Caracterización de Contribuyentes que Presentan Facturas Falsas al SII Mediante Técnicas de Data Mining

Pamela Castellón [*]
Juan D. Velásquez[**]

Resumen

En este trabajo se entregan evidencias que es posible caracterizar y pronosticar a aquellos usuarios potenciales de facturas falsas en un año determinado, en función de la información de su pago de impuestos, el comportamiento histórico y sus características particulares, utilizando para ello distintas técnicas de Data Mining. En una primera instancia se aplican técnicas de SOM, Gas Neuronal y Árboles de Decisión para identificar aquellas variables que están relacionadas con un comportamiento de fraude y/o no fraude y detectar patrones de conducta asociada a esta problemática. Posteriormente se utilizan Redes Neuronales y Redes Bayesianas para establecer en qué medida se pueden predecir casos de fraude y no fraude con la información disponible. De esta forma se contribuye a identificar patrones de fraudes y generar conocimiento que pueda ser utilizado en la labor de fiscalización que realiza el Servicio de Impuestos Internos para detectar este tipo de delito tributario.

Palabras Clave: Facturas Falsas, Fraude Tributario, Data Mining, Clusterización, Predicción.

[*]Servicio de Impuestos Internos de Chile
[**]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

1. Introducción

El fraude, en sus diversas manifestaciones, es un fenómeno del que no está libre ninguna sociedad moderna. Todas las instituciones, independiente de si son grandes o pequeñas, públicas o privadas, locales o multinacionales, se ven afectada por esta realidad que atenta gravemente contra los principios de solidaridad y de igualdad de los ciudadanos ante la Ley y pone en riesgo los negocios. De acuerdo a un estudio realizado por Ernst&Young en el año 2006 en el cual se encuestó a 150 empresas chilenas, medianas y grandes, un 41 % de ellas declaró haber sido víctima de algún tipo de fraude en los dos últimos años [8]. Esto plantea grandes desafíos en materia de detección y prevención, considerando que el fraude normalmente es mayor que lo declarado por las empresas, debido a que de alguna manera se resiente la imagen de la compañía y en muchos casos, incluso, hay empresas que no están en conocimiento de que han sido víctimas de un fraude.

La Evasión Tributaria y el Fraude Fiscal un tema que ha sido una constante preocupación de todas las administraciones tributarias, en especial de aquellas pertenecientes a países en vías de desarrollo[1]. Si bien es cierto, los impuestos no son la única fuente de financiamiento de un gobierno, es un hecho que éstos marcan una señal muy importante respecto al compromiso y la eficacia con que el Estado puede ejecutar sus funciones, y condicionar el acceso a otras fuentes de ingresos. En el caso de Chile, los ingresos tributarios proporcionan aproximadamente un 75 % de los recursos con que año a año el Estado sustenta sus gastos e inversiones, alcanzando durante el año 2010 un monto de $17,7 billones de pesos[2].

La utilización y venta de facturas falsas como mecanismo de evasión, es particularmente relevante, pues no sólo provoca una elusión de los impuestos, sino que en la mayoría de los casos implica un delito tributario. Por otra parte, junto a la generación de una merma en la recaudación, se producen efectos económicos negativos en el resto de las empresas, por el hecho de generar una competencia desleal frente a aquellas empresas que cumplen adecuadamente con sus obligaciones tributarias. Asimismo, se requiere que los recursos in-

[1] Habitualmente se habla de "elusión fiscal" cuando se hace referencias a conductas que, dentro de la Ley, evitan o reducen el pago de impuestos, mientras que la "evasión o fraude fiscal" supone un quebrantamiento de la legalidad para obtener para obtener esos mismos resultados.

[2] Información publicada en la Cuenta Pública SII 2010 de Marzo 2011, considerando los Ingresos Tributarios del Gobierno Central (sin incluir a Codelco, las Municipalidades y la Seguridad Social).

vertidos en fiscalización sean bien enfocados, detectando a aquellos de mayor riesgo de cumplimiento y no importunar ni desperdiciar tiempo y recursos en aquellos que si cumplen con sus obligaciones. Para ello, las técnicas de data mining ofrecen un gran potencial, ya que permiten extraer y generar conocimiento de grandes volúmenes de datos para caracterizar y detectar conductas fraudulentas y de incumplimiento para optimizar el uso de los recursos. Este artículo se organiza de la siguiente forma: en la sección 2 se describe la problemática e implicancias del uso de facturas falsas sobre la recaudación de los impuestos. La sección 3, describe la manera en que las técnicas de inteligencia artificial han facilitado la detección del fraude fiscal en otras administraciones tributarias. La sección 4 describe el acercamiento propuesto para caracterizar y detectar fraude en la emisión de facturas a través de las técnicas de data mining. La sección 5 presenta las principales conclusiones y las líneas de investigación futuras.

2. Necesidad de Detectar Fraude en un Institución Recaudadores de Impuestos

El Servicio de Impuestos Internos (SII) es la Institución responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario y propiciar la reducción de los costos de cumplimiento, en pos del desarrollo económico de Chile y de su gente. Para ello cuenta con 4.183 funcionarios, de los cuales el 31 % corresponde a fiscalizadores, quienes deben velar por el cumplimiento de 3.4 millones de contribuyentes, considerando los declarantes del Impuesto al Valor Agregado (IVA) y el Impuesto a la Renta.
Particularmente el IVA se ha convertido en un componente clave de la recaudación fiscal, representando durante el año 2010, el 47 % del total de los ingresos tributarios recaudados, por un monto de $8,3 billones de pesos [19]. Actualmente existen 708 mil contribuyentes que declaran IVA, de los cuales 28.000 están autorizados para emitir facturas electrónicas, lo cual ha ido aumentando progresivamente desde el año 2003, como parte de la política adoptada por el SII para modernizar su gestión y asegurar la autenticidad de los emisores de documentos tributarios. Del total de facturas emitidas, un 60 % se emite en formato papel y un 40 % en formato electrónico, generándose cerca de 400 millones de facturas al año.
El fenómeno de las facturas falsas respecto del IVA se explica por la mecánica de determinación del impuesto. Cuando una empresa recibe una factura falsa, aparenta con ello una compra que nunca existió, con lo que aumenta fraudu-

lentamente su crédito fiscal y disminuye su pago de IVA. Asimismo se produce una disminución del pago en el Impuesto a la Renta, debido al aumento de los costos y gastos declarados.

La falsedad del documento puede ser "material", si en él se han adulterado los elementos físicos que conforman la factura o "ideológica", cuando la materialidad del documento no está alterada, pero las operaciones que en ella se consignan son adulteradas o inexistentes. Ésta última es más difícil y compleja de detectar, ya que implica transacciones ficticias, en las cuales se requiere una auditoria para revisar los libros de compra y las rectificaciones o la realización de cruces de información con proveedores. Por otra parte, estos casos son más costosos para el Servicio, ya que requieren una mayor cantidad de tiempo destinado a la recopilación de antecedentes y pruebas, las cuales son más difíciles de encontrar.

Los casos más conocidos de falsedad material son la adulteración física del documento, la utilización de facturas colgadas en la que se falsifica una factura para suplantar a un contribuyente de buen comportamiento tributario, y el uso de doble juego de facturas, en la que se tiene dos facturas de igual numeración pero una de ellas ficticia y por un monto mayor. En el caso de la falsedad ideológica se encuentran las facturas utilizadas para registrar una operación inexistente o que adulteran el contenido de una operación existente. Adicionalmente existen otros delitos comúnmente relacionados, como la falsificación del inicio de actividades a través de palos blancos, con la única finalidad de adquirir facturas timbradas que posteriormente son vendidas a otros contribuyentes.

De acuerdo a un método de estimación de la evasión del IVA por concepto de facturas falsas y otros abultamientos de créditos, aplicado en el periodo 1990-2004 por el SII, la evasión por facturas falsas ha representado entre un 15 % y un 25 % de la evasión total del IVA, aumentando considerablemente en años de crisis económicas. Es así como en el año 1992, el porcentaje de participación aumentó a un 30 % y en la crisis del año 1998-1999 alcanza su punto máximo con un 38 % de participación, año en que alcanza una cifra cercana a los $317.000 millones de pesos. Esto adquiere relevancia producto que recientemente se produjo una crisis económica mundial que afectó a Chile a fines del 2008 y mediados del 2009, provocando un aumento de la tasa de evasión del IVA a un 18 %, por un monto evadido de $1,5 billones de pesos.

Asimismo, la detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera un importante costo administrativo para las áreas de fiscalización y jurídica. Durante el año 2010, el costo de recaudación de $100 fue de $0,91, es decir, aproximadamente un 1 % del valor recaudado. En el periodo 2001-2007 se han presentado

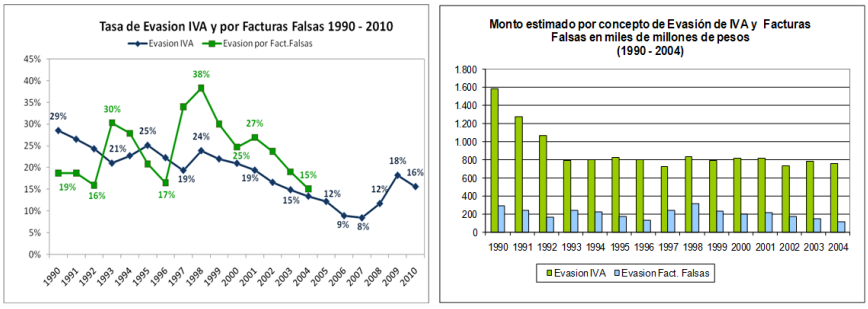


Figura 1: Tasa y Monto de Evasión en el IVA y por Facturas Falsas, Periodo 1990-2010 - Fuente: Subdirección de Estudios, SII

más de 2.300 querellas por facturas falsas y otros delitos de defensa judicial, las cuales involucraron a más de 4.000 querellados, por un monto de perjuicio fiscal cercano a los $274.130 millones de pesos.

Estadísticas SCE	2001	2002	2003	2004	2005	2006	2007	Acumulado
Cantidad de Querellas	171	394	358	407	451	306	243	2.330
Cantidad de Querellados	371	835	667	839	801	537	386	4.436
Monto Perjuicio Fiscal (MM$)	29.370	36.407	49.751	58.812	47.856	21.620	30.314	274.130
Casos SCE[3]	830	2.081	1.794	1.609	1.553	1.052	870	9.789

Tabla 1: Estadísticas de acciones legales relacionadas con facturas falsas 2001-2007 - Fuente: Cuenta Pública SII, 2005, 2006, 2007

El SII utiliza diversos métodos para seleccionar contribuyentes a ser controlados. En el caso de las fiscalizaciones masivas, los contribuyentes se determinan como resultado de un proceso de cruce de información de las declaraciones recibidas y otras fuentes de información, en la cual se detectan inconsistencias y diferencias tributarias. Las fiscalizaciones selectivas, en cambio, se generan en respuesta a determinadas figuras de evasión, ya sea a nivel nacional o local, utilizando para ello distintos ratios tributarios y condiciones, los cuales

utilizan información parcial del contribuyente. Para ello resulta fundamental, aprovechar la gran cantidad de información disponible en los sistemas respecto del comportamiento cada contribuyente en el tiempo.

3. Trabajos Relacionados

La mayor parte de las administraciones tributarias planifican su lucha contra el fraude fiscal. No obstante, existen importantes diferencias en los mecanismos, alcances, enfoque, contenido y énfasis puestos en dicha labor. Para detectar el fraude fiscal, las instituciones comenzaron aplicando auditorias de selección aleatoria o enfocándose en aquellos casos que no tuvieran fiscalizaciones en periodos anteriores recientes y seleccionando casos de acuerdo a la experiencia y conocimiento de los auditores [18]. Posteriormente, se desarrollan metodologías basadas en análisis estadísticos y en la construcción de ratios tributarios o financieros, lo cual evolucionó a la creación de sistemas basados en reglas y modelos de riesgo, que transforman la información tributaria en indicadores que permitan rankear a los contribuyentes por riesgo de cumplimiento. Durante los últimos años, las técnicas de Data Mining e Inteligencia Artificial, han sido incorporadas en las actividades de planificación de auditorías, principalmente para detectar patrones de fraude o de evasión, las cuales han sido utilizadas por las instituciones tributarias con fines específicos.
La Internal Revenue Service, institución a cargo de administrar los impuestos en Estados Unidos, ha utilizado técnicas de Data Mining con distintos fines, entre los que se encuentran la medición del riesgo de cumplimiento de los contribuyentes, la detección de la evasión tributaria y actividades financieras delictivas, la detección de fraude electrónico, la detección de abusos en impuesto de las viviendas, la detección de fraude en contribuyentes que reciben ingresos obtenidos por crédito fiscal y lavado de dinero [10]. Para ello ha utilizado modelos de regresión logística, árboles de decisión, redes neuronales, algoritmos de clustering y técnicas de visualización como Link Analysis, entre otros.
En la Administración Tributaria de Australia, el Çompliance Program" se basa en un modelo de riesgos, que utiliza estadísticas y Data Mining con el objetivo de realizar comparaciones, encontrar asociaciones y patrones mediante modelos de regresión logística, árboles de decisión y SVM cite18. Un caso de interés ha sido el enfoque utilizado por Denny, Williams y Christen [6] de descubrimiento de pequeños clusters o subpoblaciones inusuales, denominadas "Hot Spots", utilizando técnicas como el Self Organizing Map (SOM) para

explorar sus características, algoritmos de agrupación como k-means y representaciones visuales, que son fáciles de entender para usuarios no técnicos.
En el caso de Nueva Zelanda, el modelo existente asocia el grado de cumplimiento con la atención del control, el cual coincide con el utilizado por la administración australiana [18]. El Plan incluye un análisis del entorno económico, internacional, poblacional, de diversidad étnica y de estructura familiar. Por su parte, Canadá utiliza redes neuronales y árboles de decisión para distinguir las características de los contribuyentes que evaden o cometen fraude, en base a los resultados de auditorías pasadas, para detectar los patrones de incumplimiento o evasión [18].
A nivel latinoamericano, Perú fue uno de los primeros en aplicar estas técnicas para detectar evasión tributaria, incorporando al sistema de selección en la Aduana Marítima del Callao una herramienta de inteligencia artificial basada en redes neuronales [3]. Durante el año 2004, este modelo fue mejorado a través de la aplicación de reglas difusas y de asociación para el pre-procesamiento de las variables y árboles de clasificación y regresión (CART) para seleccionar las variables más relevantes. Por su parte, Brazil desarrolló el proyecto HARPIA (Risk Analysis and Applied Artificial Intelligence) de manera conjunta entre la Brazilian Federal Revenue y las universidades de ese país [7]. Este proyecto consiste en desarrollar un sistema de detección de puntos atípicos que ayude a los fiscalizadores a identificar operaciones sospechosas basado en la visualización gráfica de información de importaciones y exportaciones históricas, y un sistema de información de exportación de productos, apoyado en cadenas de markov, para ayudar a los importadores en el registro y clasificación de sus productos, evitar duplicidades y calcular para la probabilidad de que una cadena es válida en un determinado dominio.
En el caso de Chile, la primera experiencia fue desarrollada en el año 2007, utilizando SOM y K-means para segmentar contribuyentes de IVA de acuerdo a sus declaraciones de F29 y características particulares [13]. Posteriormente, siguiendo la tendencia internacional, en el año 2009 se construyen modelos de riesgos en distintas etapas del ciclo de vida del contribuyente, en los que se aplican técnicas de redes neuronales, árboles de decisión y regresión logística. Adicionalmente se desarrolla la primera experiencia para detectar potenciales usuarios de facturas falsas a través de redes neuronales artificiales y árboles de decisión, utilizando principalmente información de su declaración de IVA y Renta en micro y pequeñas empresas.

4. Aplicación de Data Mining para la Detección de Fraude en la Emisión de Facturas

A diferencia del estudio anterior desarrollado en el año 2009 relacionado con esta problemática, este trabajo busca complementar el uso de información de impuestos con variables adicionales relacionadas a su comportamiento histórico y su comportamiento en el año de análisis, así como incluir aspectos concernientes a sus relacionados directos, tales como mandatarios, socios y representantes legales. Por otra parte, se desarrolla un modelo para medianas y grandes empresas, en los que existe menor conocimiento de forma de operar respecto del uso de facturas falsas, debido a que tienen procedimientos más complejos de evasión.

4.1. Datos Utilizados

Para efectos de la caracterización se escoge el año 2006 como año de estudio. Si bien el peak de contribuyentes usuarios de facturas falsas detectados ocurre en el año 2002, se determina utilizar información más reciente, debido a que las dinámicas de evasión se van modificando en el tiempo, al igual que lo hicieron los formularios de pago de impuestos en ese periodo. Por otra parte, las auditorias se realizan hasta un periodo de 3 años atrás, lo que dificulta utilizar información más reciente, pues durante el año 2010 aún se estaban generando casos que podrían haber utilizado facturas falsas desde el año 2007 hacia adelante. De esta forma, el universo queda compuesto por todos aquellos contribuyentes que hayan presentado al menos una declaración de IVA entre el año 2005 y 2007, correspondiente a 582.161 empresas. Para caracterizar a los casos de fraude/no fraude se utiliza información de aquellas auditorias en las que existe certeza que se le revisaron sus facturas del año 2006, independiente del momento en el que fue realizada, generando un total de 1.692 empresas.

Contribuyentes del análisis	MI y PE	ME y GR	Total
Empresas activas en el periodo 2005-2007	558.319 (96 %)	23.842 (4 %)	582.161
Empresas auditadas por facturas en el 2006 con resultado de fraude o no fraude conocido	1.280 (76 %)	412 (24 %)	1.692

Tabla 2: Número de Contribuyentes Utilizados en el Análisis

Uno de los mayores inconvenientes para obtener la información de casos con fraude y no fraude se produce por la forma en la que se registra la información, pues se conoce la fecha de inicio y término de la auditoría, así como los periodos tributarios revisados y el resultado obtenido, pero la información de los periodos en los que ocurren las diferencias no está automatizada. Por lo tanto, para saber si la factura falsa detectada correspondía al año 2006 específicamente, hubo que revisar las anotaciones y comentarios efectuados por el auditor y las rectificatorias efectuadas en códigos relacionados con facturas de ese año.

Los casos de fraude y no fraude se categorizaron en tres tipos: "0" indica que el contribuyente fue auditado y no se encontraron facturas falsas en ninguno de los periodos revisados, "1" que indica que el contribuyente no utilizó facturas falsas en el año de análisis pero sí en otros periodos revisados (normalmente el año anterior o siguiente) y "2" que indica que el contribuyente utilizó facturas falsas en el año de estudio.

Para la construcción del vector de características se seleccionaron 20 códigos del Formulario de Pago Mensual de IVA (F29), 31 códigos del Formulario del Impuesto Anual de la Renta (F22) asociados a la generación de la base imponible de primera categoría y datos contables de la empresa, y 31 ratios tributarios que relacionan la información de IVA y Renta y la rentabilidad de la empresa con su liquidez, entre otros. Adicionalmente se generan 92 indicadores que pueden dar indicios de un buen o mal comportamiento en el tiempo, relacionados con su comportamiento histórico, el comportamiento de sus relacionados, sus características particulares e información generada en las distintas etapas del ciclo de vida, como se muestra en la Tabla Nº3.

4.2. Técnicas de Data Mining Implementadas

Para efecto de la caracterización e identificación de patrones, se aplican tres técnicas de data mining: el Self- Organizing Maps (SOM), el Gas Neuronal (NG) y Árboles de Decisión. Posteriormente para la predicción, se utiliza Redes Neuronales con Backpropagation y Redes Bayesianas, las que se describen a continuación:

- Self-Organizing Maps (SOM): es uno de los modelos de redes neuronales artificiales más utilizado para el análisis y visualización de datos de alta dimensión, basado en aprendizaje competitivo no supervisado. La red consiste en un conjunto de neuronas dispuestas en una grilla de dimensión a, normalmente rectangular, cilíndrica o toroidal, que genera un espacio de salida de dimensión d, con a$\leq$ d, sobre el cual se construyen relaciones de vecindad. Durante el entrenamiento de la red, las neuronas

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

Tabla 3: Tipo de Información utilizada para construir el vector de características

generan cierta actividad ante el estímulo de los datos de entrada, lo que permite determinar qué neuronas han aprendido a representar los patrones de la entrada, los cuales pueden ser agrupados dentro de una misma categoría o cluster, basándose en una medida de distancia, normalmente Euclideana. Esta herramienta usualmente es aplicada para clusterización y segmentación, generando grupos con objetos de comportamiento similar entre sí, pero diferentes a los objetos de otro grupo.

- Gas Neuronal (NG:Neural Gas): es un algoritmo relativamente nuevo de redes neuronales no supervisada, orientada a la cuantización vectorial de estructuras arbitrarias. La mayor diferencia con el SOM es que este método no define una grilla que impone relaciones topológicas entre unidades de la red y cada neurona puede moverse libremente a través del espacio de datos. Esta libertad permite al algoritmo una mejor capacidad para aproximar la distribución de los datos en el espacio de entrada, ya que las neuronas no están obligadas a tener que mantener ciertas relaciones de vecindad, sin embargo, requiere tener algunos antecedentes respecto del número de grupos que se espera obtener.

- Árboles de Clasificación: es uno de los métodos más utilizado para realizar clasificaciones, y se destaca por su sencillez y su aplicabilidad a diversas áreas e intereses. Básicamente el algoritmo consiste en formar todos los pares posibles y combinaciones de categorías, agrupando aquellas que se comportan homogéneamente con respecto a la variable respuesta en un grupo, manteniendo separadas las categorías que se comportan de forma heterogénea. Para cada posible par, se calcula el estadístico correspondiente a su cruce con la variable dependiente (estadístico chi-cuadrado en caso de campos de destino categóricos o estadístico F para salidas continuas). Para las categorías fusionadas se procede a realizar nuevas fusiones de los valores del pronosticador, pero esta vez con una categoría menos, El proceso se acaba cuando ya no pueden realizarse más fusiones porque los estadísticos entregan resultados significativos.

- Red Neuronal de Perceptrón Multicapa (MLP): es un modelo de red neuronal artificial de varias capas utilizado para la clasificación y agrupación, basado en la funcionalidad del cerebro humano a través de un conjunto de vértices interconectados. La red debe encontrar la relación existente entre los atributos de entrada y la salida deseada para cada caso. Esto lo realiza a través de un método de aprendizaje llamado "Backpropagation" o "Retropropagación del error", que minimiza el error de predicción mediante un ajuste a los pesos de la red. Este método posee dos etapas: en la primera se calculan las salidas basado en las entradas y los pesos asignados a la red inicial, para la cual se calcula el error de la predicción y en la segunda fase, se calcula el error hacia atrás a través de la red, desde las unidades de salida hacia las unidades de entrada. De esta forma se actualizan los pesos a través de un método de descenso por gradiente. Este proceso es iterativo, por lo que tras realizar varias veces el algoritmo, la red va convergiendo hacia un estado que permita clasificar todos los patrones que minimizan el error[4].

- Redes Bayesianas: son un grafo dirigido acíclico, utilizado para predecir la probabilidad de ocurrencia de diferentes resultados, sobre la base de un conjunto de hechos. La red consta de un conjunto de nodos que representan las variables del problema y de un conjunto de arcos dirigidos que conectan los nodos e indican una relación de dependencia existente entre los atributos de los datos observados. Las redes bayesianas describen la distribución de probabilidad que gobierna un conjunto de variables, especificando suposiciones de independencia condicional junto con probabilidades condicionales. Típicamente, este problema se divide en dos

[4]Normalmente se calcula el error cuadrático medio

partes: un aprendizaje estructural, que consiste en obtener la estructura de la red, y un aprendizaje paramétrico, en el que conocida la estructura del grafo, se obtienen las probabilidades correspondientes a cada nodo. Su principal ventaja es que permite obtener la probabilidad de ocurrencia de un determinado suceso en función de un conjunto de acciones, entregando una vista clara de las relaciones mediante un gráfico de red.

4.3. Pre Procesamiento de los Datos

La preparación de los datos es una parte fundamental del proceso KDD, ya que la información puede provenir de muchas fuentes, tener errores, ambigüedades o ser redundante, consumiendo gran parte del tiempo del proyecto. Por otra parte, los datos deben ser transformados de manera apropiada para realizar el análisis.

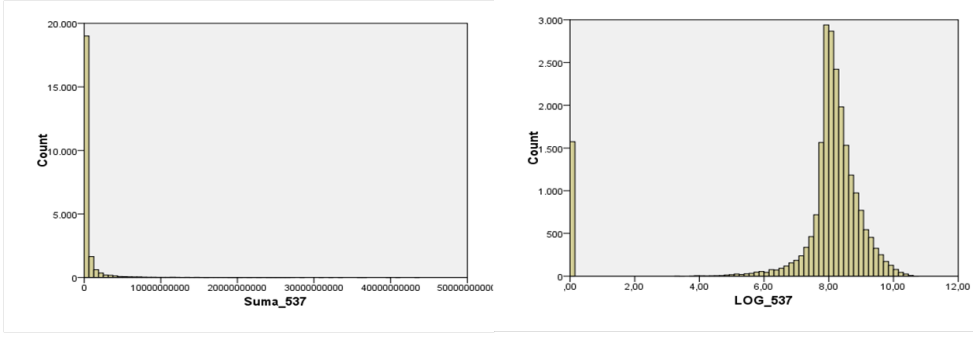
4.3.1. Limpieza

La calidad de los datos tiene una incidencia directa en los resultados, ya que si los datos no son de calidad, los resultados tampoco lo serán. Para lo anterior, se eliminan los puntos atípicos o outliers, utilizando como regla aquellos casos que superan la media más cinco veces la desviación estándar, considerando únicamente los casos con valor positivo de cada código. En la mayoría de las variables la distribución era decreciente, debido a que un gran porcentaje de contribuyentes paga montos bajos de impuestos, y sólo un pequeño grupo paga montos altos, por lo que la eliminación de datos se hizo de manera cuidadosa, considerando el juicio experto de los involucrados en el negocio, de manera de no eliminar casos que estuvieran correctos pero alejados del promedio. Lo mismo sucede con las variables de comportamiento, ya que constituyen conductas irregulares que sólo tiene un grupo pequeño de contribuyentes. Por lo tanto, al eliminar los casos con valores más altos, se elimina a aquellos contribuyentes que en general tienen un peor comportamiento, los cuales son el grupo de interés de este trabajo. Las variables de comportamiento, no tenían grandes inconsistencias debido a que fueron construidas en forma manual, sin embargo, se presentaban algunos problemas en los códigos del F29. Por ejemplo, se declaraban ventas con facturas pero no se indica una cantidad de facturas emitidas o viceversa. Dado que estos casos no eran muchos, se determina eliminarlos de la base. El mismo criterio se utilizó para el resto de los códigos de débitos y créditos.
Luego de quitar los outliers y los casos inconsistentes, el conjunto de datos final queda compuesto por 532.755 contribuyentes que son micro y pequeñas empresas, y 22.609 medianas y grandes empresas, eliminando un 4.6 % del

primer grupo y un 3.4 % del segundo.

4.3.2. Transformación y Normalización

Debido a que la declaración del pago de IVA se realiza mensualmente y la declaración de impuesto a la renta se realiza en forma anual, la primera transformación fue considerar el total anual, sumando los montos mensuales de cada código del F29 en el año para hacerlo comparable con la información de renta. Respecto de la completitud de datos nulos, la información de IVA es más completa que la de renta, debido a que los códigos del reverso del F22, sólo deben ser presentados por contribuyentes que llevan contabilidad completa. Por lo tanto, se utiliza información de débitos y créditos de IVA para completar datos de ingresos y costos del periodo, debido a la relación directa existente entre ambos.Para el resto de los campos de renta, se utiliza la mediana del código para contribuyentes del mismo tramo de ventas.Finalmente, producto de la distribución decreciente de las variables de impuesto, se aplica una transformación logarítmica para disminuir el efecto de los datos extremos como se muestra en la Figura N°2.



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

Figura 2: Ejemplo de distribución original y transformada de códigos de impuestos.

Para evitar que las variables con un mayor rango de valores le quiten importancia a otras con un rango menor, se procede a normalizar las variables de manera que sean comparables la una con la otra, utilizando la normalización “Min-Max” en el rango [0,1]. Adicionalmente, previo a la selección de las variables de utilizar en los modelos, se procede a reducir las variables de comportamiento a través del Análisis de Componentes Principales (ACP). Como resultado se generan 15 componentes principales para el grupo de las micro y pequeñas empresas, que explican un 61,3 % de la varianza de los datos. Del

mismo modo, se generan 16 componentes principales para las medianas y grandes empresas, que explican un 59,9 % de la varianza de los datos, las que se presentan en la Tabla N°4.

Micro y Pequeñas Empresas	%	Medianas y Grandes Empresas	%
(1) Nivel de facturas timbradas en los últimos años	9,7	(1) Cobertura de la empresa	9,2
(2) Delitos e irregularidades de facturas previos	7,0	(2) Fiscalizaciones previas	6,2
(3) Fiscalizaciones previas con resultado positivo	5,6	(3) Nº Actividades económicas	5,5
(4) Frecuencia de Timbraje	5,1	(4) Nivel de formalidad de la empresa y antigüedad	4,2
(5) Participación en otras empresas	4,5	(5) Clausuras y denuncios históricos	3,8
(6) Problemas de localización	4,2	(6) Verificaciones de actividad	3,4
(7) Antigüedad	3,5	(7) Giros e inconcurrencias	3,2
(8) Clausuras y denuncias históricas	3.4	(8) Representantes legales	3,2
(9) Cobertura de la empresa	3,0	(9) Delitos de los relacionados	2,9
(10) Fiscalizaciones previas con resultado negativo	2,9	(10) Irregularidades de facturas y nivel de timbraje	2,8
(11) Verificaciones de actividad	2,6	(11) Rendimiento de fiscalizaciones previas	2,8
(12) Delitos de relacionados indirectos	2,6	(12) Irregularidades recientes	2,7
(13) Irregularidades previas (pérdida facturas)	2,5	(13) Cambio de sujeto	2,6
(14) Nivel de formalidad de la empresa	2,4	(14) Antecedentes de término de giro y no ubicado	2,6
(15) Delitos de relacionados directos	2,4	(15) Antecedentes de timbraje restringido	2,5
(16) Regularización de deudas y pérdidas de rut.	2,5		

Tabla 4: Conceptos asociados a cada Componente Principal y el porcentaje de la varianza explicada

Dado que nuestro interés era generar variables de comportamiento relacionadas al uso y venta de facturas falsas y no a otros comportamientos, se seleccionan sólo aquellas variables que tienen una correlación mediana-alta con

la variable de uso de facturas falsas en el año 2006, eliminando aquellas que tienen más de un 10 % de probabilidad que el coeficiente de pearson sea cero, exceptuando algunos códigos de interés como el total de débitos, total de créditos y pago de IVA, entre otros. Igualmente, se descartan aquellas variables que tienen un gran porcentaje de valores nulos. De esta forma se seleccionan 42 variables en el segmento micro y pequeñas y 36 variables medianas y grandes para el análisis. En el primer grupo, un 35 % de las variables corresponde a códigos de la declaración de IVA, un 35 % a códigos relacionados con renta y un 30 % a variables relacionadas al comportamiento. En el segundo grupo en cambio estos porcentajes varían a un 31 %, 38 % y 31 % respectivamente, con mayor preponderancia de variables relacionadas a la renta.

4.4. Modelamiento

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

4.4.1. Caracterizando al Universo de Empresas

Inicialmente se aplica el método SOM al universo de contribuyentes, para identificar clusters o grupos de empresas de comportamiento similar. La hipótesis de trabajo suponía que al considerar sólo las variables de comportamiento relacionadas al uso de facturas falsas combinadas con variables de impuestos, era posible detectar grupos de contribuyentes que tienen un buen o mal comportamiento tributario y conocer cómo realizaban su pago de impuesto. Para ello se utiliza el paquete "som" de R, considerando una topología de red rectangular, con 3 neuronas de entrada y 24x24 neuronas de salida en el grupo de las micro y pequeñas empresas y 36x36 neuronas de salida en el grupo de las medianas y grandes empresas, con un número máximo de 100 iteraciones. En el primer grupo se considera una muestra de 100.000 empresas, debido a restricciones computacionales. En el caso de las micro y pequeñas empresas se generan 5 clusters, mientras que en las medianas y grandes se identifican 6 clusters, como se muestra en la Figura N°3.

Los clusters obtenidos en el primer grupo se diferencian principalmente

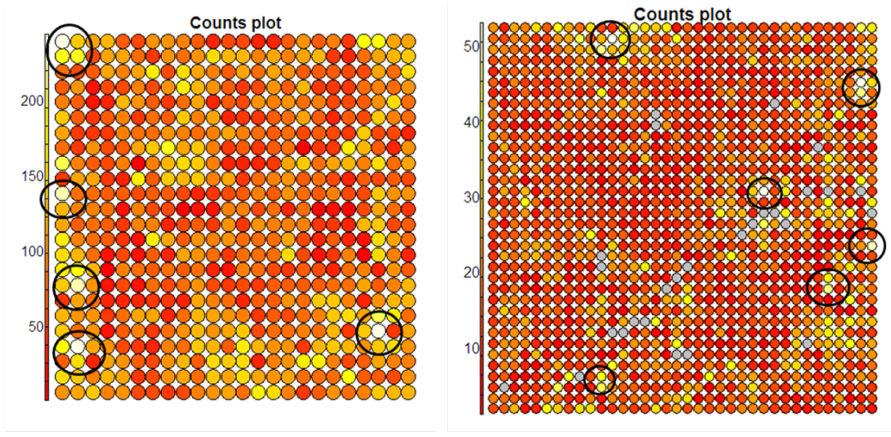


Figura 3: Mapa resultante aplicación SOM en MI y PE (izquierda) y ME y GR (derecha)

por la utilización de boletas y/o facturas, el nivel de pago de IVA, el nivel de costos declarados, el nivel de formalidad de la empresa y participación en otras empresas y algunos problemas de localización. Mientras que en las medianas y grandes, se diferencian por la utilización de boletas y/o facturas, niveles de uso de remanentes, notas de crédito y facturas de activo fijo, pasivos y activos, así como los resultados de fiscalizaciones previas y el nivel de formalidad de la empresa, como se indica en las Tablas N°5 y N°6.

Si bien se encontraron algunos patrones de comportamiento con éste método, estos no estaban relacionados específicamente a la utilización de facturas falsas, ya que los casos conocidos de fraude y no fraude se encontraban distribuidos en todo el mapa sin un patrón definido.

Posteriormente se aplica el Gas Neuronal, considerando el mismo número de clusters que el Mapa de Kohonen, utilizando el paquete "Clust" de R, el cual genera una matriz con las características de los centroides de cada variable y un vector de clasificación que señala el grupo al que pertenece cada contribuyente. En este caso, los grupos generados también se encuentran influenciados por el pago de impuestos, aunque con mayores diferencias en términos de comportamiento. Esto, permite diferenciar cuáles grupos tienen mejor y peor comportamiento, y relacionarlo con su pago de impuesto, aunque no necesariamente los casos de facturas falsas se encontraban en un mismo grupo.

De acuerdo a esto, se identificaron los siguientes patrones asociados a un mal y buen comportamiento, considerando los puntos comunes obtenidos en ambos métodos.

Cluster 1	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel alto de pago de IVA y costos altos. Con algunos problemas de localización, mayor nivel de participación en otras empresas y formalización de la contabilidad.
Cluster 2	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel intermedio-alto de pago de IVA y costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad y participación en otras empresas.
Cluster 3	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad.
Cluster 4	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara niveles altos de costos y problemas de localización.
Cluster 5	Tiene niveles altos de débitos con boletas, nivel intermedio de uso de facturas y pago de IVA, y costos altos. Relativamente joven con algunos problemas de localización y nivel intermedio de formalización.

Tabla 5: Clusters resultantes aplicación SOM en MI y PE

4.4.2. Caracterizando a los Casos con Fraude y Sin Fraude

Si bien las dos técnicas anteriores implementadas permiten caracterizar al universo de contribuyentes e identificar algunos patrones diferenciadores, considerando aquellas variables más relacionadas con el uso de facturas falsas. Éstas tienden a darle mayor importancia al pago de impuestos que a las variables de comportamiento, creando grupos que se diferencian en el tipo de operación (ventas con facturas y/o boletas), el nivel de actividad (alto-bajo nivel de ventas, costos) y pago de impuestos (alto-bajo), debido a la mayor variabilidad de estas variables en comparación a las de comportamiento.
Por otra parte, al analizar la distribución de cada variable, se observa que los casos con fraude normalmente se encuentran en los casos extremos de cada una de ellas. Por este motivo se determina aplicar árboles de decisión al conjunto de datos con resultado de auditoría conocido, ya que permite identificar el punto de corte de cada variable frente al cual se produce un cambio de comportamiento, considerar casos extremos y generar reglas que pueden ser validadas e implementadas.

Cluster 1	No utiliza boletas. Tiene nivel intermedio de remanentes y costos bajos. Presenta monto alto de créditos por factura de activo. Con un nivel alto de formalidad.
Cluster 2	No utiliza boletas. Tiene nivel intermedio de remanentes y pocas fiscalizaciones previas. Nivel intermedio de formalidad.
Cluster 3	No utiliza boletas. Tiene nivel alto de remanentes, pasivos y activos. Tiene bajo porcentaje de crédito asociado a facturas. Nivel alto de formalidad.
Cluster 4	Nivel alto de uso de boletas. Tiene nivel intermedio de remanentes y de notas de crédito. Nivel alto de formalidad.
Cluster 5	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel bajo de formalidad. Pocas fiscalizaciones previas.
Cluster 6	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel alto de formalidad. Tiene nivel intermedio de uso de notas de crédito.

Tabla 6: Clusters resultantes aplicación SOM en ME y GR

El tipo de árbol utilizado es el CHAID (Chi-square automatic interaction detection), el cual permite clasificaciones no binarias y generar un número distinto de ramas a partir de un nodo considerando tanto variables continuas como categóricas. Un punto a considerar de éste método es que se requiere disponer de tamaños de muestra significativos, ya que al dividirse en múltiplos grupos, cabe el riesgo de encontrar grupos vacios o poco representativos si no se dispone de suficientes casos en cada combinación de categorías. Adicionalmente se evalúa el método del CHAID exhaustivo, el cual es una modificación del algoritmo tradicional, que busca hacer frente algunas debilidades del CHAID tradicional.
Se realizan varios experimentos que consideran distinto número de variables y tipos de salidas (categóricas y numéricas) para identificar si se producen diferencias entre una formato de salida y otro.

Finalmente esta técnica resultó ser altamente efectiva para encontrar patrones diferenciadores entre fraude y no fraude, ya que los nodos finales estaban compuestos mayoritariamente por casos de un solo tipo, o en su defecto combinado con casos con valor de salida “1”, los cuales se aproximan más al comportamiento de los casos con fraude “2”.
Como se indica en la Tabla N°8 el número de nodos finales fue similar en ambos experimentos realizados en cada grupo, obteniéndose 33 y 36 nodos en el segmento de las micro y pequeñas empresas y 22 y 24 nodos en el segmento

Buen Comportamiento MI y PE	Declaran montos más altos de débitos (emite más boletas) y pagan más IVA. Declaran bajos niveles de créditos y de remanentes, mayor relación ingresos/costos y costos/activos. Tienen mayor cantidad de facturas timbradas y frecuencia de timbraje, menor cantidad de delitos e irregularidades previas y delitos de los relacionados indirectos. Registran pocas verificaciones de actividad.
Buen Comportamiento ME y GR	Declaran mayor nivel de costos y gastos y mayor nivel de activos y pasivos. Tienen montos más altos de créditos y remanentes. Registran un mayor nivel de formalización de su contabilidad y mayor cobertura, mayor número de representantes legales y cantidad de fiscalizaciones previas.
Mal Comportamiento MI y PE	Declaran niveles bajos de pago de IVA y una relación débito/crédito baja. Registran una mayor cantidad de créditos y acumulación de remanentes. Tienen un nivel más bajo del ratio ingresos/activo, mayor cantidad de fiscalizaciones previas con resultado positivo y un menor nivel de facturas timbradas. Registran varias verificaciones de actividad.
Mal Comportamiento ME y GR	Declaran mayores costos y remuneraciones respecto de sus activos, menor nivel de pasivos y mayor cantidad de porcentaje de débitos con boleta, aunque con un número menor de boletas. Registran mayor cantidad de anotaciones de timbraje restringido, términos de giro previos y antecedentes de no ubicado. Tienen mayor cantidad de denuncias y clausuras históricas, menor cantidad de fiscalizaciones previas y cobertura, así como un menor nivel de formalización de su contabilidad y antigüedad.

Tabla 7: Caracterización de grupos con buen y mal comportamiento

de las medianas y grandes.

A modo de ejemplo se presenta un extracto del resultado de la aplicación del experimento N°1, en el cual se identifican patrones bastante claros asociados a fraude y no fraude, debido a la preponderancia de nodos finales con casos de fraude y no fraude. Como se indica en la Figura N°4, los factores que tienen mayor incidencia fueron el resultado de las fiscalizaciones previas (ACP10) y

Exp. Nº	Segmento	Método	Nº Variables	Tipo de salida	Nº Niveles	Nº Nodos finales
1	Micro y Peq.	Árbol CHAID	30	Categórica	6	33
2	Micro y Peq.	Árbol CHAID	30	Numérica	5	36
3	Med. y Grandes	Árbol CHAID	38	Numérica	4	22
4	Med. y Grandes	Árbol CHAID	24	Numérica	6	24

Tabla 8: Caracterización de grupos con buen y mal comportamiento, según el gas neuronal

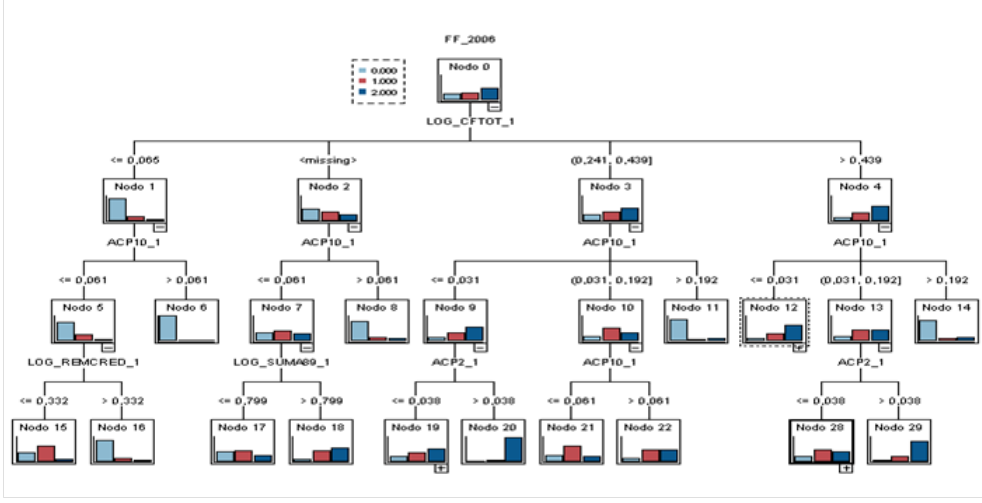


Figura 4: Clasificación resultante de la aplicación del árbol CHAID – Experimento N°1

el porcentaje de las compras sustentado en facturas (CFTOT). Esto indica que aquellos que han sido más veces fiscalizados en el pasado y no se les ha encontrado nada y sus compras no se basan principalmente en facturas, tienen menos probabilidad de utilizar facturas falsas, que aquellos que mayoritariamente registran compras con facturas y tienen fiscalizaciones productivas en el pasado. De hecho, estas dos variables por sí solas, determinan varios nodos finales con preponderancia de casos sin fraude.

Adicionalmente, la variable que indica una mayor preponderancia de delitos e irregularidades asociadas a facturas históricas combinado con la frecuencia de timbraje, genera nodos finales con preponderancia de casos con facturas falsas. Particularmente el nodo 12 que contiene casi la mitad de los casos (46 %) se descompone en varias ramas en función del valor que toma el crédito promedio por factura emitida (mientras mayor sea este indicador, mayor posibilidad

hay de que cometa fraude). De igual manera, la preponderancia de casos con fraude en cada rama depende del número de facturas emitidas, el IVA pagado, el total de débitos por boletas, la relación entre costos y activos y el nivel de participación en otras empresas.
Cómo se señala en la Figura N°5, las variables más relevantes para distinguir casos de fraude en las micro y pequeñas empresas fueron el resultado de las fiscalizaciones previas, el Total de IVA determinado, el porcentaje de crédito sustentado en facturas, la relación entre remanentes y créditos, el total de débitos por boletas y la relación entre facturas timbradas y emitidas. Mientras que en las medianas y grandes las variables corresponden a total de remanente, porcentaje de crédito respaldado en facturas, el número de representantes legales, nivel de formalización de la contabilidad, la relación entre remuneraciones y activos, entre otros.

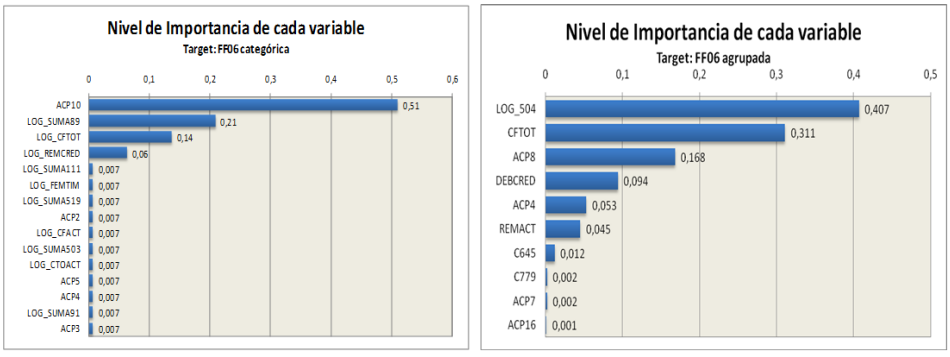


Figura 5: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

Considerando los patrones y reglas que se repiten en cada rama del árbol para diferenciar entre casos de fraude y no fraude, en la Tabla N°9 se presenta un extracto de los comportamientos asociados a cada uno de ellos en cada segmento, que resume las variables principales consideradas y las relaciones que generan nodos con y sin utilización de facturas falsas en el año de estudio.

4.4.3. Predicción del Fraude

Para la predicción, se aplicaron redes neuronales artificiales y redes bayesianas. En ambos procesos para evitar el sobreajuste de la red, los datos se dividen en dos conjuntos: uno de entrenamiento y uno de testeo, utilizando la regla 70/30. Por otra parte, ambos métodos fueron implementados utilizando la herramienta tecnológica clementine del SPSS.
Uno de las complejidades de las redes neuronales, es determinar el número de capas y nodos ocultos, así como la cantidad de épocas o iteraciones. Para

Comportamiento Asociado a Fraude	MI y PE Registran menor porcentaje de créditos asociados a facturas y más fiscalizaciones previas con resultado negativo. Emiten menor cantidad de facturas emitidas y un valor más bajo del indicador facturas emitidas/facturas timbradas. Registran un mayor monto del indicador remanentes/crédito promedio.
Comportamiento Asociado a Fraude	ME y GR Registran menor porcentaje de crédito asociado a facturas. Declaran un monto mayor de remanente acumulado del periodo anterior. Tienen valores bajos del indicador costos/activos. Registran menor cantidad de irregularidades previas asociadas a facturas y de timbraje.
Comportamiento Asociado a No Fraude	MI y PE Tienen mayor porcentaje de créditos asociados a facturas y débitos con boletas. Tienen valor alto del indicador costos/activos. Emiten una mayor cantidad de facturas y tienen valor alto del indicador facturas emitidas/facturas timbradas. Tienen montos altos de IVA determinado. Registran menos fiscalizaciones previos con resultado negativo y más fiscalizaciones previas con resultado positivo. Tienen más antecedentes de delitos e irregularidades históricas asociadas a facturas y mayor frecuencia de timbraje en los últimos dos años.
Comportamiento Asociado a No fraude	ME y GR Tienen mayor porcentaje de créditos asociados a facturas. Declaran monto menor de remanente acumulado en el mes anterior y tienen valores altos del indicador costos/activos. Tienen mayor nivel de informalidad en su contabilidad y son de menor antigüedad. Registran mayor número de actividades económicas activas e irregularidades previas asociadas a facturas y timbraje. Tienen mayor cantidad de giros e inconcurrencias a notificaciones.

Tabla 9: Caracterización de casos con y sin fraude según árbol CHAID

determinar tales parámetros se consideraron distintos números de ciclos y nodos en las capas ocultas, de manera de establecer a través de ensayo y error los valores más adecuados. Para las iteraciones se utilizaron los valores: 1.000, 5.000, 10.000 y 20.000. En el caso de los nodos se utiliza el número que el software calcula por defecto en función de los datos del modelo y otra correspondiente a la mitad del número de nodos de entrada, es decir, 3 y 20 nodos respectivamente.

En el caso de las redes bayesianas se evalúan dos métodos para construir la red: el algoritmo TAN y el algoritmo de estimación de Markov-Blanket disponibles

en el software clementine del SPSS. Adicionalmente se utiliza un preprocesamiento previo de las variables para identificar cuáles son las variables más relevantes y mejorar el tiempo de procesamiento y rendimiento del algoritmo. De igual forma se utiliza un test de independencia de máxima verosimilitud y chi-cuadrado para el aprendizaje paramétrico.
Los resultados de los experimentos se presentan en la Tabla N°10, el que contiene los siguientes indicadores obtenidos en el grupo de testeo: (1) Sensibilidad: Indica la proporción de casos con fraude clasificados en forma correcta, (2) Especificidad: Indica la proporción de casos sin fraude en los que la clasificación fue correcta, (3) Concordancia: Indica la proporción de casos con y sin fraude en los que la clasificación fue correcta y (4) Tasa de error: Indica la proporción de casos con y sin fraude que fueron asignados a una clase incorrecta.

Exp. N°	Segmento	Método	Sensitivi -dad (1)	Especifici -dad (2)	Concordan -cia (3)	Tasa Error (4)
1	Micro y Peq.	Red Neuronal	92.6 %	72.9 %	87.2 %	12.8 %
2	Micro y Peq.	Red Bayesiana	82.3 %	64.1 %	77.9 %	22.1 %
3	Med. y Grandes	Red Neuronal	84.3 %	52.2 %	65.8 %	34.2 %
4	Med. y Grandes	Red Bayesiana	73.3 %	66.7 %	70.3 %	29.7 %

Tabla 10: Experimentos realizados para predecir los casos con fraude por facturas falsas

En ambos segmentos, los mejores resultados de predicción de casos con facturas falsas se obtuvieron con la técnica de red neuronal. En el grupo de las micro y pequeñas empresas, el experimento 1 arrojó que en un 92,6 % los casos con fraude fueron asignados a la clase correcta, mientras que en el grupo de las medianas y grandes empresas la proporción de casos con fraude correctamente asignada fue de 84.3 %. Por otra parte, el poder de generalización del modelo fue bastante bueno, ya que los resultados del testeo fueron similares a los obtenidos en el entrenamiento de la red, cuya predicción fue casos con y sin fraude fue de 93.7 % y 89.6 % respectivamente.

La red neuronal generada para las micro y pequeñas empresas, indica una preponderancia de variables asociadas al pago de IVA y al comportamiento, y en menor medida, a variables relacionas a la renta. Las más relevantes corresponden a los antecedentes obtenidos de la verificación de actividades, la

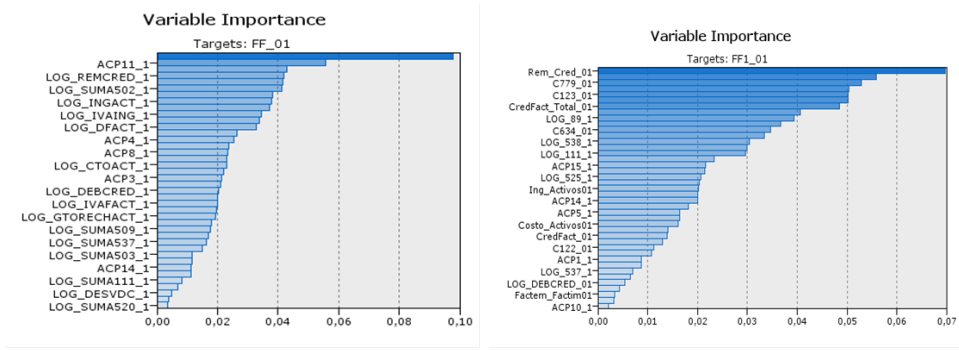


Figura 6: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

relación entre remanentes y créditos, el total de débitos por facturas emitidas, la relación entre ingresos del giro y los activos y la relación entre el IVA pagado y el Ingreso declarado. En el caso de las medianas y grandes empresas, las variables más relevantes corresponden a la relación entre remanentes y créditos, las cuentas por pagar a empresas relacionadas, el total de pasivos, la proporción de créditos asociado a facturas y el IVA determinado en el periodo.

5. Conclusión y Trabajo Futuro

La utilización y venta de facturas falsas tiene un impacto significativo en la recaudación que percibe el Estado para financiar sus proyectos. La detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera además un importante costo administrativo para el SII, lo que da cuenta de la relevancia que tiene focalizar los esfuerzos en la detección de casos de evasión y fraude fiscal.

Los métodos de clusterización y clasificación utilizados para caracterizar a los contribuyentes que tienen buen o mal comportamiento tributario asociado a la utilización de facturas falsas, demuestran que es posible identificar algunas características diferenciadoras entre un grupo y otro, las cuales hacen sentido con lo que sucede en la realidad. Particularmente el método de gas neuronal arrojó que era posible determinar algunas variables relevantes para diferenciar entre un buen o mal comportamiento, los que no necesariamente se asocian a la utilización y venta de facturas falsas. El método de kohonen, en cambio, no permitió obtener patrones de comportamiento relacionados con la utilización de facturas falsas, sino más bien, se detectaron clusters en relación al pago

de impuestos, en la que las variables con mayor cantidad de ceros y varianza resultaron ser las que más impacto tuvieron en la conformación de los grupos. Los árboles de decisión aplicados a los casos en el que el resultado de fraude y no fraude era conocido resultó ser una buena técnica para detectar variables que permiten distinguir entre casos de fraude y no fraude. Esto debido que al analizar la distribución de las variables en cada grupo, se observa que los casos con fraude tendían a tomar valores más extremos de las variables, por lo que era posible distinguir rangos a partir de los cuales, existe una probabilidad de tener o no tener fraude. Por otro lado, los resultados obtenidos fueron coherentes con lo observado en la realidad, de acuerdo a la vista experta.

Es así como en el caso de las micro y pequeñas empresas las variables que permitían distinguir entre fraude y no fraude se relacionaban principalmente con el porcentaje de créditos generado por facturas respecto del crédito total y las fiscalizaciones previas con resultado negativo. En la medida que el contribuyente fue fiscalizado más veces en el pasado y no se encontró nada, es más probable que no tenga fraude en el futuro. Por otro lado, mientras su crédito esté más asociado a otros ítemes distintos a las facturas (activo fijo u otros), es menos probable que utilice facturas para respaldar sus créditos. Otras variables relevantes fueron la cantidad de facturas emitidas en el año y su relación con las facturas timbradas en los últimos dos años, el monto de IVA total declarado, la relación entre remanentes y créditos promedio, las fiscalizaciones previas con resultado positivo y los delitos e irregularidades históricos asociadas a facturas. Mientras que en las medianas y grandes empresas, las variables más relevantes fueron la cantidad de remanente acumulado en los periodos anteriores, el porcentaje de crédito asociado a facturas, la relación entre costos y activos, el nivel de informalidad en su contabilidad y la antigüedad, así como la cantidad de irregularidades previas asociadas a facturas y la cantidad de giros e inconsciencias históricas.

En relación a los modelos predictivos, los que tuvieron mejor desempeño fueron los modelos de red neuronal de perceptrón multicapa, que para efectos del estudio contaban con una capa de entrada que contenía las variables explicativas, una capa intermedia de procesamiento y una capa de salida. En el caso de las micro y pequeñas empresas el porcentaje de casos con fraude asignado correctamente fue un 92 %, mientras que en las medianas y grandes empresas, este porcentaje fue de 84 %. Considerando que en la práctica sólo es posible fiscalizar a un grupo más bien reducido de empresas en un año, se recomienda realizar una combinación de los resultados obtenidos con las redes neuronales y las redes bayesianas, de manera de seleccionar para fiscalización a aquellos que aparecen catalogados como fraude en la red neuronal y que tienen las probabilidades más altas de cometer fraude según la red bayesiana.

En términos de recaudación, la predicción de un caso de fraude en una micro y pequeña empresa aporta un beneficio neto de $ 86.282, mientras que para una mediana y gran empresa, esta cifra aumenta a un $3.424.083, lo que permitiría reducir la evasión por concepto de IVA de manera significativa, si consideramos el total de casos auditados en un año.
De acuerdo a estudios que ha realizado el SII, se estima que aproximadamente un 20 % de los contribuyentes utilizan facturas para evadir impuesto. No existe información desagregada por tipo de contribuyente, pero suponiendo que este porcentaje se repite en cada segmento y considerando los porcentajes de clasificación de casos con fraude y no fraude de los modelos de red neuronal, se tiene que el universo de potenciales usuarios de facturas es de 116.000 micro y pequeñas empresas y 4.768 medianas y grandes empresas, que generan un ingreso por fiscalización de $21.344 millones de pesos y $80.102 millones de pesos respectivamente, generando un potencial de recaudación de $101.446 millones de pesos.
Finalmente, para probar la capacidad predictiva real del modelo desarrollado y siendo concordante con el punto anterior, resulta vital su aplicación en actividades que permitan determinar en terreno el nivel de acierto en la clasificación de los contribuyentes seleccionados en la muestra, para lo cual se recomienda la implementación de un programa piloto que estará dirigido a los dos segmentos económicos estudiados, que será concluyente en términos de la efectividad real del modelo.

Referencias

[1] Arnaiz, T., García, J. A. y López, J.M. Los Planes Integrales para la Prevención y Corrección del Fraude Fiscal. *Banco Interamericano de Desarrollo (BID)* . 2006.

[2] Bolton, R. y Hand, D. Statistical Fraud Detection: A Review. *Statistical Science*, Vol. 17- N°3. 2002.

[3] Centro Interamericano de Administraciones Tributarias. Métodos de Selección de Declaraciones sujetas al Control Concurrente ocupando Herramientas de Minería de Datos. *Programa Regional (TC-00-05-00-8-RG). Superintendencia Nacional de Administración Tributaria*, Perú. 2004.

[4] Clifton, P. y Chun, W. Investigative Data Mining in Fraud Detection. *School of Business Systems, Monash University.*. 2003.

[5] Davia, H.R., Coggins, J.W. y Kastantin, J. Accountant's Guide to Fraud Detection and Control (2da edición). 2000.

[6] Denny, Williams, G., Christe, P. (2007). Exploratory Multilevel Hot Spot Analysis: Australian Taxation Office Case Study. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. *Conferences in Research and Practice in Information Technology (CRPIT)*, Vol. 70. 2007.

[7] Digimpietri, L., Trevisan, N., Meira, L., Jambeiro, J., Ferreira, C. y Kondo, A. Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System. *Proceedings of the 9th Annual International Digital Government Research Conference.* 2008.

[8] Ernst&Young *9th Global Fraud Survey 2006: Fraud Risk in emerging markets.* Junio. 2006.

[9] Fayyad, U., Piatestky-Shapiro, G., Smyth, P. From data mining to knowledge discovery in databases. *American association for artificial intelligence* 0738-4602, 37-54. 1996

[10] Government Accountability Office (GAO), United States. Data Mining: Agencies have taken key steps to protect privacy in selected efforts, but significant Compliance Issues Remain. Mayo. 2004.

[11] Government Accountability Office (GAO), United States. Lessons Learned from Other Countries on Compliance Risks, Administrative Costs, Compliance Burden and Transition. *Report to Congressional Requesters*, Abril. 2008.

[12] Harrison, G. y Krelove, R. (2005). VAT Refunds: A Review of Country Experience. *International Monetary Fund (IMF) Working Paper.* Noviembre. 2005.

[13] Luckeheide, S. Segmentación de los Contribuyentes que declaran IVA aplicando herramientas de clustering. *Revista de Ingeniería en Sistemas. Volumen XXI.* 2007.

[14] Munoz, D.J. Proceso de Reconocimiento de Objetos asistido por computador, aplicando Gases Neuronales y técnicas de Minería de Datos. *Scientia et Technica*- Año XII, No 30, Mayo. 2006.

[15] Myatt Glenn, J. Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining. *Wiley Interscience.* 2007.

[16] OECD. Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Serie. *General Administrative Principles - GAP004 Compliance Measurement*- Junio. 1999.

[17] OECD. Compliance Risk Management, Use of Random Audit Programs. Forum on Tax Administration Compliance Subgroup. *Centre for Tax Policy and Administration*. Septiembre. 2004.

[18] OECD. Compliance Risk Management, Audit Case Selection Systems. Forum on Tax Administration Compliance Subgroup. *Centre for Tax Policy and Administration*. Octubre. 2004.

[19] Servicio de Impuestos Internos. Información de Cuenta Pública 2010. `htt://www.sii.cl/cuenta_publica/`. 2011.

[20] Superintendencia Nacional de Administración Tributaria. La Gestión de la Sunat en los últimos cinco años: Principales Avances y Desafíos. 2006.

[21] Tanzyi, V. y Shome, P. (1993). Tax Evasion: Causes, Estimation Methods, and Penalties a Focus on Latin America. *Documento elaborado para el Proyecto Regional de Política Fiscal CEPAL/PNUD*. 1993.

[22] Velasco, D. Redes Bayesianas. *Inteligencia Artificial II*. 2007

[23] Velázquez, J. y Palade, V. Adaptative Web Sites: A Knowledge Extraction from Web Data Approach". *Frontiers in Artificial Intelligence and Applications*, Volumen 170. 2008.