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

MODELO PREDICTIVO DEL NO PAGO DE GIROS RELACIONADOS CON EL FORMULARIO 29 PARA EL SERVICIO DE IMPUESTOS INTERNOS DE CHILE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

JAVIERA IGNACIA CIFUENTES MOREIRA

PROFESOR GUÍA:

LUIS ABURTO LAFOURCADE

MIEMBROS DE LA COMISIÓN:

RICHARD WEBER HAAS
BRANDON PEÑA VILLAGRA

SANTIAGO DE CHILE

2016

RESUMEN DE LA MEMORIA PARA
OPTAR AL TÍTULO DE INGENIERA
CIVIL INDUSTRIAL
Alumna: Javiera Ignacia Cifuentes Moreira
Fecha: 27/08/2016
Profesor Guía: Luis Aburto Lafourcade

MODELO PREDICTIVO DEL NO PAGO DE GIROS RELACIONADOS CON EL FORMULARIO 29 PARA EL SERVICIO DE IMPUESTOS INTERNOS DE CHILE

El Servicio de Impuestos Internos es la institución encargada de velar por el cumplimiento tributario en Chile. Uno de sus deberes es fiscalizar el pago de impuestos y emitir cobros llamados giros de impuestos, cuando el pago no se realiza en forma correcta. En el último tiempo, estos giros no han tenido buenas tasas de pago, lo que preocupa y motiva a generar acciones para mejorarlas con el fin de aumentar la recaudación fiscal.

El objetivo de este trabajo de memoria es construir un modelo que permita predecir la probabilidad de no pago de un giro específico generado por una declaración y pago incorrecto del formulario F29 para los segmentos de Micro y Pequeña Empresa, con el que el SII pueda tomar decisiones como la priorización o mayor uso de recursos en acciones preventivas y/o paliativas del no pago.

Para lograr este objetivo, primero se estudia el caso y se recopilan variables de los contribuyentes, sus negocios y sus pagos del F29, existentes en la base de datos del SII, y se generan otras nuevas, totalizando 50 variables, que se anexan a una base de giros. Luego de una limpieza de datos quedan 292.940 giros, y con el 60% de ellos se entrenan diferentes modelos predictivos de árboles de decisión, incluyendo modelos Random Forest, y regresiones logísticas. De la comparación de sus resultados, se escoge un modelo CHAID por su facilidad de interpretación y aplicabilidad. Este modelo cuenta con un Accuracy de 78,2%, una Precisión de 83,1% y 65,6% de Especificidad, mientras que las variables más importantes para describir el no pago son en mayor medida el tipo de giro que se está pagando, la cantidad de veces que el valor del giro representa el pago mensual promedio de impuestos, y luego el valor de la deuda.

Con el entendimiento de las reglas del árbol se generan propuestas acción para el SII. Para los giros con mayor probabilidad de no pago se recomienda generar estudios para verificar la existencia de malas prácticas contables, y para los de baja probabilidad, se recomienda prevenir los giros mediante campañas educativas. En relación a la cobranza de los giros, se propone que la Tesorería General de la República, institución encargada de ello, priorice los giros con mayor propensión a no ser pagados ya que el resto podrían ser pagados en los plazos estipulados. El beneficio económico de esta cobranza es positivo, sumando 1.300 millones extras a la recaudación actual. Finalmente, como trabajo futuro se proyecta la aplicación de las propuestas y la medición de su efectividad a través del desarrollo de diseños experimentales.

Dedicatoria

Dedico esta memoria a mi papá, porque la mitad de mi suerte es tuya. Te extraño todos los días viejito. Gracias por dedicar tu mundo a nosotros.

Y obviamente a mi mamá por ser la mejor, y darme la otra mitad.

Agradecimientos

Quisiera agradecer a mis papás por todo el esfuerzo y el trabajo que realizaron todos estos años para que pudiéramos disfrutar este momento. Gracias a eso, he podido disfrutar de cosas que jamás ellos hubieran pensado, y compartir mis éxitos con ellos, pero lo que más valoro es lo queda para siempre: sus valores, educación y los recuerdos juntos. Un los quiero demasiado no alcanza.

Papá, por ahí ya nos tomaremos unos tequilitas. Mamá, por mientras, nos los tomamos nosotras.

También, tengo que incluir a mis hermanitos, Jose y Nico. Jose, eres la mejor persona del mundo, siempre estás apoyando en todo, te quiero decir que fuiste parte fundamental de mi carrera. Gracias hermanito. Nico, espero que se te pegue el espíritu santo. Termina tu carrera y aprovechemos la vida juntos. Finalmente doy las gracias al resto de mi familia, a mis tres abuelitas: a mi abuelita Ester, a la Tita y a la tía Eliana. Los amo a todos.

No puedo olvidar mencionar a mi amiga de siempre Coni Vergara, a mis inseparables Coni Rojas y Mane, al Patito, Jesús y Roberto. Gracias estar conmigo en las buenas y en las más malas, y por haber sido mi grupo de la Universidad, sin ustedes nada habría sido igual.

Para que no digan que no los nombre, agradezco inmensamente al Servicio de Impuestos internos, al área de Análisis de Riesgo del Servicio de Impuestos Internos y a todos los que se preocuparon por mí, ayudándome en lo que pudieran. Fue un bonito año junto a ustedes. Especialmente agradezco a Brandon Peña, Jorge Bravo, Jorge Moreira, y Constanza Quezada por haberme recibido increíblemente y haberme apoyado en todo lo que necesite. Se pasaron!

Por último, quiero agradecer a mis dos profesores guías, Luis Aburto y Richard Weber, por todo el tiempo que me dedicaron para que pudiera terminar esta etapa de la mejor manera. Realmente lo valoro, muchas gracias a ambos.

Tabla de contenido

Índice de tablas

Índice de Ilustraciones

1. Introducción

1.1. Antecedentes Generales

1.1.1. Servicio de Impuestos Internos

El Servicio de Impuestos Internos comúnmente llamado "SII", es una institución chilena dependiente del Ministerio de Hacienda que se encarga de la aplicación y fiscalización de la mayoría de los impuestos en Chile.

En la ley orgánica que constituye al Servicio, se determina la responsabilidad de velar por el cumplimiento de la normativa tributaria vigente, sobre los impuestos actuales y los que se creasen en un futuro, si es que por ley no se especificara lo contrario, de manera justa y transparente. Para esto debe entregar las herramientas y atención necesaria para que los contribuyentes puedan responder frente a sus obligaciones rápida y efectivamente [1][2].

Para asegurar el cumplimiento de tributario, cada año la institución genera documentos y transparenta su trabajo pasado y planes de acción futuros. En el "Plan de Cumplimiento Tributario", que ha compartido la institución para el año 2015, se especifican las cuatro obligaciones tributarias esenciales que deben cumplir todos los contribuyentes, las cuales son:

- Registrarse: Cada nuevo contribuyente deberá estar registrado ante la Administración Tributaria.
- Informar: Debe proveer información propia o de terceros cuando sea requerido.
- Declarar: Presentar una correcta declaración de impuestos en las fechas estipuladas.
- Pagar: Pagar los montos de impuestos que le correspondan, los que se adeuden y multas e intereses.

Para el SII, es importante distinguir cada una de estas obligaciones, para así ejercer tareas fiscalizadoras focalizadas, ya que el incumplimiento de cada una de ellas genera una brecha tributaria que a su vez tiene repercusiones en el sistema tributario [3]. Esto se condice con una de las metas principales de la gestión del Servicio, la que es incrementar y mejorar el control de la evasión y elusión, y potenciar el cumplimiento voluntario para aportar al progreso del país maximizando la recaudación fiscal [2].

El resultado del trabajo de esta memoria, pretende apoyar la toma de decisiones del SII para la prevención de problemas relacionados con el no cumplimiento de la cuarta obligación: El pago de impuestos.

En ese sentido, es importante aclarar que el SII no está a cargo de la cobranza de impuestos sino de determinar los montos adeudados, perteneciendo la primera tarea a la Tesorería General de la República, pero si, puede proveerle de información para apoyar su trabajo. Por lo que esta memoria, se enmarca dentro los planes de trabajo en conjunto de ambas instituciones para mejorar el pago de los contribuyentes y contendrá recomendaciones para las dos. Las acciones de cobranza utilizados por Tesorería serán analizados posteriormente en el capítulo 2.

1.1.2. Clasificación de Impuestos

Los impuestos que se aplican en Chile se dividen principalmente en dos categorías: los Indirectos y Directos. Los primeros se aplican sobre algunos tipos de bienes, es decir, por el uso de la riqueza y por ende indirectamente sobre los contribuyentes, mientras que los impuestos directos se cobran directamente al titular de la renta o riqueza [4]. En la siguiente figura se encuentran los principales impuestos de Chile:

Ilustración 1: Clasificación de los impuestos en Chile

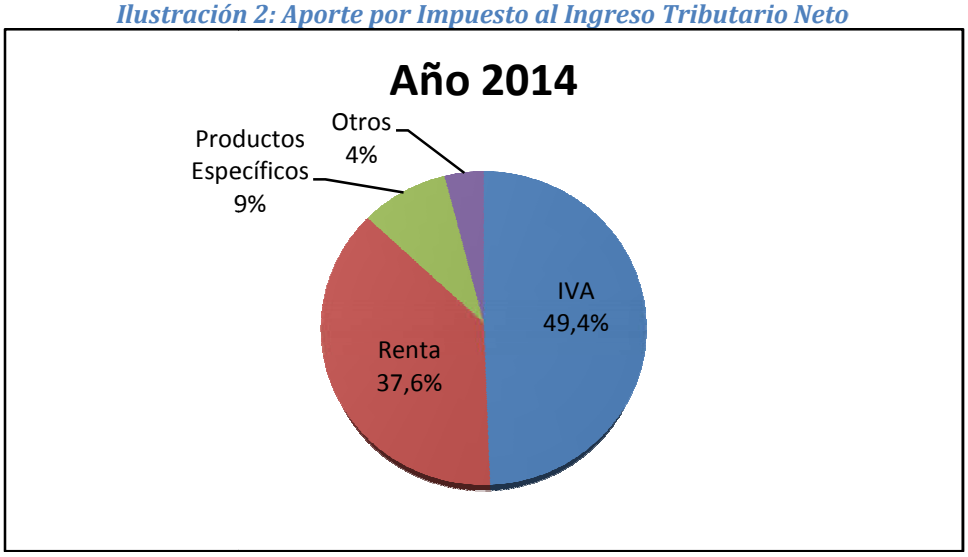
Fuente: Elaboración Propia con información de [4] y [5].

1.1.3. Importancia del IVA

Del listado de impuestos destacan, en función de su incidencia en la recaudación fiscal, el Impuesto a las Ventas y Servicios o IVA y el Impuesto a la Renta. Por ejemplo, durante el año 2014, se recaudaron $24,5 billones de pesos, de los cuales $12,1 billones de pesos fueron

aportados por el IVA y $9,3 billones por el Impuesto a la Renta, representando un 49,4% y un 37,6% de los ingresos tributarios netos del año respectivamente [6]. El 13% restante fue recaudado por Impuestos a productos específicos y otros Impuestos.

Ilustración 2: Aporte por Impuesto al Ingreso Tributario Neto



Fuente: Elaboración Propia con información de [6].

El Impuesto al Valor Agregado es un impuesto que consiste en un recargo del 19% que se aplica sobre el valor de un bien o un servicio. Este impuesto se traslada en cadena desde el vendedor al comprador ya que son los clientes finales quienes en el monto total de su compra deben pagar el valor del bien más el monto del impuesto.

El IVA se declara y paga simultáneamente de forma mensual, para esto se debe completar un documento, llamado Formulario 29 o F29, de manera electrónica a través de internet o rellenando un formulario físico (papel). Según sea la forma de declaración, se paga en línea o en alguna institución financiera autorizada para recibir su pago, presentando el formulario.

Para declarar, los contribuyentes deben anotar todos los impuestos que recuperaron en sus ventas (débito fiscal) y restar los que pagaron en sus compras (crédito fiscal). Esta diferencia de IVA es la que el contribuyente debe cancelar. Para esto se rellenan más de 100 campos o códigos del formulario, siendo el campo C89 el total de impuesto IVA determinado. En el caso de que el crédito fiscal fuera mayor al débito, se genera un remanente de crédito fiscal que le servirá al contribuyente como abono para el mes siguiente y su C89 equivalente al total del Impuesto de IVA será igual a 0. Al impuesto determinado en el campo C89, se suman las retenciones mensuales de Impuesto a la Renta con lo que finalmente se determina el código 91 equivalente al total de Impuesto a pagar en el plazo legal.

Según la forma de declaración y de los montos a pagar existen diferentes fechas para cumplir con la obligación de declaración. Las fechas establecidas para este efecto, corresponden al mes siguiente al período que se va a declarar:

- Día 12: Si se declara vía papel.
- Día 20: Si se declara vía internet y el contribuyente es facturador electrónico (con la Reforma tributaria, la mayoría de los contribuyentes de primera categoría afectos a IVA tendrán esta característica).
- Día 28: Si se declara un formulario sin movimiento o sin pago[1] [7].

La declaración posterior a estas fechas está permitida, pero se cobran intereses, reajustes y/o multas por constituir el atraso una infracción por no cumplimiento en la fecha estipulada. En Anexo 10.1, se puede encontrar una copia del F29.

1.1.4. Clasificación de las empresas según SII

Para que un contribuyente sea considerado como empresa por el Servicio de Impuestos Internos debe cumplir uno o más de los siguientes requisitos:

- Ser identificado como contribuyente de 1ra Categoría
- Presentar declaración jurada n° 1887 (Declaración de pago de sueldos)
- Presentar declaración jurada n° 1827 (Declaración de pago de honorarios)
- Ser declarante vigente de IVA

Por otra parte, las empresas son clasificadas según su tamaño, el que está determinado por el rango en que se encuentra la suma de sus ventas del año tributario anterior, en Unidades de Fomento. Según esto, se definen los siguientes segmentos de tamaños [9]:

- Micro: Ventas entre 0, 01 UF a 2.400 UF
- Pequeña: Ventas entre 2.400,01 UF a 25.000 UF
- Mediana: Ventas entre 25.000,01 UF a 100.000 UF
- Grande: Ventas entre 100.000,01 UF en adelante.

[1] Sin movimiento: Valor cero en todos los campos del F29.
Sin pago: Valor cero en el código 91 del F29. Ver Anexo10.1Formulario 29.

1.1.5. Giros de impuestos

Los giros de impuestos son órdenes de pago de impuestos y/o intereses, multas y reajustes, que emite y notifica el SII al contribuyente, remitiendo copia a la Tesorería General de la República para su cobranza [8]. Los giros pueden emitirse por 28 diferentes motivos que el SII denomina internamente como Conceptos de Giro.

Los Conceptos de Giro o razones del giro se organizan según el formulario que lo produce (F29, F50, F22), si el error fue de descuadratura del formulario, es decir, no coincidencia de la sumatoria de los campos declarados en el formulario (Giro 45) u otros (Giro 21). Los conceptos del F29 o relacionados como Auditoria IVA, serán explicados en detalle en "2.1. Flujo de los Giros Seleccionados".

Tabla 1: Origen Giros de Impuestos

Formulario	Giro	Concepto
F29	G21	Rectificatoria F29
	G45	Impuesto F29
		Multa F29
		Pago Diferido IVA
		Descuadratura Impuesto F29
F22	G21	Rectificatoria F22
	G45	Impuesto F22
		Descuadratura Multa F22
		Pago Diferido F22
F50	G21	Rectificatoria F50
	G45	ImpuestoF50
		Descuadratura F50
Otros Casos	G21	Auditoria IVA
		Auditoria Renta
		Otros impuestos

Fuente: Elaboración Propia.

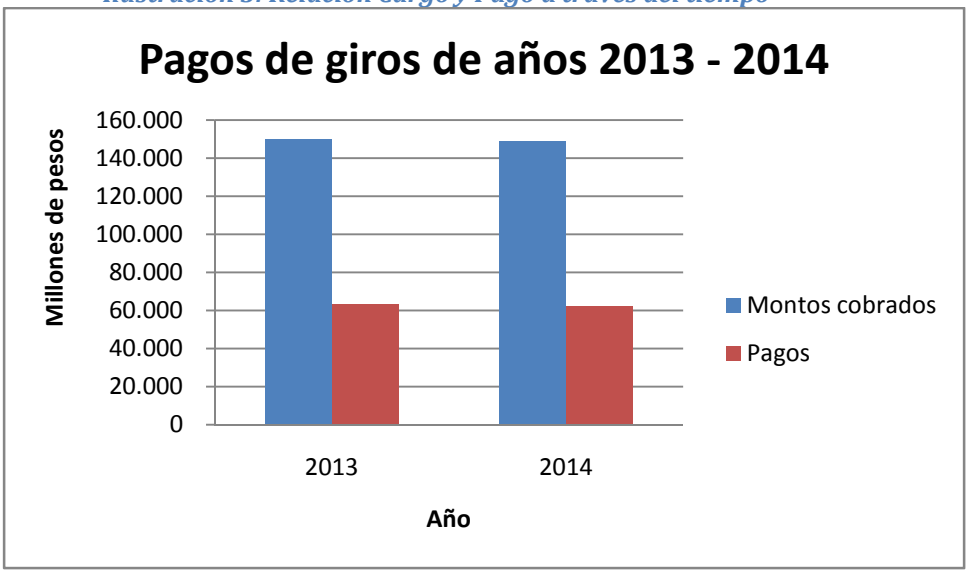
1.2. Definición y Justificación del proyecto

Explicado lo anterior, se puede indicar que este trabajo de memoria consiste en la obtención de un modelo predictivo que determine la probabilidad del no pago de giros de impuestos relacionados al Formulario 29 de contribuyentes que pertenecen a los segmentos de Micro y Pequeña Empresa estableciendo información sobre aquellos más riesgosos. Esto

permitirá en términos generales, obtener variables importantes en la predicción, la validación de reglas de negocio, y su entendimiento para el proceso de generación de giros del SII, además de realizar recomendaciones sobre la priorización de acciones de cobranza de la Tesorería General de la República según la probabilidad de no pago de los giros.

Se ha escogido trabajar el no pago de giros IVA (a excepción de pago diferido IVA) ya que en el análisis de un pequeño subconjunto de la base de datos del Servicio de Impuestos Internos, para los cuatro tipos de giros seleccionados en los años 2013 y 2014, se ha detectado un pago cercano al cuarenta y cinco por ciento, por lo que su diferencia, que constituiría el no pago, se podría revertir para aumentar la recaudación fiscal. En la figura 3, se grafican los montos de los giros cobrados en ese período y los pagos de ellos realizados hasta diciembre del 2015. Estos reflejan la existencia de una brecha que representa una oportunidad de mejora del estado actual de recaudación.

Ilustración 3: Relación Cargo y Pago a través del tiempo



Fuente: Elaboración Propia.

Por otro lado, si se evalúa el comportamiento por segmentos de empresa, de manera agregada, la micro empresa tiene más giros que la pequeña empresa, pero los giros de la pequeña son monetariamente superiores. El monto promedio de la deuda, para las pequeñas empresas es cercano a $780.000 pesos, mientras que para las micro empresas es de alrededor de $360.000. Estos valores, considerados altos para estos segmentos, podrían estar influyendo en que las empresas no puedan pagar.

Otro motivo de la realización de la memoria, es que la alta cantidad de giros y de contribuyentes relacionados al problema, no hace fácil su análisis. Según [9] y [11], para el año 2014 se reportaban 4130 funcionarios que atendían a una cantidad superior a 8 millones de contribuyentes, lo que resulta en una relación de 1 funcionario cada 1937 contribuyentes. Esta

proporción hace necesario implementar tecnologías que faciliten el trabajo masivo y afine criterios de trabajo selectivo para profundizar en casos especiales.

Además, al estudiar la distribución del tiempo de pago de los giros se verifica que existe un porcentaje de los giros que se cancelan con alta morosidad. Por ejemplo, para los giros emitidos el año 2013 existe un 16% de giros pagados luego de 6 meses y se distinguen pagos realizados hasta 24 meses después de emitido el giro. Este tiempo de pago también representa una oportunidad de mejora ya que podría reducirse al profundizar en el comportamiento de los contribuyentes.

Finalmente, el obtener un listado de variables y/o patrones que predigan el no pago permitirá tomar acciones preventivas frente a comportamientos riesgosos. El SII, como principal fiscalizador, podría generar políticas para propiciar el cumplimiento tributario, y Tesorería, conociendo de antemano la propensión a pagar de un contribuyente, podría tomar diferentes medidas para llevar a cabo su cobranza. Por ejemplo, para un deudor con altas probabilidades de no pago se podrían aplicar más acciones y/o endurecer la cobranza, mientras que para uno con probabilidad de no pago baja, se le aplicaría tratamientos menos intensivos y por ende más baratos. Este reordenamiento podría impactar en un ahorro económico en el presupuesto de Tesorería, como también, en mejoras potenciales de las tasas de recuperación de impuestos.

1.3. Objetivos

1.3.1. Objetivos Generales

- Construir un modelo que permita predecir la probabilidad de no pago de un giro específico generado por diferencias de impuestos al momento de efectuar la declaración y pago del formulario F29, para los segmentos micro y pequeña empresa, y entregar este resultado como suministro para la toma de decisiones del SII.

1.3.2. Objetivos Específicos

- Identificar las principales variables y/o patrones que influyen en el no pago de giros de impuestos.
- Analizar si el modelo generado, y las variables de este, tienen aplicación para distintos grupos de contribuyentes.
- Realizar recomendaciones de uso de las probabilidades derivadas del modelo para apoyar a Tesorería General de la República.
- Entregar un listado de probabilidades de no pago futuro para un grupo de giros "nuevos".

1.4. Alcances

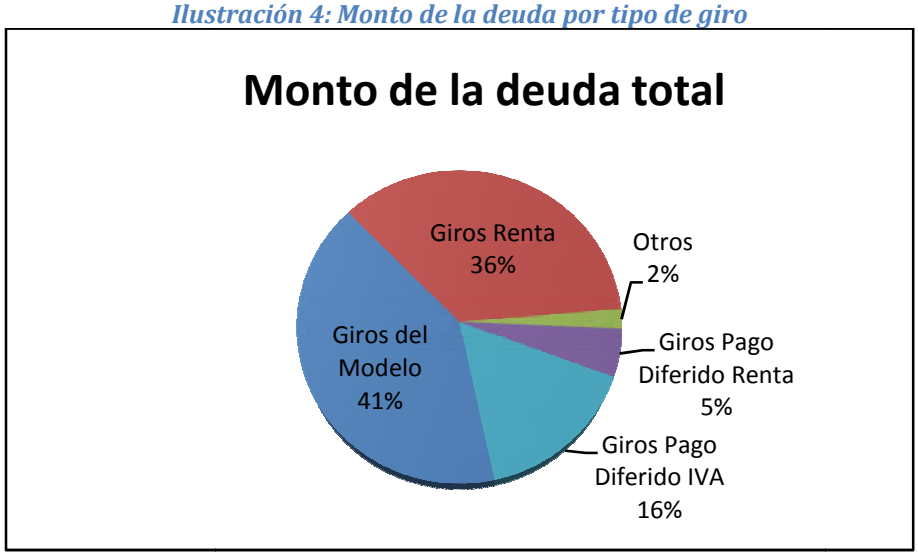
Debido al gran volumen de contribuyentes, se acota el modelo predictivo a los segmentos de micro y pequeña empresa. Se escogen estos dos grupos debido a que representan, según el Servicio de Cooperación Técnica (SERCOTEC), el 95,5% de las empresas en Chile [10], y son de interés para el SII.

Por otra parte, los giros de impuestos relacionados con el Formulario F29 que se abordarán en este trabajo son cuatro, los que son:

- Multa F29
- Impuesto F29
- Rectificatoria F29
- Auditoría IVA

La razón de su elección es que en conjunto estos giros representan un 41% de la deuda total de los segmentos con los cuales se trabajará, siendo mayor que la selección de otros grupos de giros como, por ejemplo, giros del Impuesto a la Renta (que alcanzan un 36%). La mayor deuda explicada por el IVA, valida la utilidad de trabajar con este impuesto.

Ilustración 4: Monto de la deuda por tipo de giro



Fuente: Elaboración Propia.

En el gráfico se observa un quinto tipo de giro asociado al IVA, llamado Pago Diferido IVA. Este se excluye dado que su implementación es menor a seis meses, razón por la cual al

momento del presente trabajo no existe data histórica suficiente para generar un modelo predictivo.

Por último, los datos utilizados para la construcción del modelo, son todos los giros emitidos en el periodo enero 2013 a junio 2015, para luego testear la validez del modelo sobre datos del año 2016.

1.5. Resumen metodología

El problema que preocupa relaciona varios elementos que se caracterizan por ser abundantes en datos. Por una parte, la gran cantidad de contribuyentes de micro y pequeña empresa, cada cual, con su información obligatoria, y por otra, se encuentran sus declaraciones del formulario F29 que al menos contienen 100 variables de manera mensual. Esta cantidad de datos y relaciones entre variables vuelve imposible el análisis del problema sin el uso de técnicas para el manejo de datos. Por esta razón se decide trabajar con la metodología Knowledge Discovery in Databases más conocida como KDD.

- Entendimiento del problema

La primera fase del trabajo, consiste en entender los diferentes procesos que realiza el SII para la recaudación de impuestos y los formularios de declaración existentes. Luego, a través de reuniones con diferentes áreas de la Subdirección de Fiscalización, se aprenden a manejar las bases de datos relacionadas con los giros de impuestos, para identificar la magnitud del problema en términos monetarios, y en volumen de contribuyentes y giros, de forma agregada y en detalle.

- Recopilación y selección de la data

Posteriormente, se estudian otras bases de datos. El Servicio de Impuestos Internos posee una alta cantidad de variables definidas, pero no todas son útiles para abordar este problema. Por lo tanto, se seleccionan las que sí lo son, y se crean otras nuevas, aplicando el conocimiento adquirido de estudios sobre Credit Scoring, experiencias de agencias extranjeras de impuestos y variables relevantes según juicio de expertos del propio SII.

- Pre-procesamiento de la data

En esta etapa se analizan los datos con los que se trabajará y se corrigen falencias, como datos incorrectos y/o outliers. Principalmente, se eliminaron datos inconsistentes con el

problema, como cobros y pagos negativos, o de deudas superiores a lo que por ley se puede cobrar. Además, en esta etapa se define el algoritmo de la variable objetivo a estudiar.

- Transformación de la data

Del primer conjunto de variables determinadas, se generan otras que las complementan, en particular ratios conformados por dos variables, como por ejemplo, monto de la deuda sobre ventas anuales. Dada la creación de nuevas variables, se revisan las bases de datos en búsqueda de nuevos datos inconsistentes.

- Generación de modelos

Se modela el problema con las variables predefinidas, comparando diversas técnicas para escoger la de mayor grado de predicción según métricas de rendimiento y aplicabilidad. Se utilizan distintos algoritmos del software SPSS Modeler, como árboles de decisión y regresión logística, además de Random Forest presente en el software R Studio.

- Interpretación y Evaluación de Resultados

Finalmente, se estudian patrones de comportamiento de los contribuyentes y las variables que según el modelo serían más significativas en la predicción para interpretar las situaciones que definen el problema. Se evalúan los resultados en función de los objetivos específicos del trabajo.

2. Marco Conceptual

2.1. Modelo de Gestión del Cumplimiento Tributario

2.1.1. Definición del modelo

El SII basa su trabajo de acuerdo a su modelo de Gestión del Cumplimiento Tributario, que se enfoca en el cumplimiento de las cuatro obligaciones tributarias de los contribuyentes, de registro, información, declaración y pago de impuestos, de forma tal de reducir los riesgos de incumplimiento. El modelo de gestión aborda esto desde varias perspectivas. Entre éstas está el segmento de contribuyentes a tratar, en relación a sus características de comportamiento tributario, la medición de la "brecha tributaria", correspondiente a lo que debía pagar y su nivel de cumplimiento real, posibles riesgos futuros que la puedan aumentar, las causas a estas situaciones de riesgo y, por último, según los tratamientos o acciones para reducir, eliminar,

prevenir o corregir esta situación. Es decir, el modelo que se utiliza hoy en Chile, abarca muchos aspectos que de manera complementaria logran maximizar el cumplimiento tributario.

2.1.2. Probabilidad y consecuencias de riesgo tributario

El modelo de Gestión del Cumplimiento Tributario se aplica operativamente bajo la forma de una matriz de riesgos en la cual, para cada segmento de interés, se calcula una probabilidad de ocurrencia del no cumplimiento (o acción estudiada) ponderada por la consecuencia de este. Es así, como se generan subgrupos de mayor homogeneidad representados en diversos cuadrantes. Clasificados de esta manera, se facilita la generación de políticas de tratamientos específicos para cada grupo de contribuyentes. En la figura 5 se puede observar, por ejemplo, que si un contribuyente genera graves consecuencias cuando no cumple alguna obligación tributaria y además, posee una probabilidad alta de que falle, este será clasificado con un riesgo de alto a severo. De forma contraria, si la probabilidad de fallar de una persona o empresa es baja y cuando ocurre no tiene mayores implicancias entonces será clasificado de riesgo bajo.

Ilustración 5: Matriz de riesgo Modelo Gestión del Cumplimiento Tributario

Fuente: Servicio de Impuestos Internos.

El modelo de la memoria entregará un listado de probabilidades de ocurrencia del no pago de giros de impuestos, según lo planteado en 1.3.1, como insumo para la creación de una matriz como la de la imagen, que apoye al SII en la asignación de recursos y definición de acciones para tratar a aquellos contribuyentes de menor calidad de pago.

2.2. Flujos de los giros seleccionados

Los giros de impuestos, o cobros emitidos por el SII, pueden producirse por acciones voluntarias del contribuyente para regularizar su situación, como notificar un error en su

contabilidad, o por acciones del SII como una fiscalización. A continuación se explica el origen de los cuatro giros seleccionados para su estudio y en Anexo Ejemplo de Giro10.2 se encuentra un formulario de Giro para su entendimiento.

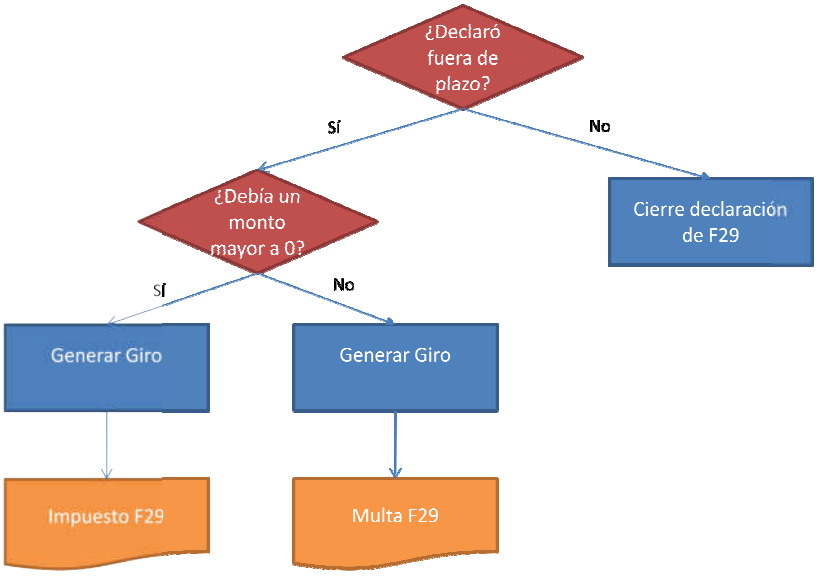
2.2.1. Multa F29

Un giro del tipo Multa F29 se emite cuando a un contribuyente le corresponde declarar un Formulario F29 de valor cero o sin pago de impuesto, pero no lo hace dentro de los plazos correspondientes, ya citados en 1.1.3. Como no hay impuestos a pagar, el valor de este giro corresponde a una multa fija determinada por ley. Esta varía entre una UTM (Unidad Tributaria Mensual) y una UTA[2] (Unidad Tributaria Anual) dependiendo de los códigos del formulario que no fueron declarados correctamente.

2.2.2. Impuestos F29

Este giro se produce para cobrar impuestos no pagados en los plazos correspondientes. Por lo que se le conoce como Giro por declaración fuera de plazo F29, y se genera por la acción voluntaria del contribuyente que declara impuestos que no había informado anteriormente o por acciones del SII.

Ilustración 6: Flujo de giros del tipo Multa F29 e Impuesto F29



Fuente: Elaboración Propia con información del Servicio.

[2] Una UTA equivale a la multiplicación del valor de una UTM del último mes del año comercial multiplicada por 12.

2.2.3. Rectificatoria F29

Luego de declarar un formulario 29, cada contribuyente tiene la posibilidad de corregir su declaración debido a que él mismo detectó un error en su contabilidad o porque ha sido informado de una diferencia de impuestos por parte del SII, luego de procesar información de terceros que entreguen dicho resultado. El trámite de rectificar se puede realizar a través de internet o en algunos casos en la unidad del SII que le corresponda al contribuyente.

Si la rectificación da como resultado un saldo de impuestos a favor del SII, es decir, que el monto imponible aumenta, se genera un giro llamado "Rectificatoria F29" en donde se cobra esta diferencia de dinero.

2.2.4. Auditoría IVA

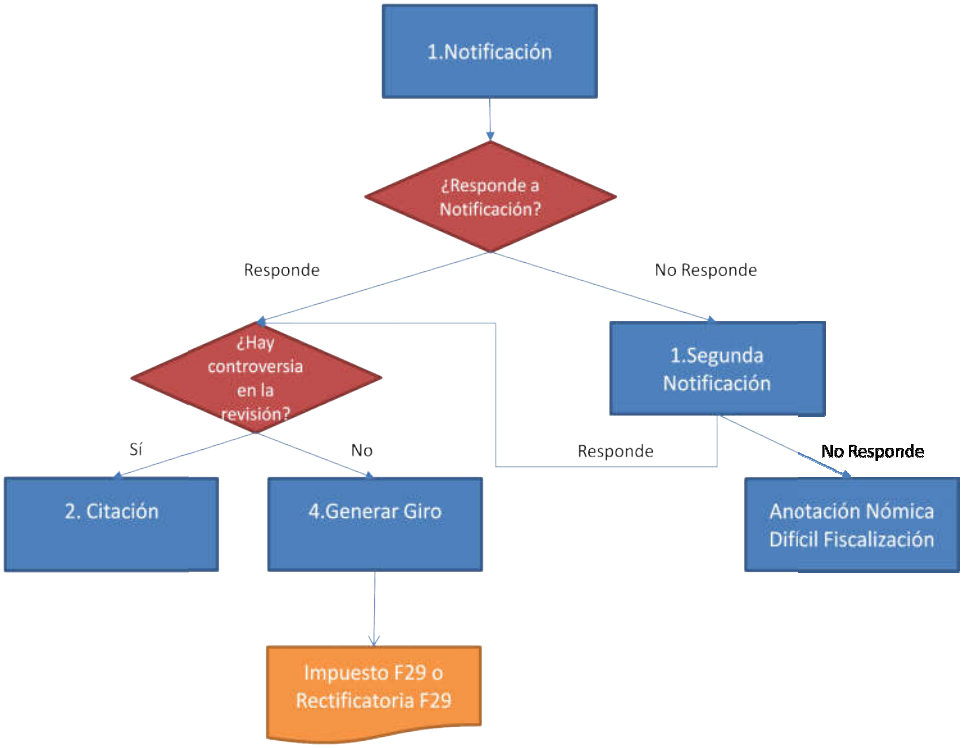
El giro Auditoría IVA se genera como resultado de una auditoría tributaria, la cual es un procedimiento de fiscalización que tiene como objetivos:

1. Verificar que las declaraciones de impuestos sean el reflejo fidedigno de todas las operaciones que se encuentran en el libro de contabilidad del contribuyente, de la documentación soportante y que refleje todas las transacciones efectuadas.
2. Establecer si las bases imponibles están correctamente determinadas y de no ser así, efectuar el cobro de ellas con las multas, intereses y reajustes que apliquen.
3. Detectar a tiempo a quienes no cumplan con sus obligaciones tributarias [8].

La auditoría tributaria IVA es un proceso complejo que consta de cuatro etapas. La primera es la de Notificación, donde el Servicio le informa al contribuyente del inicio del proceso de revisión que se llevará a cabo, y le solicita que presente la documentación necesaria. Puede ocurrir que el contribuyente no comparezca a la notificación y en ese caso se le vuelve a notificar, si el contribuyente no da respuesta una segunda vez, el Servicio está facultado para generar una anotación del contribuyente en la Nómina de Difícil Fiscalización o solicitar a la justicia ordinaria que aplique una medida de apremio según lo dispone el código tributario. Por otro lado, si el contribuyente comparece y se establece diferencias de impuestos, puede que no haya o haya controversia en el resultado de la fiscalización.

1. No hay controversia. El contribuyente acepta las diferencias de impuestos que se le adjudican, y rectifica su declaración original, emitiéndose un giro por Rectificatoria F29, o en su defecto, presenta una declaración fuera de plazo, emitiéndose un giro por Impuesto F29.
2. Hay controversia o no acepta las diferencias de impuestos. En este caso, se le citará para que presente más información que avale su postura.

Ilustración 7: Flujo de giro del tipo Auditoria IVA - Notificación



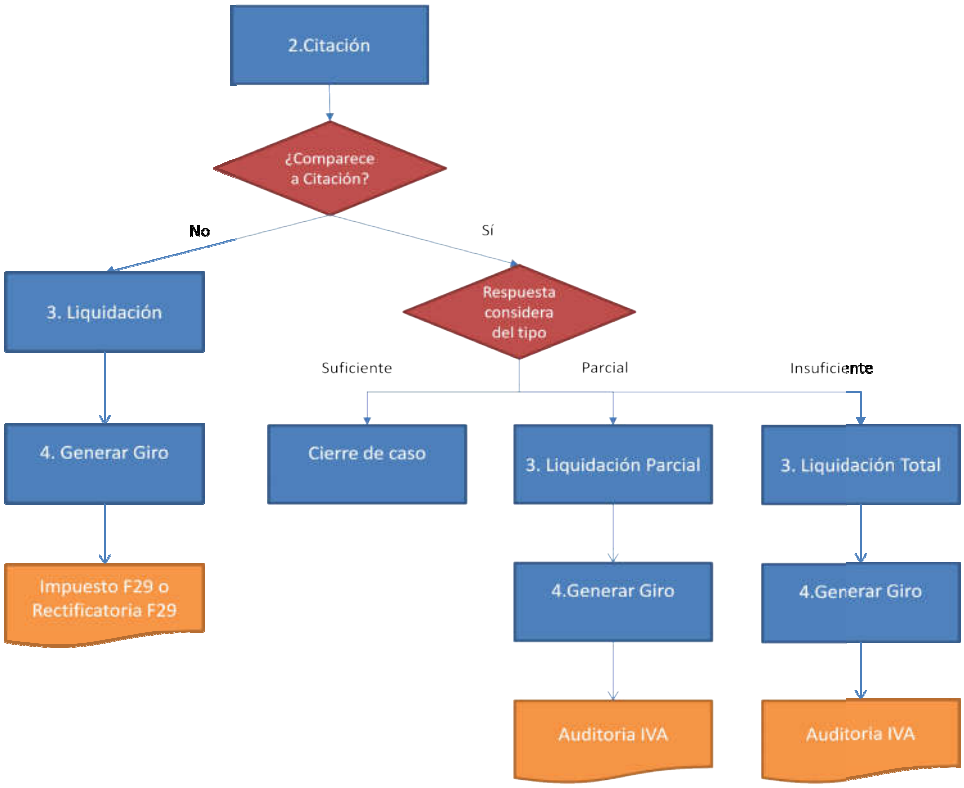
Fuente: Elaboración Propia.

En la etapa de citación se le solicita al contribuyente que presente, confirme, aclare, modifique o rectifique la declaración objeto de revisión. El citado, tiene el plazo de un mes prorrogable, por una sola vez y hasta por un mes, para presentar su respuesta a la fiscalización. Pueden ocurrir dos situaciones:

1. Que no se presente a citación con lo que habilita al SII para que liquide las diferencias de impuestos si las hubiera.
2. Que se presente a citación. Puede que la respuesta del contribuyente sea considerada suficiente, parcial o insuficiente, es decir, que aclaró todas las diferencias de impuestos por las que fue citado y no deba pagar, que deba liquidarse solo parte de los impuestos revisados o que deba liquidarse el total respectivamente.

La tercera etapa es la liquidación que consiste en la determinación de impuestos adeudados según la auditoría tributaria, que considera el valor neto de los impuestos, reajustes, intereses y multas. Finalmente, si existe una liquidación luego de una citación, se emite un giro del tipo Auditoría IVA cobrando lo anterior.

Ilustración 8: Flujo de giro del tipo Auditoria IVA – Citación



Fuente: Elaboración Propia.

Es importante destacar que para los casos de Impuesto F29, Rectificatoria F29 y Auditoria IVA, el impuesto que no se paga dentro del plazo correspondiente devenga reajustes equivalentes al I.P.C. del período e intereses de un 1,5% mensual lineal sobre los impuestos. Además de una multa que depende del valor del impuesto, diferente a la aplicada en Multa F29. El valor de la multa depende de si el giro es debido a una acción del contribuyente o a la acción del Servicio. En el primer caso la multa parte en el 10% de los impuestos adeudados aumentando 2% cada mes con tope hasta el 30%. En el caso de que exista una acción del SII, la multa va desde un 20%, y aumenta 2% cada mes con tope hasta 60%. Por último, en algunos casos, el contribuyente puede acogerse a condonación de multa e intereses [18].

Ilustración 9 : Ejemplo cuadro total a pagar de un giro

TOTAL GIRO	C91	52.890.450
I.P.C (Reajuste Art. 53 C.T)	C92	6.340.521
Intereses y Multas	C93	6.082.402
TOTAL A PAGAR	C94	65.313.373

Fuente: Elaboración propia con información del Servicio.

2.3. Tratamientos de Cobranza

Cualquiera sea el origen del giro de impuestos determinado por el SII, su cobranza recae en el ámbito de acción de la Tesorería General de la República. Como los giros de impuestos se generan mensualmente, la TGR define mes a mes las acciones, o tratamientos, de cobranza, los que se definen de acuerdo a los recursos disponibles para realizar las acciones, y según el tipo de contribuyente. Estos se dividen en segmentos de deuda y no de ventas como lo hace el SII, dividiéndose en:

- Pequeños deudores: Deuda total hasta $10 millones de pesos.
- Medianos deudores: Deuda total entre $10.000.001 - $89.999.999 pesos
- Grandes deudores: Deuda total desde $90.000.000 pesos.

Según esta segmentación, se realizan dos tipos de cobranza: la administrativa y la judicial. Para los pequeños deudores se aplica cobranza administrativa que consiste en la realización de las siguientes acciones: llamadas telefónicas, correos, mensajes de texto y cartas, que se pueden aplicar todas juntas, más de una vez, o por separado. La cobranza se realiza según juicio de expertos, dándosele prioridad a los contribuyentes de mayor deuda. Por otra parte, a los medianos y grandes deudores se les aplica una cobranza judicial, a los que se agregan pequeños deudores que no han pagado luego de cierto plazo en cobranza administrativa. Este plazo varía entre 300, 180 o 120 días dependiendo del monto de la deuda fiscal.

Ilustración 10: Tiempos de cobranza

Fuente: Elaboración propia con información de Tesorería.

2.4. Entendimiento del Problema y Generación de Variables

2.4.1. Estudio del problema dentro del Servicio

Para comprender el problema de no pago de giros, originado por Formulario F29, se estudian tres procesos por separado. Primero, el funcionamiento del impuesto IVA, luego el funcionamiento del cobro de diferencias de impuestos a través de los giros y, por último, cómo esa información se procesa dentro del SII y se registra en sus bases de datos.

2.4.2. Credit Scoring

Credit Scoring es una técnica que ayuda a las organizaciones a tomar la decisión de si se procede a prestar dinero o no, asignando un puntaje más alto a los prestatarios con mayores probabilidades de devolver el crédito que están requiriendo [19]. Para esto, a través de diversos modelos cuantitativos se calcula, según las características del prestatario, su probabilidad de pago.

Si se compara el Credit Scoring con el problema de no pago de giros de impuestos, se encuentran similitudes, dando indicios de que podrían existir variables que sean útiles para el modelo predictivo a generar. Por esta razón, se buscan casos aplicados en Chile. En base al caso descrito en [13] y [14] se obtienen atributos que, aplicados al tema, podrían explicar el no pago de impuestos y que pueden ser extraíbles de las bases de datos del Servicio. Algunos de estos son:

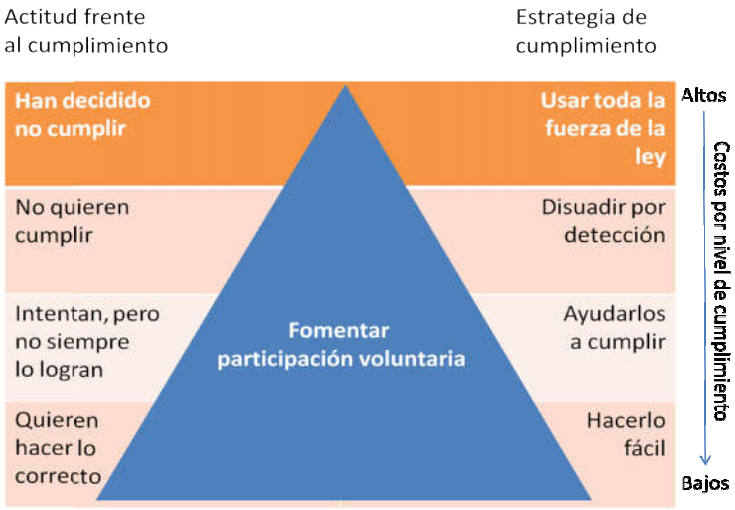
- Patrimonio, en cantidad de inmuebles y valor de ellos.
- Edad del prestatario (edad de la empresa o representante)
- Rubro de la empresa
- Región de la empresa
- Créditos pagados y vigentes (a transformar en giros pagados e impagos)
- Mora en el pago de la obligación (a transformar en antigüedad de la deuda).
- Valor del crédito (a transformar en monto girado).

Como estas variables son específicas al problema de [13], se generalizan para construir grupos de variables que apliquen al caso, por lo que se definen los siguientes grupos: Caracterización del contribuyente, que contempla variables demográficas y de descripción del tipo de empresa, y Capacidad de Pago, de variables financieras como patrimonio. Otra forma de determinar la capacidad de pago es observando los ingresos y egresos de una empresa, que se representarán utilizando variables como compras y ventas declaradas en el formulario F29. Ellas conformarán un tercer grupo de variables denominado Variables F29.

2.4.3. Estudio de otros Servicios

En la recopilación de antecedentes del problema, se estudiaron documentos públicos de Administraciones Tributarias de otros países. Del archivo "*Compliance in Focus*" [15] de la Oficina Australiana de Impuestos (ATO) se extrae el siguiente modelo de cumplimiento, en donde se categoriza a los contribuyentes según su actitud hacia el cumplimiento.

Ilustración 11 : Modelo de cumplimiento de la ATO



Fuente: Reproducción de información obtenida de [15].

Este modelo, trata en primer lugar, de la promoción de acciones para fomentar el cumplimiento voluntario, es decir, de acciones que logren que los contribuyentes se muevan desde la parte superior de la pirámide de cumplimiento hasta la base. En segundo lugar, de la priorización del uso de recursos en quienes han decidido no cumplir y tienen un mayor costo de no cumplimiento asociado. Según la ATO los tipos de contribuyentes son:

- Quienes tienen disposición a pagar
- Quienes tienen disposición a pagar pero no siempre logran hacerlo
- Quienes no quieren cumplir sus obligaciones tributarias
- Quienes han decidido no cumplir

Debido a esta forma de entender el comportamiento de los contribuyentes, se genera otro grupo de variables que intenta representar la Intención de Pago ya sea negativa, por quienes han decidido no cumplir, o positiva por los que quieren hacerlo.

2.4.4. Grupos de variables

A los grupos de variables ya descritos, se agrega un último, que se relaciona con características que definen el giro cursado, como por ejemplo, el monto a pagar, cantidad de giros adeudados o tipo de giro.

Finalmente, los grupos a utilizar serán los que se indican más abajo y las variables consideradas en cada uno se explicarán en el Capítulo 4.

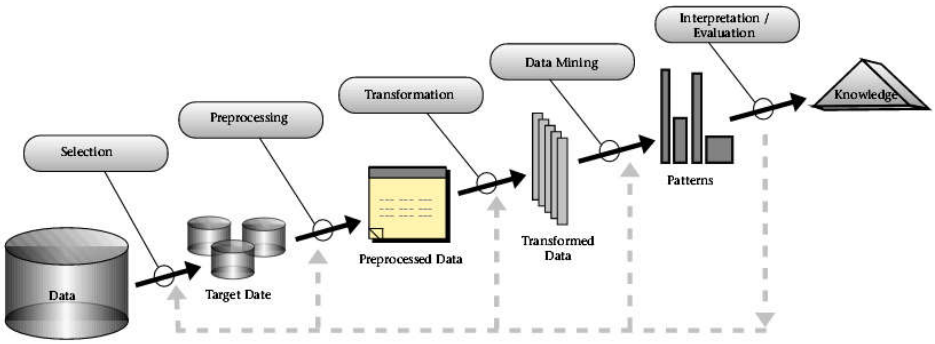
- Caracterización del Contribuyente
- Capacidad de Pago
- Variables F29
- Intención de Pago (positiva y negativa)
- Caracterización del Giro

3. Metodología

3.1. Knowkedge Discovery in Databases

En la memoria se utilizará la metodología Knowledge Discovery in Databases, o KDD, la cual es un procedimiento para la extracción de conocimiento en grandes bases de datos, consistente en una serie de etapas que permiten obtener resultados con mejor desempeño y evitar errores en la modelación. Los pasos son la consolidación de datos, selección y pre procesamiento de datos, transformación de datos, minería de datos e interpretación de resultados [13].

Ilustración 12: Esquema proceso KDD



Fuentes: Figura extraída de [20]

- Consolidación de datos: La primera etapa consiste en organizar la data y volverla accesible, para trabajar utilizando las fuentes disponibles y requeridas al objetivo.

- Selección de datos: Consiste en la elección de un conjunto de datos y variables acotadas para trabajar. Para algunos modelos es útil considerar menor cantidad de atributos ya que los resultados podrían ser más fáciles de analizar, pero esto puede repercutir en el ajuste del modelo, por lo que seleccionar una cantidad adecuada evitaría tener este problema.

- Preprocesamiento de los datos: En esta etapa se realiza la manipulación de datos para que la información contenida en ellos sea de fácil extracción. Esto requiere verificar la calidad de la data, reemplazar valores nulos o ausentes si corresponde, eliminar valores fuera de rango u outliers y otros.

- Transformación de la data: Se trabaja en generar nuevas variables a partir de las primarias, estas nuevas relaciones pueden ayudar a que la información adquiera sentido. Por ejemplo, para la memoria se trabajan variables relativas a otras como monto del giro sobre promedio declarado mensual del F29 ó monto del giro sobre el total de ventas.

- Minería de datos: A través de la aplicación de un algoritmo de aprendizaje se puede extraer conocimiento desde la base de datos trabajada. Se utilizan diferentes modelos, según el tipo de variables que se poseen y el objetivo del estudio, los que se verán en 3.2 Tipos de Modelos.

- Interpretación y Evaluación: Luego de aplicar distintas pruebas, se debe evaluar la calidad del modelo usado, midiendo su desempeño a través de métricas sobre distintas muestras del conjunto de datos. También se deben interpretar los patrones de información obtenidos y evaluar si estos son correctos según los conocimientos previos del negocio o problema.

3.2. Tipos de Modelos

Existen diferentes tipos de modelos que se agrupan en modelos supervisados y no supervisados. Los primeros son modelos que intentan descubrir la relación de un set de variables independientes con una dependiente, por lo que en general, se utilizan como modelos predictivos para lograr determinar el valor de un atributo o clase requerida. El segundo tipo de modelos, no supervisados, son de agrupación o de asociación y se utilizan para encontrar relaciones entre los atributos de la data, es decir, sin una variable objetivo. Como el objetivo de la memoria es encontrar la probabilidad de no pago de un giro, se usarán modelos del primer tipo. Las variables que ayudarán en la predicción serán las pertenecientes a los grupos de variables predefinidos en el capítulo 2.

Los modelos supervisados, a su vez, se subdividen en dos tipos, modelos de clasificación y modelos de regresión:

- Modelos de clasificación: Los modelos de clasificación permiten predecir una variable dependiente del tipo categórica, por ejemplo, una clase "buena" ó "mala" o una marca 1 ó 0. En este caso, se pueden aplicar para determinar las clases "No pago" o "Pago". Algunos algoritmos de clasificación son Árboles de Decisión, Redes Neuronales, Support Vector Machines (SVM), Regresiones Logísticas y Random Forest.

- Modelos de regresión: Los modelos de regresión permiten determinar variables independientes de carácter numérico, como por ejemplo, valores reales. Ejemplos son Regresiones Lineales, otros tipos de Árboles de Decisión como CART, y SVM [21].

De los modelos de clasificación se probarán algoritmos de Árboles de Decisión y Regresión Logística. Los modelos del tipo Redes Neuronales, Support Vector Machines y Random Forest, funcionan como "cajas negras", es decir, las reglas para determinar la predicción no son de fácil interpretación, y por lo tanto, no se puede validar sobre ellas las reglas de negocio del problema [22]. A pesar de esto, se decide probar el modelo Random Forest que no se ha utilizado con anterioridad en el SII y que podría generar predicciones con buenos resultados.

3.2.1. Árboles de Decisión

Los modelos de Árboles de Decisión son modelos muy útiles y de simple interpretación ya que generalmente se representan de forma gráfica, mostrando claramente las relaciones y el aporte de cada variable independiente a la dependiente. Son modelos jerárquicos de decisiones y sus consecuencias. Gráficamente, se muestran como un set de información, que parte desde un nodo cero o principal, y que a través de diferentes criterios se va dividiendo, obteniendo subconjuntos de mayor homogeneidad. Por cada toma de decisión se despliegan otras opciones, por lo que este proceso se conoce como ramificación, donde se llama al conjunto de decisiones "ramas", y al resultado de la clasificación final "hoja". Al conjunto de datos al cual se le aplica una decisión para crear un subconjunto, se le conoce como "nodo padre", y al subconjunto "nodo hijo". Se conoce como poda a la acción de dejar de ramificar el árbol, esto ocurre cuando los subconjuntos no permiten tomar decisiones confiables. Por último, los árboles son modelos jerárquicos de decisiones debido a que las decisiones que logran capturar más información son seleccionadas antes que las demás en el modelo.

Existen muchos tipos de Árboles de Decisión, pero a continuación se explicarán cinco que, aparte de ser reconocidos por la calidad de sus resultados, son posibles de aplicar a través de la herramienta SPSS Modeler utilizada en el SII o programando su algoritmo en softwares de licencia gratuita.

CHAID: El método CHAID (Chi-squared-Automatic-Interaction-Detection) es un algoritmo que por cada atributo independiente selecciona el par de valores del atributo que poseen menor significancia respecto a la variable dependiente. El algoritmo mide si su significancia es mayor a cierto umbral de p-valor de un test estadístico (Chi cuadrado para variables nominales) y sí lo es, junta estos dos valores en un grupo. Se itera esta acción seleccionando un nuevo valor y agregándolo al grupo sí no genera significancias importantes. Luego, de realizar esta acción con cada atributo, se mide cual de todos ellos es el que tiene mayor discriminación de la variable independiente y se selecciona como el atributo con mayor importancia para el modelo. Esta iteración se frena si es que se ha alcanzado la profundidad predeterminada (cantidad de decisiones o ramificaciones), si los nodos hijos ya alcanzaron el mínimo de casos o si los nodos padre tienen un mínimo que no les permite seguir ramificándose.

Existe un derivado de este árbol llamado CHAID Exhaustivo que explora todas las posibles fusiones de cada categoría de una variable para encontrar la más significativa.

CART: CART o Classification And Regression Trees, son árboles de construcción binaria, es decir, en cada ramificación un nodo padre se divide en solo dos grupos. Estos árboles tienen la particularidad de que pueden entregar hojas de resultados numéricos y no sólo de clasificación. El criterio de parada es el índice de Gini.

C5: Es la versión extensión del algoritmo C4.5 que a su vez es la extensión del árbol ID3. El algoritmo usa la medida de entropía para determinar cuáles atributos son más homogéneos entre sí para componer una clase. Luego, se usa la ganancia de información como criterio para priorizar la selección de atributos, parando cuando la ganancia sea igual a cero o se haya definido otro criterio de parada como poda del árbol. La ganancia de información mide la diferencia entre la entropía de los datos del conjunto inicial y la suma ponderada de las entropías cuando ya se ha dividido el conjunto.

$$Information\ Gain\ (S, A) = Entropia\ (S) \quad \sum_{v\ \in Valores\ (A)} \frac{S_v}{S} Entropia\ (S_v)$$

$$Entropia\ (S) = \sum_{i=1}^{k} p_i \quad log_2\ (\frac{1}{p_i})$$

Siendo k el número de clases de un atributo.

QUEST: El algoritmo QUEST (Quick Unbiased Efficient Statistical Tree) es un árbol que a través del estadístico F-test en un ANOVA compara cuál atributo tiene menor relación con la variable dependiente (una distribución distinta y discriminante). El árbol compara el valor F de

cada variable independiente y la que tiene el mayor valor, es decir, la que no cumple la hipótesis nula de igualdad de medias con la variable dependiente, o la cumple en menor medida, es seleccionada para particionar la data. Cuando los atributos son ordinales o continuos se utiliza un F-test, el cual se reemplaza por un estadístico Chi cuadrado cuando se trata de variables categóricas [23]. Dado que el modelo fue diseñado con el objetivo de utilizar menos tiempo de procesamiento, escogida la variable que dividirá la data, el modelo no prueba cada combinación de categorías para dividir, sino solo algunas para subdividir en dos grandes clases definidas por un criterio llamado Quadratic Discriminant Analysis. Por lo tanto, al igual que el árbol CART, QUEST es de carácter binario [29].

Random Forest: Este modelo puede ser utilizado para resolver problemas de clasificación y regresión y lo hace combinando los resultados de múltiples árboles de decisión. Al entrenar este modelo se generan árboles con los que se predice la clasificación para cada registro de la base de datos. Se dice que los árboles "votan" por una clase, luego se contabilizan los resultados de cada una, y se determina a la clasificación que más se repite como el valor predicho. Por ejemplo, para el caso de la clasificación de no pago, si la mayoría de los árboles determinan que un ID de giro tiene altas probabilidades de no pago, este se considerará como un giro que no se pagará [30].

En Random Forest se puede configurar como parámetros el número de árboles a utilizar, y la cantidad de variables que componen un subconjunto desde donde se escoge una variable para particionar el árbol y, al igual que en otros árboles, se puede escoger la cantidad de datos que componen como mínimo el nodo final. A diferencia de otros modelos, este no tiene una representación gráfica fácil ya que utiliza varios árboles juntos, pero sí posee otras ventajas que hacen que Random Forest sea escogido por sobre otros modelos. La mayoría de ellas se relaciona con cómo trabaja la data ya que aplica aleatoriedad en dos procesos para reducir el error de testeo, los que son:

- Selección de la data de entrenamiento para cada árbol: El algoritmo realiza una partición aleatoria de la data, en la que se escoge 2/3 de ésta para entrenar el modelo, y el resto, denominada "Out Of Bag", se utiliza para testear y medir el error de testeo o OOB error rate. Para evitar arrastrar errores debido a la selección de esta partición, al entrenar cada árbol la partición se recalcula y así se logran generar árboles no correlacionados.

- Selección de las variables para ramificar cada árbol: Luego de decidir la data de entrenamiento, Random Forest escoge para cada ramificación, según el número de ramificaciones predefinidas, un grupo de variables al azar (la cantidad de variables también puede ser definida por el analista) del cual se selecciona una para crear una regla de decisión. Esto logra disminuir errores debido a la correlación de variables [26] [30].

Finalmente, para evaluar la importancia de las variables existen dos medidas, MDA (Mean Decrease Accuracy) y MDG (Mean Decrease Gini). La primera mide la importancia de la

variable para determinar el impacto en el Accuracy del modelo, o como la variable influye en el error de mal clasificados, y se obtiene luego de permutar los valores de división de una variable al azar y analizar el cambio entre la predicción inicial y la predicción final. Se calcula una media de las diferencias de estos valores en los árboles donde se aplique la variable, y luego se ordenan decrecientemente. Las variables más importantes, es decir, las que alteren el Accuracy del modelo en mayor medida serán las primeras de la lista y deberían ser seleccionadas y analizadas en profundidad. La medida MDG representa el nivel de impureza de los nodos según el índice de Gini luego de seleccionar una variable. Si el índice de Gini de una variable es cercano a cero significa que cada regla de la división define una sola clase, es decir, define un subconjunto de datos homogéneos, y que la variable discrimina bien en el nodo. De lo contrario, un valor alto define subconjuntos de datos heterogéneos. El MDG se mide calculando la media de la disminución de la impureza de todos los árboles que utilicen la variable testeando diferentes subconjuntos de datos. Si la disminución es alta, la variable es más importante para el modelo [27] [32].

3.2.2. Regresión Logística

La regresión logística es un modelo que pronostica la probabilidad de ocurrencia de un suceso dicotómico en función de N variables independientes de cualquier tipo (numéricas y categóricas) asumiendo la función:

$$P(X) = \frac{1}{1 + e^{-(B_o + \sum_{i=1}^{N} B_i X_i)}}$$

Este método encuentra los parámetros de los valores Beta desconocidos $(B_o, B_1, \dots, B_N)$ que maximizan la probabilidad de obtener el conjunto de datos utilizados a través del método de la máxima verosimilitud. Con los parámetros encontrados, se puede completar y definir la función del modelo buscado.

Este algoritmo, al igual que los Árboles de Decisión, también es de sencilla interpretación ya que el aporte de cada variable independiente a la dependiente es directamente proporcional al valor de su beta asociado. Si el beta que acompaña a una variable independiente es un número positivo, entonces la variable dependiente tendrá mayor probabilidad de ocurrencia. Si el parámetro beta es negativo tendrá una relación negativa, y si el parámetro es cero significa que la variable independiente no aporta al modelo y debería excluirse.

Además de determinar el signo y la magnitud de los parámetros beta, el modelo de regresión logística evalúa la significancia de ellos con el test paramétrico de Wald de distribución Chi cuadrada. Con esto, se puede descartar variables que no aportan al modelo con una selección inversa (que en un comienzo considera todas las variables y va descartando las que no sean útiles) o se puede agregar variables que tengan una significancia considerable en una selección

hacia adelante [13]. Finalmente, de acuerdo a la probabilidad que determina el modelo y un parámetro de corte, se define desde qué probabilidad se considera a un objeto como clase uno.

3.3. Validación de Modelos

Para validar un modelo, antes de aplicar un algoritmo, se divide la data dejando una parte sin utilizar con el fin de testear sobre ella la eficacia del modelo. De todo el conjunto de datos, se crearán los modelos en un subconjunto del 60%, llamado partición de entrenamiento, luego con 20% de la data se testeará el algoritmo, y finalmente sobre el 20% restante o partición de validación se generarán métricas de desempeño del modelo.

3.3.1. Matriz de Confusión

Luego de aplicar un modelo de clasificación, se suelen organizar los resultados de la predicción hecha sobre la partición de validación, en una tabla llamada matriz de confusión en la que se compara la clase real del registro y la clase predicha.

Como se trabaja con modelos que entregan la predicción de una clasificación binaria, la matriz de confusión queda representada de la siguiente forma:

Tabla 2: Matriz de Confusión

Matriz de Confusión		
	Predicción Clase 1	Predicción Clase 0
Real Clase 1	VP	FN
Real Clase 0	FP	VN

Fuente: Elaboración propia con información de [23].

Siendo,

- VP o Verdaderos Positivos: Cantidad de casos positivos, o de clase tipo 1, que fueron predichos correctamente como tal.
- FN o Falsos Negativos: Cantidad de casos positivos que fueron predichos de manera errada, es decir, como clase tipo 0.
- FP o Falsos Positivos: Cantidad de casos negativos, o de clase tipo 0, que fueron predichos de manera errada como clase tipo 1.

- VN o Verdaderos Positivos: Cantidad de casos negativos, o de clase tipo 0, que fueron predichos correctamente como tal.
- Y la suma VP+FP+FN+VN es igual al total de la población a la que se le aplicó la predicción.

La razón de organizar los resultados de esta manera, es poder calcular de manera sencilla diferentes medidas de desempeño que buscan cuantificar la exactitud de clasificación de un modelo. Estas métricas pueden aplicarse también para comparar diferentes modelos [23]. Para la memoria, los giros predichos como que no serán pagados se considerarán la clase positiva o de tipo 1, y los que serán pagados se caracterizan con el valor 0 o clase negativa.

3.3.2. Accuracy

Mide la proporción de casos predichos correctamente sobre el total. Esta métrica asume que el costo de errar al obtener un caso FP es igual a uno del tipo FN, y se suele medir como porcentaje. Esta medida es considerada importante ya que si bien se necesita identificar a los casos de no pago para saber a quién cobrar, conocer de antemano que giros se pagarán también es valioso, porque permitirá no cobrar y ahorrar esos recursos para utilizarlos en otras deudas.

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN}$$

3.3.3. Métricas sobre los valores predichos

Precisión es la proporción entre los casos reales positivos sobre los predichos positivos. Esta medida también es importante, ya que mostrará que tan bien un modelo logra detectar el no pago según lo que ha predicho.

$$Precisión = \frac{VP}{VP + FP}$$

3.3.4. Métricas sobre los valores reales

Recall o Sensibilidad en español, es la proporción de casos predichos positivos sobre el total de casos positivos reales. Specificity o Especificidad en español, es la proporción de casos predichos negativos sobre el total de casos reales negativos.

$$Recall \text{ ó } Sensibilidad = \frac{VP}{VP + FN}$$

$$\text{Specificity ó } Especificidad = \frac{VN}{FP + VN}$$

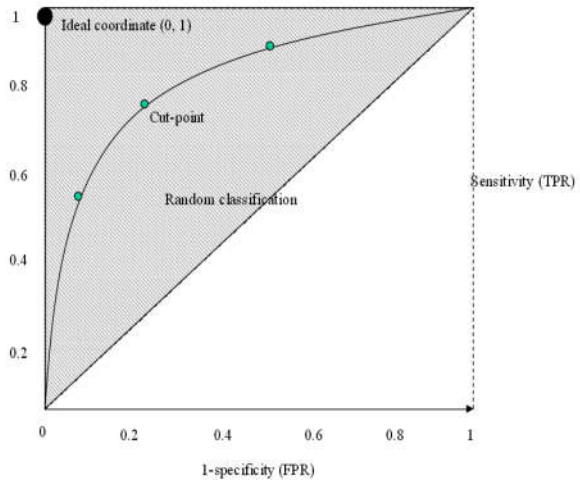
Existe una métrica que combina Precisión y Sensibilidad llamada F-Score. Es muy usada ya que promedia los efectos de cada una de manera equitativa.

$$F \quad Score = \frac{2 \quad (Precisión \quad Sensibilidad)}{Precisión + Sensibilidad}$$

3.3.5. Curva ROC y AUC

La curva Receiver Operating Characteristics, más conocida como curva ROC, es un gráfico que representa la habilidad de un clasificador de obtener casos verdaderos positivos por sobre los falsos negativos. Para esto en el eje X se grafica la Tasa de Falsos Positivos o (1 - Especificidad) y en el eje Y se grafica la Sensibilidad.

Ilustración 13: Curva ROC



Fuente: Figura extraída de [24].

A modo de ejemplo, un punto que se encuentra sobre la coordenada (0,0) representa un clasificador que tiene una tasa de equivocación al predecir un valor positivo igual a cero, pero que al mismo tiempo posee 0 por ciento de Sensibilidad o que nunca predice un valor verdadero positivo. La única forma de que ocurra esto, es no predecir un valor positivo. Al revés, si la coordenada fuera (0,1), se trata de un clasificador muy bueno, ya que de los valores positivos que predice todos en realidad los son, mientras que no se equivoca prediciendo un caso positivo siendo que son negativos. Encontrarse en un punto sobre la diagonal, es tener un clasificador con una tasa de predicción de positivos sobre reales positivos igual a la de predecir erradamente positivos sobre reales negativos, por lo que se dice que el modelo predice en un 50% de las veces bien y 50% erradamente. Por lo tanto, se busca que el gráfico ROC tenga una forma similar al gráfico de logaritmo, aumentando rápidamente en Sensibilidad sin aumentar la tasa de falsos positivos.

El AUC (Area Under the Curve) es la medida del área bajo la curva ROC, por lo que toma valores entre cero y uno.

$$AUC = \int_0^1 ROC(t)dt$$

Una medida de 0.5, sería equivalente a medir el área bajo la diagonal, por lo que este caso describiría al clasificador que predice al azar descrito anteriormente. Por eso, se busca que un modelo tenga un AUC mayor a 0.5 y que logre alcanzar un valor cercano a 1.

Ilustración 14: Clasificación modelos según AUC

Rango AUC	Clasificación
0.9 < AUC < 1.0	Excelente
0.8 < AUC < 0.9	Bueno
0.7 < AUC < 0.8	Sin valor
0.6 < AUC < 0.7	No Bueno

Fuente: Figura extraída de [24].

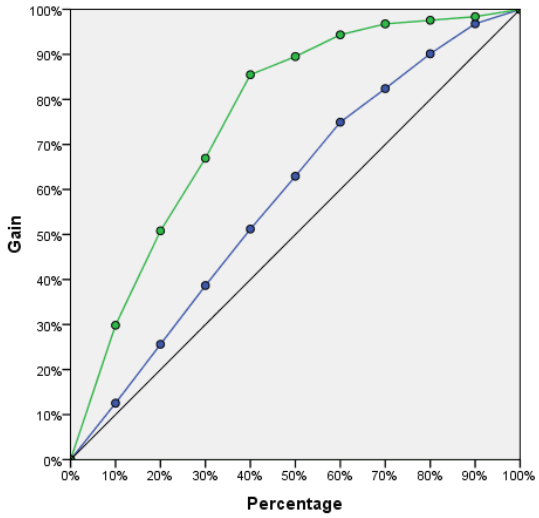
Según la figura, un modelo con un AUC mayor a 0,8 se considera de predicción buena y sobre 0,9 como excelente.

3.3.6. Curva de Ganancia de Información

Los gráficos de ganancia de información muestran el porcentaje de casos predichos positivos al tomar cierto porcentaje de los casos totales. Esto se obtiene luego de realizar una predicción y ordenar las probabilidades de obtener un caso del tipo “positivo” de mayor a menor.

Por ejemplo, como se observa en la Ilustración 15: Curva Ganancia de Información, para capturar un 30% de los casos predichos positivos, cuando estos están ordenados, solo se necesita utilizar un 10% de la data y para capturar un 50%, se necesita aproximadamente un 20% de los datos. Como los recursos para realizar tratamientos son escasos, esta técnica permitirá seleccionar un grupo que represente la mayor razón entre cantidad de giros que no serán pagados y el subconjunto de registros considerados, y así priorizar tratamientos en ellos [25].

Ilustración 15: Curva Ganancia de Información



Fuente: Figura extraída de [25]

Para finalizar, las medidas descritas en este capítulo serán las utilizadas para describir la calidad de cada modelo generado y permitirán compararlos, para escoger al que demuestre ser el más eficiente entre ellos.

4. Desarrollo del trabajo en la Base de Datos

En este capítulo se explicará el trabajo realizado en la base de datos, partiendo por cómo se encuentra almacenada la data en el Servicio, cuál será la variable objetivo y las variables independientes que se utilizarán en el modelo. Se mostrará un resumen de la data inicial y cómo luego de su pre procesamiento y transformación se obtiene la data con la que se trabajará, la que se detalla al final del capítulo.

4.1. Selección y pre procesamiento de datos

4.1.1. Descripción de la base de datos y Definición de No Pago

La base de datos del SII se compone de alrededor de 300 distintas tablas almacenadas en su Data Warehouse[3], por lo que la primera labor que se realizó fue la revisión de algunas de ellas a través del programa SPSS Modeler. Luego, a través de reuniones de trabajo con diferentes áreas de la Subdirección de Fiscalización, especialmente con el área de Control de Gestión, se comprendió el flujo de los giros y cómo la información de cada proceso es almacenada en la tabla de giros llamada "DW.TRN_RPMG_MOVIMIENTOS_GIROS" que contiene 40 atributos.

Se comienza el análisis de la base de datos, aplicando un filtro para extraer la información de giros entre los meses de enero del 2013 y diciembre del 2015. Algunas variables que definen las características de un giro y sus "movimientos" (tipo de acción sobre el giro), son:

- Rut del contribuyente que adeuda el giro
- Folio de giro (Número tipo ID que identifica el giro)
- Fecha de vencimiento del impuesto a pagar (Mes tributario del impuesto que causó el giro)
- Fecha de emisión del giro (Cuando se cursó el giro)
- Fecha del movimiento (Fecha en la que se ejecuto alguna acción sobre el registro del giro)
- Tipo de movimiento (Cargo, Descargo, Pago, Desabono)
- Subscripción a convenio de pago
- Monto de la deuda condonado

En los primeros campos de cada registro de la tabla hay datos de las características del giro, y en los últimos se explica un movimiento. Los tipos de movimientos son Cargo del giro (cantidad de impuesto cobrado), Descargo (disminución del monto cobrado), Pago realizado por el contribuyente, y Desabono (disminución de lo que ha pagado un contribuyente, por ejemplo, por error al ingresar el registro). Con estos cuatro movimientos se puede determinar el Cargo Neto o valor neto de un giro y Pago Neto o monto pagado.

$$Cargo\ Neto = \sum Cargo \quad Descargo$$

$$Pago\ Neto = \sum Pago \quad Desabono$$

Como se muestra en la Tabla 3: Ejemplo simplificado tabla giros, por cada Rut puede haber más de un giro, identificados por números de folio diferentes, y al mismo tiempo cada giro

[3] Data Warehouse es un depósito de datos organizados.

puede tener más de un movimiento. En el ejemplo, el contribuyente de RUT 2 tiene tres movimientos para un mismo giro con folio n° 12350, el cargo neto es de $12.000 pesos y su pago neto es de $12.000.

Tabla 3: Ejemplo simplificado tabla giros

CONT_RUT	FOLIO	MOV_DES	MONTO_M1	MONTO_M2	MONTO_M3	MONTO_M4
RUT 1	12345	Cargo	12400	0	0	0
RUT 1	13144	Cargo	13470	0	0	0
RUT 2	12350	Cargo	16000	0	0	0
RUT 2	12350	Descargo	0	4000	0	0
RUT 2	12350	Pago	0	0	12000	0

Fuente: Elaboración propia.

Un giro se considera pagado solo si el Cargo Neto menos el Pago Neto es menor ó igual a 0, entonces el giro del ejemplo, estaría saldado. En la memoria, sólo se considera un giro como pagado si los pagos son efectuados dentro de los primeros 6 meses posteriores a la emisión del giro. La representación matemática, de la función de pago de un giro *x*, dado un T(x) que representa el tiempo de pago, es la siguiente:

$$f(x) = \begin{cases} 0 = Pagado \quad , & Cargo\ Neto \quad Pago\ Neto \leq 0\ , T(x) \leq 6\ meses \\ 1 = No\ pagado, & Si\ no \end{cases}$$

El resultado de la función será un valor "0" cuando el giro esté pagado en un plazo máximo de seis meses, y "1" cuando no lo esté. Según lo anterior, una deuda que se pague posterior al plazo definido también se considerará no pagada.

Es importante destacar que esta función será la variable dependiente en el modelo predictivo, y que para definir cargo, sólo se considera la deuda original que no incluye reajustes, multas e intereses, pudiendo ser la deuda real aún mayor. Esto ocurre debido a que en la tabla de giros, solo quedan registrados los valores que debiesen ser pagados al último día hábil del mes al momento de la creación de un giro, y se recalculan sí es que el contribuyente paga, por lo que no se puede saber con certeza cuál es el monto de dichos conceptos para quienes tienen deuda pendiente.

Finalmente, a partir de la tabla de giros se crea una base nueva, en la cual se mantienen las características del giro; RUT, folio y fechas, pero se resumen los movimientos en los atributos Cargo Neto y Pago Neto por cada giro. Con esto se reduce la cantidad de registros y se mejora el manejo de la data.

4.1.2. Selección de variables

La base agregada de registros de giro, generada previamente, se complementa con atributos que sean útiles para explicar el no pago de giros según los grupos predefinidos en la sección 2.4.4. Grupos de variables o que al transformarlos logren el mismo objetivo. Los atributos se anexan a la base inicial de giros a través de los ID de RUT, número de folio y fecha de emisión del giro. A continuación, se explican los grupos de variables en detalle y algunas de las variables que los conforman:

- **Caracterización del contribuyente**: Para caracterizar al contribuyente se utilizan variables demográficas y del tipo de empresa. De los registros personales de los contribuyentes, se obtienen variables como año de nacimiento, que permite obtener su edad y género (cuando corresponda). De los registros de empresa se obtienen la actividad económica que realiza el contribuyente, la región donde se encuentra la unidad del SII que lo atiende, fecha de iniciación de actividades (lo que permite obtener los años de actividad), y otros.

- **Capacidad de Pago**: Este grupo de variables intenta mostrar capacidad de pago y su estabilidad en el tiempo. Para ello se seleccionan atributos como patrimonio y cantidad de trabajadores. El primer atributo, se representa como valor y cantidad de bienes raíces del contribuyente. Por la información obtenida a juicio experto (personal de la TGR), se puede inferir que a mayor monto y cantidad de patrimonio existirá un mejor pago, por embargo de una propiedad, por la posibilidad de venta de este activo con el objetivo de cancelar la deuda o por el hecho de que tener un bien podría significar que existe un respaldo económico y acceso financiero mayor. Para el presente estudio, se utilizó una base de datos de bienes raíces del año 2009, por lo que se debe indicar que algunas propiedades podrían ya no pertenecer a sus dueños. Para obtener el valor actual, se actualizan los avalúos de acuerdo al I.P.C entre los años 2009 y 2015.

- **Variables F29**: Para determinar distintos ratios de capacidad de pago mensual, se trabaja con algunos campos del formulario F29, como ventas y compras, débitos y créditos, impuesto pagado (o código 91) y tickets de venta promedio. Con las variables descritas, se puede obtener una aproximación al resultado de la empresa. Para esto, se seleccionaron todos los formularios a un año móvil a partir de la fecha de emisión del giro y se calculó la suma anual, o el promedio mensual con base anual, de los atributos escogidos.

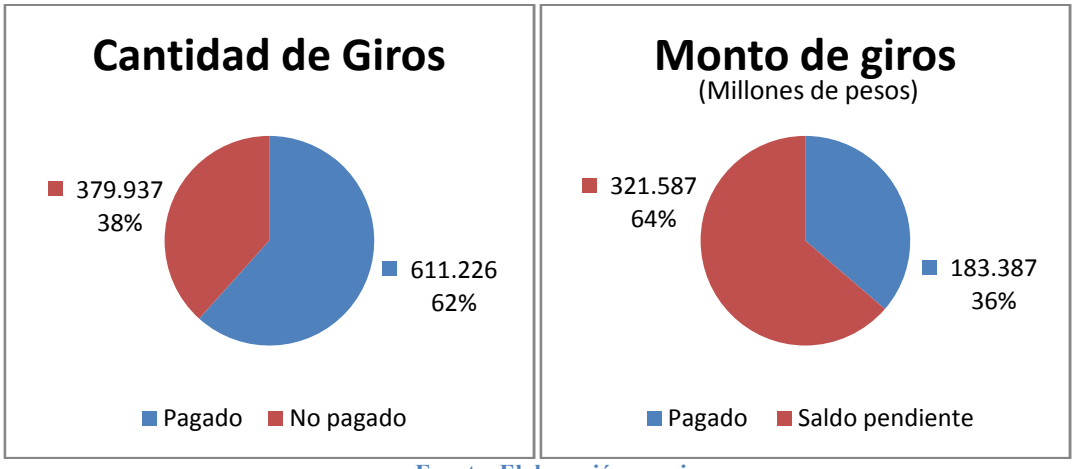
- **Intención de Pago (positiva y negativa)**: Este grupo de variables pretende demostrar la aseveración de la ATO, y del mismo SII, de que existen contribuyentes que cumplen sus obligaciones tributarias de forma voluntaria o que tienen una buena intención de pago, y que hay otros, que no quieren cumplir o que tienen una intención de pago negativa. En intención de pago positiva se encuentra la marca "tenencia de

convenio de pago" que muestra cuando un contribuyente se ha acercado a regularizar su situación comprometiéndose a pagar su deuda en una o varias cuotas. Por otro lado, la intención negativa se representa con anotaciones negativas, como anotación de fiscalización, anotación de no declarante de F29, anotación de querellado y otras.

- **Variables de Giro**: Las variables de giros son otras variables que se extraen de la base inicial de giros, no consideradas primeramente en la base agregada, estas son monto del giro, tipo de giro cobrado (Rectificatoria F29, Impuesto F29, Multa F29 o Auditoria IVA), fechas de pago, etc.

Antes del pre procesamiento, y luego de incorporar los grupos de variables, se cuenta con una data inicial sin correcciones correspondiente a 991.163 giros, para 354.917 contribuyentes diferentes, con cargos cursados que ascienden a $504.974 millones de pesos. En los gráficos siguientes, se muestra la proporción de giros pagados y no pagados, y la de montos no pagados y saldo pendiente (cargos menos pagos parciales).

Ilustración 16: Data Inicial



Fuente: Elaboración propia.

4.1.3. Pre procesamiento

En el pre procesamiento de la data se realizó una limpieza de datos para evitar modelar con datos inconsistentes con el problema. Esta limpieza se desarrolló en el siguiente orden,

- Eliminación de movimientos de giros que se hayan realizado fuera del período enero de 2013 y diciembre de 2015.

- Eliminación de giros emitidos desde junio de 2015.[4]
- Eliminación registros duplicados
- Eliminación registros de Cargo Neto negativo
- Eliminación registros de Pago Neto negativo
- Eliminación de Ruts de prueba ficticios utilizados para testeo.
- Eliminación de giros del tipo Multa F29 con cargos mayores a una UTA
- Eliminación de los giros por Rectificatorias (Explicación en 4.3.1 Análisis Univariados)

Después de eliminar registros inconsistentes, se procedió con el análisis de valores atípicos de las variables numéricas como montos de giros y variables del F29. Para el caso de la variable Cargo Neto, se decidió no eliminar valores atípicos (o mayores a la media más cuatro veces la desviación estándar) ya que se eliminarían giros que conforman un segmento de interés para la TGR, además de sumar cerca de 61.100 millones de pesos lo que corresponde al 19,4% de los montos cobrados, considerado alto. Se probó eliminar valores mayores a la media más seis veces la desviación estándar, y se obtuvo que se descartarían 14,5% de los montos por lo que tampoco se utilizó esa opción.

Para el caso de las variables del formulario F29, se decidió no eliminar valores atípicos por reglas de negocio. Se asume que los contribuyentes no declaran valores de ventas y débitos, ni pagan impuestos, más altos de los que deberían pagar, por lo tanto, aunque sean extremos no serían anormales o errados. Además, en un primer análisis de correlación todos los cargos, no solo los valores atípicos, están fuertemente correlacionados con las ventas, compras y débitos, por lo que eliminar valores en estas variables haría perder parte de la explicación de los cargos mayores.

Por último, se decide no imputar valores faltantes ni eliminar variables en esta primera etapa y se corrigen algunos registros con valores negativos. Esta etapa, sin la eliminación de registros del tipo Rectificatoria F29, concluye con un número de 669.272 giros, cerca del 67% de la data primitiva.

4.2. Transformación

Las variables primarias se transformaron para poder obtener mejor información y/o volverlas más interpretables. Mayoritariamente, se desarrollaron ratios que combinan el monto del giro ó Cargo y otras variables como ventas y compras. Ejemplos de ellas son:

[4] Se consideran solo registros hasta junio 2015, para poder definirlos como pagados en 6 meses con los movimientos hasta diciembre 2015.

- Cargo/Ventas
- Cargo/C91 (Sobre el monto de impuestos promedio mensual)
- Cargo/(Ventas+Compras)
- Cargo/(Ventas-Compras)
- Cargo/Trabajadores
- (Ventas-Compras)/Ventas

Además se realizaron otras transformaciones utilizando variables de giros, como la determinación de la cantidad total de giros anteriores, conteo de giros impagos y tasa de no pago de giros anteriores hasta el mes previo a la fecha de emisión del giro. Con las transformaciones, en total se contabilizan 50 variables. El listado final para los distintos grupos de variables se encuentra en Anexo 10.3, pero en la tabla siguiente se detallan las más importantes.

Tabla 4: Listado parcial de Variables

	Grupo de variable	Nombre variable	Tipo de Variable	Descripción
ID	ID	Folio	Numérica	Número identificador del giro
	ID	RUT	Numérica	Número identificador del contribuyente
	ID	Año_Mes_Emision	Numérica	Año y mes de emisión del giro
Variables dependientes	Caracterización del Giro	Concepto_Giro	Nominal	Tipo de Giro = {Multa F29, Impuesto F29, Rectificatoria F29, Auditoría IVA}
	Caracterización del Giro	Cargo_Neto_Cat	Numérica Ordinal	Categorías de monto de los giros
	Caracterización del Giro	Tasa_No_Pago	Numérica	Porcentaje de giros pasados al giro que se está prediciendo que no han sido pagados. Valores cercanos a 0% representan un pago total de giros anteriores.
	Caracterización del Giro	Giros_Impagos	Numérica	Cantidad de giros pasados al giro que se está prediciendo que no han sido pagados
	Caracterización del Giro	Trimestre_Pago	Numérica	Trimestre del año de emisión del giro
	Caracterización Contribuyente	Segmento	Nominal	Tamaño de empresa al año 2015 Micro o Pequeña = {SGMI,SGPM}
	Caracterización Contribuyente	Años_Actividad	Numérica	Años desde el inicio de actividades de la empresa
	Caracterización Contribuyente	Actividad_Economica	Nominal	Sector económico denominado por letras (Ej: I=Hoteles y Restaurantes)

	Caracterización Contribuyente	Regional	Nominal	Unidad regional del SII que atiende a la empresa
	Caracterización Contribuyente	Genero	Nominal	Femenino o Masculino = {F,M} (Si la empresa es formalizada como empresa natural)
	Caracterización Contribuyente	Edad_Contribuyente	Numérica	Edad en años (Si la empresa es formalizada como empresa natural)
	F29	Promedio Mensual C91	Numérica	Promedio mensual de Impuesto F29 declarado en los ult 12 anteriores a la emisión del giro
	F29	Cargo/C91	Numérica	Relación entre cargo y Promedio Mensual C91
	F29	C91/C91 Actividad	Numérica	Promedio mensual C91 sobre promedio del rubro
	F29	Ventas_u12m	Numérica	Monto de ventas de los ult 12 meses anteriores a la emisión del giro
	F29	Debito_u12m	Numérica	Monto de débito de los ult 12 meses anteriores a la emisión del giro
	F29	Credito_u12m	Numérica	Monto de crédito de los ult 12 meses anteriores a la emisión del giro
	F29	Ventas+Compras	Numérica	Promedio mensual de ventas y compras ult 12 meses
	F29	Cargo/(Ventas+Compras)	Numérica	Relación entre cargo y la suma del promedio mensual de ventas y compras
	F29	Ticket_promedio_facturas	Numérica	Promedio de valor de facturas ult 12 meses
	Intención de Pago	Convenio?	Nominal	Marca de suscripción a convenio Si o no = {1,0}
	Intención de Pago	Anotación?	Nominal	Marca de tenencia de alguna de 4 anotaciones negativas. Si o no = {1,0}
	Capacidad Pago	Trabajadores	Numérica	Cantidad de trabajadores
	Capacidad Pago	Sueldos	Numérica	Promedio de sueldo mensual del año tributario anterior al de la emisión del giro
V.OBJ.	Función Objetivo	Pagado	Marca	Función objetivo que muestra si el giro no será pagado. No pagado o Pagado = {1,0}

Fuente: Elaboración Propia.

En la siguiente tabla, se resume la cantidad de variables por grupo. La mayor cantidad pertenece al grupo de variables F29[5].

Tabla 5: Variables por Grupo

Grupos de Variables	N° Variables
Caracterización Contribuyente	8
Capacidad de Pago	5
Variables F29	20
Intención de Pago	7
Caracterización Giro	10
Total	**50**

Fuente: Elaboración Propia.

4.3. Análisis de variables

4.3.1. Análisis Univariados

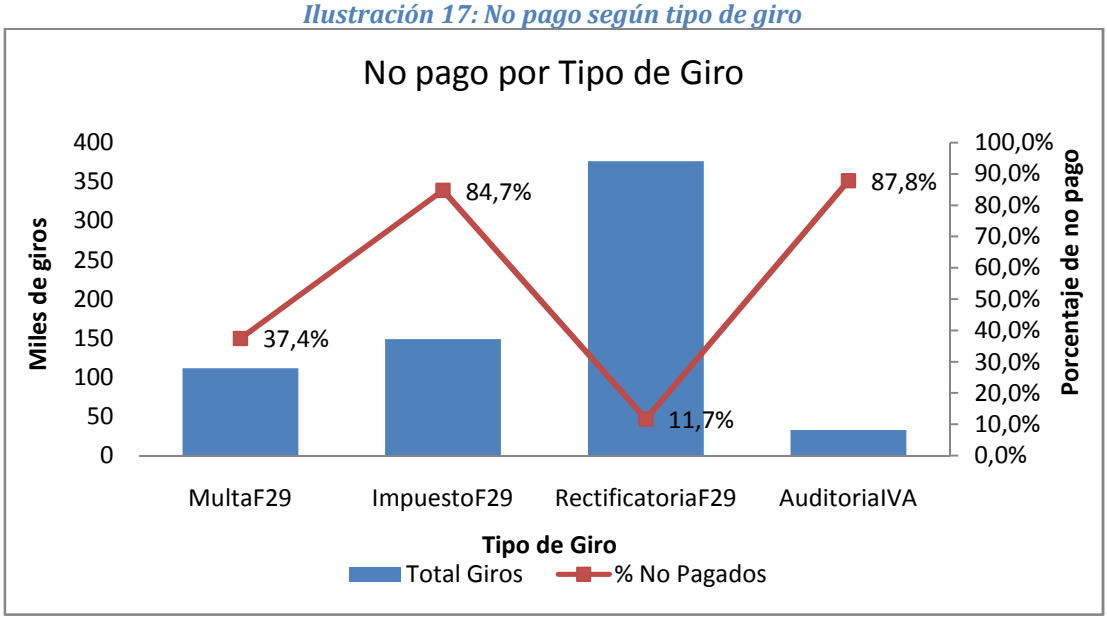
Al tener todas las variables creadas se realizó un primer análisis de variables, a continuación se muestran algunos de ellos[6],

- Concepto de Giro

Al estudiar el problema según el tipo de giro, se encuentran grandes diferencias de no pago entre ellos. En la Ilustración 17: No pago según tipo de giro, se muestra el porcentaje de no pago y el conteo de giros por categoría. Se observa que los giros por Rectificatoria F29 son el tipo de giro de mayor volumen y que se pagan en un porcentaje muy alto, que casi alcanza el 90%, seguido por las Multas F29 con un 62,6%. El buen nivel de pago de las Rectificatorias puede deberse a su carácter voluntario y a que las que se realizan por internet, se deben pagar automáticamente por este medio.

[5] En Anexo 10.4 podrá ver el detalle de las categorías de las variables nominales Cargo_Neto_Cat y Actividad_Económica.
[6] En Anexo 10.5. se encuentran otros Análisis Univariados.

Ilustración 17: No pago según tipo de giro



Fuente: Elaboración propia.

Más abajo, se detallan los montos por Concepto o tipo de giro. El giro de mayor monto promedio, corresponde a las Auditorias IVA, siendo casi 30 veces más que el promedio de las Multas F29.

Tabla 6 : Montos y giros por concepto

Concepto de Giro	Cantidad de Giros	Promedio Monto Giro
Multa F29	111.464	$ 84.444
Impuesto F29	148.889	$ 1.182.279
Auditoria IVA	32.587	$ 2.593.800
Rectificatoria F29	376.332	$ 302.567
Total	669.272	$ 1.040.773
Total s/Rectificatoria	292.940	$ 1.286.841

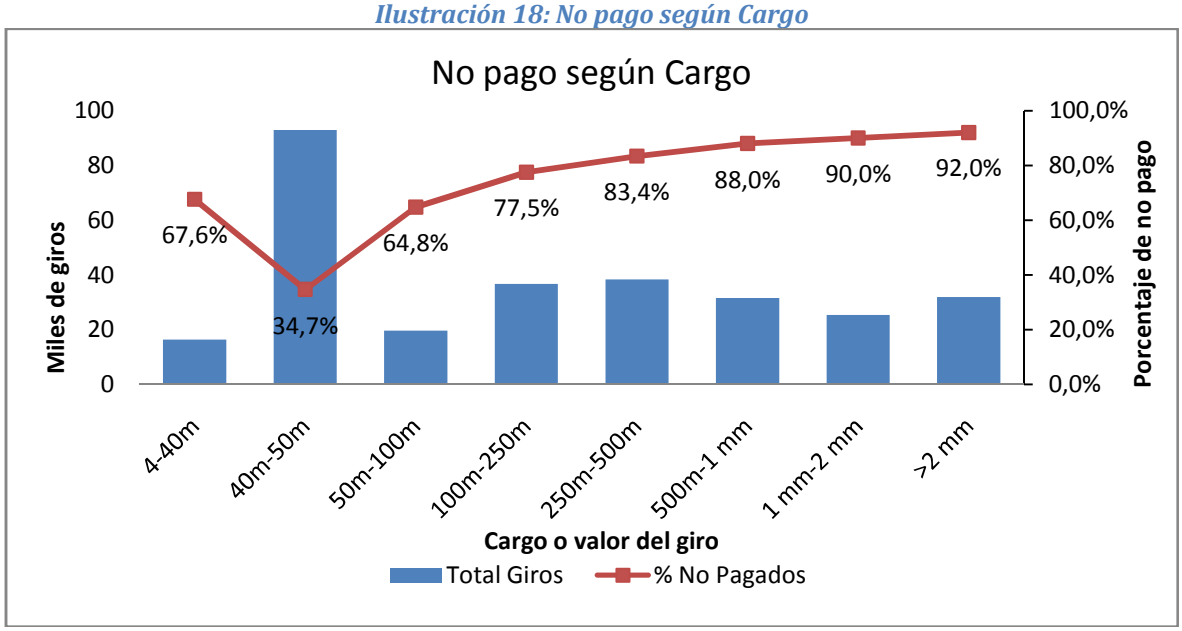
Fuente: Elaboración propia.

Del análisis de giros no pagados, entre Auditoria IVA, Impuesto F29 y Multa F29 se logra capturar más del 80% de los cargos y volumen de giros impagos. En razón de esto y el hecho de que las Rectificatorias F29 se pagan en altísimo porcentaje, se decide excluir a este tipo de giros del modelo a realizar.

- Cargo Neto

Excluidas las Rectificatorias F29, se observa el comportamiento de pago según el monto del giro o Cargo Neto. Los giros se distribuyen de la siguiente manera:

Ilustración 18: No pago según Cargo



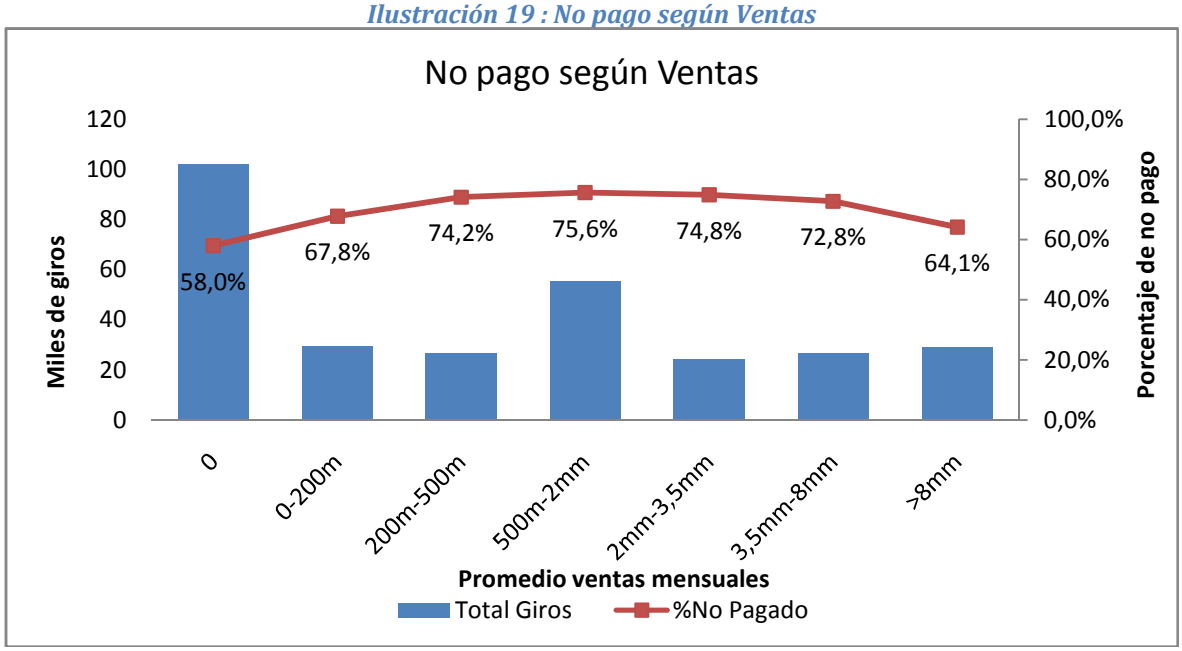
Fuente: Elaboración propia.

Del gráfico se extraen tres comportamientos de pago, uno para deudas menores a 40 mil pesos, otro para deudas entre 40 mil y 50 mil pesos que concentra la mayor cantidad de giros, y un tercero a partir de 50 mil. Para los dos primeros comportamientos de montos bajos, llama la atención, el mejor pago de los giros entre 40 y 50 mil pesos, que los de deuda menor a 40 mil pesos. Este resultado podría depender del tipo de giro, ya que el primer grupo de giros se compone en más de un 80% de giros tipo Impuesto F29, y el segundo en más de un 80% de Multa F29. Por último, desde los 50 mil pesos, se advierte que a mayor cargo mayor el porcentaje de no pago de giros, el que alcanza hasta un 92% para montos sobre 2 millones de pesos.

- Ventas:

Interesa también relacionar el no pago con el promedio mensual de las ventas con base anual (del año móvil anterior a la emisión del giro). Se esperaba encontrar que los contribuyentes de menores ventas tuvieran el porcentaje de no pago más alto ya que no tendrían dinero para pagar. No ocurre así, y son las empresas de ventas medias las que poseen un peor pago. Lo que sí se cumple es que los contribuyentes de mayores ventas pagan mejor que otros, como es el caso de los que venden sobre 8 millones de pesos.

Ilustración 19 : No pago según Ventas

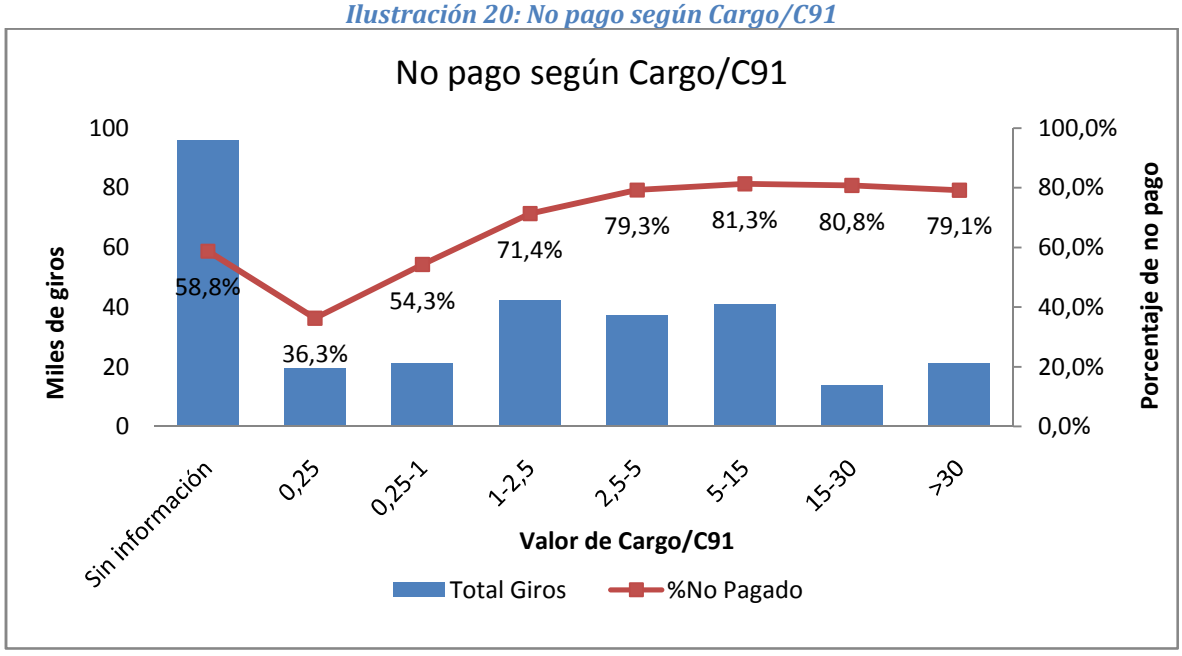


Fuente: Elaboración propia.

- Cargo/C91:

En el siguiente gráfico se muestra como el no pago aumenta conforme al número de veces que el cargo es mayor al monto declarado de impuestos. Para las categorías menores a 2,3 veces, el no pago va desde un 36,3% a un 73,8%, pero desde 2,4 veces el no pago se mantiene cercano al 80%. Es decir, desde un cargo de 2,4 veces para las empresas es muy dificultoso pagar en seis meses.

Ilustración 20: No pago según Cargo/C91



Fuente: Elaboración propia.

En esta variable, la categoría "Sin información" se origina, primero por valores faltantes y segundo por formularios con el código 91 igual a 0 (que indefine al ratio). Que el C91 sea igual a cero significa que la empresa obtuvo remanente fiscal (crédito mayor a débito) o que no tuvo movimiento comercial.

4.3.2. Análisis de Correlación

En esta sección se buscaron variables que estuvieran altamente asociadas entre sí y que pudieran entregar información redundante ya que al variar una de ellas variaría la otra. Esta asociación se mide a través del coeficiente de correlación r de Pearson, el cual al tener un valor cercano a 1 denota una relación lineal positiva perfecta entre las variables en análisis, un valor cercano a -1 muestra una relación lineal negativa perfecta, y un valor 0 un caso donde no existe correlación [28].

Debido a la etapa de transformación de variables, en la que se generaron varios ratios entre ellas, casi la mayoría están altamente correlacionadas con las variables que las formaron. Por lo tanto, se analizan variables correlacionadas (dado un coeficiente r mayor a 0,5) que no sean ratios, para no considerarlas al modelar en el capítulo 5. Las que se excluyen son las de la columna "variable 2".

Tabla 7: Variables correlacionadas

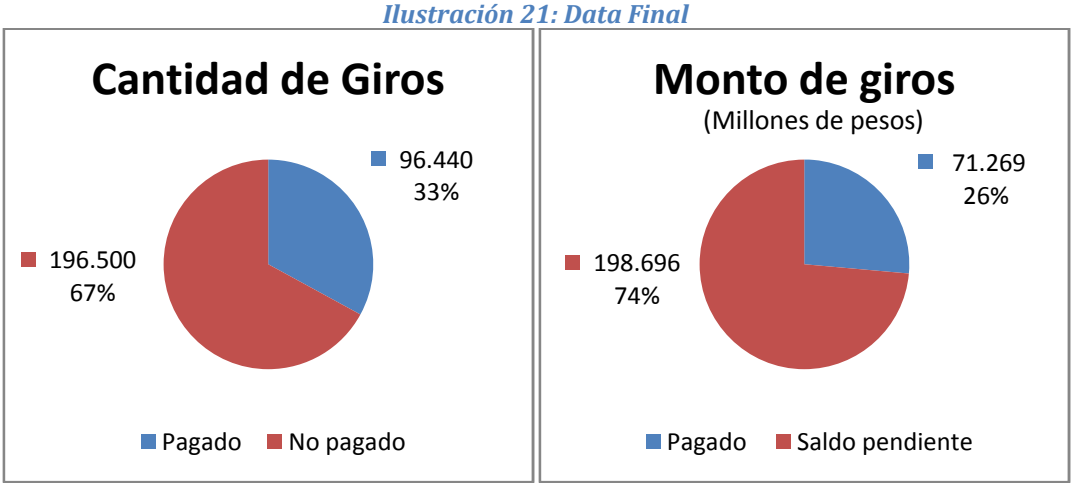
Variable 1	Variable 2	Coef. Correlación
Giros_Impagos	Giros_Total	0,91
Ventas_u12m	Compras_u12m	0,86
Trabajadores	Sueldos	0,75
Patrimonio_Estatico_Sum	Patrimonio_Estatico_Count	0,66

Fuente: Elaboración propia.

4.4. Detalle del Problema Final

Finalmente, luego del trabajo de limpieza de datos y eliminación de los giros por Rectificatoria F29, queda un 29,5% de la base original correspondiente a 292.940 giros.

Ilustración 21: Data Final



Fuente: Elaboración propia.

En relación a los contribuyentes, se tiene que el total de giros con los que se trabajará, pertenecen a 109.936 empresas diferentes. Según la clasificación del SII realizada para el año 2015, un 23,9% de ellos son pequeños empresarios y el resto micro empresa. Con esta agrupación, se obtiene que ambos grupos poseen cerca de tres giros cursados en el periodo.

Tabla 8 : Segmentos de contribuyentes

Segmentos SII	Contribuyentes	% Contribuyentes	Promedio Giros
Micro Empresa	83.652	76,1%	2,5
Pequeña Empresa	26.284	23,9%	3,1

Fuente: Elaboración propia.

Con lo realizado se generan supuestos que permitirán comenzar el trabajo de modelación, los que se detallan a continuación:

- Existen tipos de giros en los que el no pago es más alto, como Auditoria IVA e Impuesto F29, debido a que estos giros en promedio tienen cargos más altos.
- Los contribuyentes con mal historial de pago tienen menores probabilidades de pagar giros nuevos, pero, no por un mal comportamiento, sino debido a la cantidad de giros impagos o montos asociados impagos.
- Para una relación entre Cargo y C91 igual o menor a uno, existe una mejor tasa de pago que para valores mayores. La empresa podría pagar lo mismo que paga en impuesto como máximo de giro.
- El tener convenio de pago origina una mayor respuesta por parte de la empresa, pero no asegura completamente el pago.
- Las anotaciones negativas afectan, pero como las marcas están presentes en un grupo minoritario de los datos, un mal comportamiento pasado no sería la razón principal de no pago.

- Finalmente, muchos comportamientos de variables confluyen en que el cargo de giro sería la variable más importante para definir el no pago, el cual es mayor a medida que el cargo del giro aumenta. Esta visión es compartida por funcionarios del SII.

5. Resultados Modelos

Para determinar las probabilidades de no pago se han generado distintos modelos, que se dividen en tres grupos según la técnica de predicción utilizada. Todos han sido creados utilizando una partición de entrenamiento correspondiente al 60% de los datos, y las métricas que se exponen en este capítulo se han calculado sobre la partición de validación de un 20% de la data.

El primer grupo se compone de cinco modelos de árboles de decisión, el segundo de cinco regresiones logísticas y el tercero de 12 variantes de un modelo Random Forest. Se destaca que para crear los árboles y las regresiones se utilizó la herramienta SPSS Modeler, que utiliza usualmente el Servicio, mientras que los bosques Random Forest fueron programados en el software libre R Studio.

Este capítulo examina como se escoge el mejor modelo de los creados. Para esto se efectúa un análisis comparativo de métricas de desempeño (expuestas en 3.3. Validación de Modelos) y otros criterios, para escoger un modelo por cada grupo. Esos modelos se vuelven a comparar entre sí, según sus variables más importantes, las curvas de ganancia y ROC, para posteriormente seleccionar el que posee el mejor desempeño y aplicabilidad. El modelo final será explicado en detalle en el capítulo siguiente.

5.1. Elección modelo de tipo Árbol de Decisión

Para generar modelos de árboles de Decisión se elige uno de cada tipo de los siguientes algoritmos: CHAID, CHAID Exhaustivo, CART, C5 y QUEST. Para comparar sus métricas de desempeño se elabora la Tabla 9, donde se destaca en negrita los valores máximos que arrojó cada una de ellas.

Se concluye que los valores de estas medidas no son muy distintos entre sí, dándose la mayor diferencia, de casi un 6%, entre el valor mínimo de AUC de 76,6% y el máximo de 82,7% que alcanza el árbol CHAID, y que además, todos ellos son bastante buenos. El mejor desempeño lo obtiene el árbol C5 (con 3 máximos) y el peor es el modelo QUEST, que no logra tener desempeño máximo en ninguna medida y tiene cuatro valores mínimos. Posiblemente, debido a que el modelo no itera sobre todas las posibles particiones de la data, dividiendo rápidamente en ramas binarias.

Tabla 9: Desempeño de árboles SPSS

Métricas	CHAID	CHAID Exhaustivo	CART	C5	QUEST
Accuracy	78,2%	78,0%	78,4%	**78,6%**	77,2%
Precisión	83,1%	**83,4%**	82,8%	**83,4%**	81,7%
Sensibilidad	84,5%	83,6%	**85,4%**	85,0%	84,7%
Especificidad	65,6%	**66,8%**	64,4%	65,7%	62,2%
F-Score	83,8%	83,5%	84,1%	**84,2%**	83,2%
AUC	**82,7%**	82,5%	82,5%	77,1%	76,6%

Fuente: Elaboración propia.

Por otra parte, los buenos resultados del modelo C5 se contrarrestan con su pobre explicación para los giros de Auditoria IVA e Impuesto F29. El árbol luego de subdividirse según la variable Concepto de Giro, solo define que si alguien tiene uno de estos giros no los pagará con una probabilidad de 84,9%, sin más ramificaciones. Esta misma situación ocurre con los modelos CART y QUEST. Por esta razón, estos tres modelos se califican negativamente en "Explicación del problema" a través de variables relacionadas con el no pago, ya que tener información solo de Multa F29, no permite entender el problema general ni generar recomendaciones, ambos objetivos específicos del trabajo. De todas maneras, se destaca que al ser estos últimos árboles de división binaria, la explicación que entregan es sencilla de interpretar.

En relación a los árboles CHAID y CHAID Exhaustivo, estos lograron explicar en detalle pocos grupos de giros de Auditoria IVA, pero varios para Impuesto F29, por lo que se consideran superiores a los modelos C5, QUEST y CART en su explicación.

Para resumir lo anterior, se califican los modelos del uno al tres según los criterios de métricas de desempeño, identificación de variables y facilidad de interpretación. Según se muestra en la tabla siguiente, los mejores resultados son obtenidos por CHAID y CHAID Exhaustivo.

Tabla 10: Comparación árboles SPSS

Modelos	Métricas de Desempeño	Explicación del problema	Facilidad de Interpretación
CHAID	***	**	***
CHAID Exhaustivo	***	**	***
CART	***	*	***
C5	***	*	***
QUEST	**	*	***

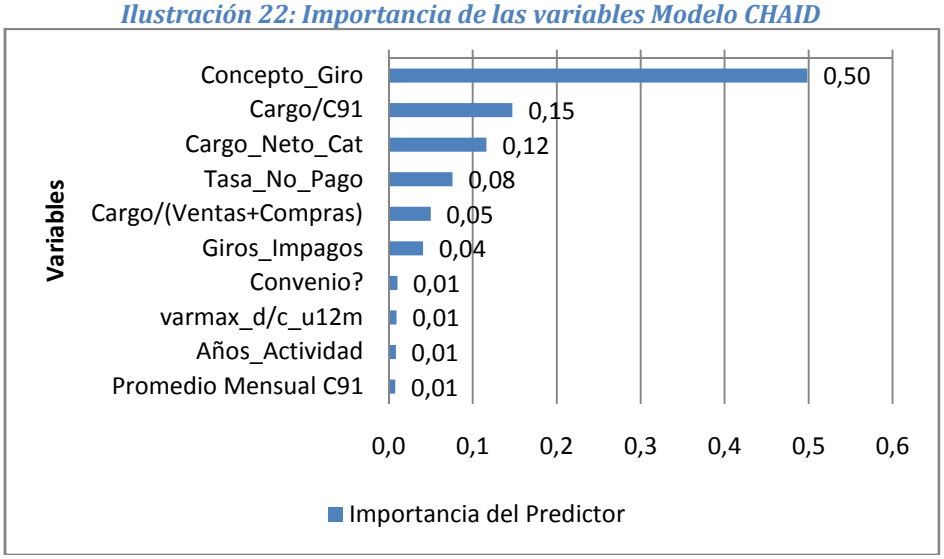
Fuente: Elaboración propia.

Para escoger uno de estos dos modelos, se vuelven a comparar sus medidas de desempeño. Por esto se destaca que el Accuracy y la Precisión son las medidas consideradas más importantes para el problema a resolver, ya que miden la capacidad de detectar No Pago y Pago, sobre el total de casos, y de predecir de buena manera el no pago sobre el total de casos predichos como No Pago. Cada modelo posee sólo un máximo para estas medidas, por lo que se dirime con el valor de AUC, y como este es mayor para el modelo CHAID, se escoge este último modelo.

5.2. Resultados Árbol CHAID

Para entrenar el árbol CHAID se utilizaron 46 variables, y un máximo de 5 ramificaciones, sin costos de error. A través del criterio Chi cuadrado, el árbol selecciona 19 variables útiles, lo que representa 40,4% del total. En la figura siguiente, se muestran las variables más importantes para el modelo según el output "Importancia del predictor" de SPSS. Este es un ranking del 0 al 1, donde las variables más importantes para la predicción, en relación a las demás, tienen valores cercanos a 1 y las que no lo son valores cercanos a 0.

Ilustración 22: Importancia de las variables Modelo CHAID



Fuente: Elaboración propia con información de SPSS Modeler.

Según esto, las variables más importantes del modelo son: Concepto de Giro con una importancia de 0,5, Cargo/C91 con 0,15 y Cargo_Neto_Cat con 0,12. Por otra parte, de la partición de validación se obtienen los siguientes resultados de clasificación del árbol, donde se analiza que la cantidad de casos predichos erradamente del tipo Falsos Negativos representan un 90% de los casos de Falsos Positivos.

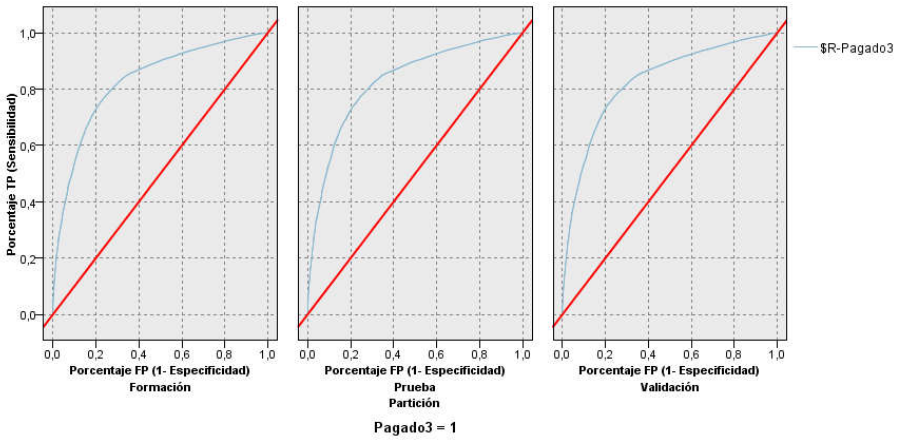
Tabla 11: Matriz de Confusión CHAID

Matriz de Confusión		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	33100	6053
Real Pago	6728	12854

Fuente: Elaboración propia.

Luego, al analizar la curva ROC, que gráfica las tasas de predecir correctamente un caso de no pago (eje y) versus la tasa de errar al predecir un pago, con un valor de 60% en la tasa de falsos positivos se obtiene un 92,8% de Sensibilidad, lo que muestra que la curva sube rápidamente su valor en el eje y, en relación al aumento de la tasa de error del eje x, que es lo deseable para un modelo. Además, el AUC de la curva ROC es 82,7%, lo que representa a un modelo con buen poder predictivo.

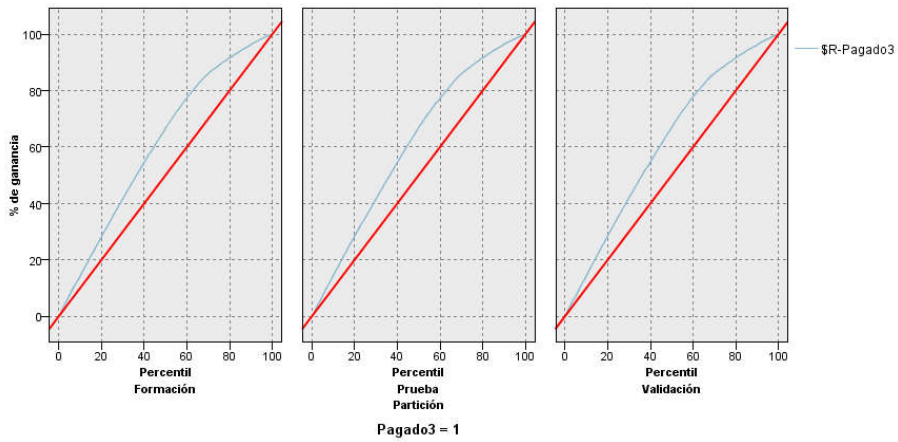
Ilustración 23: Curva ROC CHAID



Fuente: Gráfico extraído de SPSS Modeler.

Finalmente, se expone la curva de ganancia de información que muestra, con la línea azul, qué porcentaje de giros predichos como No Pago se captura al predecir sobre cierta cantidad de giros cuando se encuentran ordenados por probabilidades de mayor a menor. La línea roja es una línea base que establece las ganancias para un modelo que predice al azar. El árbol CHAID, considerando un 60% del total de empresas obtiene aproximadamente un 77,8% de las que no pagarán. Esto demuestra que la captura de casos de no pago no es tan alta en relación a los contribuyentes considerados, probablemente porque el modelo se equivoca al predecir No Pago, quedando estos con baja probabilidad de no pago, y para atender esos casos restantes se requerirá considerar a la gran mayoría de los contribuyentes.

Ilustración 24: Curva de Ganancia CHAID



Fuente: Gráfico extraído de SPSS Modeler.

5.3. Elección modelo de tipo Regresión Logística

La generación de los modelos de regresión logística involucró como primera tarea el re procesamiento de la data, al categorizar variables numéricas para que el modelo pudiera compararlas sin importar las diferentes unidades de medida de ellas. Además, existen variables con alta cantidad de datos nulos, donde el no tener información es un caso con valor propio, por lo que en vez de reemplazar esos datos, con alguna técnica estadística, también se categorizan.

Por otra parte, a diferencia de lo que ocurrió con los modelos de árboles de decisión, el modelo de regresión logística es más sensible a una alta cantidad de variables, requiriendo mucho tiempo de procesamiento y entregando resultados no satisfactorios como medidas de desempeño bajas. Por esta razón, se entrenaron varios modelos previa determinación de subconjuntos de variables a utilizar. A continuación se muestra el desempeño de cinco modelos de regresión logísticas generados:

Tabla 12: Desempeño Regresión Logística

Métrica	Modelo 1 4 variables	Modelo 2 7 variables	Modelo 3 6 variables	Modelo 4 7 variables	Modelo 5 6 variables
Accuracy	**78,4%**	78,1%	76,8%	77,1%	76,8%
Precisión	83,1%	83,6%	83,3%	83,2%	**83,9%**
Sensibilidad	**84,8%**	83,6%	81,5%	82,2%	80,6%
Especificidad	65,5%	67,1%	67,4%	66,8%	**69,2%**
F-Score	**84,0%**	83,6%	82,4%	82,7%	82,3%
AUC	**81,9%**	81,6%	80,5%	80,5%	80,5%

Fuente: Elaboración propia.

Si bien el modelo 1 posee máximos en 4 de 6 medidas de desempeño utilizadas, solo trabaja con cuatro variables, Concepto_Giro, Cargo_Neto_Cat, Tasa_No_Pago y Regional, las que no permiten generar recomendaciones más específicas para abordar el problema. Entonces se decide optar por los resultados medianamente inferiores del Modelo 2 (pero con mejor Accuracy y F-Score que el modelo 5), para incluir más variables explicativas. Las variables de este modelo son Concepto_Giro, Cargo_Neto_Cat, Regional, Trimestre de Pago, Cargo/C91 categorizado, Giros_Impagos categorizado y Convenio. Dado que los modelos tienen igual facilidad de interpretabilidad el Modelo 2 continua siendo el más apto.

5.4. Resultados Regresión Logística

Del análisis de las variables, se obtiene que la mayoría son significativas a un nivel de 95% de confianza, descartándose solo dos valores de la variable Regional y otros dos de Trimestre Pago. Del estudio de sus parámetros betas, se concluye que a mayor valor de cargo, mayor no pago, y que los giros con peor pago serían los de Auditoría IVA.

A continuación, se expone la matriz de confusión, curvas ROC y ganancia de información. De la matriz de confusión, se analiza que la relación entre la cantidad de errores del tipo falsos negativos, o predecir que un giro será pagado cuando realmente no lo será, sobre falsos positivos es 96,7%. Es decir, la regresión tiene casi los mismos errores de cada tipo. Al comparar esta matriz con la del modelo árbol CHAID de la sección 5.2, se observa que tienen casi igual cantidad de errores de clasificación, pero que la distribución cambia, teniendo el árbol CHAID menos errores falsos negativos por falsos positivos[7].

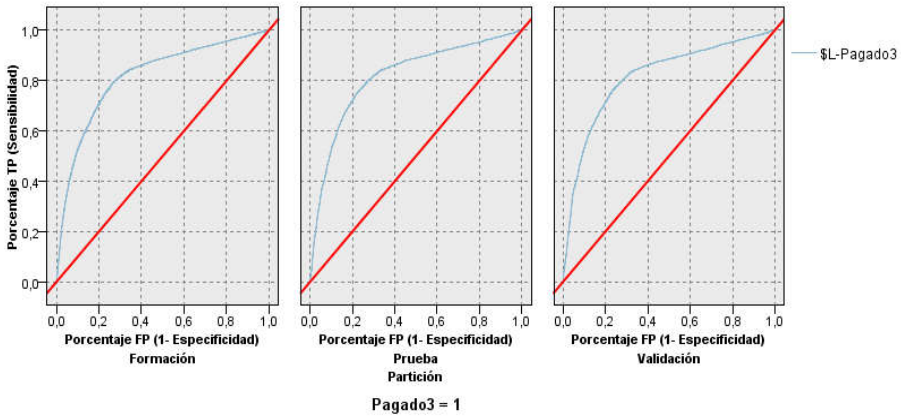
Tabla 13: Matriz de Confusión Regresión Logística

Matriz de Confusión		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	32740	6413
Real Pago	6433	13149

Fuente: Elaboración propia.

[7] En Anexo 10.6Regresión logística, podrá ver la ecuación resultante de la regresión.

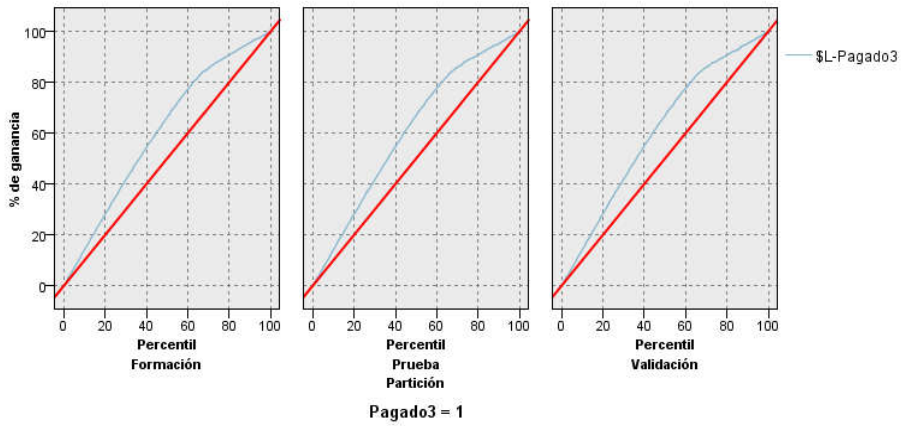
Ilustración 25: Curva ROC Regresión Logística



Fuente: Gráfico extraído de SPSS Modeler.

Con la curva ROC al considerar una tasa de falsos positivos de 60%, se obtiene una Sensibilidad de 91,1%, menor en casi 2% a la tasa de Sensibilidad obtenida por CHAID. Es decir, considerando la misma tasa de error de falsos positivos, la Regresión Logística se equivoca más al predecir no pago y su curva de Sensibilidad no aumenta tan rápido en relación a como lo hace CHAID. De todas formas, tiene buen desempeño según su AUC de 81,6%.

Ilustración 26: Curva de Ganancia Regresión Logística



Fuente: Gráfico extraído de SPSS Modeler.

De la curva de ganancia de información al tomar un 60% de los datos se captura un 77,4% de contribuyentes no pagadores, casi igual que CHAID.

5.5. Elección modelo de tipo Random Forest

Finalmente, se generan modelos con el algoritmo Random Forest. Aunque este algoritmo no se encuentra en la librería de SPSS Modeler, se decide probar dado que estudios señalan

menores problemas de sobre ajuste en relación a modelos de árboles individuales. [26]. Por otra parte, dado que la extensión "randomForest" de R Studio no permite trabajar con valores nulos, también se categorizan algunas variables.

Una de las cualidades de la técnica Random Forest es que se pueden modificar parámetros para mejorar el rendimiento del bosque. Uno de ellos es la cantidad de variables seleccionadas al azar para conformar grupos de variables dentro de los cuales se escoge la variable que mejor separa la data, y otro es la cantidad de árboles del bosque.[8] Esto permite disminuir la correlación entre las variables y árboles, lo que disminuye también el error de testeo del modelo o OOB.

Para representar más claramente lo realizado, se incluyen dos gráficos en Anexo 10.7; uno que muestra el cambio del OOB error según la cantidad de variables de cada grupo y otro, también de OOB error, versus la cantidad de árboles del bosque. Estos muestran que a mayor cantidad de variables y árboles el error disminuye, pero se observa que hay ciertos valores para los cuales el error se mantiene casi constante, estos son 2 variables por grupo y 20 árboles.

Utilizando los valores anteriores como punto de partida, se realiza un análisis de sensibilidad del modelo con 20, 80 y 150 árboles, y 2, 8 y 14 variables por cada grupo elegido al azar, con el fin de analizar cómo cambian los resultados del bosque al modificar los parámetros.

Tabla 14: Análisis de Sensibilidad Random Forest

Parámetros		*Métricas*				
Cantidad de arboles	Cantidad de variables en c/grupo	OOB Error	Accuracy	Precisión	Sensibilidad	Tiempo de ejecución
20	2	19,3%	80,8%	84,9%	86,6%	12''
20	**8**	**19,1%**	**81,9%**	**84,9%**	**88,7%**	**42''**
20	14	19,4%	81,6%	84,6%	88,4%	1'03''
80	2	18,6%	81,2%	85,5%	86,5%	41''
80	**8**	**17,8%**	**82,4%**	**85,7%**	**88,4%**	**2'46''**
80	14	18,0%	82,2%	85,7%	88,1%	5'22''
150	2	18,6%	81,2%	85,5%	86,3%	1'48''
150	**8**	**17,5%**	**82,5%**	**85,9%**	**88,3%**	**5'42''**
150	14	17,7%	82,4%	85,8%	88,2%	6'53''

Fuente: Elaboración propia.

[8] También puede cambiarse el tamaño de la hoja final, pero no se modificará en esta memoria.

Se observa en la tabla, que para cualquier combinación, todos los ratios de desempeño son buenos y que los tiempos de procesamiento son bastante cortos, menores a 7 minutos. Con esto se demuestra que Random Forest, ejecutado en R Studio, es un algoritmo eficiente para trabajar con grandes bases de datos.

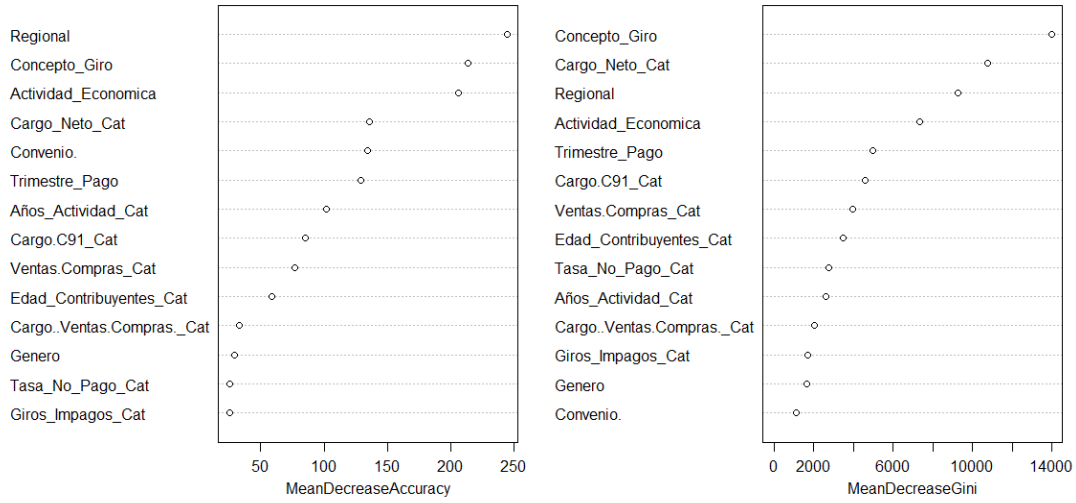
Si se analizan los bosques según la cantidad de variables por grupo, se tiene que todos los bosques de 2 variables tienen el mayor error asociado. Luego, para 8 variables este error llega a su mínimo y crece levemente cuando se utilizan grupos al azar de 14 variables. Si se analiza según la cantidad de árboles, se concluye que a mayor cantidad de árboles menor error, aunque la disminución es decreciente. Por ejemplo, para disminuir el error de 19,1% (combinación 20 árboles, 8 variables) en un 1,3% se requieren 80 árboles, pero con 70 árboles más sólo se disminuye un 0,3%.

Se selecciona entonces el modelo de 80 árboles y 8 variables ya que sus métricas de desempeño son inferiores a las del mejor modelo, el de 150 árboles y 8 variables, en tan sólo 0,1% y 0,2% para Accuracy y Precisión, y en un 0,3% en el caso del error, pero sin tener que incurrir en el cálculo de 70 árboles más, lo que ahorra casi la mitad del tiempo de procesamiento y recursos de memoria RAM (250 MB versus 150 MB).

5.6. Resultados Random Forest

A continuación se grafican dos medidas para determinar la importancia de variables utilizadas en Random Forest: Mean Decrease Accuracy (MDA) y Mean Decrease Gini (MDG),

Ilustración 27: Importancia de las variables Random Forest



Fuente: Grafico extraído de R Studio.

En el gráfico MDA, se representa en primer lugar las variables que más afectan al Accuracy del modelo al permutar los valores de división de la variable. Es decir, como al cambiar los valores de una variable para ramificar un nodo aumenta el error de clasificación general del modelo. Las variables que más afectan esté ítem son Regional, Concepto_Giro y Actividad_Economica. Por su parte, el MDG muestra las variables más importantes según el índice de Gini, es decir, cuáles reducen mayormente la impureza de un nodo al ser seleccionadas. Ambas medidas coinciden en determinar que es correcto considerar Concepto_Giro, Cargo_Neto_Cat, Regional y Actividad_Economica en el modelo.

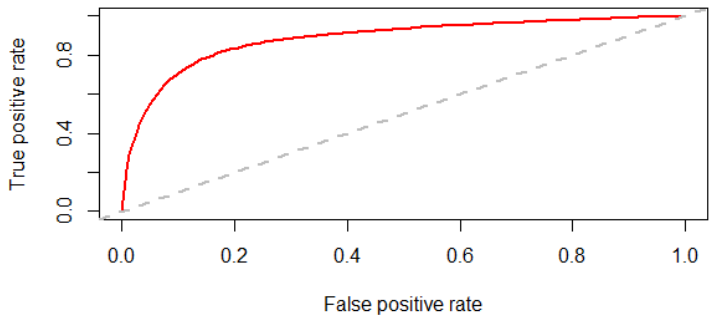
Tabla 15: Matriz de Confusión modelo Random Forest

Matriz de Confusión		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	34482	4671
Real Pago	5762	13820

Fuente: Elaboración propia.

De la matriz de confusión se nota que Random Forest predice menos casos FN y FP que los otros modelos, lo que muestra que el modelo predice mejor el Pago y el No Pago. La curva ROC por su parte, muestra que con una tasa de 60% de falsos positivos se obtiene un 90,5% de Sensibilidad.

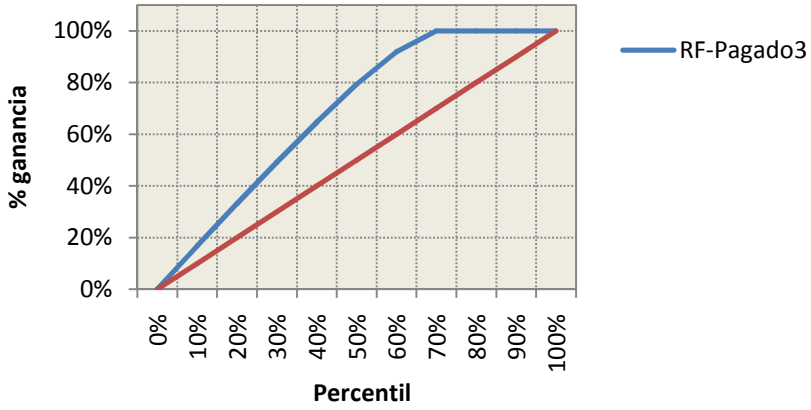
Ilustración 28: Curva ROC Random Forest



Fuente: Grafico extraído de R Studio.

Para finalizar, la curva de ganancia de información muestra que al seleccionar un 60% de la población se captura a un 92% del total de giros que no serán pagados. Esta medida es mucho más alta que la de los otros modelos, que solo logran alrededor de 77% de ganancia. Su mayor curvatura, muestra que el modelo define muy bien los casos de giros que no serán pagados, lográndose su captura antes que otros casos de pago.

Ilustración 29: Curva de Ganancia Random Forest



5.7. Comparación modelos finales

En esta última sección, se comparan los modelos pre seleccionados para escoger el más indicado para ser utilizado en la predicción de no pago. Se usará para esto el método de selección utilizado anteriormente, agregando esta vez un nuevo criterio, el de aplicabilidad del modelo.

Tabla 16: Desempeño Modelos

Métrica	Modelo CHAID	Modelo R. Logística	Modelo R. Forest
Accuracy	78,3%	78,4%	82,4%
Precision	84,6%	84,8%	88,4%
Sensibilidad	83,2%	83,1%	85,7%
Especificidad	68,0%	68,4%	75,2%
F-Score	83,9%	84,0%	87,0%
AUC	82,7%	81,9%	88,4%
N° Variables	19	7	14

Fuente: Elaboración propia.

Al comparar los modelos, Random Forest destaca al obtener mejor desempeño que los otros dos. Claramente se diferencia, en un 5,7% más de AUC y 4,2% más de Sensibilidad, aun cuando utiliza menos de la mitad de variables independientes que CHAID, el segundo mejor modelo.

Por otra parte, la explicación del problema a partir de las variables seleccionadas en los tres modelos es buena, pero visto de forma relativa, el modelo de regresión logística tiene el peor resultado en ese ítem ya que es el que está conformado por una menor cantidad de variables explicativas.

En relación a la interpretabilidad, el árbol de decisión y la regresión logística son muy fáciles de interpretar. Para ambos, el aporte directo de cada variable (o valor de la variable) al no pago, se expresa cuantitativamente, en cambio, es difícil dar esta explicación del aporte a la probabilidad de no pago en Random Forest ya que existen 80 árboles diferentes. Solo se puede llegar a una aproximación de la importancia de las variables a través de las medidas MDA y MDG.

Para la aplicabilidad de los modelos se observan dificultades para Regresión Logística y Random Forest, ya que ambos requirieron preprocesamiento extra al categorizar algunas variables. Random Forest en R Studio tiene una limitación extra: Los datos que se quieren predecir deben contener, estrictamente, la misma cantidad de categorías por cada variable con la que se entrenó el modelo. Por ejemplo, para Actividad_Economica_Cat no se tuvo registros de contribuyentes tipo M (Administración pública) para alrededor de 58.000 datos de validación, por lo que se debió imputar manualmente ese tipo a un registro y así igualar la cantidad de categorías. Este problema de clases será mayor cuando los registros sean pocos, y para solucionarlo se requerirá trabajo extra.

Más abajo, se resume los ítems de evaluación de los modelos, con la valoración 1 a 3 usada anteriormente.

Tabla 17: Comparación modelos

Modelos	Métricas de Desempeño	Identificación de Variables	Facilidad de Interpretación	Aplicabilidad del modelo
CHAID	**	***	***	***
Regresión Logística	**	*	**	***
RandomForest	***	***	*	**

Fuente: Elaboración propia.

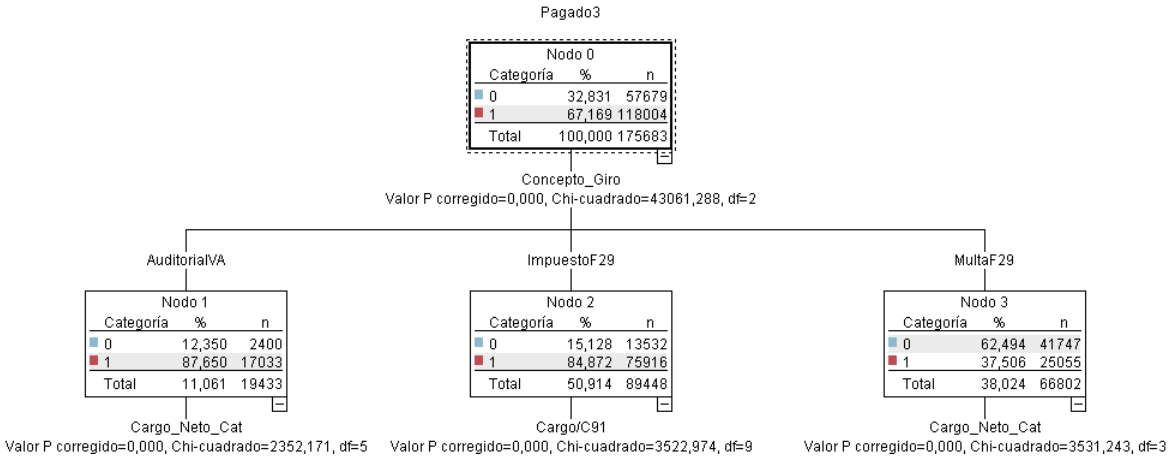
Según los criterios anteriores, el modelo CHAID es considerado como el mejor de todos, y será el modelo a desarrollar en profundidad. Por otra parte las pruebas confirman el alto poder predictivo de Random Forest, por lo que este algoritmo debería considerarse cuando se requiera predecir con foco a la determinación exacta de probabilidades por sobre otros factores, como por ejemplo, la explicación del problema tratado. Como este conocimiento es nuevo para el Servicio de Impuestos Internos, se continuará testeando la aplicabilidad de Random Forest en los capítulos siguientes.

6. Resultados específicos

6.1. Explicación de Modelo CHAID

El modelo CHAID seleccionado es entrenado con un conjunto del 60% de la data total, correspondiente a 175.683 giros, que tienen una tasa global de no pago de 67,1% (nodo 0). Este modelo es simple de interpretar ya que la primera variable discriminatoria es el tipo de giro. Esto muestra de forma inmediata que los contribuyentes que deben Multa F29, Auditoría IVA o Impuesto F29 son muy diferentes entre sí, y permite estudiar cada caso por separado. Se destaca que este resultado es igual que haber creado tres modelos de árboles, uno para tipo de giro, por lo que a futuro podría estudiarse un solo tipo sin incluir a los demás.

Ilustración 30 : Árbol CHAID, Primera ramificación



Fuente: Extraído de SPSS Modeler.

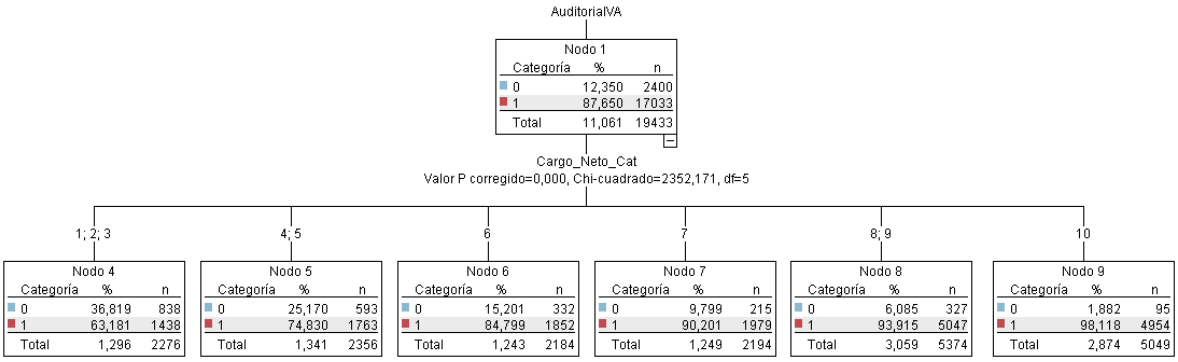
De los grupos de variables, el de caracterización del giro, con variables como tasa de no pago y cantidad de giros impagos, fue el más explicativo para Impuesto F29. Para Multa F29, lo fueron las variables de caracterización del contribuyente y las de giro, mientras que el subárbol de Auditoría IVA solo consideró 3 variables; Cargo_Neto_Cat, Promedio C91 y Ventas+Compras. Las variables que no fueron utilizadas fueron las de capacidad de pago, como patrimonio y cantidad de trabajadores, y las anotaciones negativas de los contribuyentes. Se puede inferir de acuerdo a los resultados, que estas no sirvieron para el modelo por tener gran cantidad de datos nulos lo que no les permite tener la presencia mínima necesaria para discriminar las hojas finales de cerca de 3000 giros.

6.1.1. Nodo 1. Giros por Auditoría IVA

Este nodo tiene una probabilidad de no pago de 87% y se divide primero según Cargo_Neto_Cat, formando 6 grupos de categorías. Para los primeros 4, no existe la varianza necesaria para que la data se vuelva a subdividir, pero sí lo hacen los últimos dos.

Para estos seis grupos de categorías la mínima probabilidad es de 63,1% para los montos pequeños, menores a $43.200 pesos, o categorías de cargo 1,2 y 3. Esta probabilidad es bastante alta comparada con los giros de Multa F29 del mismo valor, los que tienen un no pago de sólo un 35%. A vista experta esto se explica porque los contribuyentes de Auditoría IVA pasaron por un proceso de revisión, debido a la detección de comportamientos anómalos, con lo que desde antes tenían una tendencia a evitar el pago, aunque este no fuera tan alto. Luego, como lo indican las probabilidades, a medida que sube el monto aumenta el no pago.

Ilustración 31: Nodo 1. Auditoria IVA, Segundo nivel de profundidad.



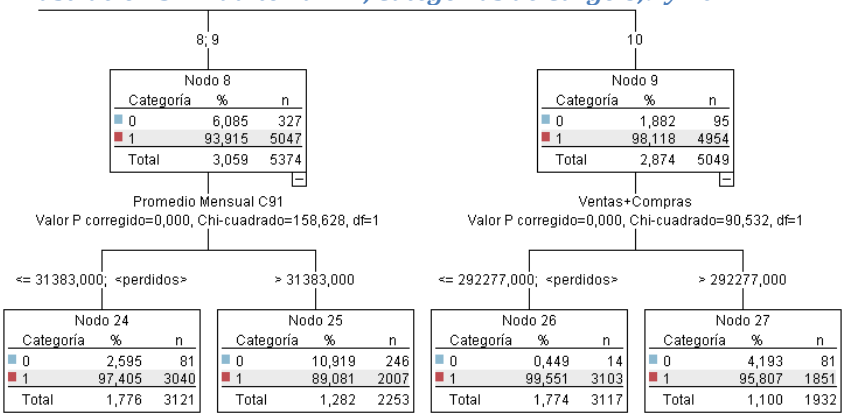
Fuente: Extraído de SPSS Modeler.

Los últimos dos grupos, nodos 8 y 9, se subdividen según Promedio mensual C91 (Cantidad promedio de impuesto mensual pagado) y Ventas+Compras respectivamente. Estas subdivisiones muestran que los giros de quienes tienen impuesto C91 menor a 31 mil pesos o Ventas+Compras menores a 292 mil, tienen mayores probabilidades de no ser pagados que los del grupo con montos mayores a esos valores (8,3% y 3,7% más probabilidad de no pago). Esto se explica porque el beneficio antes de impuesto de esas empresas, según un cálculo estimado, no alcanzaría para pagar la deuda en menos de 6 meses o tendría que ser ocupado enteramente en el pago de giros, lo que es muy difícil para una empresa considerando que además debe pagar sueldos y otros gastos operacionales.

Por ejemplo, para el nodo 8, si se toma como supuesto que el impuesto C91 representa aproximadamente un 19% de las utilidades de la empresa, pagar 31 mil pesos de impuestos significa que la empresa gana cerca de 165 mil pesos, y tendría que utilizar las ganancias completas de 3 meses para costear una deuda de 472 mil pesos (Cargo_Neto_Cat =8), o de 11 meses para una de 1,92 millones de pesos (Cargo_Neto_Cat =9).

Para los grupos de alto cargo y mayores ventas o impuesto, también existe un no pago alto, superior al 89%, lo que puede ser efecto del tipo de contribuyente, que es peor pagador de este tipo de giro que de otros y del alto valor de la deuda, pero no se puede comprender con detalle esta situación, ya que el árbol no reconoce más variables explicativas. En Anexo 10.8 Resultados específicos, puede encontrar el listado de reglas para este tipo de giro.

Ilustración 32: Auditoria IVA, Categorías de Cargo 8,9 y 10.



Fuente: Extraído de SPSS Modeler.

6.1.2. Nodo 2. Giros por Impuesto F29

En el caso de Impuesto F29, la propensión de no pago es de 84%, muy similar a la del nodo anterior, pero difiere en que existen más de 10 sub arboles con varias ramificaciones que generan más nodos finales. La primera variable que divide los datos es Cargo/C91 (o el valor del giro sobre el promedio de impuestos mensual que pagó la empresa el año móvil anterior a la emisión del giro), y a medida en que el valor de esta variable aumenta también aumenta el no pago.

En general los nodos de peor pago son los de giros con un valor de Cargo/C91 mayor o igual a 4,7, pertenecientes a empresas que deben al menos otro giro. Suponiendo que las empresas quieran cumplir con el pago de impuestos, para cualquiera de ellas, bajo el supuesto que los datos declarados son fieles a la realidad, sería difícil pagar una deuda de impuestos casi 5 veces mayor a lo que está acostumbrada a pagar, sumada a los impuestos normales que paga en el mes. Se suman a estos giros los de valor nulo en Cargo/C91 de montos mayores a 130 mil pesos. Para estos contribuyentes es aún más dificultoso pagar ya que no tienen ganancias por sus ventas, como lo es el caso Cargo/C91 igual a casos "perdidos". A continuación, se muestran los 5 grupos de giros de Impuesto F29 que tendrían la más alta probabilidad de no ser pagados en los 6 meses posteriores a su emisión,

Tabla 18: Impuesto F29. Mayores probabilidades de no pago

1ra regla	Otras reglas	N° Nodo	Prob. No pago	Giros sobre data total
Cargo/C91 = perdidos	Cargo Neto Cat = 9,10 y Tasa No Pago > 0,77	79	98,1%	1,08%
Cargo/C91 > 33,6	Giros Impagos >1 ó perdidos	46	95,6%	1,70%
Cargo/C91=(7,3; 12,9]	Giros Impagos > 0 ó perdidos y Trimestre Pago= Invierno, Septiembre	70	95,3%	1,05%
Cargo/C91=(12,9; 33,6]	Giros Impagos > 1 ó perdidos	44	95,1%	1,90%
Cargo/C91 = perdidos	Cargo Neto Cat = 6,7 y Giros Impagos > 1 ó perdidos	77	95,0%	1,12%

Fuente: Elaboración propia.

Si bien tiene sentido la poca probabilidad de pago de los nodos, lo que llama la atención es cómo una empresa puede llegar a deber tantas veces más su impuesto promedio (más de 4,7 veces). De la experiencia en terreno se estima que puede haber dos motivos; el primero es que algunas empresas posponen la declaración de periodos de alto pago de IVA, como Navidad para el sector comercio, hasta ordenar su contabilidad, o, que sub declaren varios meses por falta de dinero, para después presentar una declaración por los impuestos del mes en curso y los pasados. El problema que cualquiera de estas prácticas genera, es que la deuda aumenta y se hace difícil pagar, quedando el giro moroso.

Por otra parte, la mínima probabilidad de pago es 57,8% para giros de Cargo/C91 cercanos a 1, que no están suscritos a convenio de pago (convenio=0) de empresas de más de 11 años de actividad. El no tener convenio de pago se repite en estas reglas de mejor pago, y al comparar los nodos con el símil que sí posee convenio, se observa que los últimos pueden tener una propensión de hasta 19% más de no pagar (Ver anexos 10.8). Aunque contrariamente a lo que se podría esperar, los giros sin convenio se pagan mejor, esto podría explicarse porque los contribuyentes que saben que podrán pagar su deuda prontamente tal vez no suscriben convenio, siendo esto válido solo para giros de Cargo/C91 de hasta 2,3 veces.

Tabla 19: Impuesto F29. Menores probabilidades de no pago

1ra regla	Otras reglas	N° Nodo	Prob. No Pago	Giros sobre data total
Cargo/C91 =(0,95; 1,59]	Convenio = 0, Giros Impagos<=0 y Años Actividad > 11	92	57,8%	0,63%
Cargo/C91 <=0,95	Cargo/(Ventas+Compras) <=0,035 o perdidos	28	61,0%	0,98%
Cargo/C91= (1,59; 2,31]	Convenio = 0, Giros Impagos <= 0 y Genero = Femenino o Masculino	94	66,7%	0,98%
Cargo/C91= (0,95; 1,59]	Convenio = 0, Giros Impagos<=0 y Años Actividad <= 11	91	67,7%	0,97%
Cargo/C91<= 0,95	Cargo/(Ventas+Compras) >0,035	29	72,7%	0,88%

Fuente: Elaboración propia.

6.1.3. Nodo 3. Giros por Multa F29

Este es el tipo de giro que tiene la más baja probabilidad de no pago promedio con 37%. Al igual que para los giros de Auditoria, la primera subdivisión se realiza mediante las categorías de Cargo neto, pero a diferencia de ellos tienen una probabilidad de pago mucho mejor.

Para la mayoría de las Multas F29 menores a 130 mil pesos, la probabilidad de no pago no supera el 35%. Pero entre 130 mil y 258 mil pesos la probabilidad aumenta considerablemente a 61,5% y continúa aumentando a mayores montos. Esto muestra que para los deudores de Multa F29 comienza a ser significativa una deuda en ese rango. Cabe destacar, que este tipo de giros no posee Categorías de Cargo 9 ni 10 ya que la Multa F29 tiene un tope de hasta una UTA establecida por ley, es decir, aproximadamente 550 mil pesos.

De las probabilidades más altas de no pago se destaca la mayor de 80%, para los giros de categorías de cargo 7 u 8, es decir, entre 258 mil y 472 mil pesos, que son emitidos en los meses de "vacaciones", es decir, diciembre, enero o febrero. Si un giro con las mismas características fuera emitido en otra época del año, el no pago solo alcanzaría un 55%. Esto muestra que aunque los contribuyentes quieran cumplir, se ven imposibilitados de pagar, probablemente por los altos gastos que se generan en los seis meses posteriores a las vacaciones (patentes, época escolar, declaración de renta en abril y otros).

Entre los otros patrones encontrados, la tenencia de giros impagos influye en el no pago de deudas menores a 50 mil, aumentando en más de 15% la probabilidad y que las empresas jóvenes (3 años de actividad) de Actividades Económicas del tipo B, G, K, L y O[9] ó Pesca, Construcción, Intermediación Financiera, Actividades Inmobiliarias y Servicios Sociales y de Salud también tienen un 15% más de no pago respecto a empresas jóvenes de otros rubros.

Tabla 20: Multa F29. Mayores probabilidades de no pago

1ra regla	Otras reglas	N° Nodo	Prob. No Pago	Giros sobre data total
Cargo neto=7,8	Trimestre Pago= Vacaciones	59	80,7%	1,23%
Cargo neto = 1,3,4	Tasa no pago>0,5	53	62,4%	0,71%
Cargo neto=6	-	22	61,5%	1,29%
Cargo neto=7,8	Trimestre Pago= Invierno, Marzo, Septiembre	58	55,0%	0,85%
Cargo neto = 1,3,4	Tasa no pago<=0,5 o Perdidos, Años actividad <= 3,0 y Actividad Económica B,G,K,L,O	100	47,6%	0,51%

Fuente: Elaboración propia.

Dentro de los mejores pagos, se distingue a los giros de categorías de cargo 2 y 5 de empresas de actividad económica D, H, I y J, o rubros de Industrias Manufactureras no Metálicas, Comercio al por Mayor y Menor, Hoteles y Restaurantes, y Transporte, Almacenamiento y Comunicaciones con cerca de 18% de no pago.

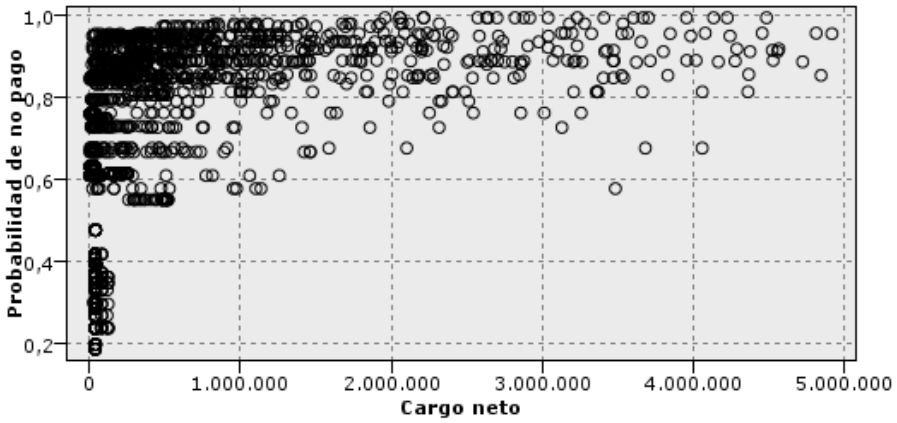
También lo son, los giros de cargo neto 1, 3 y 4, de contribuyentes con un buen pago de giros anteriores y más de 22 años de actividad. Estos tienen una probabilidad de 23,1%. A modo general en Multa F29, las empresas con más años de actividad o a nombre de personas de mayor edad tienen mejor comportamiento. En Anexo 10.8, se ejemplifica como sólo por años de actividad pueden existir diferencias de hasta 15% de no pago y se pueden ver los giros de mejor pago para Multa F29.

[9] Todas las categorías de la variable "Actividad_Economica" se encuentran en el anexo 10.4.

6.1.4. Respecto a los supuestos del capítulo 4.4

Como se había observado al finalizar el capítulo 4, existen tipos de giros que tienen un peor pago que otros, pero se ha logrado comprobar que esto no solo es debido al mayor cargo asociado que tienen. En la figura de abajo, se observa para una muestra que aun para deudas de montos bajos la probabilidad de no pago puede ser cercana a uno. De lo visto se tiene que a iguales deudas, los contribuyentes que deben Auditorías IVA e Impuesto F29 tienen una peor calidad de pago que los de Multa F29. Por lo que se infiere que tener una Auditoría IVA o una declaración fuera de plazo, de deudas altas, es una señal evidente de comportamiento negativo.

Ilustración 33: Distribución de las probabilidades de no pago según Cargo



Fuente: Figura extraída de SPSS.

Luego encontrada esta evidencia, las variables Cargo o el Cargo/C91, toman un rol preponderante en la probabilidad del no pago, donde a mayores valores disminuye el pago por parte de los contribuyentes.

La cantidad de giros impagos y una baja tasa de pago pasado, influye bastante en el no pago, pero debe corroborarse si estas variables fueron útiles porque representan una mala intención de pago o porque tener giros impagos significa tener una suma mayor de deuda asociada, que sumada al giro actual podría considerarse como un giro de mayor cargo y por ende de menor probabilidad de pago.

Respecto al convenio de pago, se observa que para algunos casos de giros no asociados a un convenio de pago se pagan mejor, pero no se logra concluir su eficacia de manera general.

6.2. Caracterización de giros

Dada la alta cantidad de nodos finales, los resultados del modelo se agrupan para comprender de mejor forma a los giros, obteniendo 5 tipos. Se agrupa a todos los del tipo Auditoría IVA, ya que no hay variables claras que dividan su comportamiento, pero, para Impuesto F29 y Multa F29, se observan dos grupos para cada uno.

Para Impuesto F29, un primer grupo de giros es el de mejor pago que el resto, con una deuda cercana al promedio y no tan alta respecto a lo que la empresa siempre paga en impuesto. Un segundo grupo, Impuesto 2, son los giros de más de 4,7 veces lo que siempre paga. En Multa F29, se dividen quienes están pagando deudas menores a 130.000 de los que no, ya que ese era el punto donde las empresas comenzaban a tener dificultades para pagar.

Tabla 21: Caracterización Giros

Promedio	Auditoría IVA	Impuesto F29 (1)	Impuesto F29 (2)	Multa F29 (1)	Multa F29 (2)
Cargo Neto	$ 2.620.456	$ 942.699	$ 1.603.793	$ 45.991	$ 347.906
Ventas ult12m	$ 39.173.931	$ 42.466.717	$ 24.831.885	$ 36.137.666	$ 32.331.158
Promedio C91	$ 206.801	$ 362.912	$ 132.987	$ 257.429	$ 162.270
Giros impagos	0,6	2,2	2,7	0,3	0,8
Giros totales	0,8	3,0	3,7	0,7	1,4
Tiempo de pago	1,8 meses	2,3 meses	2,6 meses	0,9 meses	1,4 meses
Porcentaje de pago entre 6 meses y 2 años	24,7%	29,1%	40,5%	54,6%	38,2%
Tenencia de Patrimonio	27,8%	19,2%	17,0%	24,2%	21,0%
Porcentaje de giros de persona naturales	61,3%	58,1%	45,8%	48,8%	44,7%
Porcentaje del total de la base de entrenamiento	11,1%	32,2%	18,7%	33,2%	4,9%
Probabilidad no pago	**76,8%**	**75,5%**	**88,4%**	**32,2%**	**62,7%**

Fuente: Elaboración propia.

- Auditoria IVA: Empresas con pocos giros, pero con un giro actual (en promedio) de alto valor. Aun cuando, el giro sea bajo (menor a 50 mil pesos) tiene una mediana-baja probabilidad de ser pagado. Cuando los contribuyentes que deben estos giros pagan dentro de los 6 meses plazo, se demoran tan solo 2 meses, pero de no hacerlo tienen una probabilidad de 24,7% de pagar dentro de los 2 primeros años. Calidad de pago en 6 meses: Mala.

- Impuesto F29 (1): Giros de Cargo/C91 hasta 4,7 o valores perdidos de la variable. A pesar de que las empresas dueñas de estos giros tienen las mayores ventas y el mayor pago de impuesto C91, su giro promedio es cercano al millón de pesos, menor que el de Impuesto F29 (2)

que vende en promedio menos. Con anterioridad ya han tenido otros giros, y arrastran pagos pendientes. Calidad de pago en 6 meses: Mala.

- Impuesto F29 (2): Giros con Cargo/C91 mayor a 4,7 veces, valor considerado como muy alto. Puede que este valor refleje algún mal manejo de la contabilidad. Sumado a eso, tienen más giros anteriores y deben mucho más que Impuesto F29 (1). Serían "malos" pagadores, lo que se refleja en su probabilidad de no pago en 6 meses, cercana al 90%. Además tienen menor patrimonio que el resto por lo que podrían responder menos a una cobranza. Calidad de pago en 6 meses: Muy mala.

- Multa F29 (1): Lo conforman deudas menores a 130.000 pesos. Cuando son pagadas se hace en un plazo de menos de un mes. Los contribuyentes dueños de estos giros, tienen buena conducta anterior ya que casi no tienen giros anteriores o impagos lo que da a lugar a una buena probabilidad de pago. Si no se han pagado en 6 meses, tienen una probabilidad de 55% de pago dentro de un año, mucho mejor que el resto de los casos. Calidad de Pago en 6 meses: Buena.

- Multa F29 (2): Son giros entre 130.000 y 500.000 aproximadamente. Tienen más de un giro anterior. Los contribuyentes de los giros de este tipo venden mensualmente casi 250 mil pesos menos que los de Multa F29 (1), pero deben pagar un giro 6 veces mayor, lo que podría explicar que su pago sea peor al del otro grupo. Calidad de pago en 6 meses: Regular.

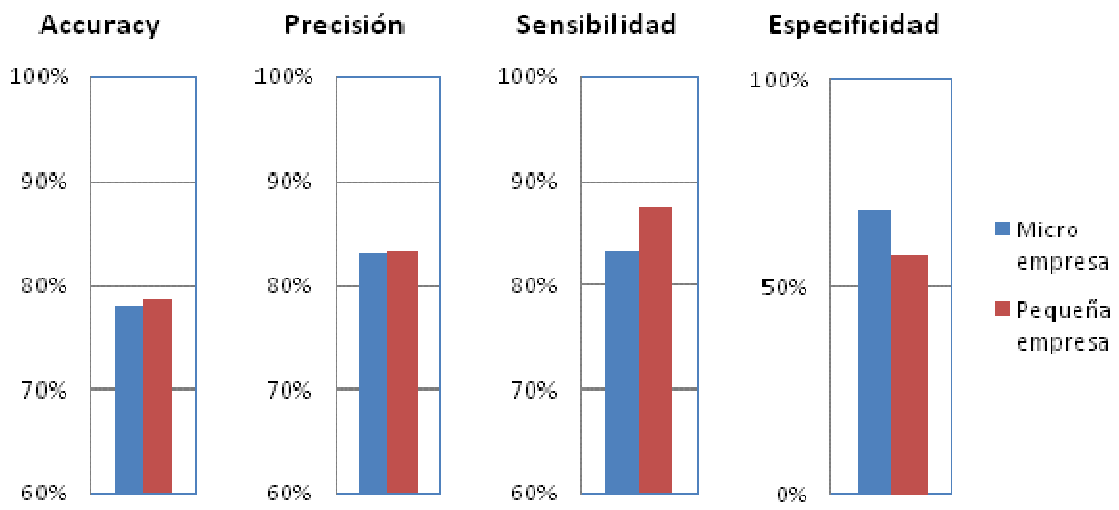
6.3. Resultados según segmentos de interés

6.3.1. Segmentos SII

Si bien el árbol tiene buenos resultados generales de clasificación, en esta sección, se testea si también los tiene para distintos segmentos. Podría suceder que la predicción de un conjunto de la base tenga tan buen poder predictivo que encubra el mal desempeño de otra predicción al promediar todos los resultados. Por eso, se medirán las métricas de desempeño del árbol sobre subconjuntos de datos conformados por los giros de impuesto pertenecientes a los contribuyentes de los segmentos utilizados por el SII, que dependen de la cantidad de ventas anuales medidas en UF, y los del segmento utilizado por la Tesorería General de la República, que clasifica a los contribuyentes según su deuda total con el fisco. Estas mediciones se realizan sobre la partición de validación.

Según la clasificación del SII, los giros de empresas del tipo Micro y Pequeña (vistas en el año 2015) tienen medidas de desempeño bastante parejas, las que se muestran a continuación:

Ilustración 34: Métricas de desempeño según Segmentos SII



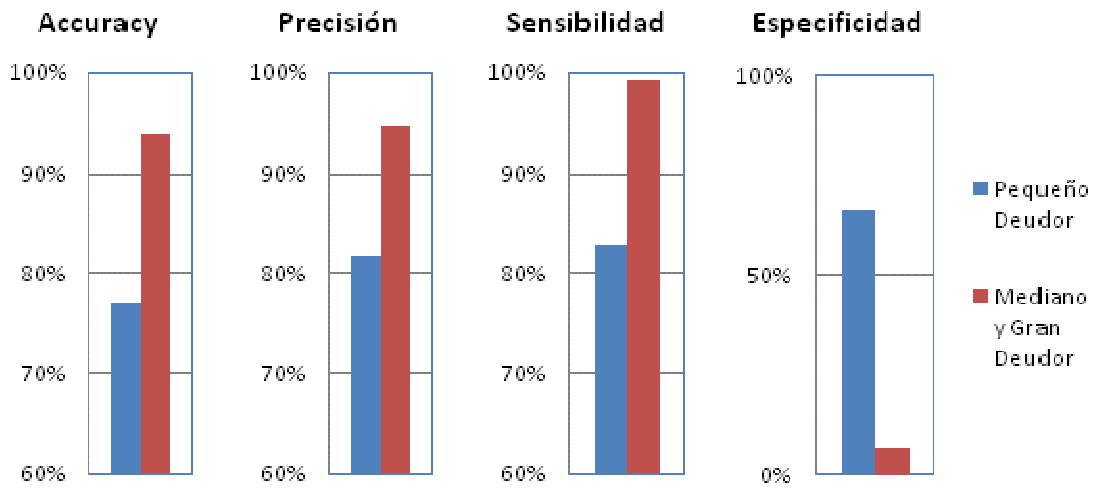
Fuente: Elaboración propia.

De los dos tipos de empresa, la Pequeña posee mejores resultados que la Micro, excepto en Especificidad, aunque las diferencias son leves, y la mayor se produce para la medida de precisión siendo de un 4%. Se concluye que los resultados para estos segmentos no varían mucho entre sí, ni en relación a las medidas globales, lo que se considera muy bueno.

6.3.2. Segmentos TGR

Para analizar los resultados de la clasificación de la TGR, no se posee la deuda fiscal total de cada contribuyente, por lo que se ha llegado a una clasificación aproximada que suma los valores de todos los giros por cada RUT (incluyendo los giros ya pagados), y se considera a un Pequeño Deudor como aquel que tiene una suma de cargos menor a 10 millones de pesos, y como Medianos y Grandes a los que posean más de ese valor.

Ilustración 35: Métricas de desempeño según Segmentos TGR



Fuente: Elaboración propia.

En este caso se puede apreciar mayores diferencias entre grupos respecto a la clasificación SII. El desempeño de la predicción de giros del grupo de Mediano y Gran Deudor es mucho mayor que el del otro segmento. Esto se da, porque casi todos los giros de los mayores deudores son de grandes montos y tienen probabilidades predichas de no pago mayores a 0,5, por lo tanto, casi todos son clasificados como no pagadores. Como hay pocos casos pagadores, son pocos los casos falsos negativos o verdaderos negativos. Con esto, las fórmulas de Accuracy, Precisión y Sensibilidad quedarían:

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN} \approx \frac{\text{VP}}{\text{VP} + \text{FP}} = Precisión$$

$$Sensibilidad = \frac{VP}{VP + FN} \approx \frac{\text{VP}}{\text{VP}} = 100\%$$

Dando como resultado que, para Mediano y Gran Deudor, como se aprecia en los gráficos, la Precisión y Accuracy son casi iguales y cercanos a 95%, y con una Sensibilidad de 99,3%. Además de lo anterior, los giros de este segmento no se están pagando en el plazo de seis meses, por lo que los casos pagados o VN son pocos. Con eso, la Especificidad teóricamente resultaría cercana a cero por ciento, lo que se corrobora con un valor de 6,5%.

$$Especificidad = \frac{VN}{FP + VN} \approx \frac{0}{FP} = 0\%$$

Finalmente, se analiza que la predicción para Mediano y Gran Deudor, que corresponden al casi 8% del total de datos, tiene buen desempeño, medido en Accuracy, Precisión y

Sensibilidad, pero a costa de un muy bajo desempeño para medir los casos predichos pagados, o Especificidad, fenómeno que se produce por la distribución de dichos casos en la muestra de entrenamiento.

6.3.3. Comparación Random Forest y CHAID

En el capítulo 5, se estableció que el modelo Random Forest podría ser muy útil cuando se requiera obtener probabilidades de mayor confianza, sin necesidad de obtener las reglas de decisión que las definan. Para confirmar esta afirmación, se vuelven a comparar las métricas de desempeño por segmentos SII y TGR.

Más abajo, se muestran los diferenciales entre los resultados del modelo Random Forest y los del árbol CHAID, partiendo por la predicción de los giros de Micro y Pequeñas empresas. Para ambos segmentos, se obtienen diferencias positivas a favor del modelo Random Forest, de casi 3% mejor en todas las métricas, y sobresale, con especificidades 4,3% y 6,5% mayores.

Tabla 22: Comparación Random Forest y CHAID SII

Micro Empresa	Diferencia RF-CHAID
Accuracy	4,6%
Precisión	2,6%
Sensibilidad	4,7%
Especificidad	4,3%
F-Score	3,7%

Pequeña Empresa	Diferencia RF-CHAID
Accuracy	3,2%
Precisión	2,5%
Sensibilidad	1,8%
Especificidad	6,5%
F-Score	2,1%

Fuente: Elaboración propia.

Para el caso de la clasificación de TGR, Random Forest predice mejor el giro del Pequeño Deudor, tanto como lo hace para los giros de la Micro Empresa del SII. Para el caso de Mediano y Gran Deudor, no hay grandes diferencias, a excepción de la Especificidad que con el modelo Random Forest es de 33,1%, es decir, 26,6 puntos porcentuales más que el modelo CHAID de 6,5%.

Tabla 23: Comparación Random Forest y CHAID TGR

Pequeño Deudor	Diferencia RF-CHAID
Accuracy	4,5%
Precisión	2,8%
Sensibilidad	4,4%
Especificidad	4,6%
F-Score	3,6%

Mediano Deudor	Diferencia RF-CHAID
Accuracy	0,6%
Precisión	1,4%
Sensibilidad	-0,9%
Especificidad	26,6%
F-Score	0,3%

Fuente: Elaboración propia.

De la comparación según segmentos de interés, se ratifica la superioridad predictiva del modelo Random Forest, al lograr predecir casos correctamente de segmentos que tengan muy poca presencia en la base de datos.

7. Propuestas

7.1. Propuestas para el Servicio de Impuestos Internos

7.1.1. Propuestas sobre variables

Dados los buenos resultados obtenidos en la predicción, se recomienda utilizar para otros estudios sobre giros de impuestos, los atributos que han sido creados y validados en este trabajo, como por ejemplo, Cargo/C91, tasa de no pago, cantidad de giros impagos, Cargo/(Ventas+Compras), trimestre de emisión del giro (que puede generalizarse a otros casos como trimestre de ocurrencia del incumplimiento) y edad del contribuyente. Además, no se debe olvidar agregar variables útiles ya existentes como Convenio de pago y Años de Actividad.

Por otra parte, como la mayoría de los contribuyentes con giros impagos tienen las peores probabilidades de no pago, se debería programar una variable que sume la deuda pendiente de esos giros impagos, ya que no es lo mismo tener una alta cantidad de ellos, con bajo monto asociado, que con alto monto asociado. Esto para cada período tributario y considerando otros tipos de giros. Es decir, obtener el cálculo de la deuda fiscal total de cada contribuyente, actualizándola mes a mes.

Además, se propone mejorar la variable de patrimonio de los contribuyentes. Como primera tarea debe realizarse una actualización de ella, y luego, se podría complementar esta variable con otro tipo de posesiones, como depósitos a plazo en bancos o vehículos motorizados, los que podrían constituir patrimonio de mayor liquidez para el pago de impuestos y/o giros.

Y por último, se debería probar una variable del tipo marca, que muestre la tenencia de contador, y su ID, en conjuntos de contribuyentes más específicos porque no todos cuentan con uno. Esta variable no pudo utilizarse en la memoria por ser una variable de acceso restringido, pero se cree que podría ser útil para mostrar cómo algunos malos comportamientos pueden trasmitirse de contribuyente en contribuyente a través de los contadores que los atienden, como por ejemplo, el no pago de un giro.

7.1.2. Propuestas sobre las predicciones de no pago

Para cada uno de los grupos de giros definidos en la sección 6.2, se generan propuestas de acción sobre ellos.

Grupo de Auditoria IVA: Las variables definidas para el problema no lograron explicar en profundidad este tipo de giros, por lo que se propone continuar estudiando este grupo, ya que tiene la mayor consecuencia monetaria asociada y un no pago considerable. Tal vez, este grupo no tenga buena disposición a pagar, por lo que podría ser más efectivo realizar un estudio para definir qué empresas podrían recibir un giro de Auditoría IVA, que permita generar medidas preventivas en vez de paliativas. Por otra parte, se podría analizar si conviene realizar auditorías de periodos contables más cortos, que generen giros más espaciados en el tiempo y de menor monto, para que se le dé opción al contribuyente de ir pagando de a poco. Quizás la tenencia de giros altos provoque una actitud de rechazo al cumplimiento.

Grupo de Impuesto F29 (1): Para este tipo de giros, el SII podría trabajar en forma conjunta con Tesorería, para que ayuden a los contribuyentes de estos giros a regularizar sus deudas anteriores, dado que todos los que poseen giros pendientes, aunque su giro no sea tan elevado respecto a lo que paga siempre en impuestos, tienen bajas probabilidades de pagar.

Grupo de Impuesto F29 (2): La recomendación para estos giros es continuar su estudio con enfoque preventivo, al igual que con Auditoría IVA, porque son los giros de mayor probabilidad de no pago y porcentaje de la base de datos (32,2%). Se propone entender porqué los cargos son tan altos respecto a los pagos normales de impuesto ya que la declaración atrasada de grandes montos de IVA y su posterior no pago podría ser producto de prácticas contables. Además, al igual que en el caso anterior podría ayudarse a las empresas a solucionar sus problemas con giros anteriores.

Grupo de Multa F29 (1): Se estableció que este grupo de contribuyentes paga bien, y que se acerca a pagar en menos de un mes. De todas maneras, este tipo de giro es evitable por los propios contribuyentes declarando que no tuvieron movimiento o que obtuvieron remanente fiscal en el F29, cuando así corresponda, dentro de la fecha estipulada. El SII puede apoyar a las empresas en su cumplimiento, implementando planes educativos, iniciando esta tarea por los grupos con mayor probabilidad de no pago definidos en Tabla 20: Multa F29. Mayores probabilidades de no pago[10]. Además de estos grupos, se podría ampliar esta campaña incluyendo avisos para las nuevas empresas en el momento de su creación, ya que son las responsables del 10% de los giros de este tipo.

Grupos de Multa F29 (2): Estos contribuyentes tienen un regular comportamiento (buen pago y casi sin giros anteriores) por lo que se les debería guiar para que no cometan los mismos errores. Además, como se observó en el capítulo 6, para los giros superiores a 250 mil pesos, una misma deuda puede tener una diferencia de pago de casi un 30% debido a la fecha del año en la que debe pagar. Por lo que se propone diseñar un plan de apoyo al pago, ya que con asistencia tributaria al momento de cursarle el giro, estos contribuyentes podrían ordenar sus finanzas y pagos, por ejemplo al ser informados de convenios de pago con la TGR.

[10] En Anexo 10.8 están los grupos con menor probabilidad de no pago que no tendrían prioridad de acción.

7.1.3. Propuestas sobre otros estudios

Al escoger trabajar la data de giros de los segmentos de Micro y Pequeña Empresa, se han considerado pocos casos de Medianos y Grandes Deudores ya que no existe una gran intersección entre los segmentos. Por esta razón, el segmento de Mediano y Gran Deudor no tiene un resultado predictivo tan bueno como el del conjunto total de la data. Por lo que se recomienda generar un modelo aparte para este subconjunto. En Anexo 10.10, se ha realizado una prueba exploratoria para determinar cuáles podrían ser las variables más importantes para este segmento y según ella, en un nuevo modelo, se deberían considerar variables como C91 / (C91 de la actividad del contribuyente), Trabajadores, Tasa de no pago y convenio de pago. En anexos, se puede encontrar la explicación del modelo exploratorio.

También se propone la realización de un nuevo modelo que considere 1 año como plazo máximo de pago, dado que para seis meses la mayoría de los casos se encuentran concentrados en altas probabilidades.

Para el estudio que se recomendaba para Auditoria IVA y su prevención, se propone utilizar variables que representen conductas negativas fuera de las anotaciones usadas, como otros incumplimientos, y comparaciones más específicas de los resultados económicos de la empresa con las de otras, que logren dar aviso de resultados y comportamientos anómalos e incluir variables del F22 (ya que las del F29 no fueron útiles). Por ejemplo, Promedio C91/ Cantidad de trabajadores promedio de la actividad económica, Promedio C91/ (Promedio C91 de la actividad del mismo segmento económico), Compras/ (Compras promedio de la actividad) ó Retiro de utilidades/Ventas.

Por último, se invita al SII a probar el algoritmo Random Forest en los estudios a realizar. Para este trabajo ha entregado buenos resultados predictivos, y quizás se logre lo mismo en otros casos.

7.2. Propuestas para la TGR

De lo obtenido en algunos nodos, se sabe con certeza que hay deudas que no se pagarán al menos en los 6 primeros meses. Por otro lado, como se ha explicado en el capítulo 2, Tesorería realiza cobranza administrativa, y si no hay pago dentro de determinado plazo se aplica una cobranza judicial. Entonces, dado que con el modelo se sabría con anticipación que a pesar de la cobranza administrativa no habrá pago (asumiendo que se ha realizado para la mayoría de los giros de los datos de entrenamiento), se podría evitar utilizar recursos en ella y adelantar la cobranza judicial, ahorrando dinero y tiempo.

Para la aplicación del modelo en la cobranza, se propone la utilización de las probabilidades de no pago para priorizar la cartera de deudores, partiendo por los giros con mayor probabilidad de no pago, en vez de cobrar otras deudas que podrían ser pagadas sin

cobranza, dado que los giros tienen probabilidad de no pago bajas. Esto logrará reducir los gastos de cobranza usando el modelo y aumentaría la recuperación atendiendo una mayor cantidad de contribuyentes con peor comportamiento. En el caso de que no se pudiese aplicar cobranza judicial se puede reforzar la cobranza administrativa, por ejemplo, al emitir más mensajes y llamados y cartas de cobranza, priorizando también los giros con mayores probabilidades de no ser pagados.

7.3. Análisis económico

La propuesta de utilización del modelo para la cobranza judicial descrita en la sección anterior, genera cuatro posibles situaciones, según los casos de la matriz de confusión, las que tienen beneficios y costos asociados al uso o no de acciones de cobranza, las que se describen a continuación:

- Caso VP: Se predice no pago y se aplica una acción de cobranza. Se supone que el giro será pagado producto de ésta, según una tasa de efectividad de la cobranza.

- Caso FP: Se predice no pago y se aplica una acción de cobranza, aunque no era necesaria ya que el contribuyente iba a cumplir con su obligación de todas formas. Entonces se recupera el valor total del giro, pero incurriendo en gastos de cobranza.

- Caso FN: Se predice pago, por lo que no se aplica cobranza, pero esta era necesaria para que el contribuyente pagara. Se estima un costo, por no cobrar, igual a lo que se habría recuperado si se hubiese cobrado.

- Caso VN: Se predice pago correctamente, sin gastar en cobranza.

La cantidad de cada uno de estos casos, ponderados por sus beneficios menos costos definen una función utilidad. Esta puede maximizarse debido a que las cantidades de cada caso y el promedio de los cargos cobrados están sujetos a una probabilidad de corte la que puede modificarse, donde las probabilidades iguales o mayores a ese punto serán casos de No pago, y para menores el caso contrario. Entonces la función utilidad es la siguiente:

$$U(x) = \sum_{i}^{I} Cantidad\ casos\ tipo\ i(x) \quad (Beneficios_{i(x)} \quad Costos_i)$$

Con,

$$i \in I = [VP, FN, FP, VN]$$
$$x = Probabilidades\ de\ Corte$$

Y beneficios menos costos, según la explicación de cada caso, de:

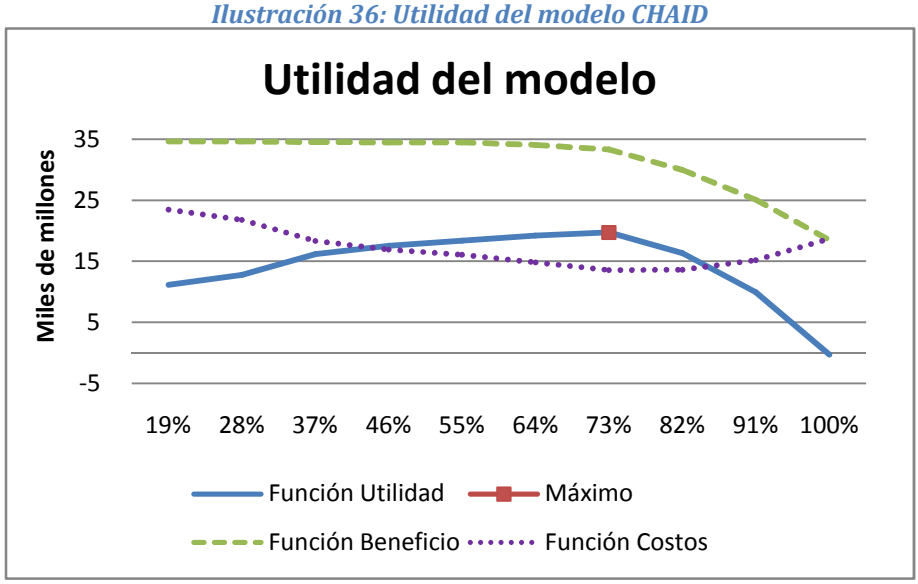
Tabla 24: Beneficios y costos del modelo

Caso	$Beneficios_{i(x)} - Costos_i$
VP	$(Cargo\ promedio_{\ vp(x)}\ \ tasa\ de\ recuperación\ de\ cobranza)\ \ Cobranza$
FP	$Cargo\ promedio_{\ fp(x)}\ \ Cobranza$
FN	$0\ \ (Cargo\ promedio_{\ fn(x)}\ \ tasa\ de\ recuperación\ de\ cobranza)$
VN	$(Cargo\ promedio_{\ vn(x)}\ )$

Fuente: Elaboración propia.

Donde se estima el valor de una cobranza judicial en 400 mil pesos y una tasa de recuperación de impuestos de 60% utilizando este tipo de cobranza. En la Ilustración 36: Utilidad del modelo CHAID, se grafica entonces las funciones de ingresos, costos y utilidad según las probabilidad de corte ordenadas de menor a mayor y considerando el conjunto de datos de la partición de validación. Esta función alcanza su máximo cuando el corte es realizado en la probabilidad de no pago de 73%, y es equivalente a recuperar 19.769 millones de pesos en impuestos, realizando acciones sobre un 52% del total de giros.

Ilustración 36: Utilidad del modelo CHAID



Fuente: Elaboración propia.

Del gráfico se observa que al considerar una probabilidad de corte cercana al 19%, es decir, considerar que desde muy bajas probabilidades de no pago un giro no será pagado, se obtienen altos ingresos ya que se considera la mayor parte de los giros, pero también tiene altos

costos. Esto es debido al aumento del error de clasificación de tipo FP, los que no logran ser compensados por la disminución del error de clasificación del tipo FN ponderados por sus costos. Por el contrario, definir un punto de corte cercano a la probabilidad de no pago de 100%, provoca cobrar muy pocos giros, y recaudar menos ingresos, además de aumentar los costos por el aumento de errores tipo FN. De todas maneras, se observa que los resultados para todos los puntos de corte son positivos, por lo que aunque no se obtenga la recuperación máxima de impuestos no se estará perdiendo dinero con el modelo.

Este resultado se compara con la situación actual de Tesorería, que realizaría cobranza en aproximadamente seis meses más solo a los giros no pagados. Su utilidad entonces se calcula, como todo el dinero de los giros pagados dentro del plazo correspondiente, más lo que recolectaría utilizando el mismo dinero de cobranza que el ocupado por el modelo, priorizando los giros de mayor valor.

Esa forma de recuperación de impuestos, tiene un beneficio económico, por cobrar solo a quienes se sabe no han pagado, ya que no incurre en gastos por cobranza, y por ende no hay cobranzas erradas, o costos por error del tipo FP, como ocurre con el modelo predictivo. Por otra parte, también existen pérdidas. La primera, es que como el valor del dinero en seis meses más no es el mismo que el de hoy, todos los montos obtenidos en el futuro deben actualizarse según la función de valor presente, por lo que hay un diferencial por cobrar después.

$$Valor\ Presente = \frac{Valor\ Futuro}{(1 + tasa\ de\ interés\ periodo)^{n°\ periodos}}$$

A esto se suma, el supuesto de que la tasa de recuperación por cobranza disminuiría de 0,6 a 0,5, ya que al no cobrar en 6 meses, el contribuyente podría adquirir otros giros y perder patrimonio. Como resultado final, se tiene que la recuperación de Tesorería sería de 18.399 millones de pesos, lo que significa que el modelo CHAID obtiene 1.370 millones más respecto al modelo actual, los que pueden ser reinvertidos en otras acciones de la TGR.

8. Conclusiones

Con el trabajo desarrollado se logra cumplir con todos los objetivos de la memoria, al elaborar un modelo predictivo que entrega probabilidades de no pago con alta precisión y desarrollar recomendaciones de su uso, a través del estudio de las variables que selecciona como importantes, de las reglas de decisión que lo definen y de su aplicación en distintos segmentos de interés. A continuación, se revisan los objetivos generales y específicos propuestos y sus resultados obtenidos.

8.1. Conclusiones generales respecto al modelo generado

El modelo CHAID escogido arroja una certeza de predicción global o Accuracy de 78,2%, 83,1% de Precisión y un 82,7% de AUC. Al comparar el modelo CHAID con el modelo Random Forest, definido como modelo alternativo, el modelo Random Forest obtiene mejores métricas de desempeño, por ejemplo, 5,7% más de AUC, pero se selecciona el modelo CHAID al ser más ventajoso en su utilización debido a que este último es de fácil interpretación, permitiendo la identificación de variables que explican el no pago y el aporte cuantitativo de las categorías de cada variable a la probabilidad de no pago (en función de la regla de decisión), por lo que serviría para apoyar al Servicio de Impuesto Internos en su tarea de generar acciones generales sobre los contribuyentes. De todas maneras, el mejor poder predictivo de Random Forest, demuestra que es un algoritmo de alto nivel, y que debería utilizarse para otros casos en los que no se requiera obtener reglas sino solo la probabilidad de un caso.

Como limitación del modelo se observa que posee muchas reglas, por lo que es difícil generalizarlas y deben analizarse por separado. Es por esta razón que para interpretar el modelo se destacan las menores y mayores probabilidades por cada tipo de giro, lo que definirían medidas de acción más o menos simples a aplicar.

8.2. Conclusiones específicas

8.2.1. Respecto a la identificación de variables y patrones

Con el árbol CHAID se obtiene una lista de variables más importantes, que son medidas en una escala de 0 al 1, en la cual destacan las variables Tipo de Giro con un valor de 0,5, Cargo/C91 con 0,15, el valor del giro o Cargo_neto_cat con 0,12, Tasa de no pago con 0,08, seguidos de Cargo/(Compras+Ventas) y Giros_Impagos. A través de esto se establece que la variable más determinante en la explicación del no pago es el Tipo de Giro. Esta afirmación se fundamenta en que el origen de los giros es diferente para cada caso, por lo que en consecuencia, el contribuyente al que está relacionado el giro y su nivel de pago también cambia. Por ejemplo, no es lo mismo un giro por Multa F29, de un contribuyente que olvidó declarar su IVA, que otro de Auditoría IVA, de un contribuyente que evadió impuestos. Los resultados de las mínimas probabilidades de pago para cada tipo de giro son 18,5%, 57,8% y 63,1% para Multa F29, Impuesto F29 y Auditoria IVA respectivamente, y las máximas son 80,7%, 98,1%, 99,5%, lo que muestra notoriamente las diferencias entre cada tipo. Con esto se recomienda que para futuros estudios, se deben generar modelos y/o estudios por cada tipo de giro.

Un resultado interesante es que antes de este trabajo se estimaba que la variable más importante era el valor del giro, pero se comprobó que incluso para giros del mismo monto, las probabilidades de no pago pueden tener variaciones de hasta 49%. Luego, bajo esta perspectiva, el no pago de cada tipo de giro tiene variables o grupos de variables más explicativas: Para Auditoría IVA es el cargo neto, para Impuesto F29 las variables del grupo de caracterización del

giro y para Multa F29, corresponden a las variables de los grupos caracterización del giro y del contribuyente.

Finalmente, los nodos finales del árbol resultante, ayudaron a entender patrones particulares de no pago de giros, y se entregan al SII en una lista ordenando los casos de mayor a menor probabilidad de default. A partir de esas probabilidades, se agrupan los casos en 5 tipos, de los cuales, los grupos que representan un mayor riesgo son los giros denominados "Impuesto F29 (2)" y los de Auditoria IVA, por ser estos de alto valor y tener las mayores probabilidades de no pago en 6 meses.

8.2.2. Respecto a los segmentos de contribuyentes

El modelo CHAID construido y sus buenas medidas de desempeño obtenidas, interpretan de buena forma los segmentos de interés del Servicio de Impuestos Internos, Micro y Pequeña Empresa, validándose así las variables encontradas.

Al realizar la misma prueba para los segmentos de trabajo de la TGR, se obtiene que las variables solo aplican satisfactoriamente en la predicción de Pequeño Deudor. Pero no así, para el segmento de Mediano y Gran Deudor ya que no logran predecir a cabalidad los casos de pago.

Para corroborar el nivel predictivo del modelo Random Forest, se probó con cada segmento de interés del SII y la TGR, el que obtuvo mejores medidas de desempeño que el modelo CHAID, incluyendo la medida de Especificidad, con lo que se concluye su efectividad para ser usado en otros estudios.

8.2.3. Respecto a la predicción de nuevos giros

Entrenado el modelo CHAID, se obtiene una predicción para un conjunto de datos reales del periodo febrero 2016 a junio 2016. Se observa que un 82,7% de los giros del mes de febrero que han sido predichos como No pago, revisados cinco meses después, aún no han sido pagados, es decir, hasta el momento se tiene una Precisión de ese valor, cercana a la calculada en la partición de validación. Además, se ha predicho un no pago para el 66% de los datos y se analiza que los grupos de Auditoria IVA e Impuesto F29 (2) han aumentado su proporción respecto al total de giros.

8.2.4. Respecto a las propuestas para el SII y TGR

Con este trabajo, se han validado variables útiles para trabajar el problema de no pago de giros. De estas, se destacan variables que no estaban programadas en el Servicio, como Cargo/C91, tasa de no pago, Cantidad de giros impagos, y Cargo/(Ventas+Compras). Se considera importante que sean agregadas en sus bases de datos. Además, se propone considerar otras variables, como por ejemplo, la suma de la deuda fiscal total (no solo la de los tres tipos de giros tratados).

Respecto a los grupos de giros, se propone el análisis específico de los giros de Auditoría e Impuesto F29 (2) para lograr prevenirlos, además, de generar planes de apoyo para que los contribuyentes que deben giros de Impuesto F29 (1) puedan sanear su deuda anterior, y planes de educación y apoyo para los contribuyentes con giros de Multa F29.

Por último, se propone que la TGR utilice las probabilidades de pago para priorizar su cobranza, iniciándola por los giros de mayor probabilidad de no pago, ya que los giros de menor probabilidad podrían ser pagados dentro del plazo de 6 meses. Para la acción de adelantar la cobranza en seis meses, se estiman beneficios positivos respecto a la actual cobranza en alrededor de 1.300 millones de pesos extras, logrando recaudar 19.769 millones de pesos.

8.2.5. Trabajos futuros

De aplicarse las medidas recomendadas, se deberá continuar el trabajo midiendo el cambio de conducta de los contribuyentes y el nivel de pago de los giros para definir si fueron eficaces o no. Esto puede lograrse a través de la realización de un diseño experimental que consiste en la comprobación de una hipótesis verificando el cambio de un resultado al modificar una variable sobre un grupo de control [33]. Las hipótesis de un diseño experimental pueden ser la efectividad de las medidas preventivas y paliativas propuestas, o la reducción del nivel de no pago. Por ejemplo, el SII puede testear la medida preventiva de educación para la disminución de la emisión giros de Multa F29 sobre un grupo de control, del subconjunto de contribuyentes con giros de mayor probabilidad de no pago, al que se le recordó la obligación de declarar en las fechas correspondientes como medida preventiva. También puede observarse en la TGR cómo aplicar más medidas en la cobranza administrativa puede mejorar el pago y evitar una cobranza judicial, o cómo cobrar anticipadamente, sabiendo que no habrá pago en los 6 meses siguientes, acelera y mejora el pago.

9. Bibliografía

[1] SERVICIO DE IMPUESTOS INTERNOS. 1980. Ley Orgánica. [Archivo Word] <www.sii.cl/pagina/jurisprudencia/legislacion/basica/ley_organica.doc>[Consulta: 4 de abril de 2016]

[2] SERVICIO DE IMPUESTOS INTERNOS. 2015. Misión, Visión y Valores del SII. [En línea] <http://www.sii.cl/sobre_el_sii/acerca/mision.htm> [Consulta: 4 de abril de 2016]

[3] SERVICIO DE IMPUESTOS INTERNOS. 2015. Plan de Gestión del Cumplimiento Tributario.
[Archivo PDF] <http://www.sii.cl/sobre_el_sii/Plan_Cumplimiento_tributario2015.pdf> [Consulta: 5 de abril de 2016]

[4] SERVICIO DE IMPUESTOS INTERNOS. 2012. Guía para Educación Superior Conoce más sobre los impuestos. [Archivo PDF] <http://www.siieduca.cl/jovenes/Documentos/92-GA-201405295939.pdf> [Consulta: 5 de abril de 2016]

[5] SERVICIO DE IMPUESTOS INTERNOS. 2016. Descripción de Impuestos. [En línea] <http://www.sii.cl/aprenda_sobre_impuestos/impuestos/descripcion.htm> [Consulta: 5 de abril de 2016]

[6] SERVICIO DE IMPUESTOS INTERNOS. 2015. Cuenta Pública 2014. [Archivo PDF] <http://www.sii.cl/cuenta_publica/cta_2014.pdf>

[7] SERVICIO DE IMPUESTOS INTERNOS. 2015. ¿Cómo se hace para? Declarar IVA, PPM y Retenciones. [Archivo PDF] <http://www.sii.cl/como_se_hace_para/declarar_imp_mensuales/Declarar_Iva.pdf> [Consulta: 6 de abril de 2016]

[8] SERVICIO DE IMPUESTOS INTERNOS. 2015. Auditoría Tributaria. [En línea] <http://www.sii.cl/principales_procesos/auditoria_tributaria.htm#4> [Consulta: 6 de abril de 2016]

[9] SERVICIO DE IMPUESTOS INTERNOS.2015. Estadísticas de Empresas por tamaño según ventas.[En línea] <http://www.sii.cl/estadisticas/empresas_tamano_ventas.htm> [Consulta: 6 de abril de 2016]

[10] SERVICIO DE COOPERACIÓN TÉCNICA. 2015. La Situación de la Micro y Pequeña y Empresa en Chile. [En línea] <http://www.sercotec.cl/Portals/0/MANUALES/situaci%C3%B3n%20de%20la%20microempresa.pdf> [Consulta: 6 de abril de 2016]

[11] SERVICIO DE IMPUESTOS INTERNOS. Número de Contribuyentes y Montos por Impuestos Personales Consolidados. [En línea]

<http://www.sii.cl/estadisticas/contribuyentes/impuestos_personales.htm> [Consulta: 6 de abril de 2016]
[12] SUPERINTENDENCIA DE BANCOS E INSTITUCIONES FINANCIERAS. 2014. [En línea]
<http://www.sbif.cl/sbifweb3/internet/archivos/publicacion_10408.pdf> [Consulta: 29 de diciembre de 2015]

[13] BRAVO, C., MALDONADO, S., & WEBER, R. 2010. Experiencias Prácticas en la Medición de Riesgo Crediticio de Microempresarios utilizando Modelos de Credit Scoring. Revista Ingeniería de Sistemas XXIV: 69-88.

[14] COLOMA, P. et al. 2006. Modelos analíticos para el manejo del riesgo de crédito. Trend Management Edición Especial V.8: 45-51.

[15] AUSTRALIAN TAXATION OFFICE. 2014. Compliance in focus 2013-2014. [Archivo PDF]
<https://www.ato.gov.au/uploadedFiles/Content/CS_C/downloads/CSC35735NAT74689.pdf>
[Consulta: 2 de marzo de 2016]

[16] RETTIG, T. 2013. Modelo de predicción de default tributario de contribuyentes del segmento de Micro y Pequeña empresa del Servicio de Impuestos Internos de Chile. Memoria de Ingeniería Industrial. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.107 p.

[17] CASTELLON, P. 2012. Caracterización y Detección de contribuyentes que presentan facturas falsas al SII mediante técnicas de Data Mining. Tesis de Magister Gestión de Operaciones y Memoria de Ingeniería Industrial. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.185 p.

[18]SERVICIO DE IMPUESTOS INTERNOS. Decreto Ley N° 830 Sobre Código Tributario. [Archivo Word] <http://www.sii.cl/pagina/jurisprudencia/legislacion/basica/dl830.doc>
[Consulta: 30 de abril de 2016]

[19] LYN, T. 2000. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting 16: 149-172.

[20] FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P.1996.From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE Fall 1996: 37-54.

[21]KOTSIANTIS, S. 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31: 249-268.

[22] LEE, H., & KIM, S. 2016. Black-Box Classifier Interpretation Using Decision Tree and Fuzzy Logic-Based Classifier Implementation. Original Article International Journal of Fuzzy Logic and Intelligent Systems Vol. 16: 27-35.

[23] ROKACH, L., & MAIMON, O. 2008.Data Mining with Decision Trees Theory and Applications. 1 ª ed. Singapore, World Scientific Publishing Co. Pte. Ltd. 244p.

[24] ZHU, W., ZENG, N., & WANG, N. 2010. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. Nesug 2010.

[25] IBM Knowledge Center. 2011. Cumulative Gains.[En línea] <http://www.ibm.com/support/knowledgecenter/es/SSLVMB_20.0.0/com.ibm.spss.statistics.cs/mlp_bankloan_outputtype_02.htm> [Consulta: 6 de junio de 2016]

[26] HORNING, N. 2010. Random Forest: An algorithm for image classification and generation of continuous fields data sets. En: INTERNATIONAL CONFERENCE on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences. 9 y 10 de diciembre de 2010. Hanoi University, Vietnam.

[27] WILLIAMS, G. 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Unites States of America. Springer. 395p.

[28] MULAKA, M. 2012. A guide to appropriate use of Correlation coefficient in medical research. Malawi Med J. Sep 24(3): 69–71.

[29] LOH, W. 2008. Classification and Regression Tree Methods. En: RUGGERI, F., KENNET, R., & FALTIN, F. Encyclopedia of Statistics in Quality and Reliability. United Kingdom. Wiley. pp: 315–323.

[30] BREIMAN, L. 2011. Random Forest. Machine learning, 45(1): 5-32.

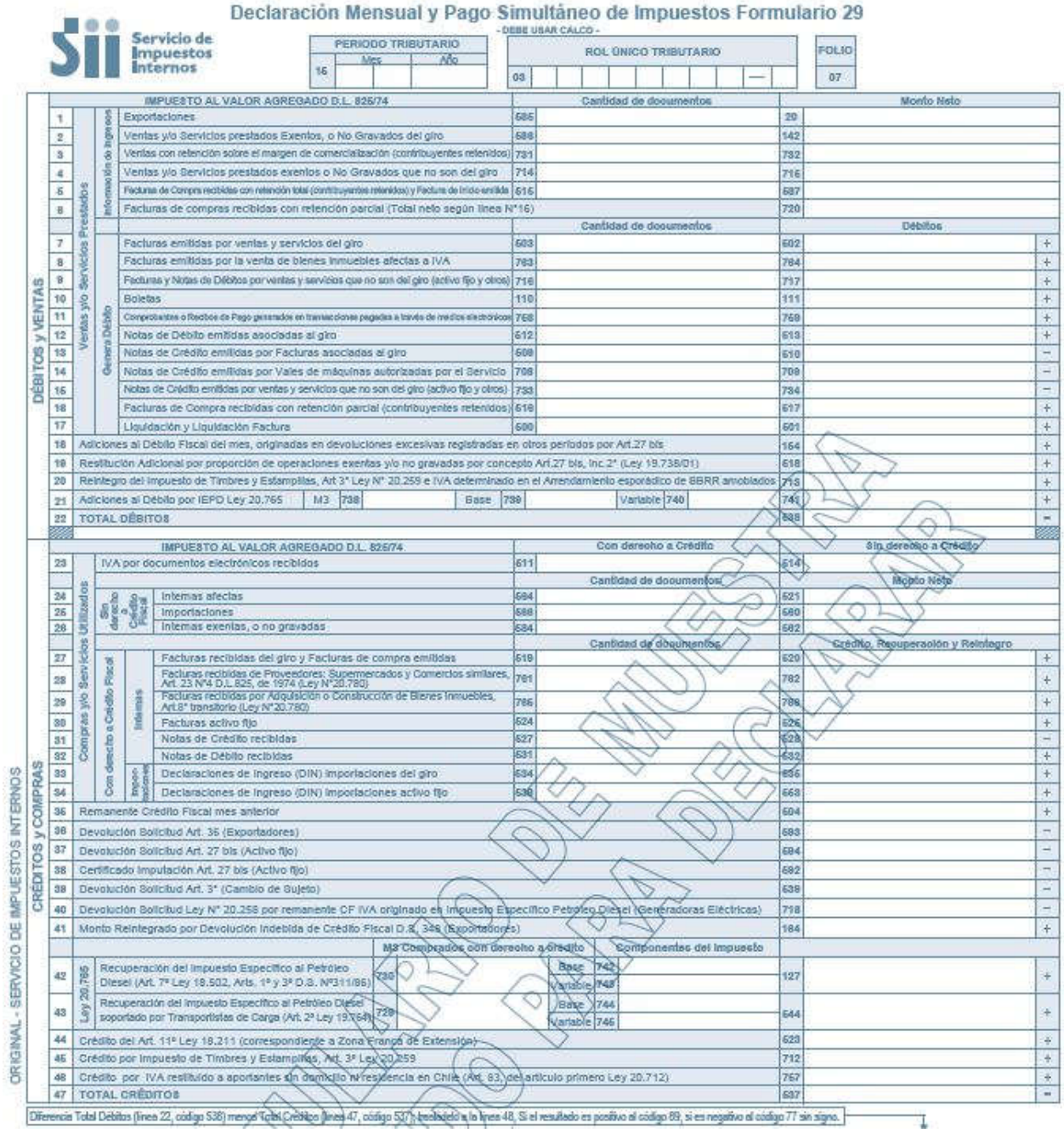
[31] BRAVO, J. 2012. Aplicación de redes neuronales artificiales en el proceso de aplicación de selección de contribuyentes a fiscalizar por utilización de facturas falsas en su contabilidad. Tesis de Magister en Ingeniería Industrial. Santiago, Universidad de Santiago de Chile, Facultad de Ingeniería.166 p.

[32] PLOS ONE. 2016. Random Forests Are Able to Identify Differences in Clotting Dynamics from Kinetic Models of Thrombin Generation. [En línea] <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153776>

[33] SOCIAL SCIENCE RESEARCH NETWORK. 2015. Field Experiments in Marketing. [Archivo PDF] <http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2658661_code617552.pdf?abstractid=2630209&mirid=1> [Consulta: 9 de Octubre de 2016]

10. Anexos

10.1. Formulario 29



Declaración Mensual y Pago Simultáneo de Impuestos Formulario 29

- DEBE USAR CALCO -

Servicio de Impuestos Internos

PERIODO TRIBUTARIO		
	Mes	Año
16		

ROL ÚNICO TRIBUTARIO
03 —

FOLIO
07

DÉBITOS y VENTAS

				Cantidad de documentos		Monto Neto	
1	Ventas y/o Servicios Prestados	Información de Ingresos	Exportaciones	585		20	
2			Ventas y/o Servicios prestados Exentos, o No Gravados del giro	586		142	
3			Ventas con retención sobre el margen de comercialización (contribuyentes retenidos)	731		732	
4			Ventas y/o Servicios prestados exentos o No Gravados que no son del giro	714		716	
5			Facturas de Compra recibidas con retención total (contribuyentes retenidos) y Factura de Inicio emitida	616		687	
6			Facturas de compras recibidas con retención parcial (Total neto según línea N°16)			720	
				Cantidad de documentos		Débitos	
7		Genera Débito	Facturas emitidas por ventas y servicios del giro	503		502	+
8			Facturas emitidas por la venta de bienes inmuebles afectas a IVA	783		784	+
9			Facturas y Notas de Débitos por ventas y servicios que no son del giro (activo fijo y otros)	716		717	+
10			Boletas	110		111	+
11			Comprobantes o Recibos de Pago generados en transacciones pagadas a través de medios electrónicos	768		769	+
12			Notas de Débito emitidas asociadas al giro	512		513	+
13			Notas de Crédito emitidas por Facturas asociadas al giro	508		510	−
14			Notas de Crédito emitidas por Vales de máquinas autorizadas por el Servicio	708		709	−
15			Notas de Crédito emitidas por ventas y servicios que no son del giro (activo fijo y otros)	733		734	−
16			Facturas de Compra recibidas con retención parcial (contribuyentes retenidos)	618		617	+
17			Liquidación y Liquidación Factura	500		501	+
18	Adiciones al Débito Fiscal del mes, originadas en devoluciones excesivas registradas en otros períodos por Art.27 bis					164	+
19	Restitución Adicional por proporción de operaciones exentas y/o no gravadas por concepto Art.27 bis, Inc.2° (Ley 19.738/01)					618	+
20	Reintegro del Impuesto de Timbres y Estampillas, Art 3° Ley N° 20.259 e IVA determinado en el Arrendamiento esporádico de BBRR amoblados					713	+
21	Adiciones al Débito por IEPD Ley 20.765	M3	738	Base	739	Variable 740	741	+
22	TOTAL DÉBITOS						538	=

CRÉDITOS y COMPRAS

					Con derecho a Crédito		Sin derecho a Crédito	
23	Compras y/o Servicios Utilizados		IVA por documentos electrónicos recibidos		611		514	
					Cantidad de documentos		Monto Neto	
24		Sin derecho a Crédito Fiscal	Internas afectas		584		521	
25			Importaciones		588		580	
26			Internas exentas, o no gravadas		584		582	
					Cantidad de documentos		Crédito, Recuperación y Reintegro	
27		Con derecho a Crédito Fiscal	Internas	Facturas recibidas del giro y Facturas de compra emitidas	519		520	+
28				Facturas recibidas de Proveedores: Supermercados y Comercios similares, Art. 23 N°4 D.L 825, de 1974 (Ley N°20.780)	781		782	+
29				Facturas recibidas por Adquisición o Construcción de Bienes Inmuebles, Art.8° transitorio (Ley N°20.780)	786		788	+
30				Facturas activo fijo	524		526	+
31				Notas de Crédito recibidas	527		529	−
32				Notas de Débito recibidas	531		532	+
33			Importaciones	Declaraciones de Ingreso (DIN) importaciones del giro	534		536	+
34				Declaraciones de Ingreso (DIN) importaciones activo fijo	536		668	+
36	Remanente Crédito Fiscal mes anterior						504	+
36	Devolución Solicitud Art. 36 (Exportadores)						593	−
37	Devolución Solicitud Art. 27 bis (Activo fijo)						594	−
38	Certificado Imputación Art. 27 bis (Activo fijo)						592	−
39	Devolución Solicitud Art. 3° (Cambio de Sujeto)						538	−
40	Devolución Solicitud Ley N° 20.258 por remanente CF IVA originado en Impuesto Específico Petróleo Diesel (Generadoras Eléctricas)						718	−
41	Monto Reintegrado por Devolución Indebida de Crédito Fiscal D.S. 348 (Exportadores)						184	+
					M3 Comprados con derecho a Crédito	Componentes del Impuesto		
42	Ley 20.765		Recuperación del Impuesto Específico al Petróleo Diesel (Art. 7° Ley 18.502, Arts. 1° y 3° D.S. N°311/86)	730		Base 742	127	+
						Variable 743		
43			Recuperación del Impuesto Específico al Petróleo Diesel soportado por Transportistas de Carga (Art. 2° Ley 19.764)	729		Base 744	644	+
						Variable 745		
44	Crédito del Art. 11° Ley 18.211 (correspondiente a Zona Franca de Extensión)						523	+
45	Crédito por Impuesto de Timbres y Estampillas, Art. 3° Ley 20.259						712	+
46	Crédito por IVA restituido a aportantes sin domicilio ni residencia en Chile (Art. 83 del artículo primero Ley 20.712)						767	+
47	TOTAL CRÉDITOS						537	=

Diferencia Total Débitos (línea 22, código 538) menos Total Créditos (línea 47, código 537) trasladado a la línea 48. Si el resultado es positivo al código 89, si es negativo al código 77 sin signo.

ORIGINAL - SERVICIO DE IMPUESTOS INTERNOS

Línea	Concepto						Código	IMPUESTO DETERMINADO	
48	Remanente de crédito fiscal para el período siguiente	77				IVA determinado	89		+
49	Restitución de devolución por concepto de Art. 27 ter D.L. 825, de 1974, inc. 2° (Ley N° 20.720)						760		+

IMPUESTO A LA RENTA D.L. 824/74

Línea		Concepto						Código	IMPUESTO DETERMINADO	
50	Retenciones	Retención Impuesto Primera Categoría por rentas de capitales mobiliarios del Art.20 N°2, según Art.73 LIR						50		+
51	Retenciones	Retención Impuesto Único a los Trabajadores, según Art. 74 N° 1 LIR	Créditos	Donación Art. 8° Ley 18.985: 751	Donación Ley 20.444/2010: 735		Impuesto Único 2da. Categoría a Pagar	48		+
52	Retenciones	Retención de Impuesto con tasa del 10% sobre las rentas del Art. 42 N°2, según Art. 74 N°2 LIR						151		+
53	Retenciones	Retención de Impuesto con tasa del 10% sobre las rentas del Art. 48, según Art. 74 N°3 LIR						153		+
54	Retenciones	Retención a Suplementeros, según Art. 74 N° 5 (tasa 0,5%) LIR						54		+
55	Retenciones	Retención por compra de productos mineros, según Art. 74 N° 6 LIR						56		+
56	Retenciones	Retención sobre cantidades pagadas en cumplimiento de Seguros Dotales del Art.17, N°3 (tasa 15%)						588		+
57	Retenciones	Retención sobre retiros de Ahorro Previsional Voluntario del Art. 42 bis LIR (tasa 15%)						589		+
			Acogido a suspensión PPM (Art. 1° bis Ley 19.420 y Ley 19.606)	Monto Pérdida Art. 90	Base Imponible	Tasa	Crédito/Tope Suspensión PPM (Arts. 1° bis Leyes 19.420 y 19.606)		PPM Neto Determinado	
58	PPM	1ra. Categoría Art. 84 a)	750	30	563	115	68	62		+
59	PPM	Mineros Art. 84 a)		565	120	542	122	123		+
60	PPM	Explotador Minero Art. 84 h)		700	701	702	711	703		+
61	PPM	Transportistas acogidos a Renta Presunta, Art. 84, e) y f) (tasa de 0,3%)						66		+
62	PPM	Crédito Capacitación, Ley 19.518/97		Crédito del Mes: 721	Remanente Mes Anterior: 722	Remanente Período Siguiente: 724	Crédito a imputar	723		−
63	PPM	2da. Categoría Art. 84, b) (tasa 10%)						152		+
64	PPM	Taller artesanal Art. 84, c) (tasa de 1,5% o 3%)						70		+

Línea	Concepto	Código		
65	SUB TOTAL IMPUESTO DETERMINADO ANVERSO. (Suma de las líneas 48 a 64, columna Impuesto y/o PPM determinado)	595		=

Si no declara tributación simplificada, Impuesto Adicional (Art. 37 o Art. 42, DL N° 825), cambio de sujeto y créditos especiales por concepto de Sistemas Solares Térmicos; Patentes por Derechos de Agua; Cotización Adicional ; Empresas Constructoras y Peajes Empresas de Transporte de Pasajeros, traslade el valor de línea 65 (código 595) a línea 119 (código 91). En caso contrario continúe al reverso.

01	Apellido Paterno o Razón Social	02	Apellido Materno	05	Nombres

Cambia datos de Domicilio 583 (Si marca con X el casillero, registre los cambios al reverso)

Declaro bajo juramento que los datos contenidos en esta declaración son la expresión fiel de la verdad, por lo que asumo la responsabilidad correspondiente.

Viene de línea 65 código 595, ó línea 113 código 547

Línea	Concepto	Código		
119	TOTAL A PAGAR EN PLAZO LEGAL	91		=
120	Más IPC	92		+
121	Más Intereses y multas	93		+
122	TOTAL A PAGAR CON RECARGO	94		=

10.2. Ejemplo de Giro

REPUBLICA DE CHILE
SERVICIO DE IMPUESTOS INTERNOS

GIRO Y COMPROBANTE DE PAGO DE IMPUESTOS

FOLIO Nº

Cod. 20	21

FORM. 21

GIRO EMITIDO POR EL SII.
FORMULARIO DEBE SER RECEPCIONADO SIN CODIGO DE BARRA.

ROL UNICO TRIBUTARIO	
03	

01	Razón Social o Apellido Paterno	02	Apellido Materno	05	Nombres
06	Calle | Nº | Of./Depto.	09	Telefono	08	Comuna
	Nombre Representante Legal			903	Rut Representante Legal

918	Unidad Giradora	900	Año - Nro Liquidación	151	Plan	115	Periodo Tributario	15	Fecha Vencimiento

Determinacion del Impuesto y Antecedentes :					
303	Rut Fiscalizador		500	Discriminante de Recargos	

IMPUESTO A LA RENTA	COD	VALOR
Rentas Capitales Mobiliarios	122	
Rtas de Bs. Rs.,Agr.,Ind.,Comer.,Min.,Finan.,Otras	123	
Tasa Adicional Ex. Art. 21	124	
Impuesto Unico a los Trabajadores	128	
Impuesto Rentas de Profesionales y Ocup. Lucrat.	129	
Impuesto Global Complementario	133	
Impuesto Adicional (anual)	135	
Impuesto Adicional (de retencion)	136	
Impuesto Unico Art. 21	120	
Impto a los Pequeños Contrib., Arts. 24 y 26 (anual)	138	
Impto Art. 20 No 5 y Art. 34 Nos. 1,2 y 3 (anual)	141	
Pagos Provisionales Mensuales	144	
Impuestos Directos Varios	165	

IMPUESTOS D.L. 825	COD	VALOR
Tasa Ventas y Servicios	175	
Tasa IVA Importaciones	178	
Tasas Especiales Arts. 37,40,41 y 42	179	
Impuesto Especial a los Combustibles	199	
Otros Impuestos	502	
Reajuste Art. 53 C.T.	170	
Intereses	259	
Multas	262	

IMPUESTOS SOBRE HERENCIAS Y DONACIONES	COD	VALOR
Impuesto Determinado Expresado (UTM 2 dec)	75	
Valor U.T.M. del mes de emisión del giro	77	

TOTAL GIRO	91	
I.P.C. (Reajuste Art. 53 C.T.)	92	
Intereses y Multas	93	
TOTAL A PAGAR	94	

GIRO EMITIDO POR EL SII.
FORMULARIO DEBE SER RECEPCIONADO SIN CODIGO DE BARRA.

Firma y timbre
Girador SII

Firma y timbre
Cajero Institucion Recaudadora autorizada

130	Rut Girador	
215	Fecha Emision Giro	

815	Fecha Liquidacion Recargos	-------

Original : Servicio de Impuestos Internos

Este giro puede ser pagado en cualquier Banco o Institución Financiera Autorizada.
Este giro es valido para ser pagado hasta el ultimo día hábil de Marzo de 2012

10.3. Listado de Variables

Grupo Caracterización del Contribuyente

N°	Variable	Tipo de Variable	Descripción
1	Actividad_Economica	Nominal	Sector económico (Ej: Hoteles y Restaurantes)
2	Segmento	Nominal	Tamaño de empresa al año 2015
3	Comuna	Nominal	Comuna de registro de la empresa
4	Regional	Nominal	Unidad regional del SII que atiende a la empresa
5	Genero	Nominal	Si es empresa de persona natural, F o M.
6	Años_Actividad	Numérica	Años desde el registro de la empresa
7	Edad_Contribuyente	Numérica	Si es empresa de persona natural, edad.
8	Difunto	Nominal	Marca de Fallecido

Grupo Capacidad de Pago

N°	Variable	Tipo de Variable	Descripción
1	Cambio_Segmento	Ordinal	Variación del tamaño de la empresa en 2 ult años
2	Trabajadores	Numérica	Cantidad de trabajadores
3	Sueldos	Numérica	Promedio de sueldo mensual
4	Patrimonio_Estatico_Sum	Numérica	Valor del patrimonio de la empresa
5	Patrimonio_Estatico_Count	Numérica	Cantidad de bienes patrimoniales de la empresa

Grupo Variables F29

N°	Variable	Tipo de Variable	Descripción
1	V007_Nat	Nominal	Aumento o disminución ventas de los últ 3 meses
2	Ventas_u12m	Numérica	Monto de ventas de los ult 12 meses
3	Compras_u12m	Numérica	Monto de compras de los ult 12 meses
4	Debito_u12m	Numérica	Monto de débito de los ult 12 meses
5	Credito_u12m	Numérica	Monto de crédito de los ult 12 meses
6	Promedio Mensual C91	Numérica	Promedio de Impuesto declarado (base anual)
7	varmax_d/c_u12m	Numérica	Variación del débito/crédito max y mínimo en relación al d/c máx de los últ 12 meses
8	Ticket_promedio_facturas	Numérica	Promedio de valor de facturas ult 12 meses
9	Ticket_promedio_boletas	Numérica	Promedio de valor boletas ult 12 meses
10	Debito/Credito	Numérica	Relación débito/crédito ult 12 meses
11	Compras/Ventas	Numérica	Relación compras/ventas ult 12 meses

12	Cargo/(Ventas-Compras)	Numérica	Relación cargo/(ventas-compras) ult 12 meses
13	Cargo/(Ventas+Compras)	Numérica	Relación cargo/(ventas+compras) ult 12 meses
14	Ventas+Compras	Numérica	Ventas más compras ult 12 meses
15	Cargo/Ventas	Numérica	Relación cargo/ventas ult 12 meses
16	Cargo/C91	Numérica	Relación cargo sobre promedio mensual C91
17	C91/C91 Actividad	Numérica	Promedio mensual C91 sobre promedio del rubro
18	C91/Trabajadores	Numérica	Promedio mensual C91 sobre cant. Trabajadores
19	Tasa_Margen	Numérica	Ventas menos compras sobre el total de compras
20	Margen	Numérica	Ventas menos compras

Grupo Intención de Pago

N°	Variable	Tipo de Variable	Descripción
1	Convenio?	Nominal	Marca de suscripción a convenio
2	Anotacion_Grave	Nominal	Marca de anotación grave
3	Anotacion_Inconcurrente	Nominal	Marca de Inconcurrente
4	Anotacion_Fiscalizacion	Nominal	Marca de fiscalización
5	Anotacion_No_Declarante	Nominal	Marca de no declarante
6	Anotacion_Querellado	Nominal	Marca de querellado
7	Anotación?	Nominal	Marca de tenencia de anotaciones descritas

Grupo Caracterización del Giro

N°	Variable	Tipo de Variable	Descripción
1	Concepto_Giro	Nominal	Tipo de Giro (Multa F29, Impuesto F29, etc.)
2	Emisión	Numérica	Fecha de emisión del giro
3	Ult_pago_emision	Numérica	Meses desde el último pago realizado
4	Papel	Nominal	Marca de declaración en papel
5	Cargo_Neto_Cat	Numérica Ordinal	Categorías de monto de los giros
6	Tasa_No_Pago	Numérica	Porcentaje de giros pasados no pagados
7	Giros_Impagos	Numérica	Cantidad de giros pasados no pagados
8	Giros_Total	Numérica	Cantidad de giros pasados
9	Mes_Pago	Numérica	Mes de emisión del giro
10	Trimestre_Pago	Numérica	Periodo del año de emisión del giro

10.4. Categorías de variables

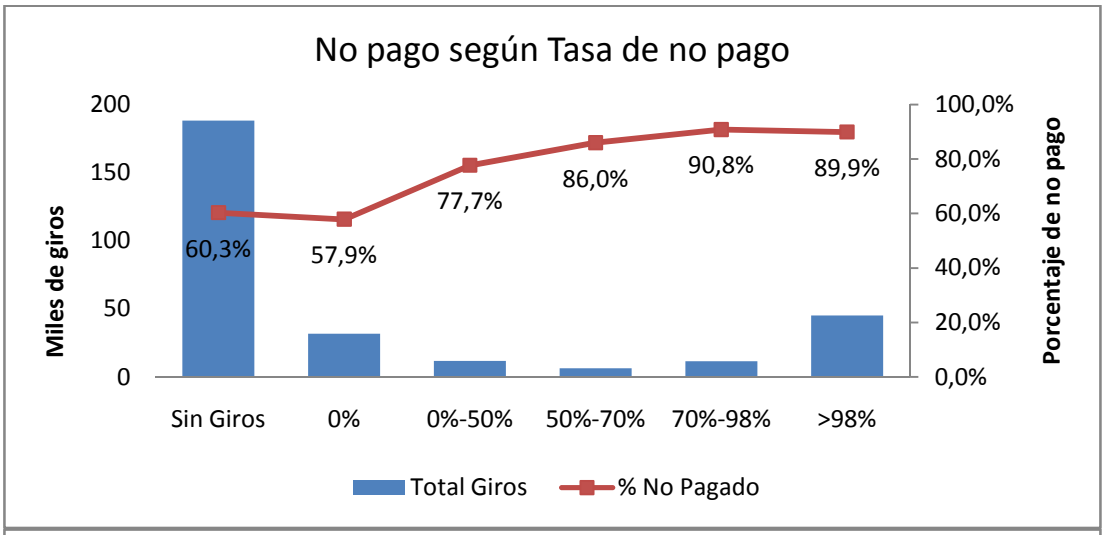
Variable Cargo_Neto_Cat

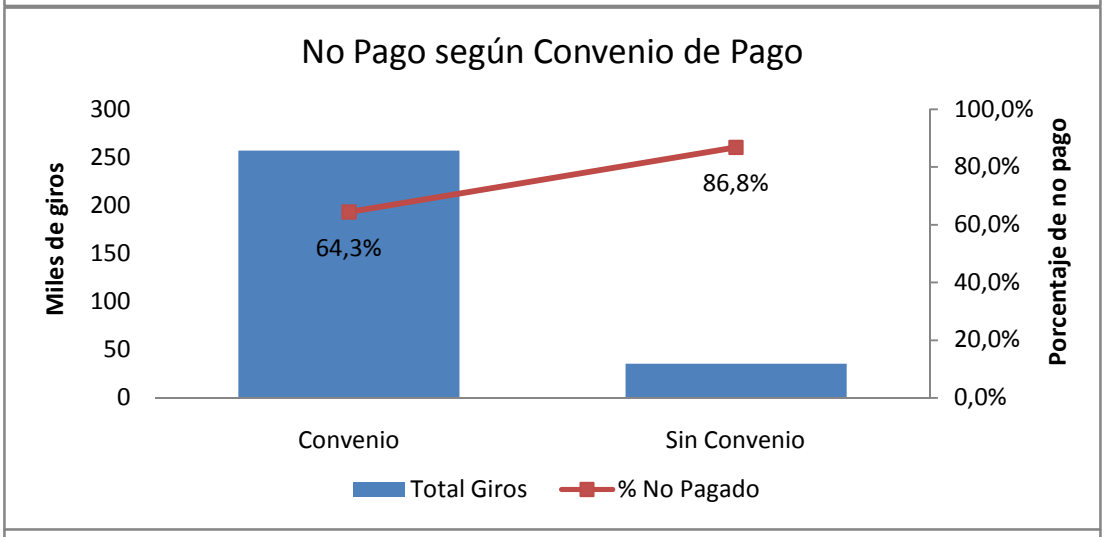
Categoría	Cargo Mínimo	Cargo Máximo
1	$ 4.000	$ 40.084
2	$ 40.085	$ 41.467
3	$ 41.468	$ 43.197
4	$ 43.198	$ 49.004
5	$ 49.005	$ 129.006
6	$ 129.007	$ 258.150
7	$ 258.151	$ 472.074
8	$ 472.024	$ 828.370
9	$ 828.370	$ 1.922.828
10	$ 1.922.829	y más

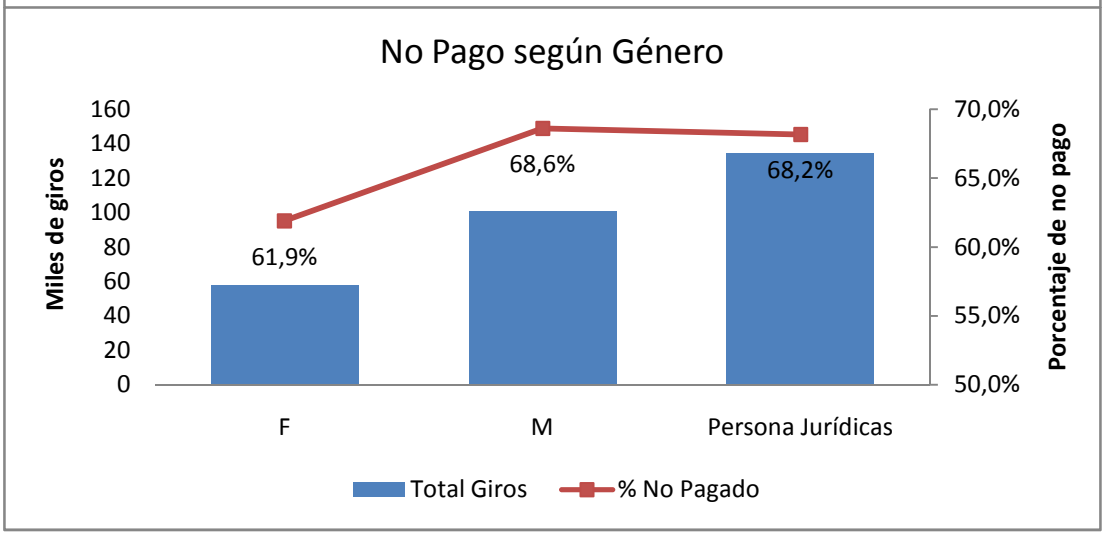
Variable Actividad Económica

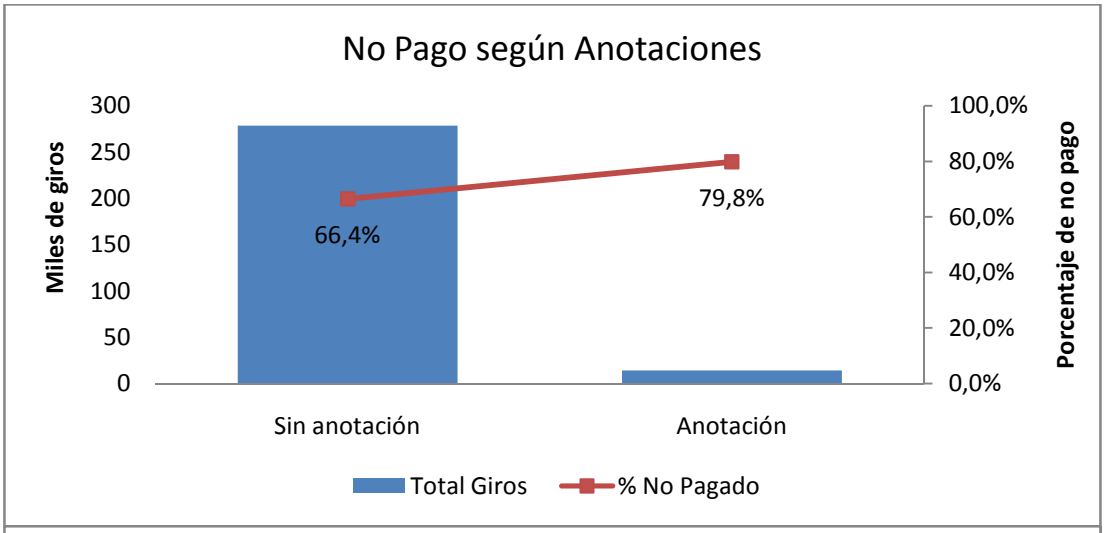
Categoría	Actividad Económica
A	Agricultura, Ganadería, Caza y Silvicultura
B	Pesca
C	Explotación de Minas y Canteras
D	Industrias Manufactureras No Metálicas
E	Industrias Manufactureras Metálicas
F	Suministro de Electricidad, Gas y Agua
G	Construcción
H	Comercio al por Mayor y Menor, Rep. Veh. Automotores/Enseres Domésticos
I	Hoteles y Restaurantes
J	Transporte, Almacenamiento y Comunicaciones
K	Intermediación Financiera
L	Actividades Inmobiliarias, Empresariales y de Alquiler
M	Adm. Pública y Defensa, Planes de Seg. Social Afiliación Obligatoria
N	Enseñanza
O	Servicios Sociales y de Salud
P	Otras Actividades de Servicios Comunitarias, Sociales y Personales
Q	Consejo de Administración de Edificios y Condominios
R	Organizaciones y órganos extraterritoriales

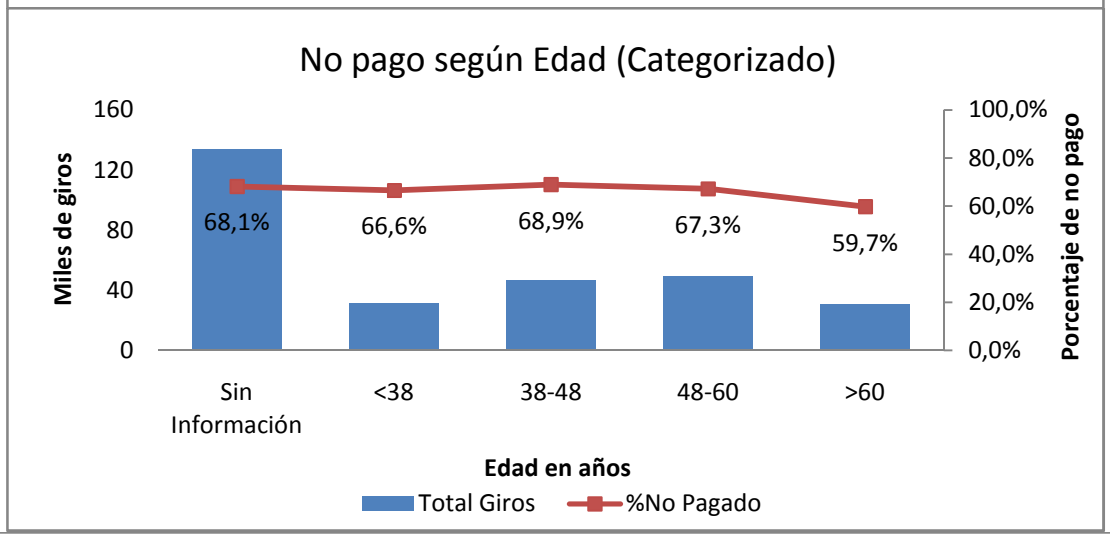
10.5. Capitulo 4. Otros análisis Univariados

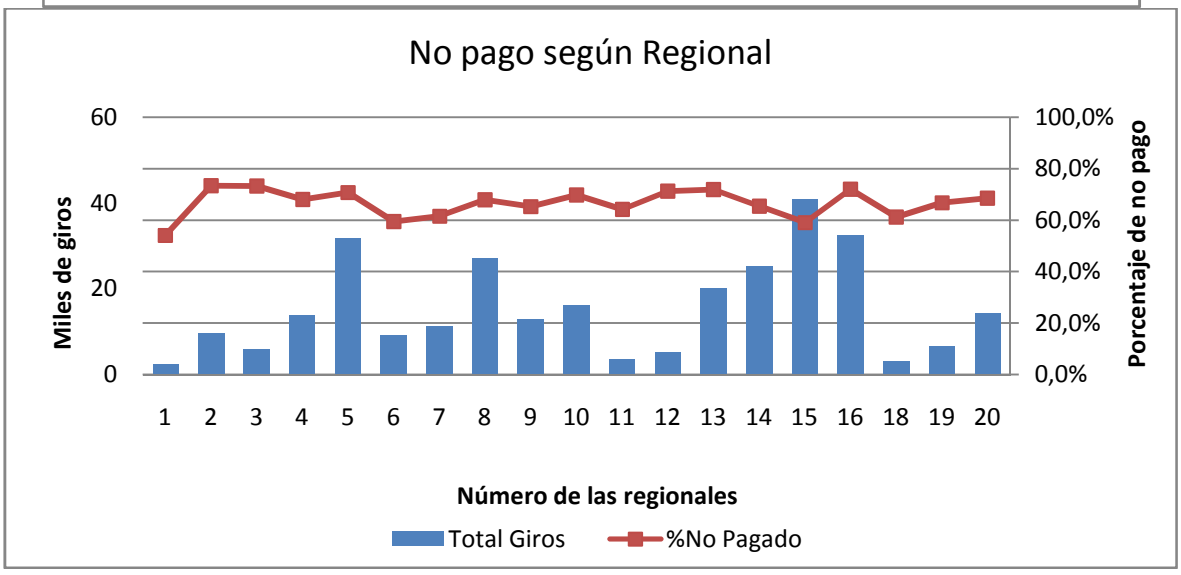


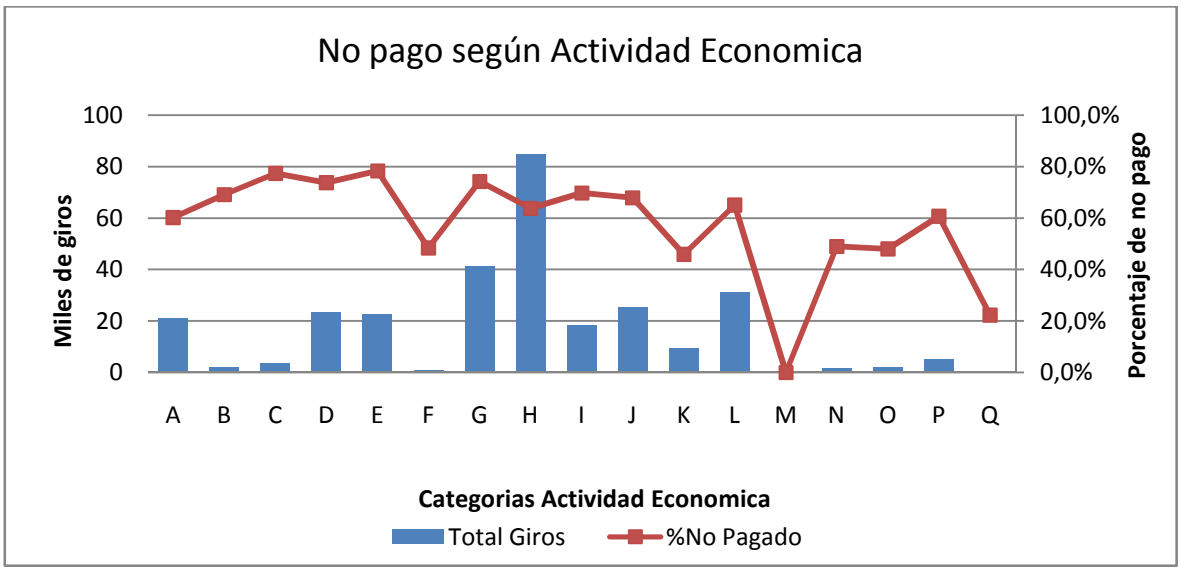





No Pago según Anotaciones
Miles de giros
300
250
200
150
100
50
0
100,0%
80,0%
60,0%
40,0%
20,0%
0,0%
Porcentaje de no pago
66,4%
79,8%
Sin anotación
Anotación
Total Giros
% No Pagado


No pago según Edad (Categorizado)
Miles de giros
160
120
80
40
0
100,0%
80,0%
60,0%
40,0%
20,0%
0,0%
Porcentaje de no pago
68,1%
66,6%
68,9%
67,3%
59,7%
Sin Información
<38
38-48
48-60
>60
Edad en años
Total Giros
%No Pagado


No pago según Regional
Miles de giros
60
40
20
0
100,0%
80,0%
60,0%
40,0%
20,0%
0,0%
Porcentaje de no pago
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 19 20
Número de las regionales
Total Giros
%No Pagado



10.6. Regresión logística

Ecuación del modelo de regresión logística seleccionado

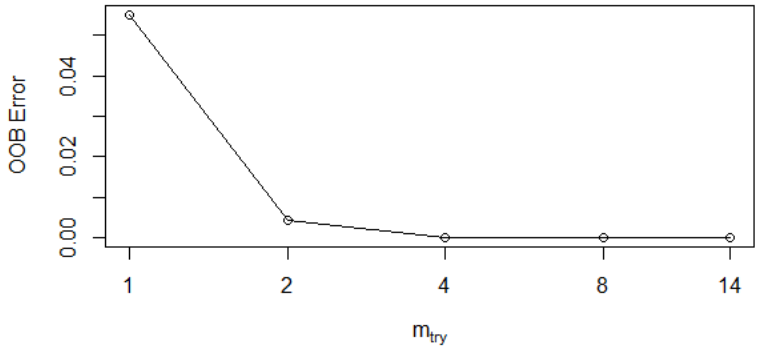
Ecuación para 1

-1,429 * [Cargo_Neto_Cat=1] +
-1,566 * [Cargo_Neto_Cat=2] +
-1,43 * [Cargo_Neto_Cat=3] +
-1,412 * [Cargo_Neto_Cat=4] +
-1,099 * [Cargo_Neto_Cat=5] +
-0,5051 * [Cargo_Neto_Cat=6] +
-0,3795 * [Cargo_Neto_Cat=7] +
-0,1556 * [Cargo_Neto_Cat=8] +
-0,1231 * [Cargo_Neto_Cat=9] +
1,747 * [Concepto_Giro=AuditorialVA] +
1,504 * [Concepto_Giro=ImpuestoF29] +
-0,3472 * [Regional=1] +
-0,03509 * [Regional=2] +
-0,2066 * [Regional=3] +
-0,1569 * [Regional=4] +
-0,1024 * [Regional=5] +
-0,2746 * [Regional=6] +
-0,3624 * [Regional=7] +
-0,146 * [Regional=8] +
-0,239 * [Regional=9] +
-0,09938 * [Regional=10] +
-0,4726 * [Regional=11] +
-0,1267 * [Regional=12] +
0,1461 * [Regional=13] +
-0,1318 * [Regional=14] +
0,0968 * [Regional=15] +
-0,01775 * [Regional=16] +
-0,4585 * [Regional=18] +
-0,2956 * [Regional=19] +
-0,02739 * [Trimestre_Pago=Invierno] +
-0,1224 * [Trimestre_Pago=Marzo] +
-0,02512 * [Trimestre_Pago=Septiembre] +
0,06354 * [Cargo/C91_Cat=>33,3] +
-0,3098 * [Cargo/C91_Cat=0,25] +
-0,4988 * [Cargo/C91_Cat=0,25-1] +
-0,7178 * [Cargo/C91_Cat=1-1,6] +
-0,5823 * [Cargo/C91_Cat=1,6-2,3] +
-0,3522 * [Cargo/C91_Cat=2,3-3,3] +
-0,1772 * [Cargo/C91_Cat=3,3-,4.8] +
0,01741 * [Cargo/C91_Cat=4,8-33,3] +
-0,2447 * [Giros_Impagos_Cat=0] +
-0,7209 * [Convenio?=0] +
+ 1,971

Fuente: Ecuación extraída de SPSS Modeler.

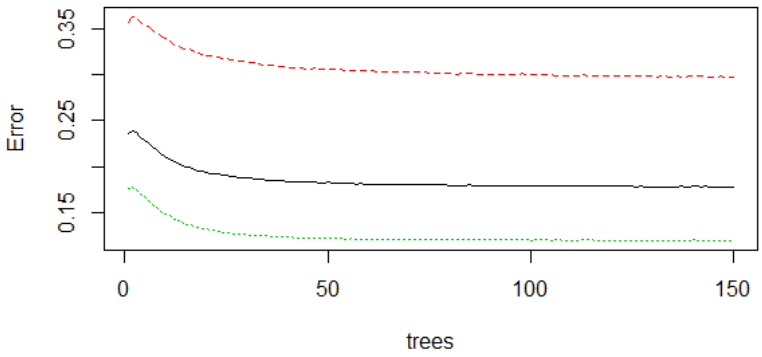
10.7. Variación OOB

OOB Error vs Número de Variables



Fuente: Gráfico extraído de R Studio.

OOB Error vs Número de Árboles



Fuente: Gráfico extraído de R Studio.

10.8. Resultados específicos

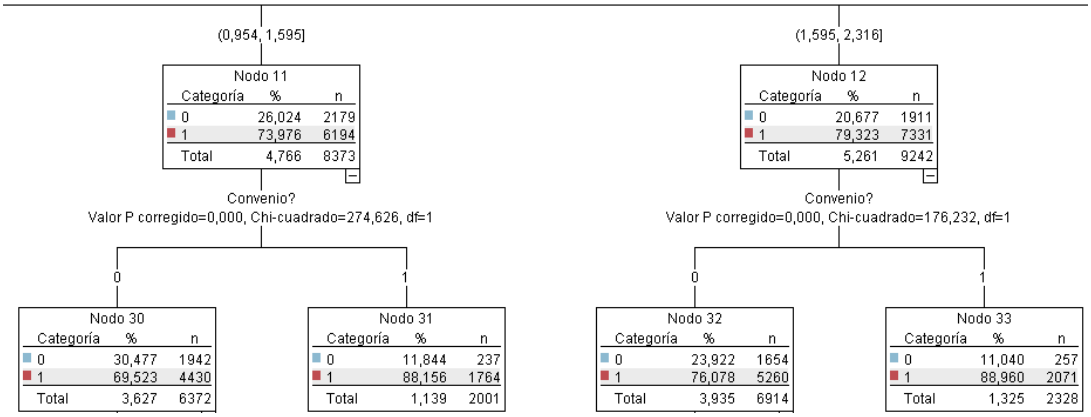
Nodo 1. Auditoría IVA

Reglas para Auditoria IVA

1ra Regla	3ra regla	N° Nodo	% No Pago	% Giros del total
Cargo neto = 10	Cargo/(Ventas + Compras) <= 292.277 o perdidos	26	99,5%	1,77%
Cargo neto = 8,9	Promedio Impuesto C91 <=31.383 o perdidos	24	97,4%	1,73%
Cargo neto = 10	Cargo/(Ventas + Compras) > 292.277	27	95,8%	1,05%
Cargo neto = 7	-	7	90,2%	1,13%
Cargo neto = 8,9	Promedio Impuesto C91 >31.383	25	89,0%	1,14%
Cargo neto = 6	-	6	84,7%	1,05%
Cargo neto = 4,5	-	5	74,3%	1,00%
Cargo neto = 1,2,3	-	4	63,1%	0,82%

Nodo 2. Impuesto F29

Efectividad convenio de pago

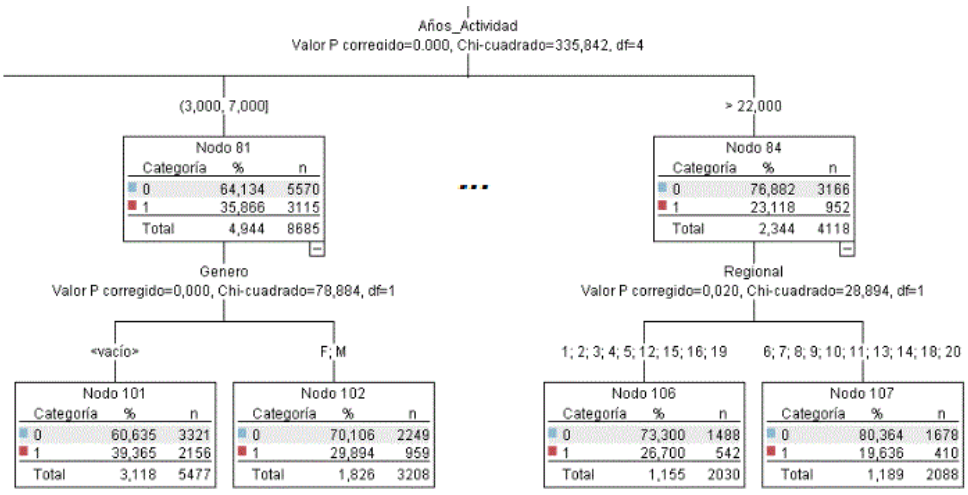


Nodo 3. Multa F29

Menores probabilidades de no pago

1ra regla	Otras reglas	N° Nodo	Prob. No Pago	Giros sobre data total
Cargo neto=2,5 y Actividad Económica=D,H,I,J	Varmax_d/c_ult12m>0 o perdidos, Edad Contribuyente>55	110	18,5%	0,22%
Cargo neto = 1,3,4 y Tasa no pago<=0,5 o Perdidos	Años actividad>22 y Regional=6,7,8,9,10,11,13,14,18,20	107	19,6%	0,23%
Cargo neto=2,5 y Actividad Económica=D,H,I,J	Varmax_d/c_ult12m>0 o perdidos y Edad Contribuyente<=42	108	23,5%	0,25%
Cargo neto=2,5 y Actividad Económica=A,F,Q	Actividad Económica=A,F,Q	54	23,7%	0,34%
Cargo neto = 1,3,4 y Tasa no pago<=0,5 o Perdidos,	Años actividad>22 y Regional=1,2,3,4,5,12,15,16,19	88	26,0%	1,30%

Variación de no pago por años de actividad



10.9. Código modelo Random Forest

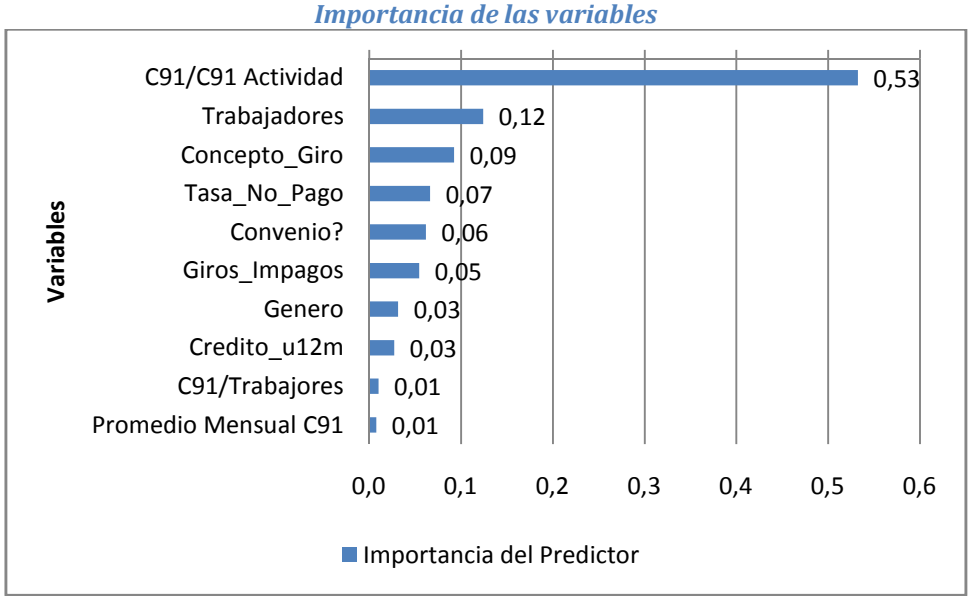
```
#Cargar datos desde la carpeta donde se encuentran
        dataselfor1 <- read.csv("C:/Users /Desktop/Modelos/Data R studio/dataselfor1", sep=";")
        dataselval1 <- read.csv("C:/Users/Desktop/Modelos/Data R studio/dataselval1", sep=";")
         View(dataselfor1)
#Cargar librerias
        library("randomForest", lib.loc="~/R/win-library/3.3")
        library("ROCR", lib.loc="~/R/win-library/3.3")
#Categorizar variables nominales que es estén leyendo como número
        dataselfor1$Pagado3 <- as.factor(dataselfor1$Pagado3)
        dataselfor1$Cargo_Neto_Cat <- as.factor(dataselfor1$Cargo_Neto_Cat)
        dataselfor1$Convenio. <- as.factor(dataselfor1$Convenio.)
        dataselfor1$Regional <- as.factor(dataselfor1$Regional)
#Folio identificador del giro
         row.names(dataselfor1) = dataselfor1$Folio_Giro
#Generar variable dependiente con las independientes
        frmla = Pagado3 ~ Cargo_Neto_Cat + Concepto_Giro + Cargo.C91_Cat +
        Tasa_No_Pago_Cat + Convenio. + Cargo..Ventas.Compras._Cat + Giros_Impagos_Cat +
        Años_Actividad_Cat + Ventas.Compras_Cat + Genero + Regional + Trimestre_Pago +
        Actividad_Economica + Edad_Contribuyentes_Cat
# Generar bosque con la variables dependiente, la data de entrenamiento y el número de árboles y
variables
        fit.rf = randomForest(frmla, data=dataselfor1, ntree=80, mtry=8)
        fit.rf
        importance(fit.rf)
         plot(fit.rf)
         plot(importance(fit.rf))
#Probando la data en partición de validación
        row.names(dataselval1) = dataselval1$Folio_Giro
# Categorizar variables nominales que es estén leyendo como número en partición de validación
        dataselval1$Pagado3 <- as.factor(dataselval1$Pagado3)
        dataselval1$Cargo_Neto_Cat <- as.factor(dataselval1$Cargo_Neto_Cat)
        dataselval1$Convenio. <- as.factor(dataselval1$Convenio.)
        dataselval1$Regional <- as.factor(dataselval1$Regional)
#Obtener probabilidades del caso positivo
        prob = predict(fit.rf,type="prob",dataselval1)[,2]
#Obtener clases según la probabilidad del caso positivo, con corte con probabilidades de 0,5.
        pred= prediction(prob, dataselval1$Pagado3)
        perf = performance(pred,"tpr","fpr")
#Graficar curva ROC
        plot(perf,main="ROC Curve for Random Forest",col=2,lwd=2)
        abline(a=0,b=1,lwd=2,lty=2,col="gray")
        auc <- performance(pred,"auc")
        auc <- unlist(slot(auc, "y.values"))
        minauc<-min(round(auc, digits = 2))
        maxauc<-max(round(auc, digits = 2))
```

```
minauct <- paste(c("min(AUC) = "),minauc,sep="")
maxauct <- paste(c("max(AUC) = "),maxauc,sep="")
#Guardar el modelo generado
save(fit.rf, file="rf80-8.RData")
load("C:/Users/Desktop/Modelos/Data R studio/Modelos RF/rf80-8.RData")
```

10.10. Modelo exploratorio Segmento Mediano y Gran Deudor

Como se sabe la data de giros de alta deuda en 6 meses esta desbalanceada, por lo que se corrige la proporción quitando no pagadores, dejándola con 40% de giros pagados y 60% de no pagados. Con esto, se genera un nuevo modelo CHAID, que con las mismas condiciones de entrenamiento, obtiene un desempeño solo un poco más bajo que el del primer árbol.

Importancia de las variables



Fuente: SPSS Modeler.

Con el gráfico se puede comparar las variables explicativas de este modelo respecto a las expuestas en la sección 5.2 del modelo CHAID general. Estas no coinciden entre sí o cambian su nivel de importancia. Ejemplos son, la variable C91/C91 Actividad, o la cantidad de impuesto promedio mensual de la empresa sobre el promedio de impuesto de todas las empresas de la misma actividad económica (Ej: Agricultura, Comercio, etc.), que no aparece en el listado del modelo general. También aparece la variable Trabajadores, que aquí tiene una importancia relativa de 12% y que en el otro modelo no considera. Entonces, se destaca que aunque el segmento puede abordarse con el modelo general, podría mejorarse en el futuro su especificidad y predicción de pagadores utilizando un modelo diferente, tomando en cuenta el listado de variables expuesto.

Matriz de Confusión		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	935	246
Real Pago	258	419

Medida	Valor
Accuracy	72,9%
Precisión	78,4%
Sensibilidad	79,2%
Especificidad	61,9%
F-Score	78,8%

10.11. Particiones modelo CHAID

Matriz de Confusión Formación		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	100089	17905
Real Pago	19823	37866

Matriz de Confusión Testeo		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	33401	5852
Real Pago	6616	12553

Matriz de Confusión Validación		
Clases	Predicción No Pago	Predicción Pago
Real No Pago	33100	6053
Real Pago	6728	12854