoría 1): Edad del representante legal es un campo nulo.** Los contribuyentes que no presentan información respecto a la edad del representante legal tienen un riesgo relativo de 0,842 con respecto a quienes si lo hacen. Este resultado permite inferir que la probabilidad de default disminuye si el campo de edad es nulo, observándose que un 45,7% de los contribuyentes caen en esta categoría.

7.3 COMPORTAMIENTO POR ZONA GEOGRÁFICA

Para calcular el default en el mes de Julio 2013 en primer lugar se obtiene una lista innominada de los contribuyentes activos para este año[20], obteniéndose un total de 2.493.441 empresarios, de los cuales un 92% corresponde a la clasificación de microempresa.

Se construyen las 15 variables indicadas en el modelo final de regresión logística. Los estadísticos descriptivos de las variables indican una alta presencia de campos nulos que son reemplazados por cero. Posteriormente se genera un archivo en el software *IBM SPSS Modeler* con los parámetros de la regresión. Esto permitirá replicar el proceso de forma automática en los siguientes meses[21].

[19] El remanente de crédito fiscal se genera por una diferencia positiva entre las compras y las ventas realizadas. Si un contribuyente tiene una mayor cantidad de compras que ventas, quedará con un abono de IVA (remanente) para el siguiente mes.

[20] Datos contenidos en la tabla DW.TMP_NOMINA_SEGMENTOS_VW.

[21] La planilla con la lista de contribuyentes y sus respectivas probabilidades de default no será publicada en este trabajo debido al acuerdo de confidencialidad que se mantiene con el SII.

Para realizar la identificación geográfica se trabaja con el campo de unidad regional operativa del SII. Descartando un 9,7% de los contribuyentes por no poseer asignación a ninguna región, se observa que el comportamiento a nivel país es relativamente similar. El promedio del default es de un 14%, tendiéndose que la diferencia entre el máximo y el mínimo es de un 7%.

Se observa que en las tres primeras regiones, en donde se concentra un 6% de los contribuyentes de micro y pequeñas empresas del país, el porcentaje de default estimado es el más alto, llegando a ser de un 20%. Por otro lado, la Región de Magallanes concentra a un 16% de los contribuyentes y cuenta con un porcentaje de default estimado de 13,87%. Para ver el mapa del resultado por zona geográfica ir al anexo 21.

Se analiza de manera independiente la Región Metropolitana por ser la que concentra la mayor cantidad de contribuyentes de micro y pequeñas empresas (cerca de un 40%). Esta región cuenta con cuatro unidades operativas que actúan sobre las distintas comunas agrupándolas de acuerdo a su orientación geográfica. El mapa de resultados de la Región Metropolitana se puede ver en el anexo 22.

La zona de Santiago Oriente es la que muestra el menor porcentaje de default estimado del país. En ella se concentra el 39% de los contribuyentes de la Región Metropolitana y cuenta con un nivel de default de 12,53%. Las zonas de Santiago Poniente y Santiago Sur, presentan un porcentaje de default estimado mayor a la media del país. En ellas se concentra el 41% de los contribuyentes de la Región Metropolitana y muestran un nivel de default de 17,52%. Se destaca la gran heterogeneidad de comportamiento encontrada entre las distintas comunas.

8. CONCLUSIONES

Los modelos de default tributario construidos tienen una buena capacidad de predicción. Las estimaciones generadas indican que para el mes de julio del 2013 el default será de un 14,3% para los contribuyentes de micro y pequeñas empresas, observándose una mayor propensión a cometer default en la zona norte que en la zona sur del país.

Los resultados muestran una precisión general que logra identificar a un 69% de los casos de default, manteniendo la sensibilidad y especificidad dentro del mismo rango porcentual. Es importante destacar que debido a la complejidad del problema, el priorizar la precisión general y el balanceo en la predicción de clases lleva a que la tasa de error tipo II fuese cercana al 90%.

Este fenómeno se explica porque el modelo está identificando como defaulters aquellos contribuyentes que tengan una gran propensión a hacerlo dadas sus características pero que, sin embargo, no cometen este acto. Se confirma la necesidad de hacer segmentos que permitan diferenciar tipos de conducta y generar modelos distintos para cada uno.

Otra posible explicación se da al analizar la calidad de datos utilizada, pues, la base de datos inicial posee una cantidad de campos nulos importante, lo que puede estar generando que las atributos definidores no logren un impacto adecuado en la predicción[22].

Siendo este trabajo la primera aproximación que se tiene para predecir el default tributario, se considera que los resultados obtenidos y los modelos presentados son un buen punto de inicio para el posterior desarrollo de un modelo más preciso y con mejor capacidad predictiva. Las propuestas de mejora que se expondrán a continuación podrán ser relevantes para futuras investigaciones.

Cabe destacar que los modelos poseen falencias que deben ser consideradas al momento de diseñar un plan de acción. Así mismo, se recalca que el modelo solo será útil por un periodo de tiempo, y que al cabo de este se deberá hacer una reevaluación. Las razones se deben a que las características relevantes para el default pueden verse modificadas en el tiempo, o bien, a que de aplicarse alguna de las medidas propuestas, el universo de contribuyentes se vea afectado.

De la experiencia de trabajo se cree que un modelo de sobrevivencia tendría mejores aplicaciones e implicancias para el SII que un modelo de *credit scoring*. Dado que la mayor parte de los contribuyentes cae alguna vez en default, es poco relevante saber si lo harán o no. Resultaría más interesante el saber cuándo lo harán para diseñar el plan de acción preventivo.

A continuación se presentan en detalle las conclusiones de la memoria, las mejoras al trabajo hecho y finalmente se hace una propuesta al SII para utilizar la información derivada de este trabajo.

[22] El SII actualmente está trabajando en el mejoramiento de calidad de datos en dos ámbitos: controles y mejoras de caracter informaticos, y la educación a los contribuyentes con el objetivo que la información proporcionada por estos tienda a contener datos de mejor calidad.

8.1 CONCLUSIONES ESPECÍFICAS

En cuanto a los objetivos iniciales

Los resultados obtenidos en esta memoria cumplen con el objetivo principal, que es el diseño del mejor modelo predictivo que balancea las dos clases de conducta, enfocándose tanto en la mejora de la precisión general como en el nivel de especificidad y sensibilidad.

Dada la complejidad del problema, esto implica que el error de tipo II fuese alto, lo que refleja la existencia de un grupo de contribuyentes cuya conducta no ha sido definido correctamente con las variables construidas. Sin embargo, como esta es una primera aproximación a la aplicación de *credit scoring* en el ámbito tributario, esta baja tasa de especificidad es esperable.

Se cumplen también los objetivos específicos, pues se identifican las variables de mayor incidencia en el comportamiento de default y se genera una segmentación por zona geográfica, permitiendo localizar las regiones más críticas en el atraso de declaración del F29.

En cuanto a los datos

Debido al gran volumen de datos nulos contenidos en las variables iniciales se hace difícil encontrar patrones de conducta a partir de ellos. El porcentaje de campos vacíos tiene un promedio de 50%, siendo las variables seleccionadas las que presentan la mayor completitud de datos con un promedio de un 26% de campos nulos.

Se debe considerar que los registros vacíos para ciertos campos puede deberse en su mayoría al alto alcance de la casuística del F29, por lo que no se requiere que todos los códigos sean llenados por los contribuyentes.

La existencia de errores en el ingreso de datos, a pesar de ser baja, también merma la calidad de las variables. Dado que la información proporcionada por los contribuyentes en algunos casos no es precisa, aún no se cuenta con una base de datos lo suficientemente madura para trabajar en *credit scoring*.

Si se desea obtener modelos de mejor capacidad predictiva, es imperante revisar la consistencia y completitud de la base de datos. Hacer un análisis de las posibles razones de este déficit con la finalidad de diseñar planes de acción que lo mitiguen es un trabajo que debe llevarse a cabo.

El SII se encuentra trabajando para generar mejoras en la calidad de su Data Warehouse. Existen diversos proyectos que siguen esta senda y que permitirán en un futuro cercano alcanzar mejoras considerables en esta materia. Los proyectos se enfocan en dos ámbitos de acción: controles y mejoras de caracter informaticos, y educación a los contribuyentes, con la finalidad de mejorar la información proporcionada por estos.

En cuanto a las variables

De los modelos obtenido se desprende que son tres variables las que entregan más información: V1 (número de veces que declara atrasado el F29 el último año), V2 (cantidad de anotaciones "no declarante" el último año) y V4 (facturas de compra emitidas el último año). Estos atributos segmentan con claridad los siguientes tipos de comportamiento:

- Casos "inactivos" y defaulters.
  Los primeros corresponden a quienes cesan las ventas sin cierre de actividades, mientras que los segundos son los que con bastante frecuencia caen en default. En este grupo se considera a los contribuyentes que tienen una o más anotaciones del tipo "no declarante" y muy probablemente constituyen a los casos observados al inicio del análisis (aquellos que caen en default y dejan de declarar por largos periodos). Cabe destacar que el SII cuenta con el registro de este último grupo de contribuyentes, por lo que su identificación resulta más sencilla de lo esperado.
- Declarantes comprometidos.
  Son los contribuyentes más activos en el SII. Se caracterizan por su responsabilidad, ya pocas veces presentan conductas reprochables. No tienen anotaciones del tipo "no declarante" y tampoco tienen atrasos en la declaración del F29 por lo que su propensión a caer en default es casi nula.
- Declarantes intermitentes.
  Corresponde a quienes tienen declaraciones estacionales o esporádicas durante el año. No tienen anotaciones del tipo "no declarante", tampoco tienen atrasos, sin embargo las facturas de compra emitidas el último año son nulas, lo que podría indicar una inestabilidad del negocio y una posible falla en el cumplimiento de sus obligaciones.

Tanto en el modelo definitivo como en el desafiante las variables del historial de comportamiento son las más relevantes al momento de predecir el default. Aquellos contribuyentes que nunca han interrumpido sus declaraciones y que no poseen anotaciones de potenciales situaciones de comportamiento tributario irregular son los que presentan una menor probabilidad de caer en default en el futuro.

Se observa que el resto de las variables del modelo sirven para predecir default en el segmento de declarantes intermitentes más que en los otros dos. En este grupo el hecho de tener contador y haber iniciado actividades hace más de 10 años, disminuye la probabilidad de no declarar. Se tiene también que a medida que se acerca el mes de declaración del F22 aumenta la probabilidad de default de estos contribuyentes.

Es posible que los datos contables sean los que entreguen mayor información en cuanto al default. Sin embargo, como presentan un gran porcentaje de campos nulos no logran reflejar su importancia en el modelo, apareciendo como variables no influyentes. Se debe considerar la inclusión de variables no consideradas en este trabajo y que tengan más influencia para predecir el default. Se discutirá sobre esto en las propuestas de mejora.

En cuanto a la distribución geográfica de los defaulters

Se puede concluir que las regiones del norte tienen una mayor propensión a no declarar el F29, llegando a tener un porcentaje de default estimado para julio 2013 de 20%. Todo lo contrario ocurre en las regiones del sur y en la división Oriente y Centro de la Región Metropolitana, donde el default estimado para el mes de julio de este año desciende a 13% aproximadamente.

Cabe destacar que la mayor cantidad de contribuyentes de micro y pequeñas empresas se encuentra en la Región Metropolitana y zona centro del país (más de un 40%), dejando una concentración de tan solo 8% a la zona norte. Esto hace que el default promedio de Chile para julio 2013 tenga una estimación de 14%.

Falta analizar si la diferencia entre zonas geográficas se debe a la variable regional, o bien al tipo de actividad económica se que desarrolla en cada zona, pues, al ser Chile un país que vive de la explotación de recursos naturales, resulta lógico suponer que las actividades económicas estarán diferenciadas por región.

8.2 PROPUESTAS DE MEJORA

De acuerdo a los resultados obtenidos y a las conclusiones generales expuestas, las recomendaciones de mejora para futuros trabajos son las siguientes.

Mejorar la calidad de datos: Se propone un plan que involucre dos líneas de acción: mejoras en los sistemas de cruce online de las bases de datos y mejoras en el proceso de toma de datos e ingreso de la información. La primera es una medida de contingencia para el corto plazo y la segunda una medida estratégica para el largo plazo.

En cuanto a la mejora de sistemas de cruce online para la construcción de variables del modelo, se propone hacer un trabajo para completar los datos nulos con información contenida en otros formularios. Por ejemplo, es posible derivar algunos campos del F29 a partir del F22, y este último presenta una mayor tasa de declaración que la del F29 por lo que es más probable poder construir las variables requeridas.

Si bien no se reflejará fielmente la información mensual, se podrá obtener un estimado de muchos de los atributos que hasta ahora están vacíos. Esto permitirá dar una mayor preponderancia a las características contables de cada contribuyente, que pueden resultar relevantes para la detección de default.

Se debiese hacer lo mismo con los campos de características demográficas, pues, de acuerdo con la teoría de *credit scoring* bancario, muchas de las características que representa estas dos categorías resultan relevantes en la predicción. Construir un registro histórico de dichas características y derivar los valores de otras tablas podría mejorar la precisión de los modelos.

En cuanto a la mejora en el proceso de toma de datos e ingreso de la información, se debiese capacitar al personal de las oficinas presenciales del SII para que revisar la completitud de los formularios entregados. A la vez se debe capacitar a quienes transcriben los formularios manuales a la base del SII.

Agregar variables de entorno: Por razones de tiempo y disponibilidad de datos no se agregó todas las variables que se consideraba pudieran influir en el comportamiento de default. Los efectos cíclicos de la economía y las características intrínsecas de cada industria podrían estar afectando la propensión de default de los contribuyentes, pues reflejarían el crecimiento o estancamiento del país, la liquidez en el mercado, el volumen de flujos por industria, entre otros. Para mejorar el trabajo realizado se propone hacer pruebas incluyendo atributos que representen esos tres factores.

Segmentar contribuyentes: De los resultados obtenidos se infiere la diferencia de comportamiento entre grupos de contribuyentes. Haciendo una analogía con el *credit scoring* bancario se propone desarrollar tres modelos distintos para cada segmento de contribuyentes, pues, en cada uno puede estar afectando variables distintas al comportamiento.

Se propone distinguir entre pagadores tardíos, pagadores intermitentes y defaulters, ya que se demostró que las variables con mejor capacidad predictiva difieren entre un grupo y otro. Generar un cuarto grupo con los etiquetados por el SII como 'no declarantes por más de 12 meses consecutivos' resulta indispensable para eliminar el sesgo de individuos que ya no están activos en el SII.

Utilizar multiclasificadores: Resultaría interesante mejorar la capacidad predictiva haciendo una fusión del modelo definitivo y el modelo desafiante debido a que, como se observó en el capítulo de iteraciones, algunos destacan en la predicción de casos malos y otros que destacan en la de casos bueno.

Una forma de conseguir mejor precisión de los modelos es recurrir a la técnica de multiclasificadores, que permite integrar los resultados de diversas fuentes para obtener resultados más robustos y eficientes. Se propone experimentar con los métodos de *bootstraping* y *bagging* disponibles en el software *IBM SPSS Modeler*.

Cambiar el objetivo del modelo: Se propone a futuro crear un modelo de alta precisión, pero poca especificidad, es decir, que logre identificar bien a los defaulters más críticos, de tal manera de minimizar la tasa de error tipo II para hacer un plan de fiscalización activa sobre el grupo detectado.

8.3 RECOMENDACIONES DE USO

Dado el alto nivel de error en los falsos negativos, no es recomendable llevar a cabo una estrategia de fiscalización activa. Se propone tomar medidas que se inclinen hacia una estrategia de prevención, considerando el siguiente plan de acción:

Contribuyentes cuya probabilidad está entre 0,6 y 0,9:

En este grupo se encuentra un 1% de los contribuyentes de micro y pequeñas empresas. El porcentaje de error tipo II es de 84,7%, pronosticando correctamente a un 15,3% de los defaulters.

La propuesta en este caso consiste en utilizar un medio masivo de comunicación para informar que se acerca una nueva fecha de declaración, evitando el default de aquellos que siempre declaran y pudiesen estarlo olvidando. Se recomienda a través de este medio insistir en la actualización de sus datos en la página web y se debiese instar

a hacer cierre de actividades a aquellos contribuyentes que no lo hayan hecho. Así mismo, se puede informar como medida disuasiva de las consecuencias tributarias que implica la no presentación del F29.

Contribuyentes cuya probabilidad está entre 0,9 y 1:

En este grupo se encuentra un 4% de los contribuyentes de micro y pequeñas empresas. El porcentaje de error tipo II es de 77,2% pronosticando correctamente a un 22,8% de los defaulters, lo que es una mejora considerable con respecto a los otros puntos de corte.

Se recomienda realizar una fiscalización presencial, pues existe la alta probabilidad de que muchos de ellos hayan cesado ventas sin hacer un correcto cierre de actividades. Se propone iniciar la aplicación de medidas disuasivas a los contribuyentes que lleven más de 12 meses sin declarar el F29, permitiendo así mantener una base de datos más limpia y evitando futuras fiscalizaciones innecesarias.

Dado que el SII no tiene la facultad de prohibir la creación de una empresa, o realizar el cierre legal de las mimas, se propone como medidas disuasivas impedir el timbraje de documentos o restringir la emisión de los mismos.

9. BIBLIOGRAFÍA

[1] **Bischop, C.M.** 1995. Neural Networks for Pattern Recognition. Editor Clarendon Press, Oxford, England.

[2] **Burke, S.** 2001. Missing values, outliers, robust statistics and Non parametric methods. RHM Technology Ltda, Europe Online Sumpplements.

[3] **Chen, K., & Chu, C.** 2001. Internal control versus external manipulation: a model of corporate income tax evasion. RAND Journal of Economics.

[4] **El Mercurio Online (EMOL).** Marzo 2012. Director del SII reconoce interés en reducir tasa de evasión del IVA hasta menos del 10%. [en línea] <http://www.emol.com/noticias/economia/2012/03/23/532359/director-del-sii-reconoce-interes-en-reducir-tasa-de-evasion-del-iva--hasta-menos-del-10.html> [consulta: 10 de mayo 2013]

[5] **Fayyad, U., Piatet Sk-Shapiro, G., Smyth, P.** 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

[6] **Gianotti, F., Mainetto, G., & Pedreschi, D.** 1999. A classification-based methodology for planning audit strategies in fraud detection. 1999. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining.

[7] **Gobierno de Chile.** Mayo 2012. Ministerio de Hacienda y director del SII anuncian virtual cumplimiento de meta oficial de reducción de evasión. [en línea] <http://www.gob.cl/informa/2012/05/10/ministerio-de-hacienda-y-director-del-sii-anuncian-virtual-cumplimiento-de-meta-oficial-de-reduccion-d.htm> [consulta: 10 de mayo 2013]

[8] **Jorrat, M., & Serra, P.** 2000. Estimación de la evasión en el impuesto a las empresas en Chile. Santiago.

[9] **Murray, B.** 2006. Practical Credit Scoring: Issues and Techniques. White Box Publishing.

[10] **Servicio de Impuestos Internos.** 2012. Cuenta Pública 2011. [en línea] <http://www.sii.cl/cuenta_publica/cta_2011.pdf> [consulta: 10 de mayo]

[11] **Servicio de Impuestos Internos.** 2013. Diccionario Básica Tributario. [en línea] <http://www.sii.cl/diccionario_tributario/dicc_a.htm> [consulta: 10 de mayo]

[12] **Servicio de Impuestos Internos.** 2013. Empresas por tamaño. [en línea] <http://www.sii.cl/cpntribuyentes/contribuyentes.htm> [consulta: 10 de mayo]

[13] **Servicio de Impuestos Internos.** 2013. Estadísticas de empresa por tramo de ventas y actividad económica. [Archivo Excel] <http://www.sii.cl/estadísticas/empresas.htm> [consulta: 10 de mayo]

[14] **Servicio de Impuestos Internos.** 2012. Legislación Tributaria. [en línea] <http://www.sii.cl/pagina/jurisprudencia/ley.htm> [consulta: 07 de mayo 2013]

[15] **Servicio de Impuestos Internos.** 2013. Misión, Objetivos y Lineamientos del SII. [en línea] <http://www.sii.cl/sobre_el_sii/acerca/mision.htm> [consulta: 10 de mayo]

[16] **Siddiqi, N.** 2006. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. John Wiley & Sons, Inc.

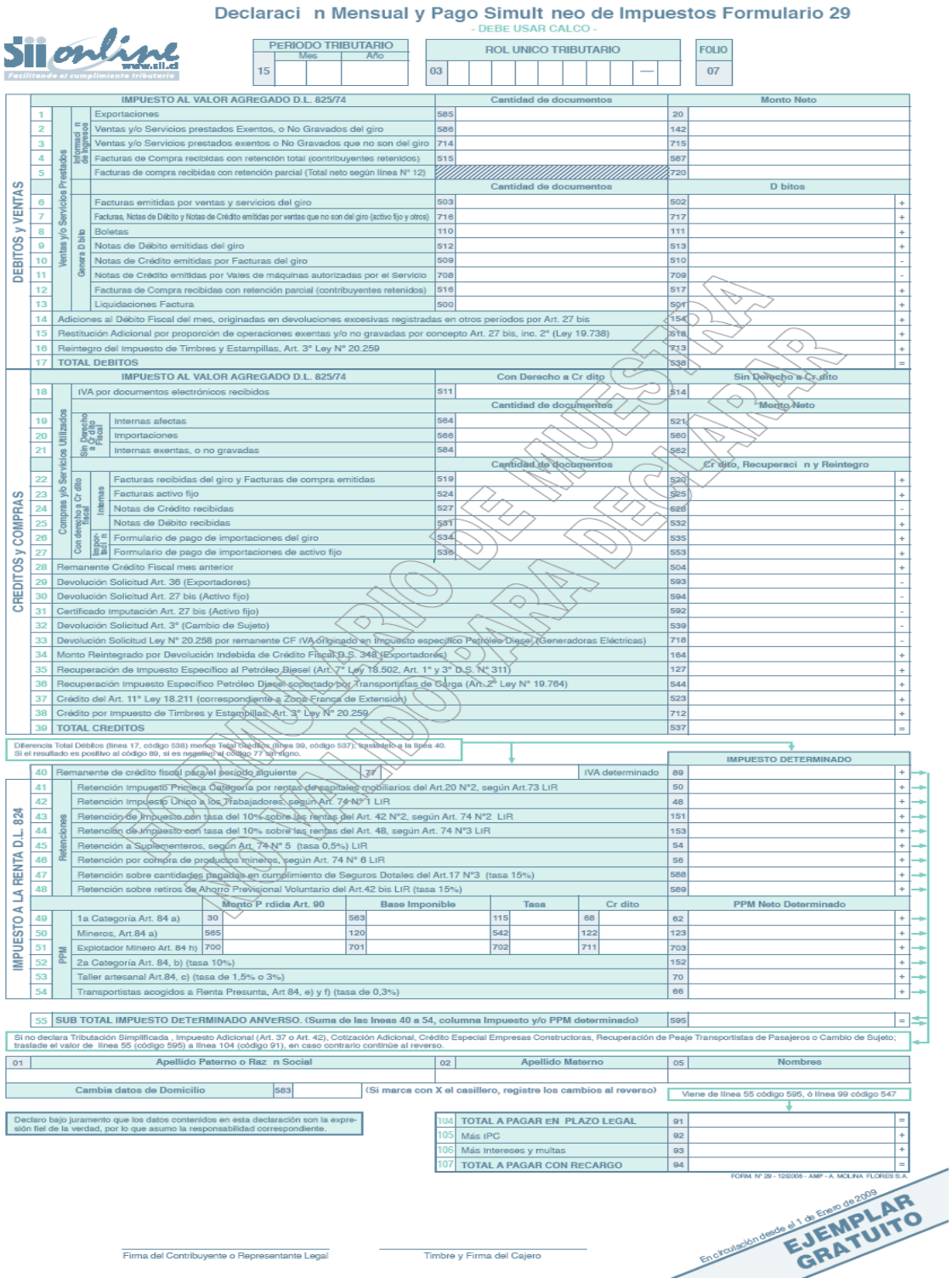
[17] **Stein, R.** 2002. Benchmarking Default Prediction Modesl: Pitfalls and Remedies in model Validation. Nueva York: Moodys KMV.

[18] **Two Crow Corporation.** 1999. Introduction to Data Mining, third edition.

[19] **Yu, F**. 2003. Data mining application issues in fraudulent tax declaration detection . Machine Learning and Cybernetics international conference, Vol.4.

10. ANEXOS

ANEXO 1: FORMULARIO 29



Declaraci n Mensual y Pago Simult neo de Impuestos Formulario 29

- DEBE USAR CALCO -

SII online — Facilitando el cumplimiento tributario

PERIODO TRIBUTARIO		
	Mes	Año
15		

ROL UNICO TRIBUTARIO
03 —

FOLIO
07

DEBITOS y VENTAS

			IMPUESTO AL VALOR AGREGADO D.L. 825/74	Cantidad de documentos		Monto Neto	
1	Ventas y/o Servicios Prestados — 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	Compras y/o Servicios Utilizados		IVA por documentos electrónicos recibidos	511		514	
				Cantidad de documentos		Monto Neto	
19		Sin Derecho a Cr dito Fiscal	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		Con derecho a Cr dito fiscal — Internas	Facturas recibidas del giro y Facturas de compra emitidas	519		520	+
23			Facturas activo fijo	524		525	+
24			Notas de Crédito recibidas	527		526	-
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	Devolución Solicitud Ley N° 20.258 por remanente CF IVA originado en Impuesto específico Petróleo Diesel (Generadoras Eléctricas)					718	-
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 Extensión)					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 DETERMINADO	
40	Remanente de crédito fiscal para el período siguiente	77	IVA determinado	89		+

IMPUESTO A LA RENTA D.L. 824

41	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		+

PPM	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)	565		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	+

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 - AMP - A. MOLINA FLORES S.A.

Firma del Contribuyente o Representante Legal

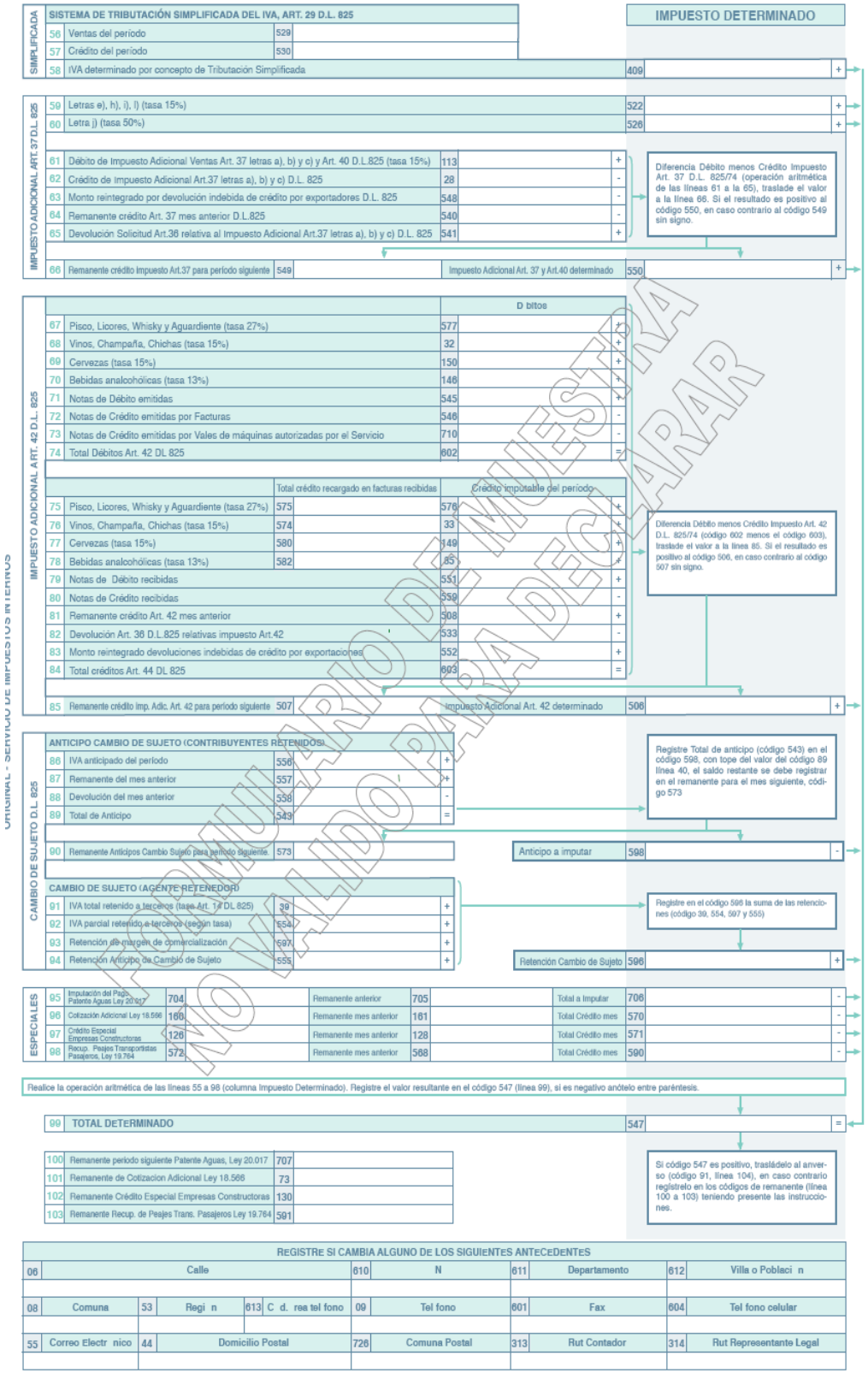
Timbre y Firma del Cajero

En circulación desde el 1 de Enero de 2009

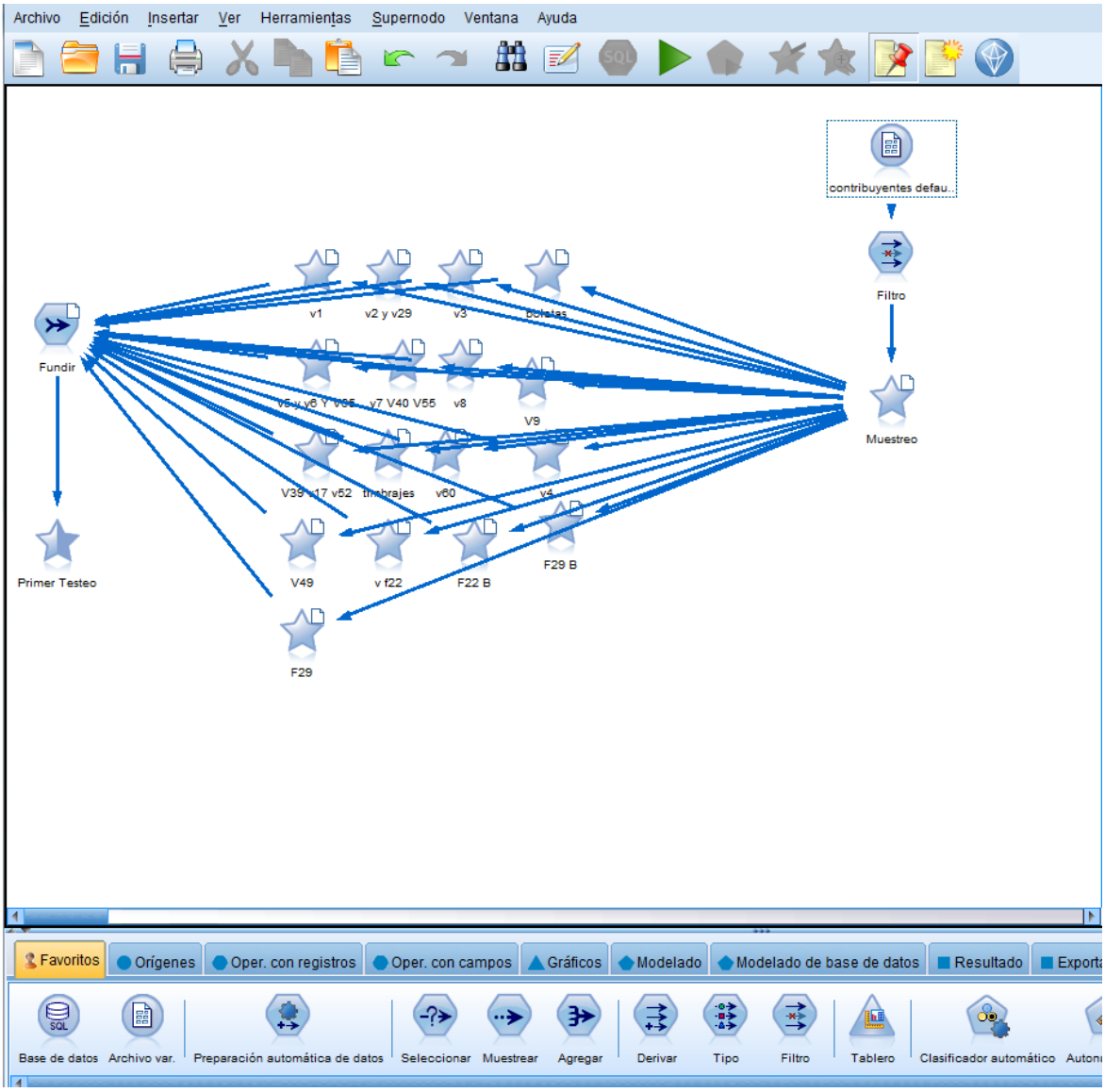
EJEMPLAR GRATUITO

ORIGINAL

- DEBE USAR CALCO -



ANEXO 2: INTERFAZ DE IBM SPSS MODELER



ANEXO 3: CAMPOS SELECCIONADOS

Tabla	Nombre del Campo
DW_TRN_CONTRIBUYENTES_HISTORICOS	TICO_SUB_TPO_CONTR
DW_TRN_CONTRIBUYENTES_HISTORICOS	CONH_RAZON_SOCIAL_VO
DW_TRN_DIRECCIONES	DIRE_COD_COMUNA_VO
DW_TRN_DIRECCIONES	DIRE_COD_TIPO_PROPIETARIO_VO
DW_TRN_F22	F22_C_101
DW_TRN_F22	F22_C_102
DW_TRN_F22	F22_C_113
DW_TRN_F22	F22_C_120
DW_TRN_F22	F22_C_122
DW_TRN_F22	F22_C_123
DW_TRN_F22	F22_C_129
DW_TRN_F22	F22_C_167
DW_TRN_F22	F22_C_18
DW_TRN_F22	F22_C_181
DW_TRN_F22	F22_C_187
DW_TRN_F22	F22_C_188
DW_TRN_F22	F22_C_19
DW_TRN_F22	F22_C_195
DW_TRN_F22	F22_C_226
DW_TRN_F22	F22_C_227
DW_TRN_F22	F22_C_231
DW_TRN_F22	F22_C_232
DW_TRN_F22	F22_C_284
DW_TRN_F22	F22_C_318
DW_TRN_F22	F22_C_36
DW_TRN_F22	F22_C_365
DW_TRN_F22	F22_C_366
DW_TRN_F22	F22_C_373
DW_TRN_F22	F22_C_382
DW_TRN_F22	F22_C_384
DW_TRN_F22	F22_C_628
DW_TRN_F22	F22_C_629
DW_TRN_F22	F22_C_630
DW_TRN_F22	F22_C_631
DW_TRN_F22	F22_C_632
DW_TRN_F22	F22_C_633
DW_TRN_F22	F22_C_634
DW_TRN_F22	F22_C_635
DW_TRN_F22	F22_C_636
DW_TRN_F22	F22_C_640

DW_TRN_F22	F22_C_641
DW_TRN_F22	F22_C_642
DW_TRN_F22	F22_C_643
DW_TRN_F22	F22_C_645
DW_TRN_F22	F22_C_646
DW_TRN_F22	F22_C_647
DW_TRN_F22	F22_C_648
DW_TRN_F22	F22_C_651
DW_TRN_F22	F22_C_74
DW_TRN_F22	F22_C_761
DW_TRN_F22	F22_C_768
DW_TRN_F22	F22_C_77
DW_TRN_F22	F22_C_773
DW_TRN_F22	F22_C_777
DW_TRN_F22	F22_C_778
DW_TRN_F22	F22_C_779
DW_TRN_F22	F22_C_783
DW_TRN_F22	F22_C_784
DW_TRN_F22	F22_C_785
DW_TRN_F22	F22_C_807
DW_TRN_F22	F22_C_816
DW_TRN_F22	F22_C_817
DW_TRN_F22	F22_C_82
DW_TRN_F29	F29_C_111
DW_TRN_F29	F29_C_142
DW_TRN_F29	F29_C_154
DW_TRN_F29	F29_C_20
DW_TRN_F29	F29_C_39
DW_TRN_F29	F29_C_409
DW_TRN_F29	F29_C_501
DW_TRN_F29	F29_C_502
DW_TRN_F29	F29_C_510
DW_TRN_F29	F29_C_513
DW_TRN_F29	F29_C_517
DW_TRN_F29	F29_C_518
DW_TRN_F29	F29_C_520
DW_TRN_F29	F29_C_521
DW_TRN_F29	F29_C_525
DW_TRN_F29	F29_C_528
DW_TRN_F29	F29_C_529
DW_TRN_F29	F29_C_530
DW_TRN_F29	F29_C_532
DW_TRN_F29	F29_C_535

DW_TRN_F29	F29_C_543
DW_TRN_F29	F29_C_553
DW_TRN_F29	F29_C_554
DW_TRN_F29	F29_C_555
DW_TRN_F29	F29_C_556
DW_TRN_F29	F29_C_557
DW_TRN_F29	F29_C_558
DW_TRN_F29	F29_C_560
DW_TRN_F29	F29_C_562
DW_TRN_F29	F29_C_587
DW_TRN_F29	F29_C_596
DW_TRN_F29	F29_C_597
DW_TRN_F29	F29_C_709
DW_TRN_FISCALIZACION_SELECTIVA	FSE_DEV27BIS_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_SOLI27BIS_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_EN_PARTE_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_IVA_VO
DW_TRN_NEGOCIOS	NEGO_FECHA_CREACION_VO
DW_TRN_NEGOCIOS	NEGO_FECHA_INICIO_VO
DW_TRN_NEGOCIOS	CONT_RUT
DW_TRN_NEGOCIOS	CONT_RUT_ALIAS
DW_TRN_NEGOCIOS	NEGO_CAPITAL_ENTERADO_VO
DW_TRN_NEGOCIOS	NEGO_CAPITAL_POR_ENTERAR_VO
DW_TRN_NEGOCIOS	NEGO_NRO_FACTURAS_6MESES_VO
DW_TRN_NEGOCIOS	NEGO_RUT_VO
DW_TRN_NEGOCIOS	NEGO_RUT_VO_ALIAS
DW_TRN_NEGOCIOS	NEGO_OBLIGADO_IVA_AGNO_4_IC
DW_TRN_NEGOCIOS	NEGO_OBLIGADO_IVA_AGNO_ANT_IC
DW_TRN_NEGOCIOS	NEGO_OBLIGADO_IVA_ANTEANTERIOR_IC
DW_TRN_NEGOCIOS	NEGO_OBLIGADO_IVA_ENCURSO_IC
DW_TRN_NEGOCIOS	NEGO_CLA_CONTROL_DOCTO_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_FECHA_INICIO_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_FECHA_TERMINO_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_FECHA_VIGENCIA
DW_TRN_REPRESENTANTES	REPR_FECHA_INICIO_VO
DW_TRN_REPRESENTANTES	REPR_FECHA_TERMINO_VO
DW_TRN_REPRESENTANTES	CONT_RUT
DW_TRN_REPRESENTANTES	CONT_RUT_REPRESENTANTE
DW_TRN_CONTRIBUYENTES	CONT_FECHA_CREACION_VO
DW_TRN_CONTRIBUYENTES	CONT_FECHA_NACIMIENTO_VO
DW_TRN_CONTRIBUYENTES	TICO_SUB_TPO_CONTR
DW_TRN_CONTRIBUYENTES	CONT_COD_PAIS_VO
DW_TRN_CONTRIBUYENTES	CONT_RUT

DW_TRN_TIMBRAJES	TIMB_FECHA_LEGALIZACION_VO
DW_TRN_TIMBRAJES	CONT_RUT
DW_TRN_TIMBRAJES	TIMB_COD_TIPO_DOCTO_VO
DW_TRN_TIMBRAJES	TIMB_CODIGO_VO
DW_TRN_TIMBRAJES	TIMB_NUMERO_FINAL_VO
DW_TRN_TIMBRAJES	TIMB_NUMERO_INICIAL_VO
DW_TRN_TIMBRAJES	TIMB_DESC_TIPO_DOCTO_VO
DW_TRN_TIMBRAJES	TIMB_ESTADO_TIMBRAJE_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_RUT_VO
DW_TRN_ACTIVIDAD_ECONOMICA	CONT_RUT
DW_TRN_ACTIVIDAD_ECONOMICA	ACEC_COD_ACTECO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_AFECTA_IVA_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_CATEGORIA_TRIBUTARIA_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_COD_ACTECO_VO
DW_TRN_ACTIVIDAD_ECONOMICA	ACTECO_VIGENCIA
DW_TRN_FISCALIZACION_SELECTIVA	FSE_FECHA_NOTIF_VO
DW_TRN_ALERTAS	ALER_DESC_TIPO_ALERTA_VO
DW_HEC_CONT_COMPORTAMIENTO	PERI_AGNO_MES_ANALISIS
DW_HEC_CONT_COMPORTAMIENTO	TRRE_COD_TMO_RTA
DW_HEC_CONT_COMPORTAMIENTO	TRVE_COD_TMO_VTA
DW_TRN_FISCALIZACION_SELECTIVA	FSE_OBSERVACI3_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_OBSERVACI4_VO
DW_HEC_CONT_COMPORTAMIENTO	COMU_COD_COMUNA_PRINCIPAL
DW_HEC_CONT_COMPORTAMIENTO	PERI_AGNO_TRIBUTARIO_RENTA
DW_TRN_ALERTAS	ALER_FECHA_DESACTIV_VO
DW_HEC_CONT_COMPORTAMIENTO	ACEC_COD_ACTECO_PRINCIPAL
DW_TRN_ALERTAS	ALER_FECHA_ACTIV_VO
DW_HEC_CONT_COMPORTAMIENTO	COCO_AGNO_COMERCIAL
DW_TRN_ALERTAS	ALER_CODIGO_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_OBSERVACI1_VO
DW_TRN_FISCALIZACION_SELECTIVA	FSE_OBSERVACI2_VO
DW_HEC_CONT_COMPORTAMIENTO	COCO_MTO_VENTAS
DW_TRN_FISCALIZACION_SELECTIVA	FSE_FECHA_CITACION_VO
DW_HEC_CONT_COMPORTAMIENTO	TICO_SUB_TPO_CONTR
DW_TRN_ALERTAS	ALER_COD_TIPO_ALERTA_VO
DW_HEC_CONT_COMPORTAMIENTO	COCO_COD_ACTECO_F22

ANEXO 4: VARIABLES INDEPENDIENTE

N°	Variable	Categoría
1	N° de veces que declara atrasado el F29 en el último año	7
2	N° de anotaciones no declarante de F29 teniendo que declarar en último año	7
3	N° de facturas timbradas en el último año	5
4	N° de facturas de compra emitidas en el último año	3
5	N° de anotaciones graves y muy graves MIPE en los últimos 2 años	8
6	Tiempo desde la última anotación grave o muy grave MIPE	8
7	N° de ampliaciones de giro a otro rubro	4
8	Meses de antigüedad desde inicio de actividad	1
9	N° de boletas timbradas	3
10	(Boletas emitidas/boletas timbradas) del último año	6
11	N° de boletas emitidas	3
13	Promedio de débito de los últimos 12 meses	3
14	Promedio de crédito de los últimos 12 meses	3
15	Promedio de (débito/crédito) de los últimos 12 meses	3
16	Meses desde el último período con pérdida tributaria	2
17	Tramo de renta al que pertenece el contribuyente el último año	5
18	Variación de los montos de renta imponible penúltimo y último año	5
19	Tiene contador el último año	2
20	Declara renta efectiva último año	2
21	% renta efectiva sobre renta total en el último año	2
22	Tiene contabilidad completa	2
23	N° de rectificaciones del F29 en los últimos dos años	8
25	Presenta reorganización empresarial en los últimos 2 años	1
26	Variación de la pérdida tributaria en los últimos dos años	3
27	N° de períodos con pérdida tributaria en los últimos dos años	3
28	N° de períodos desde el último período con remanente fiscal (desde el 2007)	3
29	Promedio (compra/venta) de los últimos 12 meses	5
31	Monto de renta declarada último año	5
32	Declaración F22 del año pasado fue presencial	2
33	Períodos tributarios desde la última declaración F22	2
37	% variación de las ventas máximas y mínimas de los últimos 6 meses	5
40	Cantidad de ACTECOS vigentes	4
42	Nivel de ingresos del último F22	5
47	N° de notas de débito timbradas el último año	2
49	Cantidad de fiscalizaciones que ha tenido desde el 2007	8
51	N° de períodos con remanente de crédito fiscal en los últimos dos años	3
52	Nivel de ventas anuales del último año	5

54	Promedio de nivel de ingresos en los últimos dos años	5
55	Tiene algún ACTECO tipo "saco"	4
58	N° Notas de crédito timbradas último año	2
59	Promedio del nivel de pago del F29 en el último año	5
60	Edad del representante legal	1
61	Promedio de nivel de pago del F22 en los últimos 2 años	5
62	N° de facturas de compra timbradas el último año	3
63	N° de boletas de honorario timbradas el último año	3
64	% variación del nivel máximo y mínimo del F29 en el último año	5
65	N° de anotaciones de comercio informal vigente	8
66	Monto total de rectificaciones	7
69	(N° de rectificaciones/N° declaraciones) del F29 en los últimos 2 años	7
70	Monto de renta imponible último año	5
71	Declara renta presunta el último año	2
72	Tiene contabilidad simplificada	2
29.a	Promedio de compra de los últimos 12 meses	5
29.b	Promedio del % de variación (compra/venta) de los últimos 12 meses	5
37.a	Promedio de ventas de los últimos 6 meses	5

ANEXO 5: ESTADÍSTICOS DESCRIPTIVOS PREVIO A LA LIMPIEZA DE DATOS

Variable	Medida	Media	Desviación	Mínimo	Máximo
1	discreta	2	2	1	12
2	discreta	3	2	1	12
3	discreta	157	2.332	1	999.999
4	discreta	1.244	126.136	1	49.218.848
5	discreta	3	5	1	84
6	discreta	3	4	2	27
7	discreta	2	1	1	37
8	discreta	131	73	1	1.342
9	discreta	10.946	18.940	1	1.050.000
10	continua	9	903	0	312.366
11	discreta	9.391	159.053	1	48.359.880
13	continua	522.708	1.025.708	-467.918	42.307.392
14	continua	933.363	5.273.417	-53.274	982.746.406
15	continua	9	2.045	-221	958.134
16	discreta	2	1	2	27
18	discreta	389.898	5.411.416	-383.229.603	314.720.305
21	continua	99	8,54	1	100
23	discreta	2	2,44	1	40
26	discreta	26	2.854	-1	601.547
27	discreta	7	8	1	24
28	discreta	11	13	1	69
29	continua	3	838	-3.196	467.372
31	discreta	5.653.555	10.832.105	1	1.386.708.530
33	discreta	7	4	1	23
37	continua	0,43	0,41	0	27
40	discreta	3	2	1	42
42	discreta	21.525.747	69.767.899	-592.344	10.429.660.742
47	discreta	39	370	1	11.000
49	discreta	1	5	1	358
51	discreta	10	8	1	24
52	discreta	34.310.957	65.249.275	1	1.881.258.632
54	continua	37.628.420	69.952.375	-592.344	5.214.830.371
58	discreta	28	163	1	10.200
60	discreta	52	14	8	111
61	continua	1.236.528	6.244.976	0,5	1.337.450.589
62	discreta	49	198	1	5.850
63	discreta	369	1628	1	59.500
64	continua	0,23	0,82	-1	1
66	discreta	39.771.659	177.539.531	89	4.049.902.245
69	continua	0,09	0,09	0	1

70	discreta	5.722.200	10.909.766	1	1.386.708.530
29.a	continua	1.948.211	4.085.724	-2319550	112.368.469
29.b	continua	127.423	72.172.550	-16660344343	36.004.600.443
37.a	continua	3.311.172	6.362.587	-3716636	203.405.508
59	continua	209.359	5.394.486	0	3.675.979.800

Variable	Medida	Moda	Mediana
17	ordinal	1	1
19	dummy	1	1
20	dummy	0	0
22	dummy	1	1
25	dummy	0	0
32	dummy	0	0
55	dummy	0	0
71	dummy	0	0
72	dummy	0	0
65	discreta	1	1

ANEXOS 6: RESUMEN DE TRATAMIENTO DE DATOS NULOS

Variable	Tratamiento	Cantidad de Nulos	%
1	reemplazar por 0	452.661	79,8%
2	reemplazar por 0	454.192	80,1%
3	reemplazar por 0	338.643	59,7%
4	reemplazar por 0	184.310	32,5%
5	reemplazar por 0	445.225	78,5%
6	nada	0	
7	reemplazar por 0	194.523	34,3%
8	reemplazar por -1	1.531	0,3%
9	reemplazar por 0	414.198	73,0%
10	reemplazar por 0	416.385	73,4%
11	reemplazar por 0	334.787	59,0%
13	reemplazar por 0	180.073	31,7%
14	reemplazar por 0	153.059	27,0%
15	reemplazar por 0	184.894	32,6%
16	nada	0	
18	reemplazar por 0	296.308	52,2%
21	reemplazar por 0	352.122	62,1%
23	reemplazar por 0	480.330	84,7%
26	reemplazar por 0	487.290	85,9%
27	reemplazar por 0	532.497	93,9%
28	reemplazar por -1	214.276	37,8%
29	reemplazar por 0	191.700	33,8%
31	reemplazar por 0	264.000	46,5%
33	reemplazar por -1	28.656	5,1%
37	nada	0	
40	nada	0	
42	nada	0	
47	reemplazar por 0	535.384	94,4%
49	reemplazar por 0	534.420	94,2%
51	reemplazar por 0	262.401	46,3%
52	reemplazar por 0	147.368	26,0%
54	reemplazar por 0	248.230	43,8%
58	reemplazar por 0	517.243	91,2%
60	reemplazar por 0	148.227	26,1%
61	reemplazar por 0	470.006	82,9%
62	reemplazar por 0	534.082	94,2%
63	reemplazar por 0	531.174	93,6%
64	reemplazar por 0	30.624	5,4%
66	reemplazar por 0	537.101	94,7%
69	reemplazar por 0	480.330	84,7%

70	reemplazar por 0	262.232	46,2%
29.a	reemplazar por 0	178.087	31,4%
29.b	reemplazar por 0	203.242	35,8%
37.a	reemplazar por 0	192.923	34,0%
59	nada	0	
17	nada	0	
19	nada	0	
20	nada	0	
22	nada	0	
25	nada	0	
32	nada	0	
55	nada	0	
71	nada	0	
72	nada	0	
65	nada	0	

ANEXOS 7: RESUMEN DE TRATAMIENTO DE OUTLIERS

Variable	Atípicos	Extremos	Tratamiento
1	2.237	0	nada
2	2.672	0	nada
3	184	77	borrar
4	32	99	borrar
5	998	814	nada
6	14.518	4.478	nada
7	4.859	799	nada
8	284	46	borrar
9	1.338	747	borrar
10	36	19	borrar
11	49	89	borrar
13	6.887	2.948	borrar
14	893	996	borrar
15	5	10	borrar
16	385	3.472	nada
17	0	0	nada
18	3.081	1.237	borrar
19	0	0	nada
20	0	0	nada
21	1.090	2.406	nada
22	0	0	nada
23	896	390	borrar
25	0	0	nada
26	3	5	borrar
27	0	0	nada
28	5.432	0	nada
29	2	15	borrar
31	3.003	990	borrar
32	0	0	nada
33	1.440	0	nada
37	48	13	borrar
40	8.576	1.207	nada
42	7.444	4.224	borrar
47	1	17	borrar
49	1	2	borrar
51	0	0	nada
52	7.965	3.094	borrar
54	6.333	1.736	borrar
55	0	0	nada
58	28	62	borrar

59	62	94	borrar
60	344	0	nada
61	173	101	borrar
62	26	22	borrar
63	15	20	borrar
64	0	0	nada
65	0	0	nada
66	6	10	nada
69	1.018	421	nada
70	3.131	987	borrar
71	0	0	nada
72	0	0	nada
29.a	6.154	3.112	borrar
29.b	3	13	borrar
37.a	6.758	2.532	borrar

ANEXO 8: IMPORTANCIA DE VARIABLES

Variable	Importancia	Eliminada
66	0,031	si
29	0,0443	si
29.b	0,12	si
7	0,329	si
47	0,39	si
25	0,523	si
26	0,765	si
16	0,89	no
49	0,958	no
18	0,976	no
10	0,992	no
63	0,998	no
4	0,999	no
62	0,999	no
1	1	no
2	1	no
3	1	no
5	1	no
6	1	no
8	1	no
9	1	no
11	1	no
13	1	no
14	1	no
15	1	no
21	1	no
23	1	no
27	1	no
28	1	no
31	1	no
33	1	no
37	1	no
40	1	no
42	1	no
51	1	no
52	1	no
54	1	no
58	1	no
60	1	no
61	1	no

64	1	no
69	1	no
70	1	no
29.a	1	no
37.a	1	no
59	1	no
17	1	no
19	1	no
20	1	no
22	1	no
32	1	no
55	1	no
71	1	no
72	1	no
65	1	no

ANEXO 9: RESUMEN DE MATRIZ DE CORRELACIÓN

Variable	42	54	52	70	20	31	17	21	23	69	37.a	29.a	13	40	7
42	1,00														
54	0,89	1,00													
52	0,75	0,85	1,00												
70	*	*	*	1,00											
20	*	*	*	*	1,00										
31	*	*	*	1,00	*	1,00									
17	*	*	*	0,83	*	0,83	1,00								
21	*	*	*	*	0,97	*	*	1,00							
23	*	*	*	*	*	*	*	*	1,00						
69	*	*	*	*	*	*	*	*	0,94	1,00					
37.a	*	*	0,71	*	*	*	*	*	*	*	1,00				
29.a	*	*	*	*	*	*	*	*	*	*	0,83	1,00			
13	*	*	0,73	*	*	*	*	*	*	*	0,89	0,88	1,00		
40	*	*	*	*	*	*	*	*	*	*	*	*	*	1,00	
7	*	*	*	*	*	*	*	*	*	*	*	*	*	0,90	1,00

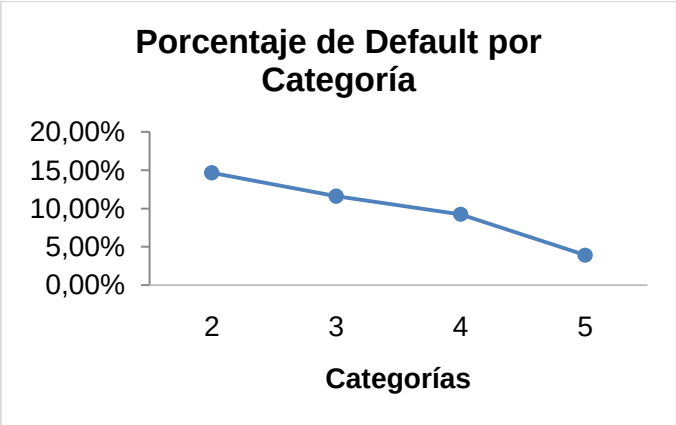
ANEXO 10: VARIANZA TOTAL EXPLICADA POR MÉTODO ACP

Componente	Total	% de la Varianza	% Acumulado
1	6,640	12,072	12,072
2	3,088	5,614	17,686
3	2,279	4,143	21,829
4	2,191	3,983	25,812
5	2,141	3,893	29,705
6	2,007	3,649	33,354
7	1,640	2,981	36,336
8	1,569	2,852	39,187
9	1,482	2,695	41,882
10	1,451	2,639	44,520
11	1,281	2,329	46,849
12	1,271	2,311	49,161
13	1,165	2,118	51,279
14	1,094	1,990	53,268
15	1,076	1,956	55,224
16	1,063	1,933	57,158
17	1,029	1,870	59,028
18	1,014	1,844	60,873
19	1,004	1,826	62,699
20	1,001	1,820	64,520
21	0,996	1,811	66,331
22	0,991	1,802	68,132
23	0,982	1,785	69,918
24	0,974	1,771	71,689
25	0,953	1,733	73,422
26	0,947	1,721	75,143
27	0,926	1,683	76,826
28	0,866	1,574	78,400
29	0,853	1,550	79,950
30	0,831	1,510	81,460
31	0,808	1,468	82,928
32	0,774	1,408	84,336
33	0,754	1,371	85,707
34	0,735	1,336	87,043
35	0,728	1,323	88,365
36	0,718	1,305	89,671
37	0,710	1,291	90,962
38	0,680	1,236	92,198
39	0,611	1,112	93,309
40	0,531	0,966	94,275

41	0,517	0,941	95,215
42	0,447	0,812	96,027
43	0,374	0,680	96,708
44	0,311	0,565	97,273
45	0,287	0,522	97,794
46	0,246	0,447	98,242
47	0,232	0,422	98,664
48	0,220	0,400	99,064
49	0,150	0,272	99,336
50	0,098	0,179	99,515
51	0,091	0,165	99,680
52	0,082	0,150	99,829
53	0,069	0,125	99,954
54	0,023	0,043	99,997
55	0,002	0,003	100,000

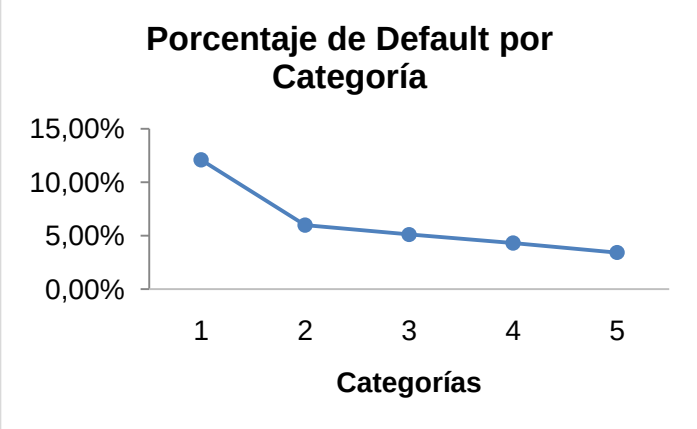
ANEXO 11: CATEGORIZACIÓN DE VARIABLES.

Variable 6



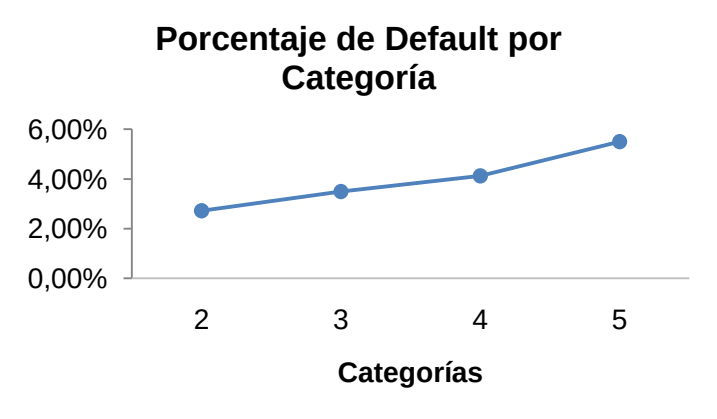
Categoría	Reglas	
1		< 5
3	>= 5	< 6
4	>= 6	< 10
5	>= 10	Y =0

Variable 8



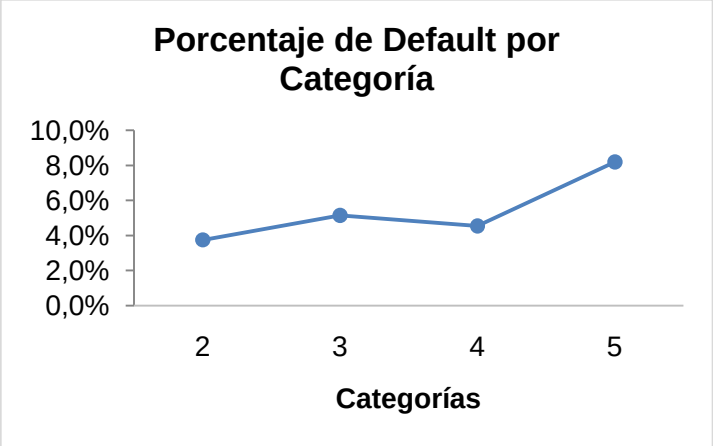
Categoría	Reglas	
1		< 1
2	>= 1	< 55
3	>= 55	< 154
4	>= 154	< 194
5	>= 194	

Variable 28



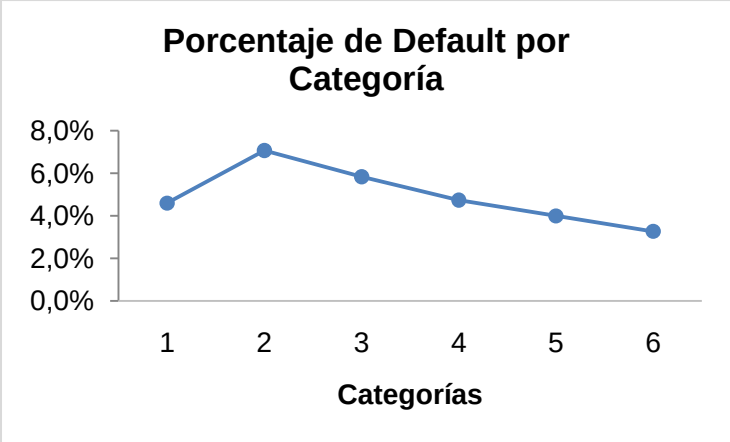
Categoría	Reglas	
2	>= 1	< 2
3	>= 2	< 4
4	>= 4	< 8
5	>= 8	Y <1

Variable 33



Categoría	Reglas	
2	>= 1	< 8
3	>= 8	< 10
4	>= 10	< 13
5	>= 13	Y <1

Variable 60



Categoría	Reglas	
1		nulos
2	>= 19	< 34
3	>= 34	< 43
4	>= 43	< 54
5	>= 54	< 66
6	>= 66	

ANEXO 12: DETALLE DE TRANSFORMACIÓN DE VARIABLES.

El tratamiento 1 corresponde a las transformaciones de escala aplicadas para la prueba general de los árboles de decisión. El tratamiento 2 corresponde al tratamiento aplicado posterior el tratamiento 1, para cumplir con los requerimientos de las técnicas de redes neuronales y regresión logística.

Variable	Tratamiento 1	Tratamiento 2
1	min-max	
2	min-max	
3	logN	min-max
4	logN	min-max
5	min-max	
6	categorizar	dummy
7	min-max	
8	categorizar	dummy
9	logN	min-max
10	min-max	
11	logN	min-max
13	logN	min-max
14	logN	min-max
15	min-max	
16	min-max	
18	min-max	
21	min-max	
23	min-max	
26	logN	min-max
27	min-max	
28	categorizar	dummy
29	min-max	
31	logN	min-max
33	categorizar	dummy
37	inversa	min-max
40	min-max	
42	min-max	
47	min-max	
49	min-max	
51	min-max	
52	min-max	
54	logN	min-max
58	min-max	
60	categorizar	dummy
61	logN	min-max
62	min-max	

63	min-max	
64	min-max	
66	min-max	
69	min-max	
70	min-max	
29.a	min-max	
29.b	min-max	
37.a	min-max	
59	logN	min-max
19		
20		
22		
25		
32		
55		
71		
72		
65	min-max	

ANEXO 13: MATRIZ DE COSTOS PARA LA TÉCNICA DE ÁRBOLES DE DECISIÓN.

La inclusión de una matriz de costos para el análisis de precisión de los árboles de decisión se realiza sobre la configuración que trae por defecto el software *IBM SPSS Modeler*. Lo que se testea es el cambio del ratio costo de error tipo II versus costo de error tipo I.

	Precisión Global	Especificidad	Sensibilidad	Error II
Sin Matriz de Costos	68,39%	67,88%	68,41%	90,18%
Ratio				
1	68,40%	67,87%	68,43%	90,34%
1,1	70,42%	66,22%	70,62%	90,12%
1,2	70,42%	66,22%	70,62%	90,12%
1,3	73,88%	61,19%	73,88%	89,55%
1,4	77,09%	56,38%	77,09%	88,87%
1,5	82,78%	47,05%	84,52%	87,12%
2	85,28%	41,49%	87,39%	86,07%
3	91,27%	23,49%	94,57%	82,61%

ANEXO 14: ITERACIONES PARA LA TÉCNICA DE ÁRBOLES DE DECISIÓN.

	Precisión Global	Especificidad	Sensibilidad	Error II	Elección
Algoritmo de crecimiento					
CHAID	68,39%	67,88%	68,41%	90,18%	X
EXHAUSTIVE CHAID	69,57%	67,75%	69,66%	90,20%	
Máxima profundidad					
5	68,39%	67,88%	68,41%	90,18%	X
6	65,76%	71,37%	65,48%	90,10%	
7	70,39%	67,23%	70,55%	90,16%	
8	70,53%	67,44%	70,68%	90,16%	
9	70,53%	67,44%	70,68%	90,16%	
Chi-cuadrado para objetivos categóricos					
Pearson	68,39%	67,88%	68,41%	90,18%	X
Verosimilitud	68,39%	67,88%	68,41%	90,18%	
Nivel de significancia para la división					
0,02	67,55%	69,89%	67,43%		
0,05	68,39%	67,88%	68,41%	90,18%	X
0,1	68,39%	67,88%	68,41%	90,18%	
Nivel de significancia para la combinación					
0,02	68,39%	67,88%	68,41%	90,18%	
0,05	68,39%	67,88%	68,41%	90,18%	X
0,1	70,80%	66,52%	71,01%	90,15%	

ANEXO 15: ITERACIONES PARA LA TÉCNICA DE REDES NEURONALES.

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II	Elección
Modelo							
Perceptrón multicapa	71,89%	63,32%	72,31%	0,371	0,741	89,88%	X
Función base radial	54,72%	71,91%	53,87%	0,264	0,674	90,36%	
Tiempo de iteraciones							
10 minutos	71,12%	61,27%	71,61%	0,347	0,746	89,03%	
15 minutos	68,76%	68,33%	68,78%	0,378	0,748	89,87%	X
20 minutos	71,12%	61,27%	71,61%	0,347	0,747	89,03%	
Conjunto de prevención de sobreajuste							
25%	61,39%	69,80%	60,97%	0,370	0,745	90,01%	
30%	68,76%	68,33%	68,78%	0,378	0,748	89,87%	X
35%	69,53%	65,76%	69,71%	0,377	0,748	89,87%	

ANEXO 16: ITERACIONES PARA LA TÉCNICA DE REGRESIÓN LOGÍSTICA.

En las tablas de Backward y Forward se puede ver el análisis de los indicadores de acuerdo al cambio de los otros parámetros. No se expuso la tasa de error tipo II pues su varianza es muy baja. Solo se expone este resultado en la tabla Resumen.

BACKWARD

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Tiempo	N° variables
Método Stepwise							
Backward Condicional	68,89%	67,40%	68,96%	0,369	0,745	45 seg	33
Backward RL	no converge						
Backward Wald	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
Significancia para entrada							
0,01	68,84%	67,60%	68,90%	0,369	0,745	42 seg	30
0,05	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
0,1	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
Significancia para salida							
0,1	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
0,2	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
0,3	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
N° máximo de iteraciones							
2	64,45%	68,81%	64,24%	0,367	0,743	38 seg	31
5	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
10	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33
20	68,89%	67,40%	68,96%	0,369	0,745	40 seg	33

FORWARD

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Tiempo	N° variables
Método Stepwise							
Forward Condicional	68,43%	68,01%	68,45%	0,369	0,745	42 seg	29
Forward RL	no converge						
Forward Wald	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29
Significancia para entrada							
0,01	68,16%	67,35%	68,20%	0,367	0,742	38 seg	27
0,05	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29
0,1	67,26%	68,39%	67,21%	0,367	0,743	35 seg	31
Significancia para salida							
0,1	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29
0,2	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29
0,3	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29
N° máximo de iteraciones							
2	68,36%	67,51%	68,40%	0,366	0,744	34 seg	27
5	69,17%	66,98%	69,27%	0,367	0,744	34 seg	27
10	68,47%	67,88%	68,50%	0,367	0,745	35 seg	28
20	68,43%	68,01%	68,45%	0,369	0,745	39 seg	29

RESUMEN

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II	Elección
Selección de variables							
Enter	70,59%	65,47%	70,84%	0,368	0,746	90,06%	
Forward Condicional	68,43%	68,01%	68,45%	0,369	0,745	90,04%	
Forward RL	no converge						
Forward Wald	68,43%	68,01%	68,45%	0,369	0,745	90,04%	X
Backward Wald	68,89%	67,40%	68,96%	0,369	0,745	90,03%	
Backward RL	no converge						
Backward Wald	68,89%	67,40%	68,96%	0,369	0,745	90,03%	
Significancia para entrada							
0,01	68,16%	67,35%	68,20%	0,367	0,742	90,03%	
0,05	68,43%	68,01%	68,45%	0,369	0,745	90,04%	X
0,1	67,26%	68,39%	67,21%	0,367	0,743	90,04%	
Significancia para salida							
0,1	68,43%	68,01%	68,45%	0,369	0,745	90,04%	X
0,2	68,43%	68,01%	68,45%	0,369	0,745	90,04%	
0,3	68,43%	68,01%	68,45%	0,369	0,745	90,04%	
N° máximo de iteraciones							
2	68,36%	67,51%	68,40%	0,366	0,744	90,03%	
5	69,17%	66,98%	69,27%	0,367	0,744	90,05%	
10	68,47%	67,88%	68,50%	0,368	0,745	90,03%	
20	68,43%	68,01%	68,45%	0,369	0,745	90,04%	X

ANEXO 17: ITERACIÓN DE CANTIDAD DE NEURONAS PARA EL MODELO FINAL DE REGRESIÓN LOGÍSTICA.

	Precisión Global	Especificidad	Sensibilidad	KS	AUC	Error II	Elección
Máxima cantidad de nodos							
8	65,36%	65,04%	71,98%	0,366	0,740	88,75%	
12	68,41%	69,03%	68,38%	0,371	0,745	90.79%	
15	68,25%	68,47%	68,24%	0,372	0,745	90,29%	
20	71,44%	65,05%	71,76%	0,369	0,746	89,81%	
25	69,03%	68,16%	69,07%	0,372	0,748	90,20%	X
30	71,34%	65,83%	71,08%	0,368	0,745	89,83%	

ANEXO 18: REGLAS DEL MODELO FINAL DE ÁRBOLES DE DECISIÓN.

El siguiente es un diagrama de las reglas obtenidas del modelo árboles de decisión. Los cortes para cada variable han sido modificados de acuerdo a la función inversa aplicada en la etapa de transformación.

Paso 1	Paso 2	Paso 3	Paso 4	Paso 5	Predicción
v2 <= 0	V1 <= 0	v4 <= 0	v33 = 2	v51 <= 13	0
				v51 > 13	0
			v33 = cat 3 o v33= cat 4	v51 <= 9	1
				v51 > 9	0
			v33 = cat 5		1
		0 < v4 <= 11	v37.a <= 81139	v51 <= 4	0
				v51 > 4	0
			v37.a > 81139		0
		11 < v4 <= 35	v8 = cat 1 or v8 = cat 2		0
			v8 = cat 3		0
			v8 = cat 4 o v8 = cat 5	v37.a <= 81139	0
				v37.a > 81139	0
		35 < v4 <= 318	v8 = cat 1 o v8 = cat 2		0
			v8 = cat 3 o v8 = cat 4	v37.a <=670118	0
				v37.a > 670118	0
			v8 = cat 5		0
		v4 > 318	v8 = cat 1 o v8 = cat 2		0
			v8 = cat 3		0
			v8 = cat 4 o v8 = cat 5		0
	0 < V1 <= 1	v8 = cat 1 o v8 = cat 2			1
		v8 = cat 3 o v8 = cat 4 o v8 = cat 5	v28 = cat 2 o v28 = cat 4		0
			v28 = cat 3 o v28 = cat 5		1
	1 < V1 <= 2				1
	V1 > 1				1
0 < v2 =< 1	V1 <= 1	v37.a <= 81139	v33 = cat 2 o v33 = cat 4		1
			v33 = cat 3 o v33 = cat 5		1
		v37.a > 81139	V1 <= 0		0
			V1 > 0		1
	1 < V1 <= 2				1
	V1 > 2				1
1 < v2 <= 2	V1 <= 0				1
	1 < V1 <= 2				1
	V1 > 2				1
v2 > 2	v51 <= 4	v33 = 2	V1 <= 2		1
			V1 > 2		1
		v33 = cat 3 o v33 = cat 4 o v33 = cat 5			1
	v51 > 4				1

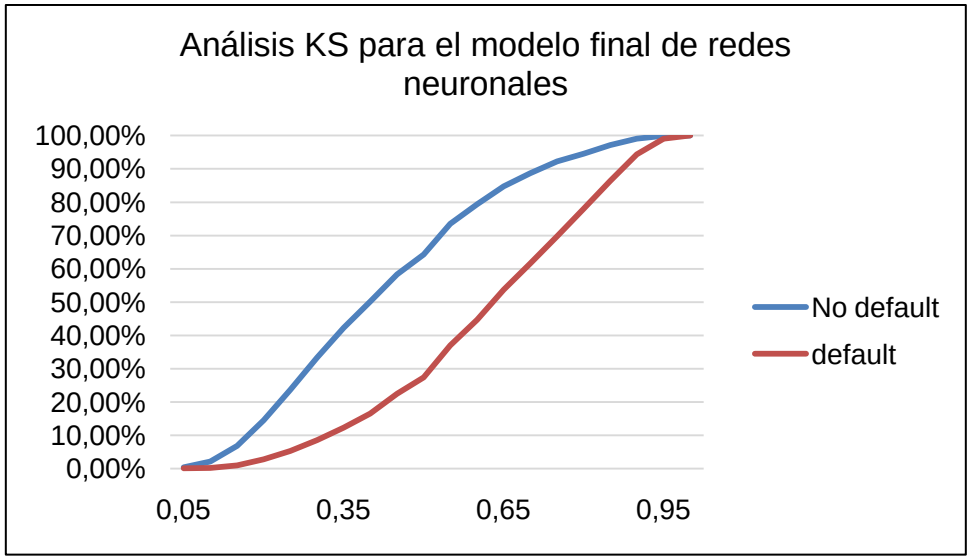
ANEXO 19: OUTPUT DEL MODELO FINAL DE REGRESIÓN LOGÍSTICA.

Variable	B	S.E.	Sig.	Exp(B)
4	-1,598	0,109	0,000	0,202
11	-0,394	0,074	0,000	0,674
19	-0,322	0,046	0,000	0,724
71	-0,578	0,055	0,000	0,561
51	-0,256	0,066	0,000	0,774
22	-0,194	0,034	0,000	0,823
55	0,094	0,028	0,001	1,098
2	1,503	0,143	0,000	4,494
1	4,016	0,151	0,000	55,493
6 (categoría 5)	-0,559	0,046	0,000	0,572
8 (categoría 2)	0,215	0,036	0,000	1,240
8 (categoría 5)	-0,253	0,033	0,000	0,776
28 (categoría 2)	-0,180	0,052	0,001	0,835
33 (Categoría 2)	-0,250	0,029	0,000	0,778
60 (categoría 1)	-0,172	0,033	0,000	0,842
Constante	1,150	0,062	0,000	3,158

ANEXO 20: ANÁLISIS PARA LA OBTENCIÓN DEL KS DE LOS MODELOS FINALES.

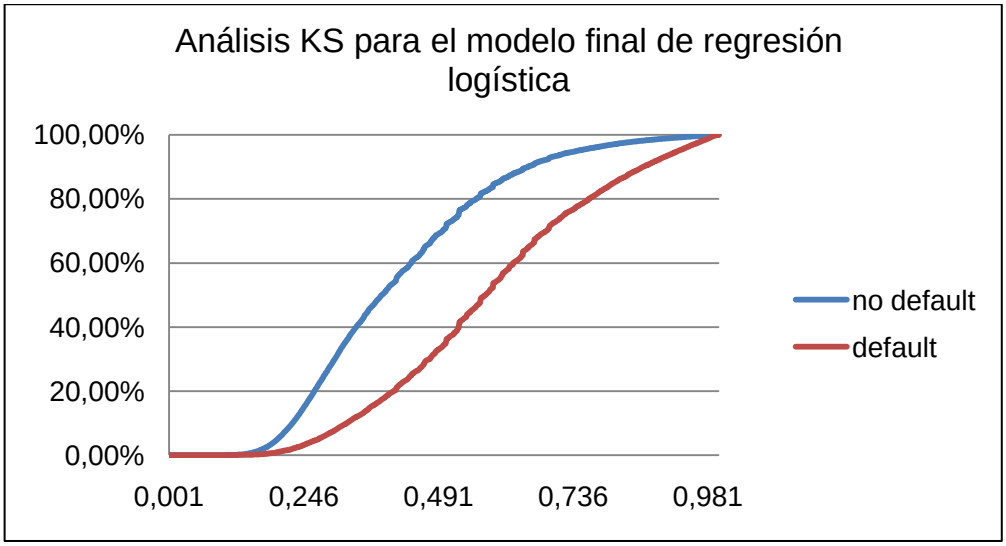
Redes Neuronales

KS	Corte
0,36922142	0,481

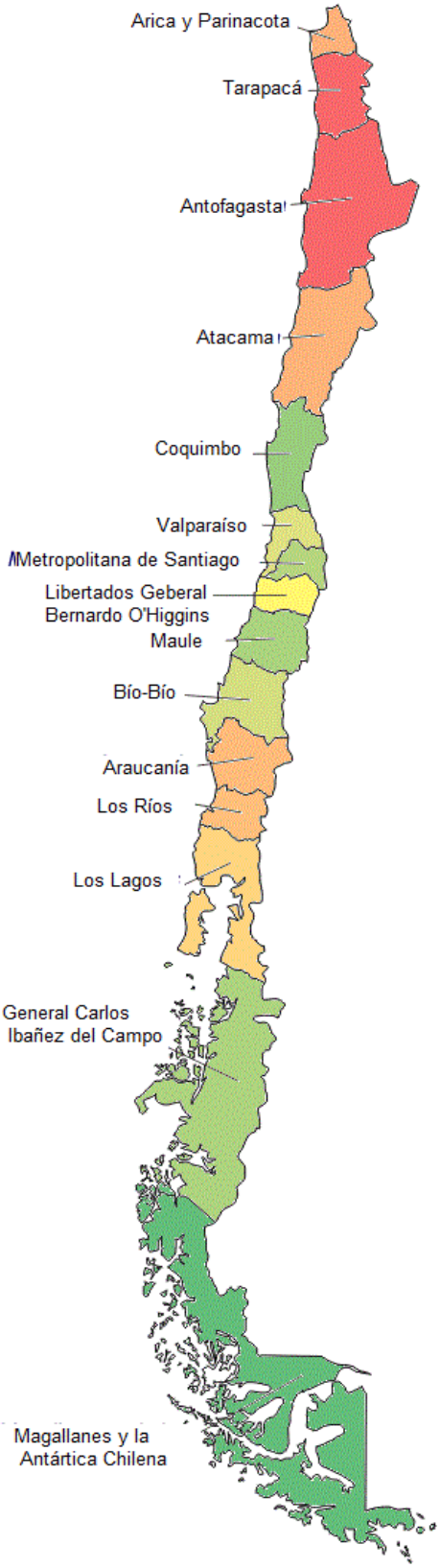


Regresión Logística

KS	Corte
36,35%	0,488

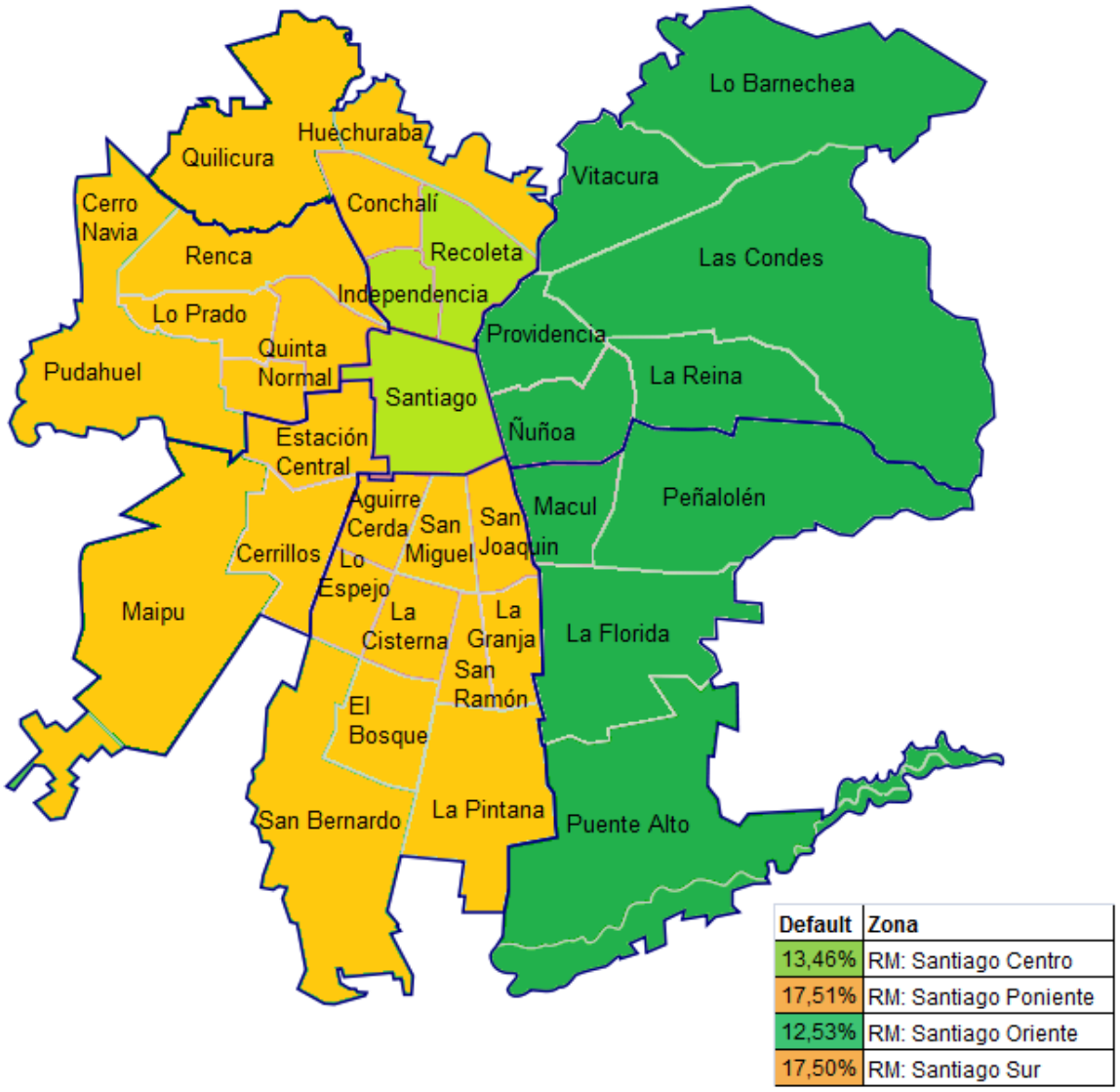


ANEXO 21: MAPA DE PREDICCIÓN DE DEFAULT POR REGIÓN.



Default	Región
19,55%	Tarapacá
19,48%	Antofagasta
17,76%	Atacama
14,76%	Coquimbo
15,91%	Valparaiso
16,41%	Libertador General Bernardo Ohiggins
15,06%	Maule
15,73%	BioBio
17,57%	Araucanía
16,95%	Los Lagos
15,10%	General Carlos Ibañez del Campo
13,87%	Magallanes y la Antártica Chilena
15,25%	Metropolitana de Santiago
17,79%	Arica y Parinacota
17,46%	Los Rios

ANEXO 22: MAPA DE PREDICCIÓN DE DEFAULT EN LA RM.



Default	Zona
13,46%	RM: Santiago Centro
17,51%	RM: Santiago Poniente
12,53%	RM: Santiago Oriente
17,50%	RM: Santiago Sur