Universidad de Buenos Aires

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

---

Tesis de Maestría

Análisis de riesgos de evasión de impuestos mediante la aplicación de técnicas de Minería de Datos

María Eugenia Andrade Trujillo

Director: Waldo Hasperué

Diciembre 2019

Agradecimientos

A mi hija Isabela y a Santiago, por su amor y apoyo incondicional para culminar el desarrollo de esta tesis.

A Waldo, por su guía, valiosas sugerencias y comentarios.

Índice

Contenido

Listado de Ilustraciones

Listado de Tablas

RESUMEN

Uno de los principales ingresos de una nación son los ingresos tributarios que son recaudados por las Administraciones Tributarias a nivel mundial; siendo la problemática principal en este ámbito, los diferentes mecanismos de evasión y elusión de impuestos que se emplean para reducir el pago de impuestos.

En este sentido, se plantea como objetivo general de esta investigación el uso de técnicas de Minería de Datos para explotar masivamente la información y extraer conocimiento sobre perfiles de riesgo y anomalías que presentan los contribuyentes de los sectores económicos Comercio, Manufactura y Servicios Profesionales, por ser las actividades que concentran mayor número de contribuyentes y generan el mayor aporte a la recaudación del impuesto al valor agregado en Ecuador; de tal forma que mediante la generación de modelos de Minería de Datos se puedan detectar potenciales riesgos de evasión, siendo esta una herramienta que hoy en día se está aplicando de manera generalizada para incrementar la eficiencia de la gestión en el control y auditoría de contribuyentes; lo que a su vez incide en el mejoramiento del cumplimiento tributario.

Para la consecución del objetivo propuesto en esta tesis se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). El modelo de referencia está conformado por 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, implementación.

En la fase de modelado, se realizó un análisis de segmentación mediante el empleo de mapas auto-organizados de Kohonen (SOM por sus siglas en inglés), que es una de las técnicas no supervisadas para detección de anomalías. Específicamente, se aplicaron los algoritmos multi-SOM y SOM e, índices de validación interna para realizar la evaluación del número óptimo de grupos.

A partir de estos modelos, se obtuvo la segmentación de los datos de cada uno de los sectores económicos; donde hay un patrón común, en el sentido en que hay grupos minoritarios, que presentan características para ser catalogados como potenciales contribuyentes riesgosos; como el registro de mayores variaciones en términos de ventas y otras variables del negocio, en relación al promedio del sector.

En este estudio, se obtuvieron óptimos resultados a partir del uso de los mapas autoorganizados de Kohonen, dado la naturaleza de los datos de la Administración Tributaria, en términos de volumen y valores perdidos que son características propias del negocio. Adicionalmente, con los resultados obtenidos se evidenció que, mediante el uso de estas herramientas, efectivamente, se pueden detectar anomalías; lo cual es un aporte primordial para un control tributario más efectivo.

CAPÍTULO I

1. EVASIÓN DE IMPUESTOS

Uno de los problemas que adolecen las Administraciones Tributarias a nivel mundial es la evasión de impuestos que se manifiesta de diversas formas como la omisión de ingresos, la inclusión de costos y gastos inexistentes, la no presentación de declaraciones tributarias, la generación de errores aritméticos voluntarios, la corrección sucesiva de declaraciones; todo esto con el fin de disminuir los impuestos a pagar.

Bajo este contexto, para las Administraciones es una prioridad combatir la evasión, ya que ésta trae consigo problemas subyacentes [Ameur&Tkiouat, 2012] como son:

- La pérdida de recursos fiscales, al disminuir la recaudación de impuestos, particularmente del impuesto al valor agregado (IVA) y del impuesto a la renta, por el subregistro de ventas. Esto se traduce en un costo social, ya que el Estado deja de percibir los recursos necesarios para destinar a la inversión pública y gasto social.
- Costos en los que debe incurrir la Administración y asignación de recursos humanos para la detección de evasores que, sin embargo, resultan limitados para controlar y auditar al universo de contribuyentes.
- La injusticia e inequidad horizontal[1] entre los contribuyentes que cumplen y quienes no lo hacen.
- El incentivo de la economía informal o subterránea.
- El fomento de la competencia desleal entre empresas y, por lo tanto, el desaliento de la inversión.

[1]El principio de equidad horizontal hace referencia al tratamiento igualitario de los individuos que se encuentran en circunstancias similares.

- El bloqueo de las acciones del gobierno.
    - El fomento del incumplimiento tributario, de quienes cumplen y al percibir que la evasión es generalizada, se ven menos comprometidos; lo cual provoca la auto-reproducción de la evasión.

Las Administraciones Tributarias han adoptado varias medidas para reducir la evasión, que van desde el fomento de la cultura tributaria, fortalecimiento institucional, reformas legales, hasta el uso intensivo de herramientas tecnológicas para consolidar la información proveniente de fuentes propias (declaraciones presentadas por el contribuyente) y, de terceros (información comercial, financiera, laboral, entre otros); que sirvan, entre otros, para la construcción de modelos para el análisis y detección de riesgos de evasión de impuestos.

Para la Administración Tributaria del Ecuador, el panorama no es distinto; el combate a la evasión de impuestos es una de las tareas prioritarias, en cumplimiento de los principios constitucionales que rigen el sistema tributario como son la eficiencia, equidad, transparencia y suficiencia recaudatoria. En este contexto, la Administración Tributaria elaboró el Plan de Control y Lucha contra el Fraude Fiscal (Servicio de Rentas Internas, 2018); cuyas acciones principales son:

1. El fortalecimiento del Modelo de Gestión de Riesgos Tributarios, que contempla las fases de identificación, priorización, tratamiento y evaluación sistemática de riesgos tributarios.
2. La innovación a través implementación de la Plataforma de Análisis del Comportamiento Tributario, dentro de la cual se consideran los modelos predictivos, desarrollados con la aplicación de técnicas de minería de datos e inteligencia de información; con el fin de que las acciones del SRI sean más eficientes.
3. La transparencia mediante la publicación del resultado de acciones de control, acciones judiciales y de cobro; de información tributaria de grandes

contribuyentes, contribuyentes especiales y; el desarrollo de estudios de evasión tributaria.

Las fases del Modelo de Gestión de Riesgos Tributarios, se describen a continuación.

1.1 Modelo de Gestión de Riesgos Tributarios

El modelo fundamentado en la teoría de Matthijs H. Jacob Alink comprende 6 fases de la administración de riesgos: la identificación, clasificación y valoración de las posibles formas de evasión y elusión tributaria; la reducción, la detección, la selección, la cobertura de riesgos tributarios y, la evaluación de los mismos (Servicio de Rentas Internas, 2016) (Ilustración 2.1).

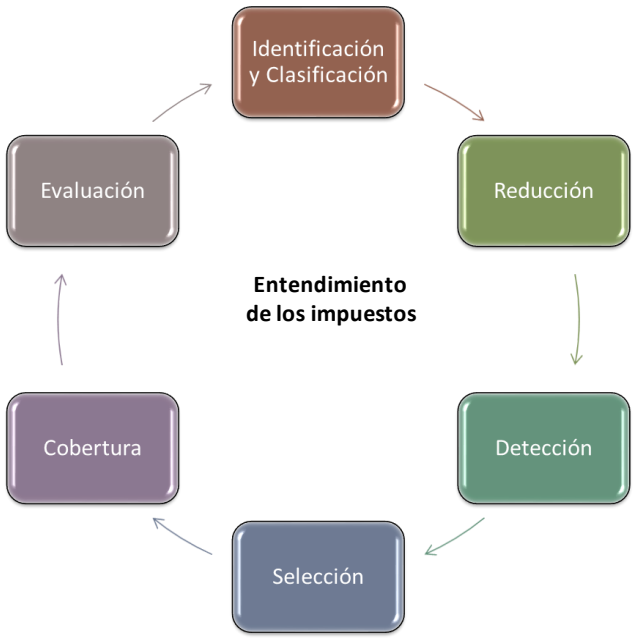

Ilustración 2.1. Modelo de Gestión Integral de Riesgos Tributarios.
Fuente: Plan Estratégico Institucional 2016-2019.

En cada una de estas fases se realizan las siguientes actividades:

1.1.1 Identificación y Clasificación de Riesgos Tributarios

En esta etapa se recopila información de eventos relacionados con posibles riesgos de incumplimiento tributario; para la confirmación de los mismos, registro y clasificación de los riesgos tributarios identificados.

La identificación de riesgos tributarios puede efectuarse mediante diversos análisis sectoriales, de estructuras de negocio de grupos económicos, de tramas de planificación tributaria, de análisis previos de riesgos tributarios específicos, investigación de fraudes tributarios, diferencias identificadas entre la normativa contable vigente y la normativa tributaria vigente y/o de la información declarada por los mismos contribuyentes o por terceros.

1.1.2 Reducción de Riesgos Tributarios

En esta etapa se realiza principalmente el análisis de las causas de los riesgos de incumplimiento tributario, la definición del alcance de las medidas y determinación de las estrategias para la mitigación de los riesgos tributarios. Las estrategias de reducción de riesgos tributarios comprenden, la facilitación del cumplimiento de las obligaciones tributarias, la generación de estrategias persuasivas, la generación de cambios normativos, aplicación de las sanciones respectivas, entre otros.

1.1.3 Detección de Riesgos Tributarios

El principal insumo para la detección, es la información proveniente del propio contribuyente y de terceros (declaraciones de impuestos, anexos de información, comprobantes electrónicos, otras fuentes de información que se obtienen

mediante convenios interinstitucionales). En esta etapa se construyen los modelos de Minería de Datos para la debida detección de riesgos tributarios, para la obtención de grupos de sujetos pasivos o contribuyentes para su respectivo tratamiento.

El objetivo principal de esta etapa es determinar si se ha producido un riesgo real o pronosticar la ocurrencia de un riesgo potencial.

1.1.4 Selección de Riesgos Tributarios

Esta etapa asegura la eficacia del sistema de control fiscal, ya que optimiza la utilización de los recursos que dispone la Administración Tributaria, con el propósito de controlar el incumplimiento, combatir la evasión e incrementar la percepción de riesgo tributario en los contribuyentes; ya sea mediante auditorías, comunicaciones o, acercamientos con el contribuyente.

En esta etapa se analizan los resultados obtenidos de la detección, para identificar segmentos de contribuyentes con mayor concentración de riesgos tributarios, y sobre los mismos definir criterios técnicos de selección, así como establecer las estrategias de tratamiento que correspondan. La selección se puede apoyar en matrices o modelos de riesgo, o en análisis previos, para aumentar su precisión.

1.1.5 Cobertura de Riesgos Tributarios

Las estrategias de cobertura están encaminadas a reducir, combatir o prevenir riesgos tributarios, que se enmarcan en tres líneas de tiempo: proactivas, concurrentes y, reactivas. El fin último de estas estrategias es buscar el cambio de comportamiento tributario de los contribuyentes, disminuyendo así las brechas tributarias.

1.1.6 Evaluación de Riesgos Tributarios

El proceso de evaluación contempla el seguimiento al plan y programas de control; evaluación de aspectos como: impacto de las estrategias en el cambio del comportamiento tributario de los contribuyentes y en la recaudación tributaria, efectividad y calidad de las acciones realizadas, evolución del proceso.

Para ello, se utilizan tres tipos de herramienta: evaluación de impacto, monitoreo y, herramientas de análisis y; en cada una de éstas se aplican varias técnicas de análisis.

2. MODELOS DE MINERÍA DE DATOS PARA LA DETECCIÓN DE EVASIÓN

2.1. METODOLOGÍA CRISP-DM

Según Chapman, et al. (2000), la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) es un modelo jerárquico que consta de cuatro niveles de abstracción: fases, tarea genérica, tarea especializada, e instancia de procesos. Son seis fases las que componen esta metodología:

a. *Comprensión del negocio.-* Esta fase se basa en el entendimiento del problema que se quiere resolver, para identificar los objetivos y requerimientos de un negocio. Las principales tareas de esta fase son:
    - Definición de los objetivos del negocio.
    - Evaluación de la situación actual.
    - Determinación de los objetivos de la minería de datos.
    - Definición del plan del proyecto.

b. *Comprensión de los datos.-* Esta fase comienza con la recopilación inicial de datos y continúa con la revisión de la calidad de los mismos, el descubrimiento de los primeros conocimientos en los datos para la definición de hipótesis. Es una de las fases que más tiempo toma dentro del proyecto. Esta fase consta de las siguientes tareas:
    - Recopilación inicial de datos.
    - Descripción de los datos.
    - Exploración de los datos.
    - Verificación de la calidad de datos.

c. *Preparación de los datos.-* Esta fase incluye todas las actividades necesarias para construir el conjunto final de datos, con el cual se generará el modelo. Las tareas, que probablemente se realicen varias veces, consisten en:
    - Selección de los datos.
    - Limpieza de datos.
    - Construcción de datos.
    - Integración de datos.
    - Formateo de datos.

d. *Modelado.-* En esta fase, se seleccionan y aplican las técnicas de modelado de acuerdo al problema que se quiere resolver y, se calibran sus parámetros a valores óptimos. Algunas técnicas tienen requerimientos específicos sobre la estructura de los datos; por ello, generalmente se regresa a la fase de preparación de datos. Las tareas son:
    - Selección de la técnica de modelado.
    - Diseño de la evaluación.
    - Construcción del modelo.
    - Evaluación del modelo.

e. *Evaluación*.- Una vez que se han construido uno o varios modelos que han alcanzado la calidad suficiente desde una perspectiva de análisis de datos y

antes de realizar la implementación final, es importante evaluarlo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Es clave identificar si hay algún aspecto del negocio que no haya sido considerado. En esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos. Las tareas son:

- Evaluación de resultados.
- Revisión del proceso.
- Establecimiento de las siguientes acciones.

f. *Implementación*.- En esta fase se explota la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización. Dependiendo de los requerimientos del negocio, esta fase consistir simplemente en la generación de un informe o en la automatización y ejecución periódica del proceso de análisis de datos, que es algo más complejo. Las tareas que se desarrollan en esta fase son:

- Implementación del modelo.
- Monitoreo y mantenimiento del modelo.
- Generación del informe final
- Revisión del proyecto

Esta metodología, que es útil para el desarrollo de proyectos de minería de datos, se resume en la ilustración 2.2.

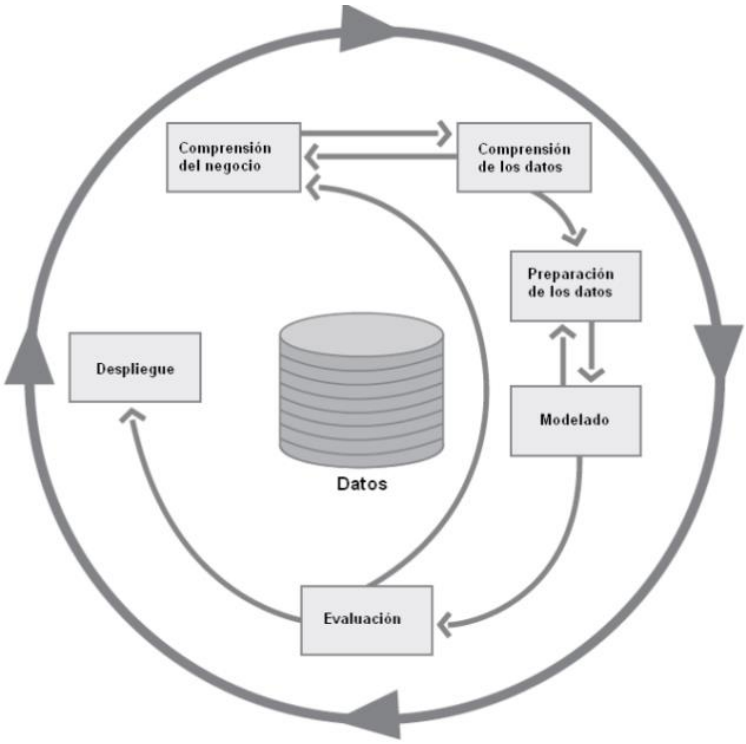

Ilustración 2.2. Fases metodología CRISP-DM

2.2. MODELOS PARA DETECCIÓN DE ANOMALÍAS

En la actualidad, es de alta relevancia la construcción de modelos que permitan detectar anomalías; es decir, patrones que se alejan del comportamiento normal esperado. Dichas anomalías se producen en diferentes ámbitos como el financiero, salud, informático, ambiental, entre otros.

Como señalan Lotfi Shahreza et al. (2011); se pueden distinguir 3 categorías de las técnicas para detección de anomalías: las técnicas supervisadas que aprenden a clasificar, utilizando la variable que contiene las observaciones categorizadas en normal o anómala. Las técnicas semi-supervisadas que generan un modelo que representa el comportamiento normal de un conjunto de datos de entrenamiento, para luego evaluar la probabilidad de que una observación o instancia sea gene-

rada por este modelo; y, finalmente están las técnicas no supervisadas que detectan anomalías en un conjunto de datos sin la clase, bajo el supuesto de que la mayoría de los casos en el conjunto de datos son normales.

Dentro de las técnicas no supervisadas para detección de anomalías están, por ejemplo, los mapas auto-organizados de Kohonen (SOM, por sus siglas en inglés) y la optimización de partículas (PSO, por sus siglas en inglés).

2.3. REDES NEURONALES ARTIFICIALES: APRENDIZAJE NO SUPERVISADO

Las características principales de las redes neuronales de aprendizaje no supervisado son que descubren en los datos de entrada y de forma autónoma, las características, regularidades, categorías de los datos y, es capaz de obtenerlas de forma codificada en la salida; por lo tanto, muestran cierto grado de auto-organización.

Este tipo de red neuronal, conformada por una arquitectura simple, genera resultados útiles si existe algún tipo de redundancia, a partir de lo que se puede identificar patrones. Suelen requerir menores tiempos de entrenamiento que las supervisadas. Estos modelos son los más cercanos a estructuras neurobiológicas; ya que tienden a imitar su comportamiento (Isasi Viñuela & Galván León, 2004).
Haciendo uso de este tipo de redes, se pueden resolver varios problemas; entre ellos están:

a. Análisis de componentes principales para identificar las componentes que representan en mayor medida al conjunto de datos y, eliminar aquellas que no.
b. Agrupamiento de un conjunto de datos, para identificar a qué grupo pertenecen y cuáles son sus características; a partir de los patrones detectados por la red.

c. Identificación de un prototipo de la clase a la que pertenece un dato de entrada.
d. Caracterización de un conjunto de datos, mediante un mapa topológico.

2.3.1. MAPAS AUTO-ORGANIZADOS DE KOHONEN

Los mapas auto-organizados (SOM, por sus siglas en inglés) fueron diseñados por Teuvo Kohonen (1982); los mismos que consisten en una red neuronal de 2 capas, una capa de entrada y una capa competitiva.

Esta red realiza un aprendizaje no supervisado competitivo, siendo su principal objetivo realizar la agrupación de datos similares en una misma neurona; es decir, los datos de entrada son agrupados en las neuronas, en función de las correlaciones existentes entre ellos. La aplicación más usual de este tipo de red es para realizar agrupamientos o extracción de características; relacionando además las agrupaciones entre sí (Isasi Viñuela & Galván León, 2004).

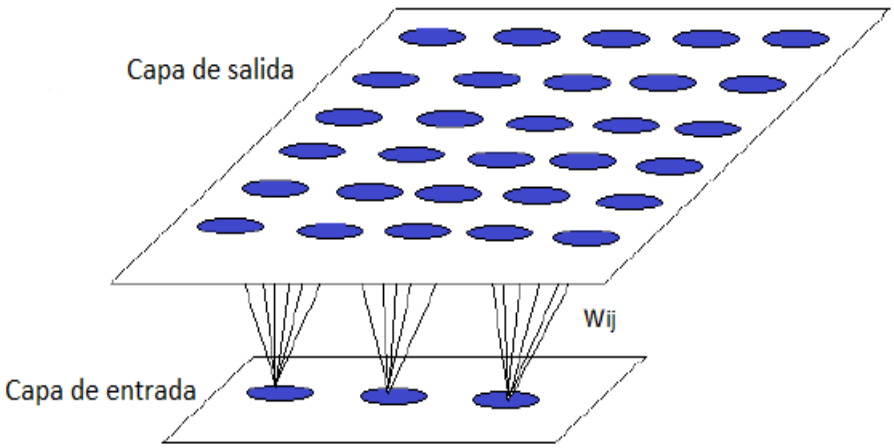

Ilustración 2.3. Estructura del mapa auto-organizado de Kohonen.

La capa de entrada está formada por *i* neuronas, una por cada atributo; esta capa recibe y transmite la información externa a la capa competitiva, donde se procesa la información y se forma el mapa de características. A su vez, cada neurona de entrada *(E_i)* se conecta a todas las neuronas de la capa competitiva *(C_j)*, mediante un peso W_{ij}. Por lo tanto, las neuronas de la capa competitiva tienen asociado un vector de pesos W_{ij} denominado vector de referencia, porque constituye el vector prototipo de la categoría representada por la neurona de salida C_j. De esta manera, la red define una proyección desde un espacio de datos de alta dimensión a un mapa bidimensional de neuronas.

Las neuronas de la capa competitiva ejercen cierta influencia sobre sus vecinas, a través de la aplicación de la función denominada vecindad, que genera la topología del mapa, que puede ser hexagonal o rectangular:

Las neuronas adyacentes pertenecen a una vecindad N_j de la neurona C_j. La topología y el número de neuronas permanece fijo desde el principio. El número de neuronas determina la suavidad de la proyección, lo cual influye en el ajuste y capacidad de generalización de la red SOM. Durante la fase de entrenamiento, el SOM forma una red elástica que se pliega dentro de la nube de datos originales.

El algoritmo para generar un mapa auto-organizado (SOM) funciona de la siguiente manera:

1. Se selecciona el tamaño y el tipo del mapa, que puede ser hexagonal o rectangular, dependiendo de la estructura que se requiera. Generalmente, se utiliza el tipo hexagonal.
2. Los pesos de las conexiones se inicializan aleatoriamente y se van modificando durante el proceso de esta fase.
3. Un vector *x* es seleccionado al azar del conjunto de datos y se calcula su distancia a los vectores de referencia de cada neurona, haciendo uso de una función de distancia (τ_j) predeterminada, para calcular la salida de las

neuronas de la capa competitiva. Todas las salidas se comparan entre sí, para seleccionar a aquella que genere la salida más pequeña, es decir a la neurona ganadora.

4. Una vez que se ha encontrado el vector más próximo, el resto de vectores de referencia es actualizado. La neurona ganadora y sus vecinas se mueven cerca del vector *x* en el espacio de datos. La magnitud de dicha atracción está regida por la tasa de aprendizaje (α). Mientras se va produciendo el proceso de actualización y nuevos vectores se asignan al mapa, la tasa de aprendizaje decrece gradualmente hacia cero. Junto con ella también decrece el radio de vecindad.
5. Una vez que se ejecuta este proceso, la red debe responder de forma semejante a estímulos parecidos; para ello, se aplica la regla de actualización para el vector de referencia dado *i*, es decir, se refuerzan aquellas unidades que hayan respondido en mayor grado, de forma proporcional al valor de entrada:

$$\frac{dW_{ij}}{dt} = \alpha(t)\tau_j(t)\big(E_i(t) - W_{ij}(t)\big)$$

Esta ecuación corresponde a un aprendizaje hebbiano, donde el incremento del valor del peso de cada conexión es proporcional al valor de activación de las células que conecta $(\tau_j \, . E_i)$. El esquema final de aprendizaje está definido como:

$$\frac{dW_{ij}}{dt} = \begin{cases} \alpha(t)\left(E_i(t) - W_{ij}(t)\right) & si\ C_i\ es\ ganadora \\ 0 & caso\ contrario \end{cases}$$

Los pasos 3 y 4 se van repitiendo hasta que el entrenamiento termina. El número de iteraciones para el entrenamiento se debe fijar a priori, para calcular la tasa de convergencia de la función de vecindad y de la tasa de aprendizaje. Una vez terminado el entrenamiento, el mapa ha de ordenarse

en sentido topológico: n vectores topológicamente próximos se aplican en n neuronas adyacentes o incluso en la misma neurona.

Como se mencionó anteriormente, en el método de Kohonen la tasa de aprendizaje decrece a medida que avanza el proceso de entrenamiento de la red, lo que se denomina función “de olvido” que hace que los patrones ya incluidos previamente sean aprendidos con mayor intensidad por la red neuronal. Hay dos esquemas de aprendizaje para el decrecimiento de α:

1. En el primer esquema, con cada iteración (ciclo completo de los patrones de aprendizaje), el valor de α disminuye en una cantidad constante pequeña β.

   $$\alpha\,(t+1) = \alpha(t) - \beta$$

   El número de iteraciones o ciclos de aprendizaje, para el cual el valor de α sería 0 y la red estaría estabilizada, se puede determinar mediante el parámetro β de la siguiente manera:

   $$Iteraciones = \frac{\alpha(0)}{\beta}$$

2. En el segundo, la disminución del valor de α sigue un esquema logarítmico; esta disminución es elevada en las primeras iteraciones y se va reduciendo paulatinamente hasta alcanzar valores muy pequeños cercanos a cero. Dado este esquema, las primeras iteraciones son las que tienden a formar la estructura de la red; ajustándose con pequeños desplazamientos de las neuronas hasta alcanzar el equilibrio.

Así, el funcionamiento de la red cuando se introduce un patrón, en términos geométricos es igual al cálculo de la distancia entre dicho patrón y cada una de las neuronas, que hacen las veces de prototipos de los patrones de entrenamiento. Entonces, la neurona ganadora será aquella que tenga la menor distancia respecto al patrón.

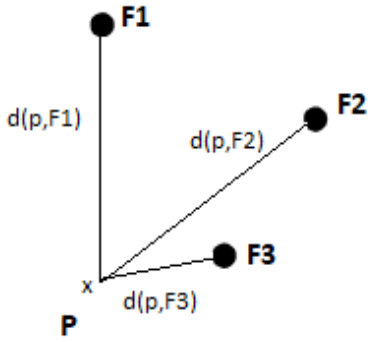

Ilustración 2.4. Célula ganadora

Por lo tanto, el entrenamiento se expresa mediante la siguiente fórmula:

$$W_{ij} = W_{ij} + \frac{dW_{ij}}{dt}$$

Cabe señalar, que el incremento de la conexión es proporcional (α) a la distancia entre el patrón y la neurona ganadora. Geométricamente, la modificación de los valores de las conexiones se traduce en un desplazamiento del prototipo al que representa la neurona ganadora, acercándose más al ejemplo de entrada; cuánto más grande sea el valor de α, mayor será el desplazamiento.

Es precisamente este esquema el indicado para resolver problemas de agrupamiento; ya que los prototipos se desplazan hacia el centro de gravedad de “nubes” más o menos compactas de patrones de entrenamiento.

Otro parámetro del sistema que se define como parte de la arquitectura de la red es el vecindario. El mecanismo del vecindario, que es usado durante el aprendizaje, consiste en definir para cada neurona de la capa competitiva, mediante una estructura espacial, un conjunto de neuronas denominadas “vecinas”. Cuando se

introducen vecindarios, el aprendizaje se produce para la neurona ganadora y para sus vecinas más cercanas.

La proximidad del vecindario está definida por el número de neuronas que se atraviesan para llegar desde el origen hacia el destino; es decir se tiene la distancia del vecindario.

El efecto del vecindario es muy útil, porque al término del aprendizaje ocurrirá que las neuronas cercanas en un vecindario ocuparán posiciones cercanas en el espacio. Por lo tanto, se podrá analizar qué prototipos están relacionados entre sí; además de saber los patrones que son representados por un cierto prototipo.

Las principales ventajas de un mapa auto-organizado son su eficiencia en el manejo de grandes volúmenes de datos, siendo robusto ante la presencia de ruido en los datos. Reducen grandes volúmenes de información, conservando al máximo las relaciones topológicas relevantes del conjunto de datos en un plano bidimensional. Se entrenan sin necesidad de tener un conocimiento previo y realizar supuestos sobre la clase de pertenencia de los datos (Sharma & Omlin, 2009).

En tanto que, las desventajas es que necesitan que el número de grupos sea especificado, lo cual no es trivial ya que no se conoce previamente las características de la información; para lo cual se puede ejecutar el algoritmo con varias configuraciones. Adicionalmente, se debe realizar una inspección manual o aplicar los algoritmos tradicionales, para definir los límites de los agrupamientos obtenidos.

2.3.2. ALGORITMO MULTI-SOM

Tal como lo mencionan Lu y Segall (2013), el rendimiento de una red neuronal SOM se basa en la fijación de los distintos parámetros como el tamaño del mapa, la tasa de aprendizaje, los pesos iniciales. Es por ello, que un tamaño de mapa

grande, no quiere decir necesariamente que sea el mejor y; dado que, el método SOM tiene la limitación de determinar un número específico de grupos, ya que su función principal es visualizar datos en la forma de un mapa; se propuso el enfoque multi-SOM que es una extensión del modelo SOM y, devuelve un número óptimo de grupos. Según varios autores (Ghouila, y otros, 2009), (Lu & Segall, 2013) y (Khanchouch, Charrad, & Limam, 2015), este algoritmo es eficiente y confiable para agrupar un conjunto de datos.

El algoritmo funciona de la siguiente manera:

1. Primero se entrena el modelo con el algoritmo SOM (mapa original).
2. A partir del mapa original, mediante un proceso de generalización en línea; de forma iterativa se generan otros niveles de agrupación, disminuyendo gradualmente el número de neuronas, hasta que la dimensión del mapa es de 4 neuronas. De forma general, la dimensión del mapa en cada nivel será *q x p, (q, p ≥ 2)*; que a su vez en el nivel próximo generará un mapa de dimensión (*q – 1) x (p – 1).*
3. En cada nivel, el perfil de la neurona se calcula con la siguiente fórmula:

$$W_q^{P+1} = \frac{1}{4} \sum_{q_k \in V_q^P} W_{q_k}$$

Donde,

V_q^P representa la vecindad cuadrada en el mapa *P* asociado a la neurona *q* del nuevo mapa sintético *P + 1.*

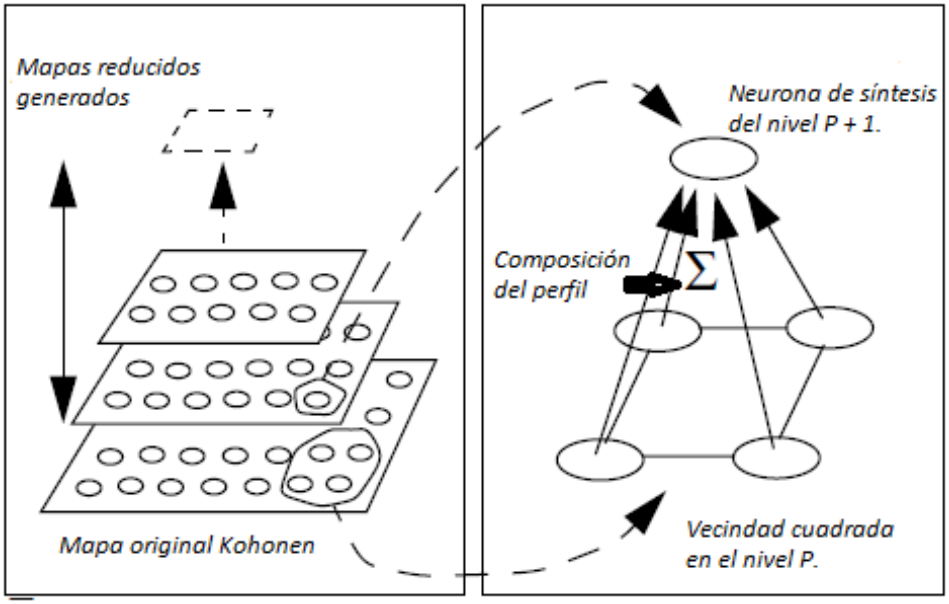

Ilustración 2.5. Arquitectura algoritmo multi-SOM.
Fuente: (Lamirel, 2002).

Esta solución tiene la ventaja de preservar la estructura de la vecindad en los nuevos niveles generados; conservando las propiedades de la topografía del mapa de neuronas y, la cercanía de las áreas de las neuronas en los mapas generalizados. De esta forma, se construye una jerarquía de mapas auto-organizados basados en la reciprocidad de la vecindad, mediante la superposición entre varios mapas SOM (Khanchouch, Charrad, & Limam, 2015). Cada mapa en multi-SOM representa un punto de vista y la información en cada mapa es representado por nodos (clases) y áreas lógicas (grupo de clases).

El paquete ‘multisom’ del lenguaje R, incluye el algoritmo Multi-SOM que como ya se ha mencionado permite realizar la segmentación de un conjunto de datos, determinar el mejor número de agrupamientos y, además obtener el mejor esquema de segmentación a partir de diferentes resultados. En el paquete hay dos versiones del algoritmo Multi-SOM: estocástico y por lotes. Cabe señalar que, en cada nivel se calculan los índices de validación; a partir de los cuales se puede evaluar el número de grupos óptimo.

2.4. CRITERIOS DE EVALUACIÓN DE UN ANÁLISIS DE SEGMENTACIÓN

Cuando se realiza un modelo de segmentación, el procedimiento para llevarlo a cabo es el siguiente:

1. Preprocesamiento de los datos.
2. Evaluación de la tendencia de agrupamiento.
3. Selección y ejecución del algoritmo.
4. Evaluación de la calidad de los grupos obtenidos.

Como ya se mencionó en secciones anteriores; el paso 1 y 3, se realizan en cualquier desarrollo de un modelo de Minería de Datos. En cuanto a la evaluación de la tendencia de agrupamiento y la calidad de los grupos obtenidos, existen criterios estadísticos e índices que permiten realizar este análisis.

2.4.1. EVALUACIÓN DE LA TENDENCIA DE AGRUPAMIENTO

Previo a la aplicación de un método de agrupación en un conjunto de datos, es importante evaluar si este conjunto contiene agrupaciones; es decir, estructuras no aleatorias y, por lo tanto, que el análisis de segmentación será válido. A este procedimiento se denomina evaluación de tendencia de agrupamiento. Para ello, el enfoque más común, es el uso de pruebas estadísticas para aleatoriedad espacial; una de las métricas que se utiliza para realizar esta evaluación es el estadístico de Hopkins.

a. Estadístico de Hopkins.- En este método se generan *p* puntos que son distribuidos aleatoriamente en el espacio y, se obtiene una muestra *p* del conjunto de datos original; para estos dos conjuntos de datos, se obtiene las distancias al vecino más cercano del conjunto de datos original. Si los

datos no están distribuidos en grupos, las distancias de los dos conjuntos serán similar en promedio (Tan, Steinbach, Karpatne, & Kumar, 2019). El estadístico está definido por la siguiente fórmula:

$$H = \frac{\sum_{i=1}^{p} W_i}{\sum_{i=1}^{p} U_i + \sum_{i=1}^{p} W_i}$$

Donde,

U_i es igual a las distancias al vecino más cercano del conjunto de datos generados aleatoriamente.

W_i es igual a las distancias al vecino más cercano de la muestra *p* obtenida del conjunto de datos original.

Considerando que la hipótesis nula es que los datos no son agrupables, el estadístico sigue una distribución beta con los dos parámetros igual al número de puntos seleccionados de la muestra *p*. Esta hipótesis se cumple si los dos conjuntos de datos tienen distancias similares y, por lo tanto, el estadístico *H* será cercano a 0,5 (Tan, Steinbach, Karpatne, & Kumar, 2019).

2.4.2. VALIDACIÓN DE AGRUPAMIENTOS

Considerando que, la premisa de un análisis de segmentación es que las observaciones pertenecientes a un grupo sean muy similares y, que sean altamente diferentes con observaciones de otros grupos; el principal problema en la agrupación es determinar el número de grupos ideal, bajo estas condiciones. Por lo tanto, es necesario evaluar la calidad de los grupos. Para ello, existe la validación externa y, validación interna (Khanchouch, Charrad, & Limam, 2015); (Tan, Steinbach, Karpatne, & Kumar, 2019).

b. Validación externa

Consiste en comparar los resultados de un análisis de agrupamiento y un resultado externo conocido, como la clase objetivo; para medir el grado de coincidencia de los grupos con las clases conocidas. Por ello, este enfoque se utiliza generalmente para seleccionar el algoritmo de agrupamiento correcto para un conjunto de datos específico. Existen varios criterios de validación como la pureza, entropía y medida-F.

c. Validación interna

En este tipo de validación se utiliza la información interna de los grupos, para evaluar la calidad de dichos agrupamientos. De esta forma, se puede estimar el número óptimo de agrupaciones y, seleccionar el algoritmo de agrupamiento apropiado. Las métricas de validación interna se basan generalmente en los criterios de cohesión (distancia mínima entre los miembros de un mismo grupo) y separación (distancia máxima entre miembros de diferentes grupos).

Existen múltiples índices de validación interna (Charrad, Ghazzali, Boiteau, & Niknafs, 2014), entre ellos están:

c1. *Davies-Bouldin (DB).*- Valores pequeños de este índice de validación interna indican mejor calidad del agrupamiento. Se calcula de la siguiente manera:

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max i \neq j \left\{\frac{d(Xi) + d(Xj)}{d(Ci, Cj)}\right\}$$

Donde,

c es el número de grupos.

$i\ y\ j$ son los grupos.

$d(Xi) y\, d(Xj)$ son las distancias entre todas las observaciones en los grupos *i* y *j* a sus respectivos centroides.

$d(Ci, Cj)$ es la distancia entre estos dos centroides.

c2. Dunn.- Este índice es igual a la relación entre la distancia más pequeña entre las observaciones de distintos grupos y la máxima distancia entre las observaciones de un grupo (C_k). El índice toma un valor entre 0 e infinito; valores altos de este índice indican mejor calidad del agrupamiento. La fórmula de cálculo es:

$$DI = \min 1 \leq i \leq c \left\{ min \left\{ \frac{d_{(Ci,Cj)}}{\max 1 \ \leq k \ \leq c \ (d_{Xk})} \right\} \right\}$$

Donde,

$d_{(Ci,Cj)}$ denota la distancia entre dos grupos.

d_{Xk} representa la distancia dentro de un grupo *k.*

c es el número de grupos de un conjunto de datos.

c3. Silhouette.- Este índice se define como:

$$S = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Donde,

$a(i)$ es la distancia promedio de la observación i al resto de observaciones que conforman el grupo.

$b(i)$ es la distancia promedio mínima, que se obtiene de las distancias promedio calculadas entre la observación i y todas las observaciones de cada uno de los grupos que no contienen a esta observación.

Este índice puede tener valores entre -1 y +1; un valor alto indica mejor calidad del agrupamiento.

c4. *Índice C.*- Valores mínimos del índice muestran una buena calidad del agrupamiento; este índice toma valores entre 0 y 1. Se estima de la siguiente forma:

$$Índice\ C = \frac{S_c - S_{min}}{S_{max} - S_{min}}, S_{min} \neq S_{max}$$

Donde,

c = es el número de grupos de un conjunto de datos.

S_c = es la suma de las distancias de todos los pares de observaciones de un mismo grupo.

S_{min} = es la suma de las distancias más pequeñas de todos los pares de observaciones en el conjunto de datos.

S_{max} = es la suma de las máximas distancias entre todos los pares de observaciones en el conjunto de datos.

c5. *Calinski and Harabasz (CH).*- El índice está definido por:

$$CH(c) = \frac{B(c)/(c-1)}{W(c)/(n-c)}$$

Donde,

c es el número de grupos.

$B(k)$ es la suma de cuadrados entre grupos.

$W(k)$ es la suma de cuadrados intra-grupos.

El número de grupos óptimo será aquel que maximice el valor del índice CH.

c6. *Ball.-* Este índice fue propuesto por Ball and Hall (1965) y, se basa en la distancia promedio entre las observaciones y el centroide en cada uno de los grupos. La fórmula es:

$$Ball = \frac{W_q}{q}$$

El número óptimo de grupos se obtiene con la máxima diferencia entre los niveles de jerarquía del índice.

c7. Hartigan.- El índice Hartigan es una regla general heurística con un considerable éxito. Al igual que el índice Calinski and Harabasz, este índice se basa en la suma de los cuadrados intra-grupo (distancia Euclídea) y considera el cambio cuando se incrementa k. La fórmula para el cálculo de este índice es:

$$Hartigan = \left(\frac{W_k}{W_{k+1}} - 1\right)(n - k - 1)$$

Donde,

$k \in \{1, \dots, n-2\}$

Para determinar el número óptimo de grupos, se toma la máxima diferencia entre los niveles de jerarquía del índice (Hartigan, 1975).

c8. Índice SDbw.- La definición de este índice se basa en los criterios de cohesión dentro de los grupos y separación entre los grupos. Este índice se calcula con la ecuación:

$$SDbw(q) = Scat(q) + Densidad.bw(q)$$

Para calcular el término $Scat(q)$, se utiliza la siguiente fórmula:

$$Scat(q) = \frac{\frac{1}{q}\sum_{k=1}^{q}\left\|\sigma^{(k)}\right\|}{\|\sigma\|}$$

Donde,

σ es el vector de varianzas para cada variable en el conjunto de datos.

$\sigma^{(k)}$ = es el vector de varianza para cada grupo $C^{(k)}$.

El término Densidad.bw(q) es la densidad entre grupos, que evalúa la densidad promedio en la región entre grupos en relación a la densidad de los grupos y, se calcula de la siguiente forma:

$$Densidad.bw(q) = \frac{1}{q(q-1)} \sum_{i=1}^{q} \left(\sum_{j=1, i \neq j}^{q} \frac{\sum_{l=1}^{n_{ij}} f(x_l, u_{ij})}{\max\left(densidad(C_i), densidad(C_j)\right)} \right)$$

Donde,

(u_{ij}) es el punto medio de la línea de segmento definido por los centroides de los grupos C_i y C_j.

n_{ij} es el número de tuplas que pertenece a los grupos C_i y C_j.

$f(x_l, u_{ij})$ = es igual a 0 si $d(x, u_{ij}) > Desvest$ y 1 caso contrario.

$Desvest$ es la desviación estándar promedio de los grupos.

Se considera el número de grupos óptimos, aquel que tenga el mínimo valor de SDbw.

3. RESUMEN

En este capítulo se ha reseñado la problemática que genera la evasión de impuestos y, los retos que frente a ello han asumido las Administraciones Tributarias a nivel mundial. Por otro lado, se revisó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que es la metodología utilizada para desarrollar

esta investigación y, que consta de 6 fases que van desde la comprensión del negocio hasta la implementación del modelo. Se revisaron también los modelos para detección de anomalías y dentro de ellos, las redes neuronales de aprendizaje no supervisado; específicamente, los mapas auto-organizados de Kohonen (SOM, por sus siglas en inglés) y el algoritmo multi-SOM. Finalmente, se detallaron los índices de validación interna que sirvan para evaluar el número óptimo de grupos.

En el siguiente capítulo, se desarrolla la metodología para la construcción del modelo como tal y, se obtienen los índices de validación para evaluar los grupos obtenidos.

CAPÍTULO II

Para el desarrollo del modelo de Minería de Datos, se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). El modelo de referencia está conformado por 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, implementación.

1. COMPRENSIÓN DEL NEGOCIO

Considerando la relevancia que el análisis de la evasión de impuestos tiene para una Administración Tributaria, por las afectaciones que genera a los ingresos del fisco; se ha definido como objetivo principal de esta investigación realizar el análisis de potenciales riesgos de evasión del impuesto al valor agregado (IVA); ya que es uno de los impuestos que mayor recaudación genera. Es así que, mediante el uso de técnicas de Minería de Datos para explotar masivamente la información y extraer conocimiento sobre patrones de comportamiento, perfiles de riesgo y anomalías que presentan los contribuyentes; se generó un modelo para analizar potenciales riesgos de evasión.

En la definición de este objetivo se han considerado varios aspectos como son la estructura del impuesto, los distintos segmentos de contribuyentes, la transaccionalidad, entre otros.

2. COMPRENSIÓN DE LOS DATOS

La Administración Tributaria dispone de un almacén de datos que se alimenta de las distintas fuentes de información, que provienen de las declaraciones de impuestos, anexos transaccionales de compras, ventas, importaciones, exportaciones presentados por los contribuyentes; así como, de las transacciones financieras e información reportada por otras instituciones, tanto del sector público como del sector

privado; esto se traduce en millones de registros administrativos de distinta naturaleza. Para el desarrollo de esta investigación, se obtuvieron los datos de las siguientes fuentes de información:

Tabla 3.1. Descripción de fuentes de información

Fuente de información	Descripción
Registro Único de Contribuyentes (RUC)	Características de los contribuyentes: tipo contribuyente, clase de contribuyente, actividad económica, fecha de inscripción, entre otras.
Declaración del impuesto al valor agregado	Información de ventas, compras, impuesto generado.
Declaración del impuesto a la renta	Información de inventarios, patrimonio, entre otros.
Estadísticas de declaraciones	Número de retrasos en la presentación de declaraciones y anexos y, declaraciones y anexos no presentados.
Devolución de impuesto al valor agregado	Número de solicitudes realizadas, monto solicitado y monto devuelto.
Anexo transaccional	Proveedores locales y del exterior; clientes locales y del exterior.

Fuente: SRI.

A continuación, se detalla el número de registros administrativos de las principales fuentes de información utilizadas:

Tabla 3.2. Registros administrativos

Año	RUC	Impuesto al valor agregado		Impuesto a la renta sociedades	Impuesto a la renta personas naturales
		Declaraciones mensuales	Nro. Contribuyentes que declaran		
2014		10.452.348	724.313	132.183	827.911
2015		10.729.390	810.106	134.428	821.419
2016		10.546.024	885.809	128.581	765.105
Total	4.270.736	31.727.762	2.420.228	395.192	2.414.435

Fuente: SRI.

Cabe señalar que, el universo de contribuyentes en estado activo sobrepasa los 2,2 millones, entre sociedades y personas naturales; tal como se muestra en la Tabla 2.3.

Tabla 3.3. Tipo de contribuyente

Tipo de contribuyente	Número
Personas naturales	2.035.075
Sociedades	174.952
Total	2.210.027

Fuente: SRI.

Para definir la información a utilizar de estas fuentes de información, se tomaron en cuenta 2 aspectos: la revisión de experiencias internacionales en el desarrollo de modelos de Minería de Datos para el análisis de riesgos de evasión de impuestos (Alm et al., 2004; Arias, 2004; Gupta & Nagadevara (2007); Denny et al., 2007; Pisani & De Sisti (2007/8); Wu et al., 2012; Khwaja et al., 2012; Castellón González & Velásquez, 2013) y, el criterio de expertos del negocio.

3. PREPARACIÓN DE LOS DATOS

En primera instancia, se aplicaron varios filtros para realizar una primera depuración de la información, considerando la estructura del impuesto al valor agregado y las características de los contribuyentes; conforme al siguiente detalle:

a. Exclusión de contribuyentes en estado pasivo a la fecha; para obtener únicamente el universo de contribuyentes que se encontraban realizando alguna actividad económica en el periodo de análisis.
b. Exclusión de contribuyentes que pertenecen al sector público; porque se asume que en este grupo de contribuyentes no existe evasión.
c. Exclusión de organismos y misiones internacionales. De acuerdo a la normativa vigente, el impuesto al valor agregado pagado por misiones diplomáticas, consulares, organismos internacionales y sus funcionarios rentados de nacionalidad extranjera, acreditados en el Ecuador, se devuelve de acuerdo con los convenios internacionales y otros instrumentos diplomáticos; por lo que no es objeto de este estudio.
d. Exclusión de contribuyentes con declaración semestral, que la realizan aquellos contribuyentes que transfieren bienes o prestan servicios gravados con tarifa 0% o no gravados, así como aquellos que están sujetos a la retención total (100%) del IVA causado. Dadas las características de estos contribuyentes, no habría espacio para que se genere evasión; por lo tanto, no son considerados para el análisis.

Es importante mencionar que, con el paso de los años se han ido perfeccionando los formularios para las declaraciones de impuestos y anexos; se han ido incorporando distintas validaciones para mitigar la inclusión de datos erróneos; lo que ha permitido que haya un mejoramiento paulatino de la calidad y consistencia de la información reportada por los contribuyentes.

Para la preparación de datos; así como, para el modelado y evaluación se utilizó el lenguaje de programación R. La principal fuente de información utilizada son las

declaraciones del impuesto al valor agregado para el periodo 2014-2016. Con los filtros mencionados anteriormente, la base extraída consta de 25.392.692 registros; cada registro corresponde a la declaración mensual de un contribuyente. Esta información fue agregada para disponer de los datos anuales relacionados con las ventas, compras e impuesto causado. Por otro lado, se tiene la información de la declaración anual del impuesto a la renta de sociedades y personas naturales; la información descriptiva de los contribuyentes; las estadísticas anuales de presentación de declaraciones, devoluciones, entre otros.

Dado que se quiere realizar un análisis transversal y longitudinal, para detectar patrones de comportamiento de los contribuyentes en relación a sus similares y, en función de su propio comportamiento en el tiempo; se excluyó de la base de datos a aquellos contribuyentes con menos de 12 declaraciones mensuales del IVA y, a aquellos que no registraron ingresos durante el periodo de estudio; para tener suficiente información histórica. De igual forma, se excluyeron a los grandes contribuyentes, que es un grupo conformado por menos de 200 empresas; a los cuales se les realiza auditorías extensivas y, presentan características particulares distintas al resto de contribuyentes; por lo cual, los potenciales riesgos tributarios podrían diferir completamente de los riesgos de negocios más pequeños.

Una vez depurada la base de datos, se realizó la selección de variables y, a partir de éstas se generaron nuevas variables:

Tabla 3.4. Descripción de variables seleccionadas

i. Características del contribuyente	**Tipo de variable**	**Dominio**
a. Actividad económica	Nominal	Real
b. Tipo de contribuyente	Nominal	Real
c. Clase de contribuyente	Nominal	Real
d. Tamaño	Ordinal	Calculada

e. Antigüedad del contribuyente	Continua	Calculada
f. Provincia	Nominal	Real
g. ¿Está obligado a llevar contabilidad?	Dicotómica	Real
h. ¿Es exportador habitual?	Dicotómica	Real
i. Contador coincidente (2 o más contribuyentes)	Dicotómica	Calculada
j. Representante legal coincidente (2 o más contribuyentes)	Dicotómica	Calculada
k. ¿Es cliente de empresa fantasma?	Dicotómica	Calculada
ii. Nivel de cumplimiento en presentación de declaraciones	**Tipo de variable**	**Dominio**
a. Número de declaraciones del IVA presentadas	Discreta	Calculada
b. Número de retrasos en la presentación de declaraciones del IVA	Discreta	Calculada
c. Número de declaraciones del IVA no presentadas	Discreta	Calculada
d. Número de retrasos en la presentación de anexos transaccionales (ATS)[2].	Discreta	Calculada
e. Número de anexos transaccionales (ATS) no presentados	Discreta	Calculada
iii. Indicadores de rendimiento de la actividad económica y sus variaciones	**Tipo de variable**	**Dominio**
a. Impuesto generado en las ventas / Ventas totales	Continua	Calculada
b. Impuesto a liquidar en el mes por las ventas realizadas / Ventas totales	Continua	Calculada
c. Utilidad bruta (porcentaje de la Ventas totales)	Continua	Calculada
d. Ventas tarifa 0% / Ventas totales	Continua	Calculada
e. Inventario / Ventas totales[3]	Continua	Calculada
f. Compras / ventas	Continua	Calculada
g. Tipo impositivo efectivo (TIE)[4]	Continua	Calculada
h. Variación de las ventas	Continua	Calculada

[2] La construcción de variables a partir del anexo transaccional, aplica sólo para aquellos contribuyentes que están obligados a presentar dicho anexo.

[3] Aplica sólo para las personas obligadas a llevar contabilidad.

[4] Calculado como la relación entre el impuesto causado (IVA ventas – IVA compras) y las ventas totales.

i. Variación de los impuestos	Continua	Calculada
j. Variación de la participación de las ventas locales y exportaciones	Continua	Calculada
k. Variación de la participación de las ventas tarifa 12% y 0%	Continua	Calculada
l. Variación de la rentabilidad	Continua	Calculada
m. Variación de la participación compras / ventas	Continua	Calculada
n. Brecha del tipo impositivo efectivo[5]	Continua	Calculada
iv. Indicadores de devolución del impuesto al valor agregado	**Tipo de variable**	**Dominio**
a. Número de solicitudes de devolución realizadas	Discreta	Calculada
b. Monto de devolución anual solicitada; aquella cuyo monto sea mayor a 10 USD	Continua	Calculada
c. Monto promedio solicitado por contribuyente	Continua	Calculada
d. Monto máximo solicitado por contribuyente	Continua	Calculada
e. Monto mínimo solicitado por contribuyente	Continua	Calculada

Fuente: SRI.

Para calcular estas variables, se generó la programación mediante el siguiente procesamiento de los datos:

1. Depuración de datos atípicos.
2. Generación de variables dicotómicas, que indiquen si un contribuyente cumple o no con algún atributo y, discretización de variables.
3. Generación de ratios.
4. Estimación de variaciones y desviaciones estándar de ventas, compras, impuestos, entre otros.
5. Fusión de las distintas fuentes de información.

[5] Calculado como la diferencia entre el TIE del contribuyente y el TIE del sector al que pertenece, según la actividad económica (Clasificación Internacional Industrial Uniforme, CIIU 4.0-nivel 3) y tamaño (según nivel de ingresos vigente para el periodo 2014-2016).

Para resumir la información tributaria para cada uno de los contribuyentes en el periodo 2014 – 2016; se calcularon las variaciones de ventas, impuestos, rentabilidad, entre otros. Dada la naturaleza de los registros administrativos que incluyen un porcentaje importante de valores igual a 0 en las distintas variables; se utilizó la fórmula de las variaciones relativas medias simétricas, para evitar una significativa pérdida de datos. La fórmula es igual a:

$$\Delta X\% = \frac{X_{t+1} - X_t}{\frac{X_{t+1}+X_t}{2} + 0{,}01}$$

Donde,

X_{t+1} = La variable seleccionada en el año t + 1.

X_t = La variable seleccionada en el año t.

Después del procesamiento realizado, finalmente se obtuvo una base agregada de 686.736 registros; a partir de la cual, se realizó el análisis exploratorio y el desarrollo del modelo.

4. MODELADO Y EVALUACIÓN

Para realizar el modelo de segmentación, se realizó el siguiente procedimiento:

1. Análisis exploratorio de los datos.
2. Evaluación de la tendencia de agrupamiento.
3. Generación del modelo de segmentación.
4. Evaluación de la calidad de los grupos obtenidos.

4.1 ANÁLISIS EXPLORATORIO DE LOS DATOS

Como ya se mencionó en apartados anteriores, existen diferentes patrones de comportamiento en los grupos de contribuyentes; por lo que es pertinente reali-

zar el análisis y la modelación por segmentos. Es así que, para ejecutar el análisis exploratorio, se dividió a la base por tipo de contribuyente, esto es, sociedades y personas naturales. El ejercicio se realizó para sociedades.

La base de sociedades consta de 84.439 registros. En la tabla 2.5, se presenta el resumen de los estadísticos de las variables cuantitativas.

Tabla 3.5. Resumen de estadísticos

Variables	Unidad	Mínimo	1er Cuartil	Mediana	Promedio	3er Cuartil	Máximo	Datos perdidos
Ventas_2014	USD	-	4.757	71.208	1.198.631	414.540	477.729.887	8.375
Ventas_2015	USD	-	2.600	56.309	1.040.409	348.983	466.681.819	545
Ventas_2016	USD	-	1.433	48.632	969.831	312.129	404.076.418	903
IVA_ventas_2014	USD	-	-	1.984	85.360	22.791	53.396.556	8.375
IVA_ventas_2015	USD	-	-	1.388	72.748	18.726	56.001.818	545
IVA_ventas_2016	USD	-	-	1.145	71.231	17.633	38.829.602	903
Impuestogen_totalvtas14	%	0%	0%	12%	8%	12%	13%	18.030
Impuesto_liquidar_totalvtas14	%	0%	0%	12%	24%	12%	1050000%	18.030
Vtas0af_totalvtas14	%	0%	0%	0%	35%	99%	100%	18.030
Vtas0sinaf_totalvtas14	%	0%	0%	0%	33%	92%	100%	18.030
Compras_ventas14	%	0%	0%	100%	202700%	100%	4339995700%	18.030
IVAreten_IVAcompras14	%	0%	1%	10%	29%	32%	302220%	18.823
IVApagar_IVAcompras14	%	0%	5%	25%	775%	68%	10379900%	18.823
Impuestogen_totalvtas15	%	0%	0%	12%	8%	12%	20%	13.481
Impuesto_liquidar_totalvtas15	%	0%	0%	12%	10%	12%	120012%	13.481
Vtas0af_totalvtas15	%	0%	0%	0%	35%	99%	100%	13.481
Vtas0sinaf_totalvtas15	%	0%	0%	0%	33%	93%	100%	13.481
Compras_ventas15	%	0%	0%	100%	5068000%	100%	322800000000%	13.481
IVAreten_IVAcompras15	%	0%	1%	12%	38%	36%	914212%	13.344
IVApagar_IVAcompras15	%	0%	5%	26%	442%	70%	2000000%	13.344
Impuestogen_totalvtas16	%	0%	0%	12%	9%	13%	20%	15.827
Impuesto_liquidar_totalvtas16	%	0%	0%	12%	9%	13%	100%	15.863
Vtas0af_totalvtas16	%	-1687%	0%	0%	35%	99%	100%	15.820
Vtas0sinaf_totalvtas16	%	-1687%	0%	0%	34%	95%	100%	15.820
Compras_ventas16	%	0%	0%	100%	81500%	100%	2150076600%	15.820
IVAreten_IVAcompras16	%	0%	1%	12%	30%	36%	236756%	15.542
IVApagar_IVAcompras16	%	0%	5%	27%	765%	71%	10800000%	15.542
Variacion_sim_vtas15_14	%	-200%	-37%	0%	-5%	23%	200%	8.416
Variacion_sim_vtas16_15	%	-200%	-67%	-4%	-13%	29%	200%	1.426
Varianza_ventas16_14	USD	-	5.749	32.801	286.713	136.522	275.817.479	526
Exportaciones_Vtastot14	%	0%	0%	0%	2%	0%	100%	18.030
Exportaciones_Vtastot15	%	0%	0%	0%	28%	0%	1382213%	18.894
Exportaciones_Vtastot16	%	0%	0%	0%	1080%	0%	52120930%	23.514
Varianza_export16_14	%	0%	0%	0%	580%	0%	30091970%	19.682
Vtas12_Vtastot14	%	0%	1%	100%	65%	100%	100%	18.030
Vtas12_Vtastot15	%	0%	1%	100%	65%	100%	100%	13.481
Vtas12_Vtastot16	%	0%	0%	100%	64%	100%	101%	15.820
Brecha_TIE	%	-10%	-3%	0%	0%	3%	96%	5
TIE_IRC_14sinatip	%	0%	0%	4%	4%	8%	97%	18.037
TIE_IRC_15sinatip	%	0%	0%	4%	4%	8%	83%	13.490
TIE_IRC_16sinatip	%	0%	0%	4%	5%	9%	100%	15.852
TIEpromedio_sector14	%	0%	3%	4%	5%	7%	12%	18.030
TIEpromedio_sector15	%	0%	3%	4%	5%	6%	11%	13.481
TIEpromedio_sector16	%	0%	3%	5%	5%	7%	12%	15.820
Brecha_TIE14	%	-9%	-3%	0%	0%	2%	91%	18.037
Brecha_TIE15	%	-9%	-3%	0%	0%	2%	79%	13.490
Brecha_TIE16	%	-10%	-3%	-1%	0%	3%	96%	15.852
Compras_2014	USD	-	3.036	44.884	939.685	285.071	477.729.659	8.375
Compras_2015	USD	-	2.003	35.724	798.610	236.632	348.453.558	545
Compras_2016	USD	-	1.167	29.197	725.523	202.341	312.801.010	903
Margen_utilidad14	%	-4339995600%	0%	0%	-202600%	100%	100%	18.030
Margen_utilidad15	%	-322800000000%	0%	0%	-5068000%	100%	100%	13.481
Margen_utilidad16	%	-2150076500%	0%	0%	-81400%	100%	100%	15.820
Variacion_sim_vtas16_14	%	-200%	-123%	-12%	-22%	42%	200%	9.278
Variacion_sim_imp16_14		-200%	-87%	0%	-17%	27%	200%	9.278

Fuente: SRI.

Adicionalmente, se realizó un análisis exploratorio multivariado, específicamente un Análisis de Componentes Principales (ACP) con el objetivo de: identificar variables que se encuentran correlacionadas entre si y, por lo tanto, aportan la misma información; identificar a priori la existencia de agrupaciones naturales e; identificar datos atípicos.

Considerando la naturaleza de la información tributaria y, como se evidencia en el análisis descriptivo univariado, la base consta de un porcentaje significativo de valores perdidos. En este sentido y dado que en el programa R no se puede realizar un ACP cuando hay presencia de datos perdidos; se realizó el análisis con aquellas observaciones que tienen los datos completos. No se optó por implementar un algoritmo para imputación de datos; ya que, en este caso, no se trata de información faltante, sino que responde a las condiciones propias de los registros tributarios, como por ejemplo que hay contribuyentes que no registran declaraciones para todo el periodo de estudio, porque son relativamente nuevos.

Al realizar el ACP y analizar el gráfico de los componentes 1 y 2, que conjuntamente explican el 46,1% de la variabilidad de los datos (Ilustración 3.1); se puede observar claramente que existen variables altamente correlacionadas de forma positiva (relación directa) o negativa (relación inversa), lo que se confirma con la matriz de correlaciones incluida en el anexo 1.

Esto responde a las características de los datos, ya que se relacionan con la actividad económica que ejercen los contribuyentes; es decir, es un resumen del resultado económico en un determinado periodo y, por ende, las variables son linealmente dependientes. Un ejemplo de ello son las ventas y compras. De igual forma se tiene una fuerte correlación negativa entre el indicador impuesto generado/ventas totales y el indicador ventas con tarifa 0%/ventas totales. Finalmente, otras correlaciones se dan porque en la base se tienen variables temporales para el periodo 2014-2016.

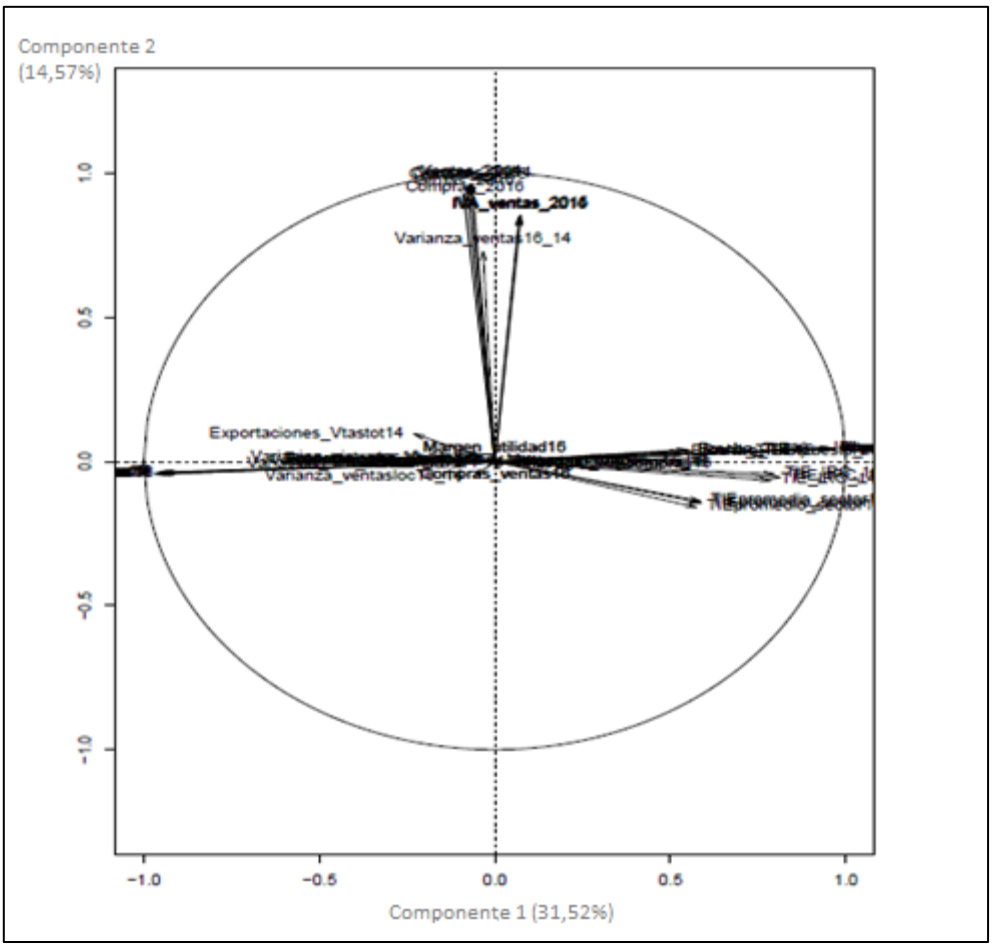

Ilustración 3.1. ACP - Gráfico de variables.

Por otro lado, al analizar el gráfico de individuos (contribuyentes) no se distinguen agrupaciones naturales a simple vista; se observa una sola nube de puntos, con varios individuos que se alejan y un valor atípico extremo. Esto sugiere la necesidad de realizar una mayor segmentación por otras características de los contribuyentes como actividad económica, para elaborar el modelo. También se evidencia que, los contribuyentes se aglomeran a lo largo del componente 1, por lo que su comportamiento estaría explicado principalmente por las ventas y su variabilidad.

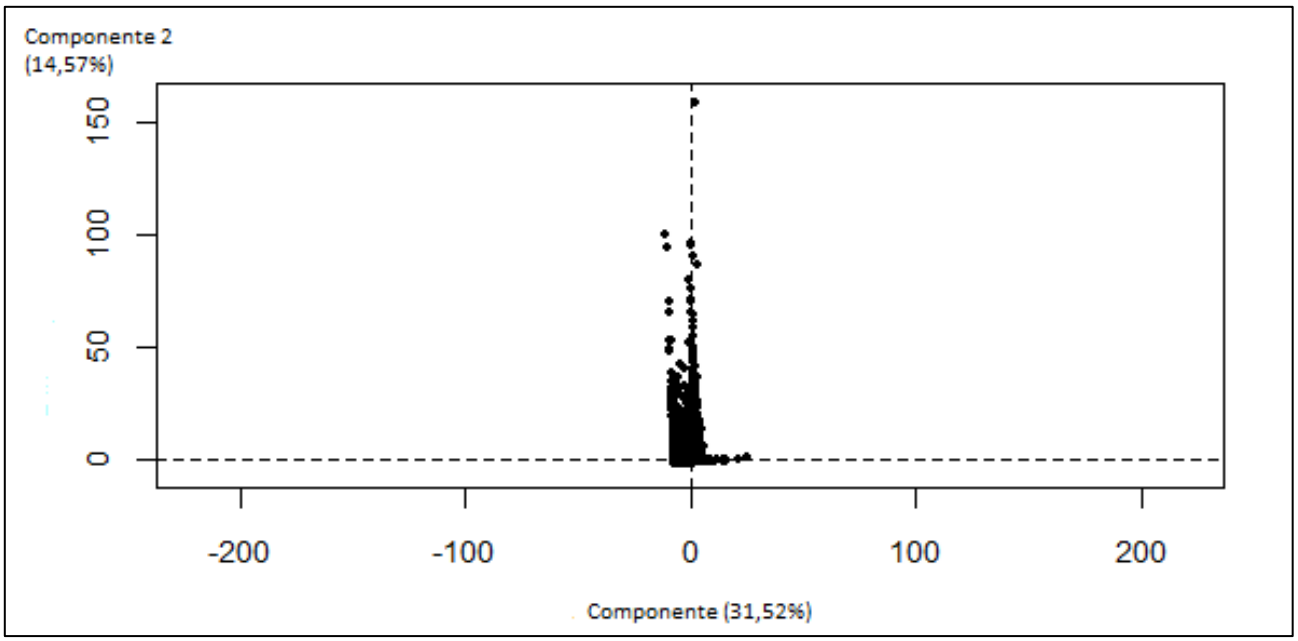

Ilustración 3.2. ACP - Gráfico de individuos.

Dado que, la multicolinealidad afectaría al modelo que se genere para segmentar a los contribuyentes; se realizó un clúster jerárquico de variables con el paquete ClustOfVar[6] en R (Chavent et al, 2017), para realizar una segunda selección de variables.

En este caso se utilizó el dendrograma, que muestra cómo se agrupan las variables de acuerdo a su nivel de similitud y; la matriz de correlación, verificando a aquellas que están fuertemente correlacionadas de forma lineal ($\rho \geq 0{,}75$ ó $\rho \leq -0{,}75$). Otro parámetro que se consideró para esta nueva selección fue la cantidad de datos perdidos que tiene cada variable.

[6] ClustOfVar es un paquete de R para realizar el agrupamiento de variables que están fuertemente relacionadas; las variables pueden ser cuantitativas, cualitativas o, mixtas. Por ello, este paquete es útil para la selección de variables y reducción de la dimensión de un conjunto de datos.

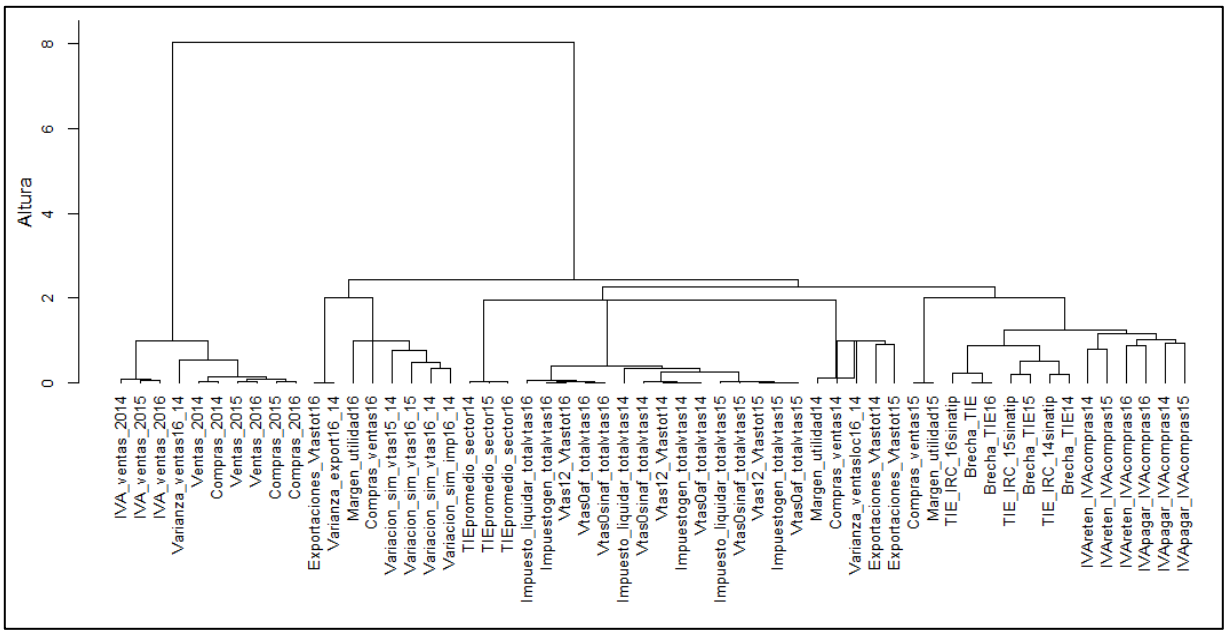

Ilustración 3.3. Dendograma de variables.

En este proceso se obtuvieron 18 variables para generar el modelo:

Tabla 3.6. Variables seleccionadas para el modelo

Variables	Tipo de variable	Unidad	Mínimo	1er Cuartil	Mediana	Promedio	3er Cuartil	Máximo
Ventas_2015	Continua	USD	-	40.279	177.312	1.617.715	727.515	466.681.819
IVA_ventas_2015	Continua	USD	-	215	7.429	113.121	42.682	56.001.818
Impuestogen_totalvtas15	Continua	%	0,0%	0,3%	11,9%	7,8%	12,0%	12,0%
Variacion_sim_vtas15_14	Continua	%	-200,0%	-23,8%	0,0%	2,0%	24,5%	200,0%
Variacion_sim_vtas16_15	Continua	%	-200,0%	-50,0%	-10,3%	-19,5%	12,7%	200,0%
Varianza_ventas16_14	Continua	USD	-	12.793	52.541	364.758	192.147	126.014.288
Brecha_TIE	Continua	%	-10%	-3%	0%	0%	3%	96%
Margen_utilidad14	Continua	%	-299900%	9,0%	31,5%	-171,4%	58,7%	100,0%
Margen_utilidad15	Continua	%	-275652%	10,8%	33,2%	-104,4%	59,9%	100,0%
Margen_utilidad16	Continua	%	-292594%	11,5%	35,5%	-164,3%	62,4%	100,0%
Variacion_sim_vtas16_14	Continua	%	-200,0%	-61,5%	-8,7%	-13,3%	29,4%	200,0%
Variacion_sim_imp16_14	Continua	%	-200,0%	-52,4%	0,0%	-8,7%	28,7%	200,0%
Contador_repetido	Dicotómica							
Represent_repetido	Dicotómica							
ATS_no_present	Dicotómica							
ATS_tardios	Dicotómica							
IVA_tardios	Dicotómica							
IVA_no_present	Dicotómica							

Fuente: SRI.

Cabe señalar que se tiene como referencia que, para la generación de un modelo de detección de fraude, se utiliza en promedio entre 10 y 15 variables según Baesens et al (2015). Adicionalmente, se excluyeron los datos atípicos para las variables margen de utilidad, quedando una base de 51.954 registros.

Tal como lo señalan algunos autores (York, 2011), para el análisis de riesgos de evasión de impuestos, el primer paso es la segmentación del universo de contribuyentes; lo cual permite realizar a la Administración Tributaria un tratamiento más adecuado de los riesgos y desarrollar estrategias de cumplimiento, para dar respuestas más efectivas a las necesidades de los contribuyentes.

Considerando esto y el análisis exploratorio efectuado, se segmentó la base preparada para generar el modelo según la actividad económica que se registra de acuerdo a la Clasificación Industrial Internacional Uniforme (CIIU versión 4.0). En la práctica, los contribuyentes deberían tener comportamientos similares en función del sector económico al que pertenecen. Los datos se distribuyen de la siguiente manera:

Tabla 3.7. Distribución de datos según la actividad económica.

Actividad económica	Código CIIU 4.0	Número de contribuyentes
Comercio	G	12.407
Actividades Profesionales, Científicas y Técnicas	M	5.702
Transporte y Almacenamiento	H	5.075
Manufactura	C	4.058
Actividades Inmobiliarias	L	3.772
Construcción	F	3.024
Agricultura, Ganadería, Silvicultura y Pesca	A	2.649
Salud y Asistencia Social	Q	1.840
Enseñanza	P	1.735
Información y Comunicación	J	1.528
Actividades Financieras y Seguros	K	1.313
Actividades Alojamiento y Alimentación	I	1.150

Otras actividades	7.701
Total	**51.954**

Fuente: SRI.

Para generar los modelos de Minería de Datos, se seleccionaron las 3 principales actividades económicas que cumplen con los siguientes parámetros: mayor concentración de contribuyentes y mayor aporte a la recaudación del impuesto al valor agregado. Las actividades que cumplen con estos criterios son: Comercio; Manufactura y; Actividades Profesionales, Científicas y Técnicas; siendo excluido Transporte y Almacenamiento.

Con las tres actividades económicas y las variables definidas; se escaló cada uno de los subconjuntos.

4.2. EVALUACIÓN DE LA TENDENCIA DE AGRUPAMIENTO

Previo al entrenamiento de los modelos de segmentación, se realizó la evaluación de la tendencia de agrupamiento, para confirmar que los datos son agrupables; es decir, que las estructuras no son aleatorias y, por lo tanto, que el análisis de segmentación será válido. Para ello, con el paquete clustertend (YiLan & RuTong, 2015), se realizó la prueba de aleatoriedad espacial de los datos, haciendo uso del estadístico de Hopkins (Tan, Steinbach, Karpatne, & Kumar, 2019). Los valores obtenidos de este estadístico para cada uno de los subconjuntos son:

Tabla 3.8. Estadístico de Hopkins

Actividad Económica	**Estadístico de Hopkins**
Comercio	0,0044
Manufactura	0,0275
Actividades Profesionales, Científicas y Técnicas	0,0171

Considerando que la hipótesis nula de esta prueba es que los datos no son agrupables y, que se cumple si los dos conjuntos de datos (datos generados y datos originales) tienen distancias promedio similares al vecino más cercano del conjunto original de datos y, por lo tanto, el estadístico *H* será cercano a 0,5. En este sentido, el valor del estadístico de Hopkins obtenido para cada uno de los sectores económicos, confirma que contienen algún tipo de agrupación.

4.3. GENERACIÓN DEL MODELO Y EVALUACIÓN DE LA CALIDAD DE LOS GRUPOS OBTENIDOS

Una vez que se confirmó la validez de realizar el análisis de segmentación, se realizó el entrenamiento de los mapas auto-organizados de Kohonen con el algoritmo multi-SOM (Chair & Charrad, 2017), para cada actividad económica. De esta manera, se pueden analizar patrones de comportamiento, para detectar posibles riesgos de evasión tributaria.

El paquete 'multisom' del lenguaje R, incluye el algoritmo multi-SOM que permite realizar la segmentación de un conjunto de datos, determinar el mejor número de agrupamientos y, además obtener el mejor esquema de segmentación a partir de diferentes resultados.

En primera instancia se entrena el modelo con el algoritmo SOM (mapa original). A partir del mapa original, mediante un proceso de generalización en línea; de forma iterativa se generan otros niveles de agrupación, disminuyendo gradualmente el número de neuronas, hasta que la dimensión del mapa es de 4 neuronas.

En el paquete hay dos versiones del algoritmo multi-SOM: estocástico y por lotes. Cabe señalar que, en cada nivel se calculan los índices de validación; a

partir de los cuales se puede evaluar el número de grupos óptimo. Para el análisis se ejecutaron las 2 versiones de este algoritmo.

Para el entrenamiento del conjunto de datos con el algoritmo multi-SOM estocástico se emplearon los siguientes parámetros:

a. Tamaño del mapa o vecindad

    - Filas x Columnas = [10x10, 9x9, 8x8, 7x7, 6x6, 5x5, 4x4, 3x3, 2x2]

b. Topología del mapa = Hexagonal
c. Velocidad de aprendizaje = [0,05 0,01 0,075 0,1]
d. Número máximo de iteraciones = 500
e. Índices de validación = Dunn; Silhouette; Davies y Bouldin (DB); Índice C, Calinski y Harabasz (CH), Ball, Hartigan, Índice SDbw.

En tanto que, para el entrenamiento con el algoritmo multi-SOM por lotes se emplearon los siguientes parámetros:

f. Tamaño del mapa o vecindad

    - Filas x Columnas = [10x10, 9x9, 8x8, 7x7, 6x6, 5x5, 4x4, 3x3, 2x2]

g. Topología del mapa = Hexagonal
h. Radio mínimo de la vecindad = 0,00013
i. Radio máximo de la vecindad = 0,002
j. Número máximo de iteraciones = 500
k. Índices de validación = Dunn; Silhouette; Davies y Bouldin (DB); Índice C, Calinski y Harabasz (CH), Ball, Hartigan, Índice SDbw.

Cabe mencionar que, en este estudio se realizó únicamente la validación interna; dado que, no se cuenta con la información de la clase y, por lo tanto, no es posible realizar la validación externa.

En esta validación se utiliza la información interna de la agrupación para evaluar la bondad de una estructura de agrupamiento y, estimar el número de agrupaciones óptimo.

Para realizar el análisis e interpretación de los valores de cada uno de los índices, se consideraron los criterios de la tabla 2.9.

Tabla 3.9. Índices de validación interna

Índice	Número óptimo de grupos	Rango
Dunn	Máximo valor del índice	[0 , + α)
Silhouette	Máximo valor del índice	[-1 , + 1]
Davies y Bouldinv (DB)	Mínimo valor del índice	
Índice C	Mínimo valor del índice	[0 , + 1]
Calinski y Harabasz (CH)	Máximo valor del índice	
Ball y Hall	Máxima diferencia entre niveles de jerarquía del índice	
Hartigan	Máxima diferencia entre niveles de jerarquía del índice	
Índice SDbw	Mínimo valor del índice	

Fuente: (Tan, Steinbach, Karpatne, & Kumar, 2019); (Khanchouch, Charrad, & Limam, 2015).

1.1.1. Resultados actividad económica Comercio

Con el algoritmo multi-SOM estocástico; se obtuvieron los siguientes resultados:

Tabla 3.10. Resultados multi-SOM estocástico Comercio

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,001	0,004	3,258	0,036	165,022	1.234,340	0,119	1,142
9x9	0,183	0,098	0,490	0,350	17,777	0,189	3,351	0,028
8x8	0,235	0,102	0,560	0,319	26,118	0,122	3,213	0,024
7x7	0,229	0,139	0,483	0,201	31,118	0,040	3,470	0,013
6x6	0,182	0,119	0,393	0,117	30,782	0,015	3,510	0,011
5x5	0,164	0,154	0,479	0,079	26,915	0,010	3,121	0,019
4x4	0,315	0,185	0,483	0,056	19,867	0,006	2,653	0,029
3x3	0,370	0,202	0,398	0,033	12,713	0,003	1,850	0,043
2x2	0,947	0,601	0,384	0,015	22,224	0,001	1,155	0,146

En tanto que, al ejecutar el algoritmo multi-SOM por lotes se obtuvieron estos resultados:

Tabla 3.11. Resultados multi-SOM por lotes Comercio

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,003	- 0,013	1,144	0,078	560,674	444,290	1,455	0,722
9x9	0,175	- 0,143	2,075	0,262	61,861	13,757	4,024	0,007
8x8	0,008	NA	1,281	0,312	40,697	30,967	3,656	0,012
7x7	0,160	- 0,234	0,684	0,366	23,738	31,901	3,520	0,016
6x6	0,218	- 0,238	0,648	0,405	17,703	48,901	2,957	0,029
5x5	0,161	- 0,278	0,619	0,570	22,512	41,016	3,171	0,025
4x4	0,085	- 0,290	0,844	0,350	11,411	127,443	2,268	0,050
3x3	0,537	- 0,375	0,486	0,500	31,649	64,427	3,049	0,025
2x2	0,501	- 0,093	0,561	0,696	16,349	158,563	2,283	0,076

Considerando los criterios que se deben verificar en los índices (tabla 2.9), para determinar el número de grupos óptimos y, los rangos de valores que toman; al comparar, los resultados obtenidos tanto con el algoritmo multi-SOM estocástico, como con el algoritmo multi-SOM por lotes; se tiene que, en 4 de los 8 índices (Dunn, Silhouette, DB, Índice C), 2 es el número óptimo para segmentar el sector Comercio (Tabla 2.12); en tanto que, 2 de los 8 índices (Ball y Hartigan) determinan un número de 57 grupos.

Tabla 3.12. Número de grupos óptimos Comercio

Algoritmo	Resultados	Dunn	Silhouette	DB	Índice C	CH	Ball	Harti-gan	SDbw
Multi-SOM estocástico	Número de grupos	**2**	**2**	**2**	**2**	85	**57**	**57**	15
	Valor del índice	**0,947**	**0,601**	**0,384**	**0,015**	165,022	**1.234,151**	**3,232**	0,011
Multi-SOM por lotes	Número de grupos	6	95	6	95	**95**	45	45	**45**
	Valor del índice	0,537	- 0,013	0,486	0,078	**560,674**	430,533	2,569	**0,007**

1.1.2. <u>Resultados actividad económica Manufactura</u>

En el caso del sector Manufactura, con el algoritmo multi-SOM estocástico; se obtuvieron los siguientes resultados:

Tabla 3.13. Resultados multi-SOM estocástico Manufactura

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,002	0,015	2,320	0,051	75,394	435,690	0,326	0,753
9x9	0,083	0,013	0,594	0,196	12,362	0,506	2,575	0,035
8x8	0,236	0,168	0,427	0,347	26,771	0,069	3,590	0,016
7x7	0,235	0,056	0,484	0,191	29,375	0,025	3,799	0,012
6x6	0,366	0,193	0,429	0,087	41,041	0,013	3,798	0,011
5x5	0,465	0,239	0,417	0,067	34,234	0,009	3,476	0,014
4x4	0,536	0,221	0,417	0,041	21,954	0,007	2,753	0,028
3x3	0,460	0,201	0,553	0,019	11,570	0,004	1,755	0,088
2x2	0,518	0,170	0,499	0,008	7,062	0,001	0,856	0,171

Mientras que, al ejecutar el algoritmo multi-SOM por lotes se obtuvieron estos resultados:

Tabla 3.14. Resultados multi-SOM por lotes Manufactura

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,010	- 0,023	1,359	0,105	145,145	175,120	1,260	0,525
9x9	0,145	- 0,191	0,841	0,286	42,251	5,758	3,764	0,010
8x8	0,140	- 0,262	1,246	0,436	46,538	9,127	3,790	0,010
7x7	0,243	NA	0,883	0,539	50,175	5,491	4,334	0,006
6x6	0,171	- 0,138	0,763	0,297	23,733	24,496	3,084	0,025
5x5	0,141	- 0,278	0,839	0,301	23,759	28,603	3,111	0,017
4x4	0,281	- 0,320	0,668	0,346	21,467	32,676	2,900	0,031
3x3	0,295	- 0,284	0,838	0,612	7,125	78,944	1,830	0,110
2x2	0,344	- 0,333	1,032	0,613	3,502	179,342	0,743	0,504

Los resultados son diversos en cuanto a número de grupos óptimos, como se puede evidenciar en la tabla 2.15. El índice DB y silhouette coinciden en que el número óptimo es 8, comparando los dos tipos de algoritmo y los valores de los índices obtenidos.

Tabla 3.15. Número de grupos óptimos Manufactura

Algoritmo	Resultados	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
Multi-SOM es-	Número de grupos	**4**	**8**	**8**	**2**	74	**40**	40	14
tocástico	Valor del índice	**0,536**	**0,239**	**0,417**	**0,008**	75,394	**435,184**	2,250	0,011
Multi-SOM	Número de grupos	2	97	6	97	**97**	49	**49**	**20**
por lotes	Valor del índice	0,344	-0,023	0,668	0,105	**145,145**	169,362	**2,503**	**0,006**

1.1.3. Resultados actividad económica Servicios Profesionales

Para la actividad de Profesionales, Científicas y Técnicas, los resultados generados por el algoritmo multi-SOM estocástico son:

Tabla 3.16. Resultados multi-SOM estocástico Servicios Profesionales

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,001	0,011	2,443	0,043	97,458	508,137	0,355	1,022
9x9	0,172	0,053	0,507	0,289	15,433	0,181	3,340	0,025
8x8	0,330	0,101	0,440	0,263	23,528	0,059	3,669	0,016
7x7	0,315	0,136	0,428	0,153	30,520	0,030	3,706	0,013
6x6	0,361	0,103	0,427	0,109	28,993	0,017	3,534	0,012
5x5	0,434	0,253	0,452	0,058	31,923	0,013	3,176	0,018
4x4	0,396	0,186	0,383	0,056	19,635	0,007	2,641	0,027
3x3	0,559	0,130	0,530	0,014	10,444	0,003	1,941	0,070
2x2	0,494	0,480	0,351	0,008	15,298	0,001	0,782	0,143

En tanto que, al ejecutar el algoritmo multi-SOM por lotes se obtuvieron los siguientes resultados:

Tabla 3.17. Resultados multi-SOM por lotes Servicios Profesionales

Tamaño mapa	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
10x10	0,009	- 0,017	1,200	0,128	301,891	184,437	1,603	0,537
9x9	0,185	- 0,151	1,394	0,431	54,531	10,921	3,776	0,006
8x8	0,161	- 0,212	1,100	0,334	32,225	13,966	3,573	0,016
7x7	0,241	- 0,224	0,541	0,509	32,595	7,116	4,107	0,010
6x6	0,304	- 0,306	0,798	0,464	26,078	20,176	3,345	0,019
5x5	0,130	- 0,169	0,641	0,429	10,331	38,226	2,623	0,034
4x4	0,125	- 0,237	0,561	0,476	12,739	31,745	2,881	0,039
3x3	0,212	- 0,254	0,667	0,461	16,759	77,989	2,413	0,056
2x2	0,521	- 0,222	0,507	0,713	15,404	122,331	2,224	0,050

En esta actividad económica, analizando los criterios para determinar el número de grupos óptimos para cada índice y comparando los dos tipos de algoritmo, se tiene que 3 de los 8 índices (Silhouette, DB, Índice C) proponen que 2 es el número óptimo y; 2 índices (Ball y Hartigan) proponen que 57 es el mejor número de grupos (Tabla 2.18).

Tabla 3.18. Número de grupos óptimos Servicios Profesionales

Algoritmo	Resultados	Dunn	Silhouette	DB	Índice C	CH	Ball	Hartigan	SDbw
Multi-SOM	Número de grupos	**4**	**2**	**2**	**2**	83	**57**	**57**	18
estocástico	Valor del índice	**0,559**	**0,48**	**0,351**	**0,008**	97,458	**507,957**	**2,985**	0,012
Multi-SOM	Número de grupos	3	93	3	93	**93**	41	41	**41**
por lotes	Valor del índice	0,521	-0,017	0,507	0,128	**301,891**	173,517	2,173	**0,006**

4.4. VISUALIZACIÓN DE AGRUPACIONES MEDIANTE MAPAS AUTO-ORGANIZADOS DE KOHONEN

Una vez, que se han ejecutado los algoritmos y se ha evaluado el número de grupos obtenidos; se tomó como referencia estos resultados para visualizar los datos mediante los mapas auto-organizados de Kohonen. Para ello, se utilizaron los paquetes kohonen (Wehrens & Kruisselbrink, 2018) y, RColorBrewer (Neuwirth, 2015).

Para el entrenamiento del modelo SOM se emplearon los mismos parámetros aplicados en los algoritmos multi-SOM estocástico y por lotes. La configuración es:

a. Tamaño del mapa o vecindad, considerando el número de grupos óptimos que se generaron para cada actividad económica.
   - Comercio = 5x4
   - Manufactura = 5x5
   - Servicios Profesionales = 4x4

b. Velocidad de aprendizaje = [0,05 0,01 0,075 0,1]

c. Número máximo de iteraciones = 500.

4.4.1. Progreso del entrenamiento de la red neuronal

A medida que avanzan las iteraciones del entrenamiento del mapa auto-organizado, disminuye la distancia de los pesos de cada neurona a las observaciones representados por dicha neurona. Lo adecuado es que esta distancia alcance una meseta mínima.

Cabe señalar que, con la ejecución de las funciones disponibles en el paquete kohonen, en base a las distintas combinaciones de la configuración de parámetros del algoritmo mencionados anteriormente, respecto al tamaño del mapa por sector, los distintos valores de la velocidad de aprendizaje y, el número máximo de iteraciones que es igual a 500; se obtiene como un solo resultado la distancia total que es un promedio ponderado de las distancias de capa, que se generan.

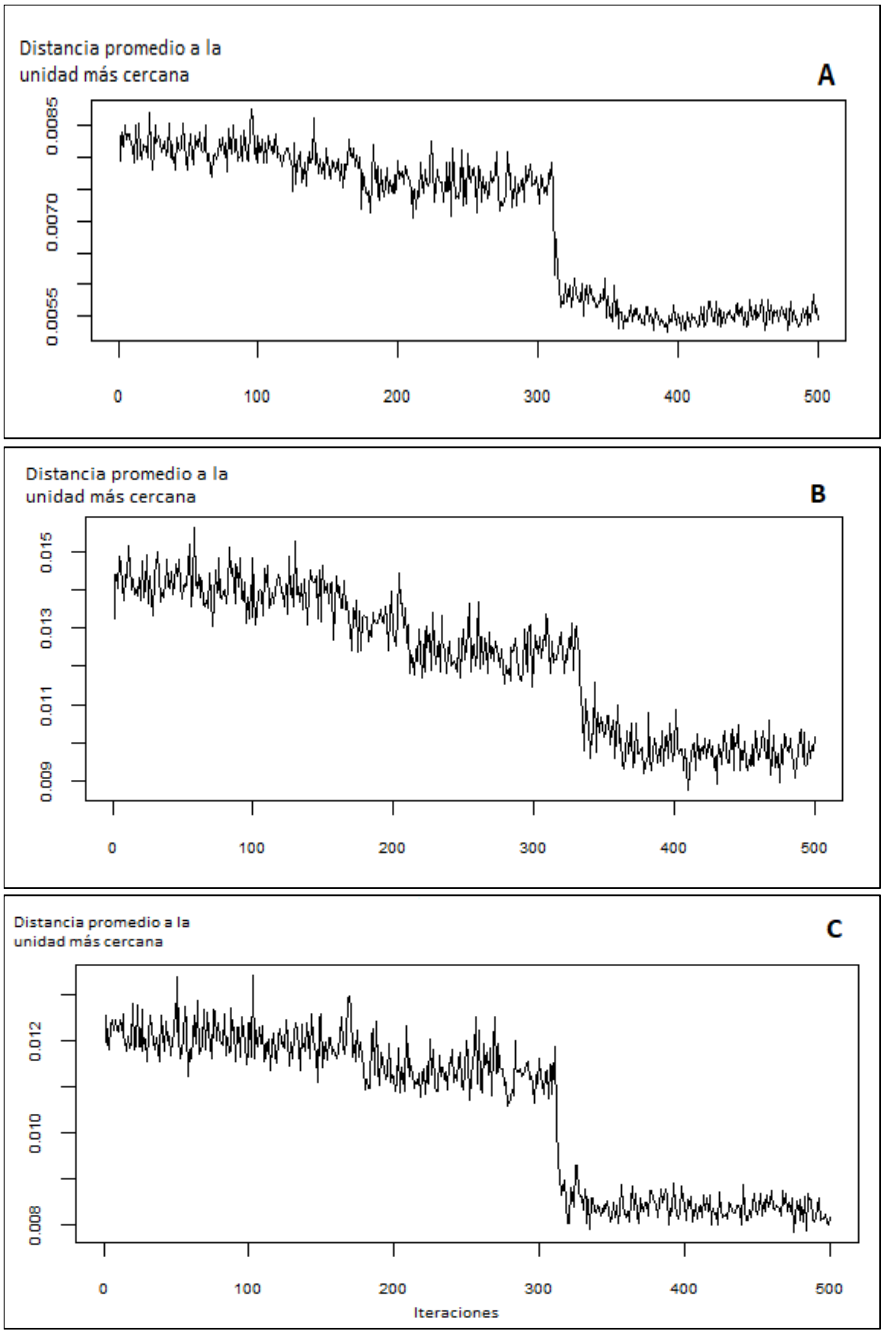

Ilustración 3.4. Progreso del entrenamiento. (A) Comercio; (B) Manufactura; (C) Profesionales, Científicas y Técnicas.

Como se observa en la ilustración 2.4; en cada uno de los entrenamientos, la distancia disminuye continuamente hasta estabilizarse después de las 300 iteraciones; lo que significa que no se requiere de un mayor número de iteraciones para el entrenamiento.

4.4.2. Conteo de individuos

En esta visualización (ilustración 2.5), se puede evidenciar el número de individuos que se ubican en cada neurona del mapa auto-organizado para cada una de las actividades económicas; de acuerdo a la similitud de sus características.

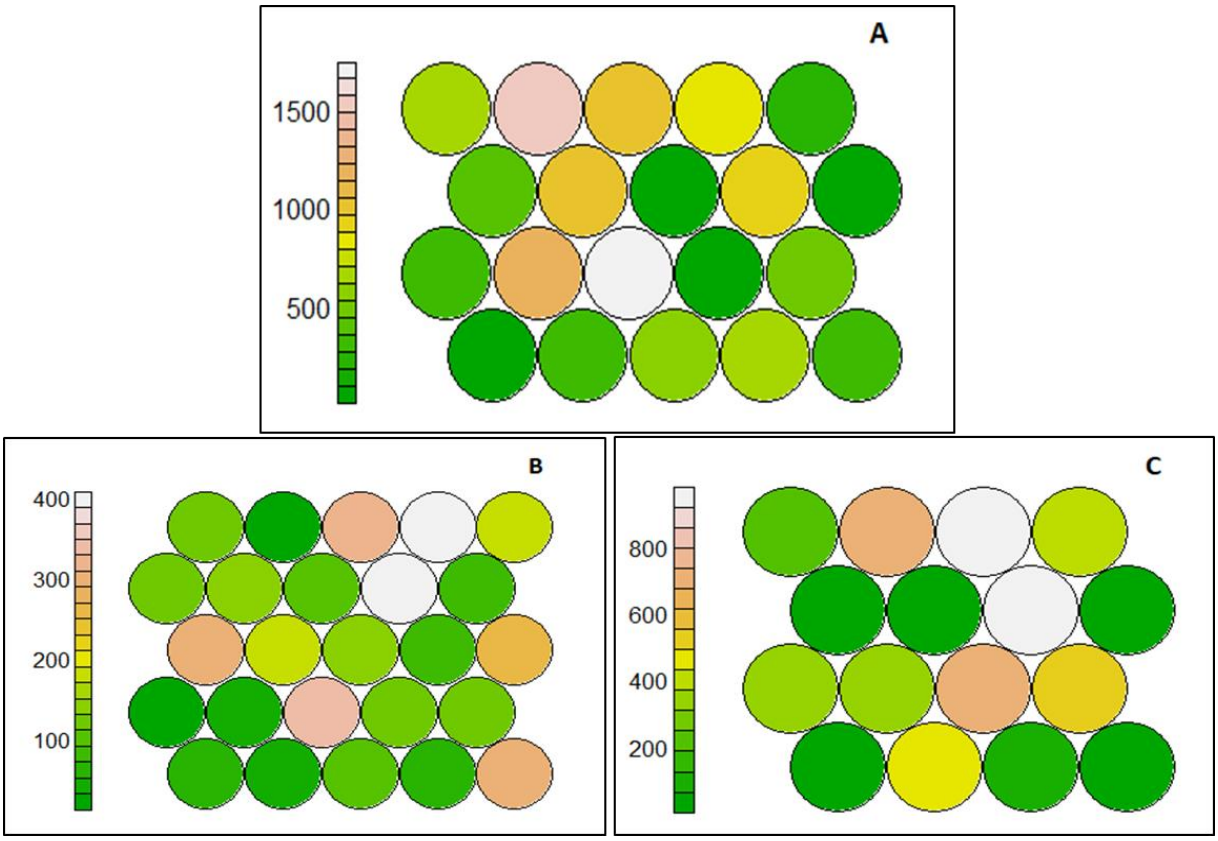

Ilustración 3.5. Conteo de individuos. (A) Comercio; (B) Manufactura; (C) Profesionales, Científicas y Técnicas.

4.4.3. Distancia de la vecindad

La matriz de distancias, a menudo referida como matriz-U (Stefanovic & Kurasova, 2011), es una visualización de los mapas auto-organizados que representa la distancia entre cada neurona y sus vecinas. Por lo tanto, a partir de

esta representación se puede analizar las similitudes entre los datos, las áreas de menor distancia entre neuronas vecinas indican que hay similitud; mientras que, las áreas con grandes distancias indican disimilitud. En este sentido, se evidencian agrupaciones naturales de los datos.

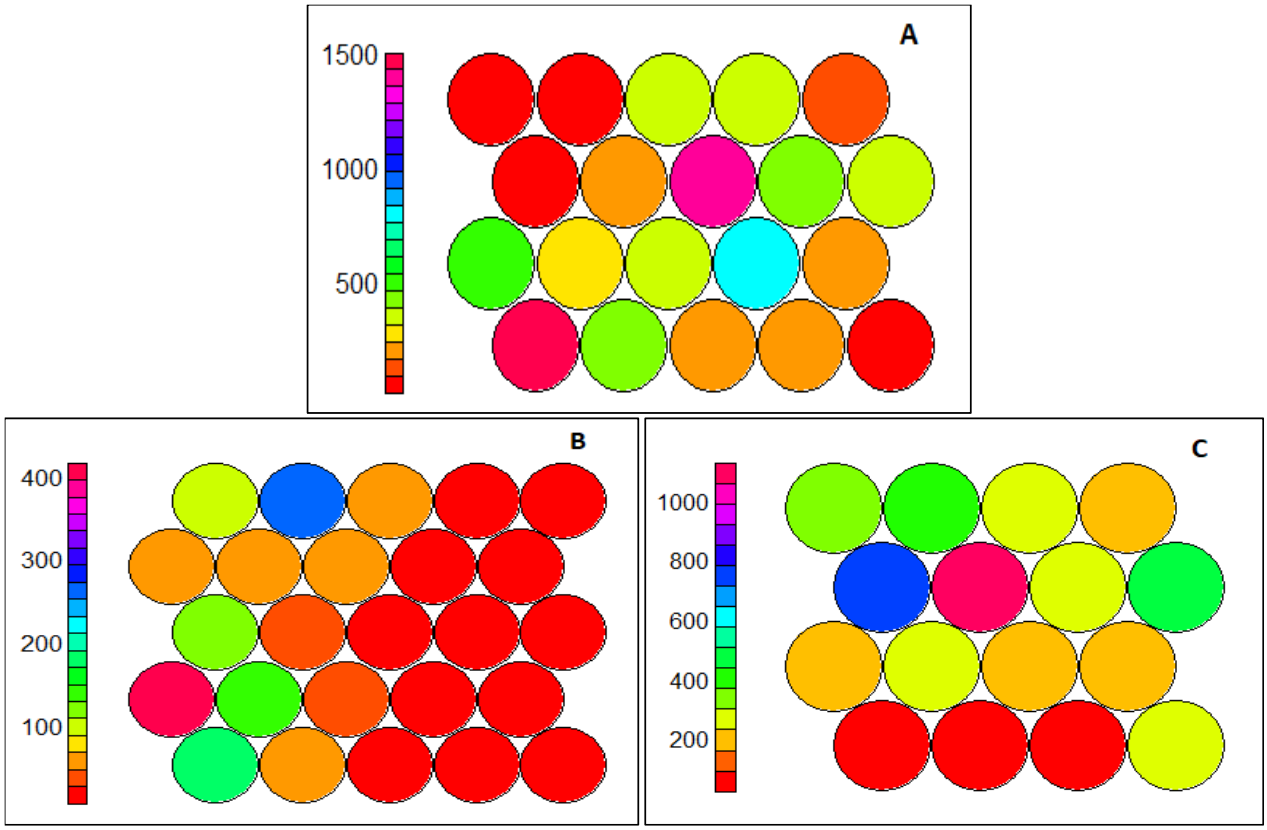

Ilustración 3.6. Distancias de la vecindad. (A) Comercio; (B) Manufactura; (C) Profesionales, Científicas y Técnicas.

4.4.4. Visualización de agrupaciones

Estas visualizaciones fueron generadas con el número de grupos obtenidos a partir de los algoritmos multi-SOM estocástico y por lotes. Para el sector Comercio, con la ejecución del algoritmo multi-SOM estocástico y según los índices

Dunn, Silhouette, Davies y Bouldin e, Índice C, el número de grupos óptimo es 2. Mientras que, con el algoritmo multi-SOM por lotes y según los índices Dunn y, Davies y Bouldin el mejor número es 6 (ilustración 2.7).

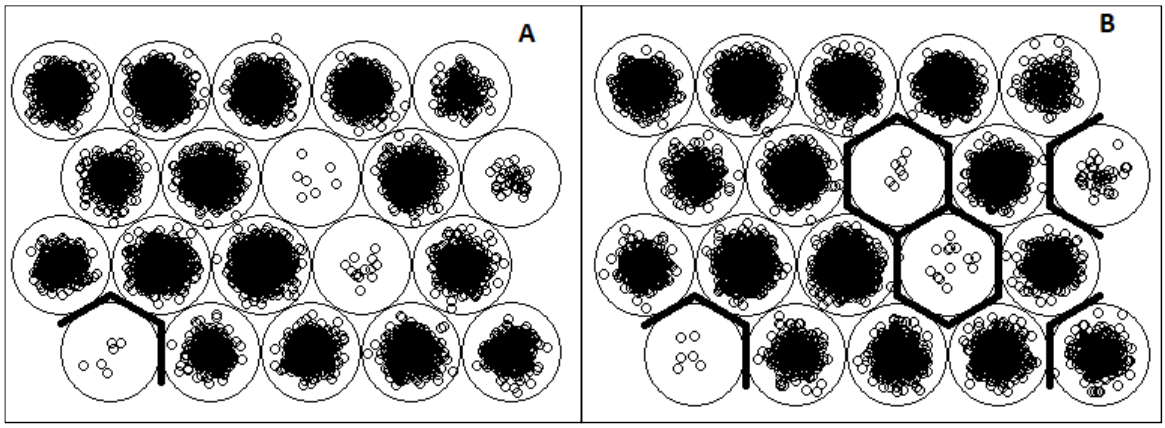

Ilustración 3.7. Agrupaciones sector Comercio. (A) Dos grupos; (B) Seis grupos.

Para el sector Manufactura, el mejor número de grupos según los índices Dunn (4 grupos), Silhouette (8 grupos) y, Davies y Bouldin (8 grupos) se obtuvo con el algoritmo multi-SOM estocástico (ilustración 2.8).

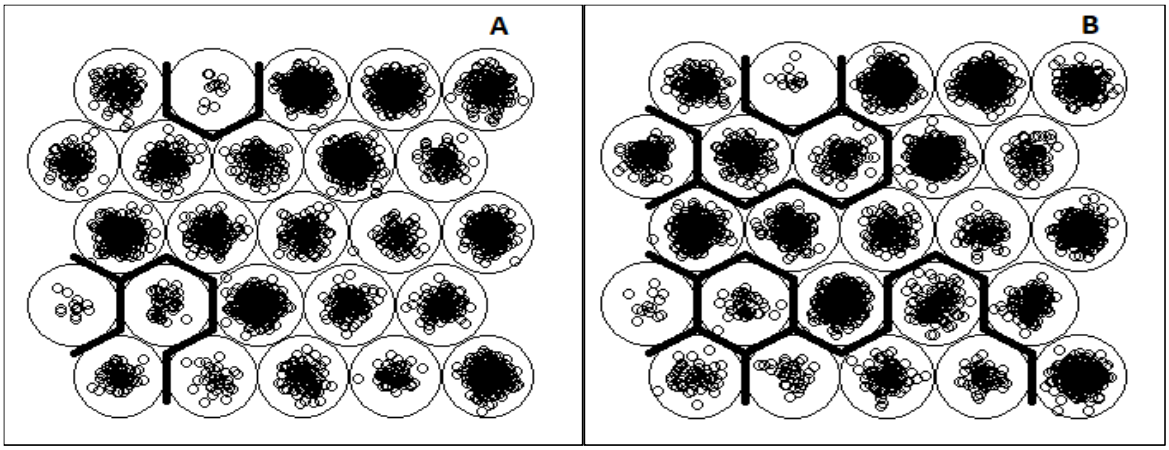

Ilustración 3.8. Agrupaciones sector Manufactura. (A) Cuatro grupos; (B) Ocho grupos.

Para el sector Servicios Profesionales, el mejor número de grupos según los índices Silhouette, Davies y Bouldin e, Índice C es 2 y, según el índice Dunn es 4 (ilustración 2.9).

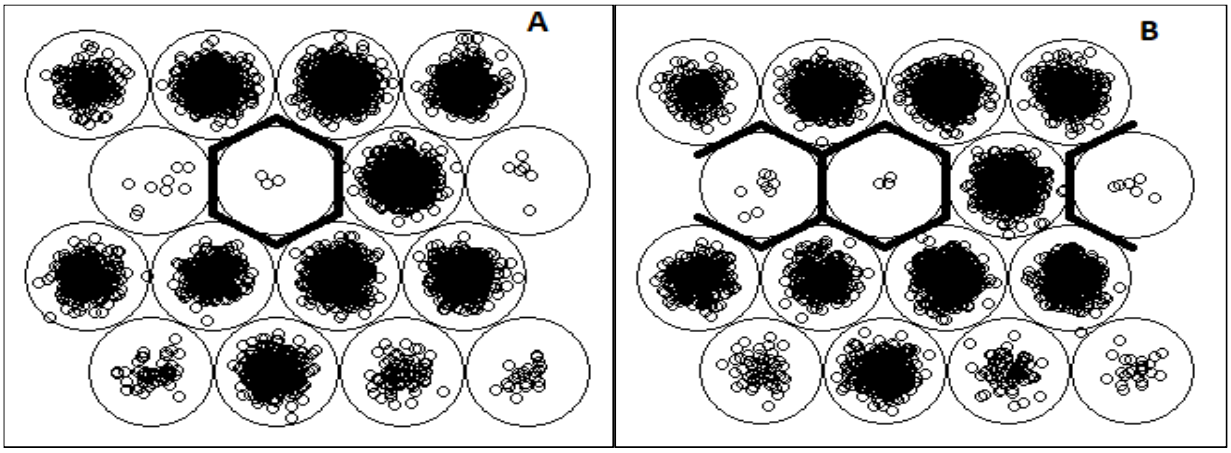

Ilustración 3.9. Agrupaciones sector Servicios Profesionales. (A) Dos grupos; (B) Cuatro grupos.

Considerando el mayor número de grupos mostrados en las visualizaciones anteriores; los contribuyentes se distribuyen de la siguiente manera (tabla 2.19).

Tabla 3.19. Distribución de contribuyentes

Actividad económica	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Total
Comercio	11,977	43	7	6	349	12			12,394
Manufactura	62	324	3,102	13	39	117	379	12	4,048
Servicios Profesionales	5,668	9	3	6					5,686

5. RESUMEN

En este capítulo se generó el modelo de segmentación para detección de potenciales contribuyentes evasores del impuesto al valor agregado, mediante el desarrollo de la metodología CRISP-DM, que va desde la comprensión del negocio hasta la evaluación del modelo, considerando el alcance de esta investigación.

Es así que, una vez que se revisó la información del sistema tributario, la estructura del impuesto al valor agregado, los segmentos de contribuyentes, las distintas fuentes de información, experiencias internacionales y criterios de experto; se obtuvieron millones de registros administrativos de diversa naturaleza. A partir de ello, se realizaron los filtros y depuraciones respectivas, para seleccionar las variables y

generar otros indicadores como variaciones y ratios. Una vez que se tuvo lista la base para el análisis, se efectuó el siguiente procedimiento: análisis exploratorio de los datos univariado y multivariado mediante análisis de componentes principales y clúster jerárquico de variables; evaluación de la tendencia de agrupamiento haciendo uso del estadístico de Hopkins; generación del modelo de segmentación mediante el entrenamiento de mapas auto-organizados de Kohonen con el algoritmo multi-SOM y; evaluación de la calidad de los grupos obtenidos, con los índices de validación como Dunn; Silhouette, entre otros. Por último, se realizó un ejercicio de visualización del número de grupos óptimo obtenidos con mapas auto-organizados de kohonen.

5 CAPÍTULO III

1. DISCUSIÓN Y CONCLUSIONES

Al analizar cada uno de los grupos y los contribuyentes que los conforman, es interesante evidenciar que, con el modelo de segmentación, efectivamente se identificaron potenciales contribuyentes evasores del impuesto al valor agregado, que es el objetivo de esta investigación. Los hallazgos encontrados son los siguientes:

En el sector Manufactura, se identificó potenciales riesgos en el grupo 1 y 5, que son empresas grandes, cuya antigüedad promedio es de 28 años; para las cuales se evidencian los siguientes aspectos:

- Presentan disminuciones significativas de las ventas de 2016 en relación a las ventas de 2014, lo que no sucede con otras empresas del mismo tamaño y sector económico.
- La varianza promedio de las ventas, de igual forma, es mucho más alta que el resto de empresas grandes que se ubican en otros grupos.
- En el grupo 1, hay una brecha promedio negativa del tipo impositivo efectivo (impuesto/ingresos) cercana al -1%, en relación al tipo impositivo efectivo promedio del sector, siendo esta medida un indicador referencial de cuál debería ser la contribución tributaria esperada de las empresas.
- Cabe señalar que, en estos grupos no se observa mayor cambio en el margen de utilidad.

En tanto que, en los grupos 4, 6, 7 y 8, en los que se ubican micro y pequeñas empresas, se observa el siguiente comportamiento:

- Registran una disminución drástica de ventas y, por ende, del impuesto al valor agregado.

- Son empresas con una antigüedad promedio de 13 años.
- Varias empresas no presentaron anexos transaccionales y, declaraciones de impuestos al valor agregado.
- Se observa una importante disminución del margen de utilidad y, en diversas empresas siempre se registran una rentabilidad negativa.

Similar comportamiento se identificó en el sector Servicios Profesionales (conformado por 4 grupos). En este sector, los grupos 2, 3 y 4 están conformados por micro y pequeñas empresas; en los que se registra:

- Una reducción significativa de las ventas.
- Una brecha promedio del tipo impositivo de -3% respecto al tiempo impositivo efectivo de este sector económico.
- Rentabilidades negativas.

Mientras que, en el grupo 1, se concentra la mayoría de datos; por lo que para efectos de aplicación; se podría realizar un mayor desglose de número de grupos de esta información, de acuerdo a los resultados obtenidos a partir de los índices de validación.

En el caso del sector Comercio, conformado por 6 grupos; el grupo 2 está constituido por empresas grandes con las siguientes características:

- En promedio registran una mayor disminución de las ventas.
- Se observa una importante disminución del margen de utilidad en 2016, en relación a 2014.

En tanto que, los grupos 3, 5 y 6 están conformados por micro y pequeñas empresas; en los que se observa principalmente una importante reducción de sus ventas, anexos y declaraciones no presentadas. De igual manera, el grupo 1 es el que tiene el mayor número de contribuyentes, por lo que se podría realizar una mayor desagregación de grupos.

En primera instancia, en el sector de Manufactura se obtuvieron más casos para revisión por potenciales riesgo de evasión. Del modelo de segmentación generado para la detección de potenciales contribuyentes evasores del impuesto al valor agregado y, del análisis de los diferentes grupos, se han identificado que hay características propias de las empresas grandes y medianas que permiten evidenciar este potencial riesgo como, por ejemplo, que no se observa mayor cambio en el margen de rentabilidad, pero si en otras variables. En tanto que, en las micro y pequeñas empresas, una de las principales características es la disminución significativa en el margen de utilidad y la no presentación de declaraciones y anexos.

En función de las agrupaciones obtenidas para cada uno de los tres sectores económicos y, las características identificadas se tienen listados de potenciales contribuyentes riesgosos, que pueden ser parte de un proceso de revisión por parte de la Administración Tributaria. De igual forma, esta información puede ser utilizada para perfilamiento de contribuyentes.

Se puede realizar un análisis más extensivo con los distintos resultados obtenidos con la aplicación del algoritmo multi-SOM y la evaluación del número de grupos óptimos con los índices de validación. En este caso, el análisis se realizó considerando el mayor número de coincidencias en el número de grupos óptimo en función de los índices utilizados y, verificando que se cumplan los criterios de cada índice. Sin embargo, otro aspecto que se puede considerar es que de los índices utilizados, SDbw en comparación con los índices Dunn, Silhouette, Davies y Bouldin, Calinski y Harabasz es la única medida que tiene un buen funcionamiento ante los siguientes aspectos: monotonicidad, ruido, densidad, subgrupos y distribuciones sesgadas, según evidencias comprobadas en otro estudio (Liu, Li, Xiong, Gao, & Wu, 2010).

Respecto a las técnicas utilizadas, según varios autores (Ghouila, y otros, 2009), (Lu & Segall, 2013) y (Khanchouch, Charrad, & Limam, 2015), el algoritmo multi-

SOM en comparación con otros algoritmos de segmentación clásico es más eficiente y confiable para determinar el número de grupos óptimo; además de tener la ventaja de funcionar bien con conjuntos de datos de alta dimensionalidad. En este caso, el procesamiento fue rápido y su rendimiento óptimo.

Con el algoritmo multi-SOM por lotes, se evidenció que no funciona bien el coeficiente Silhouette; en todos los casos es negativo y, esto no es deseable porque quiere decir que la distancia promedio de los puntos dentro de un grupo, es mayor a la distancia promedio mínima a los puntos de otro grupo.

La siguiente etapa de este trabajo, es proporcionar el listado de potenciales contribuyentes evasores del impuesto al valor agregado al área de Control de la Administración Tributaria de Ecuador y, el perfilamiento realizado, para la verificación y controles respectivos; así como, la identificación de futuras líneas de investigación que se pueden desprender de este estudio.

Una investigación futura puede enfocarse en un análisis de segmentación difuso, análisis de redes o, la aplicación de otros algoritmos para detección de anomalías.

A. REFERENCIAS BIBLIOGRÁFICAS

1. Alm, J., Blackwell, C., & McKee, M. (2004). Audit Selection and Firm Compliance with a Broad–based Sales Tax. *National Tax Journal*, 209-227.
2. Arias, R. (2004). Reglas de selección para la fiscalización de Impuestos a lasVentas. *Revista de Economía y Estadística*, Vol. 42, No. 2, pp. 29-62.
3. Castellón González, P., & Velásquez, J. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications 40*, 1427–1436.
4. Chair, S., & Charrad, M. (2017). multisom: Clustering a Data Set using Multi-SOM Algorithm. R package version 1.3.
5. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0. Step-by-step data mining guide.*
6. Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*.
7. Chavent, M., Kuentz, V., Liquet, B., & Saracco, J. (2017). ClustofVar: Clustering of Variables. R package version 1.1.
8. Denny, D., Graham J., W., & Peter, C. (2007). *Exploratory Multilevel Hot Spot Analysis: Australian Taxation Office Case Study.*
9. Ghouila, A., Yahia, S. B., Malouche, D., Jmel, H., Laouini, D., Guerfali, F. Z., & Abdelhak, S. (2009). Application of Multi-SOM clustering approach to macrophage gene expression analysis. *Infection, Genetics and Evolution*, 328–336.
10. Gupta, M., & Nagadevara, V. (s.f.). Audit Selection Strategy for Improving Tax Compliance – Application of Data Mining Techniques. 378-387.
11. Hartigan, J. (1975). *Clustering algorithms.*
12. Isasi Viñuela, P., & Galván León, I. (2004). *Redes de Neuronas Artificiales. Un enfoque práctico.* Madrid: Pearson Educación.

13. Kassambara, A., & Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package versión 1.0.5.

14. Khanchouch, I., Charrad, M., & Limam, M. (Junio de 2015). A Comparative Study of Multi-SOM Algorithms for Determining the Optimal Number of Clusters. *International Journal of Future Computer and Communication, 4*(3), 198-202.

15. Khwaja, M. S., Awasthi, R., & Loeprick, J. (2012). *Risk-Based Tax Audits: Approaches and Country Experiences.* Washington D.C.

16. Kohonen, T. (1981). Automatic formation of topological maps of patterns in a self-organizing system. *in Proc. the 2SCIA, Scand. Conference on Image Analysis*, (págs. 214-220).

17. Lamirel, J.-C. (2002). MultiSOM : a multimap extension of the SOM model. Application to information discovery in an iconographic context.

18. Lotfi Shahreza, M., Moazzami, D., Moshiri, B., & Delavar, M. (2011). Anomaly detection using a self-organizing map and particle swarm optimization. *Scientia Iranica*, 1460–1468.

19. Lu, S., & Segall, R. (2013). Multi-SOM: an Algorithm for High-Dimensional, Small Size Datasets. *Systemics, Cybernetics and Informatics*, 41-46.

20. Mitsa, T. (2010). *Temporal Data Mining.*

21. Neuwirth, E. (2015). RColorBrewer: ColorBrewer Palettes. R package versión 1.1-2.

22. Pisani, S., & De Sisti, P. (2007/8). *Risk Analysis applied to Tax Evasion using data mining methodology.*

23. Servicio de Rentas Internas. (2016). *Servicio de Rentas Internas - Plan Estratégico Institucional.* Obtenido de https://www.sri.gob.ec/web/guest/plan-estrategico-institucional

24. Servicio de Rentas Internas. (2018). *Servicio de Rentas Internas*. Obtenido de https://www.sri.gob.ec/web/guest/plan-de-control-y-lucha-contra-el-fraude-fiscal

25. Sharma, A., & Omlin, C. (2009). Performance Comparison of Particle Swarm Optimization with Traditional Clustering Algorithms used in Self-Organizing Map. *International Journal of Computational Intelligence*, 32–41.

26. Stefanovic, P., & Kurasova, O. (2011). Visual analysis of self-organizing maps. *Nonlinear Analysis: Modelling and Control, Vol. 16*(No. 4), 488–504.

27. Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining.* New York: Pearson.

28. Wehrens, R., & Kruisselbrink, J. (2018). kohonen: Supervised and Unsupervised Self-Organising Maps. R package versión 3.0.8.

29. Wickham, H., Chang, W., & Henry, L. (2018). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package versión 3.1.0.

30. Wu, R.-S., Ou, O., Lin, H.-y., Chang, S.-I., & Yen, D. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications 39*, 8769–8777.

31. YiLan, L., & RuTong, Z. (2015). Clustertend. R package versión 1.4.

ANEXO 1. Matriz de correlaciones de sociedades

Variables	Ventas_2014	Ventas_2015	Ventas_2016	IVA_ventas_2014	IVA_ventas_2015	IVA_ventas_2016	Impuestogen_totalvtas14	Impuesto_liquidar_totalvtas14	Vtas0af_totalvtas14	Vtas0sinaf_totalvtas14	Compras_ventas14	IVAreten_IVAcompras14	IVApagar_IVAcompras14
Ventas_2014	1.00	0.95	0.92	0.76	0.73	0.71	-0.03	-0.02	0.03	0.03	0.00	0.00	0.00
Ventas_2015	0.95	1.00	0.97	0.70	0.74	0.72	-0.04	-0.03	0.04	0.04	0.00	0.00	0.00
Ventas_2016	0.92	0.97	1.00	0.63	0.67	0.70	-0.05	-0.04	0.05	0.05	0.00	0.00	0.00
IVA_ventas_2014	0.76	0.70	0.63	1.00	0.95	0.91	0.11	0.08	-0.11	-0.10	0.00	-0.01	0.00
IVA_ventas_2015	0.73	0.74	0.67	0.95	1.00	0.96	0.11	0.08	-0.11	-0.10	0.00	-0.01	0.00
IVA_ventas_2016	0.71	0.72	0.70	0.91	0.96	1.00	0.10	0.08	-0.10	-0.10	0.00	-0.01	0.00
Impuestogen_totalvtas14	-0.03	-0.04	-0.05	0.11	0.11	0.10	1.00	0.81	-1.00	-0.98	-0.01	-0.02	0.01
Impuesto_liquidar_totalvtas14	-0.02	-0.03	-0.04	0.08	0.08	0.08	0.81	1.00	-0.80	-0.79	-0.01	-0.02	0.01
Vtas0af_totalvtas14	0.03	0.04	0.05	-0.11	-0.11	-0.10	-1.00	-0.80	1.00	0.99	0.00	0.02	-0.01
Vtas0sinaf_totalvtas14	0.03	0.04	0.05	-0.10	-0.10	-0.10	-0.98	-0.79	0.99	1.00	0.00	0.02	-0.01
Compras_ventas14	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	1.00	0.00	0.00
IVAreten_IVAcompras14	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.02	-0.02	0.02	0.02	0.00	1.00	0.00
IVApagar_IVAcompras14	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	-0.01	-0.01	0.00	0.00	1.00
Impuestogen_totalvtas15	-0.03	-0.04	-0.05	0.10	0.11	0.10	0.94	0.76	-0.94	-0.93	0.00	-0.02	0.01
Impuesto_liquidar_totalvtas15	-0.03	-0.04	-0.05	0.10	0.10	0.10	0.93	0.74	-0.93	-0.92	0.00	-0.02	0.01
Vtas0af_totalvtas15	0.03	0.04	0.05	-0.10	-0.11	-0.10	-0.94	-0.76	0.94	0.93	0.00	0.02	-0.01
Vtas0sinaf_totalvtas15	0.03	0.04	0.05	-0.10	-0.10	-0.10	-0.93	-0.75	0.94	0.93	0.00	0.02	-0.01
Compras_ventas15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00
IVAreten_IVAcompras15	0.01	0.01	0.01	0.00	-0.01	-0.01	-0.04	-0.03	0.04	0.04	0.00	0.22	0.00
IVApagar_IVAcompras15	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	-0.01	-0.01	0.00	0.00	0.07
Impuestogen_totalvtas16	-0.03	-0.04	-0.05	0.10	0.11	0.11	0.91	0.73	-0.91	-0.90	0.00	-0.02	0.01
Impuesto_liquidar_totalvtas16	-0.03	-0.03	-0.05	0.10	0.10	0.10	0.88	0.71	-0.88	-0.87	0.00	-0.02	0.01
Vtas0af_totalvtas16	0.03	0.04	0.05	-0.10	-0.10	-0.10	-0.90	-0.73	0.90	0.89	0.00	0.02	-0.01
Vtas0sinaf_totalvtas16	0.03	0.04	0.05	-0.10	-0.10	-0.10	-0.89	-0.72	0.90	0.89	0.00	0.02	0.00
Compras_ventas16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
IVAreten_IVAcompras16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
IVApagar_IVAcompras16	-0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.02	0.02	-0.02	-0.02	0.00	0.00	0.06
Variacion_sim_vtas15_14	-0.02	0.01	0.01	-0.04	0.00	0.00	-0.04	-0.03	0.04	0.03	0.02	0.00	0.00
Variacion_sim_vtas16_15	0.00	0.01	0.05	-0.01	-0.01	0.04	-0.08	-0.07	0.08	0.08	0.01	0.01	0.00
Variacion_sim_vtas16_14	-0.03	0.00	0.04	-0.05	-0.01	0.02	-0.08	-0.06	0.08	0.08	0.02	0.02	0.00
Variacion_sim_imp16_14	-0.02	0.01	0.04	-0.04	-0.01	0.03	-0.11	-0.09	0.11	0.11	0.02	0.01	0.00
Varianza_ventas16_14	0.73	0.66	0.59	0.66	0.61	0.55	-0.01	-0.01	0.01	0.01	0.00	0.00	0.00
Exportaciones_Vtastot14	0.12	0.12	0.12	-0.01	-0.01	-0.01	-0.22	-0.18	0.22	0.23	0.02	0.02	0.00
Exportaciones_Vtastot15	0.01	0.01	0.01	0.00	0.00	0.00	-0.02	-0.02	0.02	0.02	0.00	0.00	0.00
Exportaciones_Vtastot16	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.01	0.01	0.00	0.00	0.00
Varianza_export16_14	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.01	0.01	0.00	0.00	0.00
Vtas12_Vtastot14	-0.03	-0.04	-0.05	0.11	0.11	0.10	1.00	0.80	-1.00	-0.98	0.00	-0.02	0.01
Vtas12_Vtastot15	-0.03	-0.04	-0.05	0.10	0.11	0.10	0.94	0.76	-0.94	-0.93	0.00	-0.02	0.01
Vtas12_Vtastot16	-0.03	-0.04	-0.05	0.10	0.10	0.10	0.91	0.73	-0.91	-0.90	0.00	-0.02	0.01
Brecha_TIE	-0.01	-0.01	-0.01	0.02	0.03	0.02	0.42	0.34	-0.42	-0.42	0.00	0.00	0.01
TIE_IRC_14sinatip	-0.09	-0.10	-0.10	-0.02	-0.02	-0.02	0.72	0.61	-0.72	-0.71	-0.01	0.01	0.02
TIE_IRC_15sinatip	-0.08	-0.09	-0.10	-0.01	-0.01	-0.01	0.70	0.55	-0.70	-0.69	0.00	0.01	0.02
TIE_IRC_16sinatip	-0.07	-0.08	-0.09	0.00	0.00	0.00	0.65	0.52	-0.65	-0.65	0.00	0.01	0.02
TIEpromedio_sector14	-0.15	-0.15	-0.15	-0.06	-0.06	-0.06	0.51	0.41	-0.51	-0.51	0.00	0.01	0.01
TIEpromedio_sector15	-0.13	-0.14	-0.14	-0.04	-0.04	-0.04	0.52	0.42	-0.53	-0.53	0.00	0.01	0.01
TIEpromedio_sector16	-0.12	-0.13	-0.14	-0.03	-0.03	-0.04	0.52	0.42	-0.52	-0.52	0.00	0.01	0.01
Brecha_TIE14	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.49	0.43	-0.49	-0.48	-0.01	0.00	0.02
Brecha_TIE15	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.47	0.37	-0.46	-0.46	0.00	0.00	0.01
Brecha_TIE16	-0.01	-0.01	-0.01	0.02	0.03	0.02	0.42	0.34	-0.42	-0.42	0.00	0.00	0.01
Compras_2014	0.97	0.94	0.91	0.75	0.73	0.71	-0.02	-0.02	0.02	0.03	0.00	0.00	0.00
Compras_2015	0.93	0.97	0.95	0.68	0.70	0.70	-0.03	-0.03	0.04	0.04	0.00	0.00	0.00
Compras_2016	0.89	0.92	0.96	0.61	0.62	0.68	-0.04	-0.04	0.05	0.05	0.00	0.00	0.00
Margen_utilidad14	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	-1.00	0.00	0.00
Margen_utilidad15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
Margen_utilidad16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Variables	Impuestogen_totalvtas15	Impuesto_liquidar_totalvtas15	Vtas0af_totalvtas15	Vtas0sinaf_totalvtas15	Compras_ventas15	IVAreten_IVAcompras15	IVApagar_IVAcompras15	Impuestogen_totalvtas16	Impuesto_liquidar_totalvtas16	Vtas0af_totalvtas16	Vtas0sinaf_totalvtas16	Compras_ventas16	IVAreten_IVAcompras16	IVApagar_IVAcompras16
Ventas_2014	-0,03	-0,03	0,03	0,03	0,00	0,01	0,00	-0,03	-0,03	0,03	0,03	0,00	0,00	-0,01
Ventas_2015	-0,04	-0,04	0,04	0,04	0,00	0,01	0,00	-0,04	-0,03	0,04	0,04	0,00	0,00	-0,01
Ventas_2016	-0,05	-0,05	0,05	0,05	0,00	0,01	0,00	-0,05	-0,05	0,05	0,05	0,00	0,00	-0,01
IVA_ventas_2014	0,10	0,10	-0,10	-0,10	0,00	0,00	0,00	0,10	0,10	-0,10	-0,10	0,00	0,00	0,00
IVA_ventas_2015	0,11	0,10	-0,11	-0,10	0,00	-0,01	0,00	0,11	0,10	-0,10	-0,10	0,00	0,00	-0,01
IVA_ventas_2016	0,10	0,10	-0,10	-0,10	0,00	-0,01	0,00	0,11	0,10	-0,10	-0,10	0,00	0,00	-0,01
Impuestogen_totalvtas14	0,94	0,93	-0,94	-0,93	0,00	-0,04	0,01	0,91	0,88	-0,90	-0,89	0,00	0,00	0,02
Impuesto_liquidar_totalvtas14	0,76	0,74	-0,76	-0,75	0,00	-0,03	0,01	0,73	0,71	-0,73	-0,72	0,00	0,00	0,02
Vtas0af_totalvtas14	-0,94	-0,93	0,94	0,94	-0,01	0,04	-0,01	-0,91	-0,88	0,90	0,90	0,00	0,00	-0,02
Vtas0sinaf_totalvtas14	-0,93	-0,92	0,93	0,93	-0,01	0,04	-0,01	-0,90	-0,87	0,89	0,89	-0,01	0,00	-0,02
Compras_ventas14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
IVAreten_IVAcompras14	-0,02	-0,02	0,02	0,02	0,00	0,22	0,00	-0,02	-0,02	0,02	0,02	0,00	0,02	0,00
IVApagar_IVAcompras14	0,01	0,01	-0,01	-0,01	0,00	0,00	0,07	0,01	0,01	-0,01	0,00	0,00	0,00	0,06
Impuestogen_totalvtas15	1,00	0,98	-1,00	-0,99	-0,01	-0,04	0,02	0,94	0,91	-0,93	-0,92	0,00	0,00	0,02
Impuesto_liquidar_totalvtas15	0,98	1,00	-0,98	-0,97	-0,01	-0,04	0,02	0,92	0,88	-0,91	-0,90	0,01	0,00	0,02
Vtas0af_totalvtas15	-1,00	-0,98	1,00	0,99	-0,01	0,04	-0,02	-0,94	-0,91	0,93	0,92	0,00	0,00	-0,02
Vtas0sinaf_totalvtas15	-0,99	-0,97	0,99	1,00	-0,01	0,04	-0,02	-0,93	-0,90	0,92	0,92	0,00	0,00	-0,02
Compras_ventas15	-0,01	-0,01	-0,01	-0,01	1,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,01	0,00	0,00
IVAreten_IVAcompras15	-0,04	-0,04	0,04	0,04	0,00	1,00	0,02	-0,03	-0,03	0,03	0,03	0,00	0,06	0,01
IVApagar_IVAcompras15	0,02	0,02	-0,02	-0,02	0,00	0,02	1,00	0,01	0,01	-0,01	-0,01	0,00	0,00	0,11
Impuestogen_totalvtas16	0,94	0,92	-0,94	-0,93	0,00	-0,03	0,01	1,00	0,96	-0,98	-0,97	-0,01	0,00	0,02
Impuesto_liquidar_totalvtas16	0,91	0,88	-0,91	-0,90	0,00	-0,03	0,01	0,96	1,00	-0,95	-0,94	-0,01	0,00	0,02
Vtas0af_totalvtas16	-0,93	-0,91	0,93	0,92	-0,01	0,03	-0,01	-0,98	-0,95	1,00	0,99	0,00	0,00	-0,02
Vtas0sinaf_totalvtas16	-0,92	-0,90	0,92	0,92	-0,01	0,03	-0,01	-0,97	-0,94	0,99	1,00	0,00	0,00	-0,02
Compras_ventas16	0,00	0,01	0,00	0,00	0,01	0,00	0,00	-0,01	-0,01	0,00	0,00	1,00	0,00	0,00
IVAreten_IVAcompras16	0,00	0,00	0,00	0,00	0,00	0,06	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,14
IVApagar_IVAcompras16	0,02	0,02	-0,02	-0,02	0,00	0,01	0,11	0,02	0,02	-0,02	-0,02	0,00	0,14	1,00
Variacion_sim_vtas15_14	0,06	0,05	-0,06	-0,06	-0,02	-0,01	0,01	0,04	0,04	-0,04	-0,04	0,00	0,00	0,00
Variacion_sim_vtas16_15	-0,08	-0,08	0,08	0,08	0,02	0,00	0,00	-0,09	-0,11	0,11	0,11	-0,02	-0,01	0,00
Variacion_sim_vtas16_14	-0,10	-0,11	0,10	0,10	0,00	0,00	0,01	-0,10	-0,11	0,12	0,11	-0,02	-0,01	0,00
Variacion_sim_imp16_14	-0,05	-0,06	0,05	0,05	0,00	0,00	0,00	0,03	0,01	-0,01	-0,01	-0,02	-0,01	0,00
Varianza_ventas16_14	-0,01	-0,01	0,01	0,01	0,00	0,01	0,00	-0,01	0,00	0,01	0,01	0,00	0,00	-0,01
Exportaciones_Vtastot14	-0,21	-0,21	0,21	0,22	0,00	0,06	0,00	-0,20	-0,19	0,20	0,20	0,00	0,01	-0,01
Exportaciones_Vtastot15	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00
Exportaciones_Vtastot16	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00
Varianza_export16_14	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00	-0,01	-0,01	0,01	0,01	0,00	0,00	0,00
Vtas12_Vtastot14	0,94	0,93	-0,94	-0,93	0,01	-0,04	0,01	0,91	0,88	-0,90	-0,89	0,00	0,00	0,02
Vtas12_Vtastot15	1,00	0,98	-1,00	-0,98	0,01	-0,04	0,02	0,94	0,91	-0,93	-0,92	0,00	0,00	0,02
Vtas12_Vtastot16	0,94	0,92	-0,94	-0,93	0,01	-0,03	0,01	0,99	0,96	-0,98	-0,97	0,00	0,00	0,02
Brecha_TIE	0,45	0,43	-0,45	-0,44	-0,01	0,01	0,02	0,51	0,56	-0,50	-0,49	-0,01	-0,01	0,05
TIE_IRC_14sinatip	0,68	0,66	-0,68	-0,67	0,00	0,01	0,04	0,66	0,63	-0,65	-0,64	-0,01	0,00	0,06
TIE_IRC_15sinatip	0,74	0,76	-0,74	-0,73	-0,01	0,01	0,05	0,69	0,66	-0,69	-0,68	0,00	0,00	0,05
TIE_IRC_16sinatip	0,68	0,66	-0,68	-0,67	0,00	0,01	0,03	0,73	0,76	-0,71	-0,70	-0,01	0,00	0,06
TIEpromedio_sector14	0,51	0,50	-0,51	-0,51	0,01	0,01	0,02	0,50	0,49	-0,50	-0,50	0,00	0,00	0,03
TIEpromedio_sector15	0,53	0,52	-0,53	-0,53	0,01	0,01	0,02	0,52	0,50	-0,51	-0,51	0,00	0,00	0,03
TIEpromedio_sector16	0,53	0,52	-0,53	-0,53	0,01	0,01	0,02	0,52	0,50	-0,51	-0,52	0,01	0,00	0,03
Brecha_TIE14	0,44	0,43	-0,44	-0,44	0,00	0,00	0,03	0,43	0,41	-0,42	-0,41	-0,01	-0,01	0,05
Brecha_TIE15	0,52	0,55	-0,52	-0,50	-0,02	0,01	0,04	0,47	0,44	-0,46	-0,45	0,00	-0,01	0,04
Brecha_TIE16	0,45	0,43	-0,45	-0,44	-0,01	0,01	0,02	0,51	0,56	-0,50	-0,49	-0,01	-0,01	0,05
Compras_2014	-0,03	-0,02	0,03	0,03	0,00	0,01	0,00	-0,02	-0,02	0,02	0,03	0,00	0,00	-0,01
Compras_2015	-0,04	-0,03	0,04	0,04	0,00	0,01	-0,01	-0,03	-0,03	0,03	0,04	0,00	0,00	-0,01
Compras_2016	-0,05	-0,04	0,05	0,05	0,00	0,01	0,00	-0,04	-0,04	0,04	0,05	0,00	0,00	-0,01
Margen_utilidad14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Margen_utilidad15	0,01	0,01	0,01	0,01	-1,00	0,00	0,00	0,00	0,00	0,01	0,01	-0,01	0,00	0,00
Margen_utilidad16	0,00	-0,01	0,00	0,00	-0,01	0,00	0,00	0,01	0,01	0,00	0,00	-1,00	0,00	0,00

Variables	Variacion_sim_vtas15_14	Variacion_sim_vtas16_15	Variacion_sim_vtas16_14	Variacion_sim_imp16_14	Varianza_ventas16_14	Exportaciones_Vtastot14	Exportaciones_Vtastot15	Exportaciones_Vtastot16	Varianza_export16_14	Vtas12_Vtastot14	Vtas12_Vtastot15	Vtas12_Vtastot16	Brecha_TIE	TIE_IRC_14sinatip
Ventas_2014	-0,02	0,00	-0,03	-0,02	0,73	0,12	0,01	0,00	0,00	-0,03	-0,03	-0,03	-0,01	-0,09
Ventas_2015	0,01	0,01	0,00	0,01	0,66	0,12	0,01	0,00	0,00	-0,04	-0,04	-0,04	-0,01	-0,10
Ventas_2016	0,01	0,05	0,04	0,04	0,59	0,12	0,01	0,00	0,00	-0,05	-0,05	-0,05	-0,01	-0,10
IVA_ventas_2014	-0,04	-0,01	-0,05	-0,04	0,66	-0,01	0,00	0,00	0,00	0,11	0,10	0,10	0,02	-0,02
IVA_ventas_2015	0,00	-0,01	-0,01	-0,01	0,61	-0,01	0,00	0,00	0,00	0,11	0,11	0,10	0,03	-0,02
IVA_ventas_2016	0,00	0,04	0,02	0,03	0,55	-0,01	0,00	0,00	0,00	0,10	0,10	0,10	0,02	-0,02
Impuestogen_totalvtas14	-0,04	-0,08	-0,08	-0,11	-0,01	-0,22	-0,02	-0,01	-0,01	1,00	0,94	0,91	0,42	0,72
Impuesto_liquidar_totalvtas14	-0,03	-0,07	-0,06	-0,09	-0,01	-0,18	-0,02	-0,01	-0,01	0,80	0,76	0,73	0,34	0,61
Vtas0af_totalvtas14	0,04	0,08	0,08	0,11	0,01	0,22	0,02	0,01	0,01	-1,00	-0,94	-0,91	-0,42	-0,72
Vtas0sinaf_totalvtas14	0,03	0,08	0,08	0,11	0,01	0,23	0,02	0,01	0,01	-0,98	-0,93	-0,90	-0,42	-0,71
Compras_ventas14	0,02	0,01	0,02	0,02	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01
IVAreten_IVAcompras14	0,00	0,01	0,02	0,01	0,00	0,02	0,00	0,00	0,00	-0,02	-0,02	-0,02	0,00	0,01
IVApagar_IVAcompras14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,02
Impuestogen_totalvtas15	0,06	-0,08	-0,10	-0,05	-0,01	-0,21	-0,01	-0,01	-0,01	0,94	1,00	0,94	0,45	0,68
Impuesto_liquidar_totalvtas15	0,05	-0,08	-0,11	-0,06	-0,01	-0,21	-0,01	-0,01	-0,01	0,93	0,98	0,92	0,43	0,66
Vtas0af_totalvtas15	-0,06	0,08	0,10	0,05	0,01	0,21	0,01	0,01	0,01	-0,94	-1,00	-0,94	-0,45	-0,68
Vtas0sinaf_totalvtas15	-0,06	0,08	0,10	0,05	0,01	0,22	0,01	0,01	0,01	-0,93	-0,98	-0,93	-0,44	-0,67
Compras_ventas15	-0,02	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	-0,01	0,00
IVAreten_IVAcompras15	-0,01	0,00	0,00	0,00	0,01	0,06	0,00	0,00	0,00	-0,04	-0,04	-0,03	0,01	0,01
IVApagar_IVAcompras15	0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,01	0,02	0,04
Impuestogen_totalvtas16	0,04	-0,09	-0,10	0,03	-0,01	-0,20	-0,01	-0,01	-0,01	0,91	0,94	0,99	0,51	0,66
Impuesto_liquidar_totalvtas16	0,04	-0,11	-0,11	0,01	0,00	-0,19	-0,01	-0,01	-0,01	0,88	0,91	0,96	0,56	0,63
Vtas0af_totalvtas16	-0,04	0,11	0,12	-0,01	0,01	0,20	0,01	0,01	0,01	-0,90	-0,93	-0,98	-0,50	-0,65
Vtas0sinaf_totalvtas16	-0,04	0,11	0,11	-0,01	0,01	0,20	0,01	0,01	0,01	-0,89	-0,92	-0,97	-0,49	-0,64
Compras_ventas16	0,00	-0,02	-0,02	-0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01
IVAreten_IVAcompras16	0,00	-0,01	-0,01	-0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	0,00
IVApagar_IVAcompras16	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,00	0,00	0,02	0,02	0,02	0,05	0,06
Variacion_sim_vtas15_14	1,00	-0,05	0,42	0,61	-0,02	0,01	0,00	0,00	0,00	-0,03	0,06	0,04	0,02	-0,05
Variacion_sim_vtas16_15	-0,05	1,00	0,66	0,46	-0,03	0,02	0,00	-0,01	-0,01	-0,08	-0,08	-0,10	-0,03	-0,03
Variacion_sim_vtas16_14	0,42	0,66	1,00	0,67	-0,04	0,00	-0,01	-0,01	-0,01	-0,08	-0,10	-0,11	-0,06	-0,06
Variacion_sim_imp16_14	0,61	0,46	0,67	1,00	-0,03	0,03	0,00	0,00	0,00	-0,11	-0,05	0,01	0,03	-0,08
Varianza_ventas16_14	-0,02	-0,03	-0,04	-0,03	1,00	0,11	0,01	0,00	0,00	-0,01	-0,01	-0,01	0,00	-0,07
Exportaciones_Vtastot14	0,01	0,02	0,00	0,03	0,11	1,00	0,09	0,03	0,03	-0,22	-0,21	-0,20	-0,08	-0,17
Exportaciones_Vtastot15	0,00	0,00	-0,01	0,00	0,01	0,09	1,00	0,00	0,00	-0,02	-0,02	-0,02	-0,01	-0,02
Exportaciones_Vtastot16	0,00	-0,01	-0,01	0,00	0,00	0,03	0,00	1,00	1,00	-0,01	-0,01	-0,01	-0,01	0,00
Varianza_export16_14	0,00	-0,01	-0,01	0,00	0,00	0,03	0,00	1,00	1,00	-0,01	-0,01	-0,01	-0,01	0,00
Vtas12_Vtastot14	-0,03	-0,08	-0,08	-0,11	-0,01	-0,22	-0,02	-0,01	-0,01	1,00	0,94	0,91	0,42	0,72
Vtas12_Vtastot15	0,06	-0,08	-0,10	-0,05	-0,01	-0,21	-0,02	-0,01	-0,01	0,94	1,00	0,94	0,45	0,68
Vtas12_Vtastot16	0,04	-0,10	-0,11	0,01	-0,01	-0,20	-0,02	-0,01	-0,01	0,91	0,94	1,00	0,50	0,65
Brecha_TIE	0,02	-0,03	-0,06	0,03	0,00	-0,08	-0,01	-0,01	-0,01	0,42	0,45	0,50	1,00	0,48
TIE_IRC_14sinatip	-0,05	-0,03	-0,06	-0,08	-0,07	-0,17	-0,02	0,00	0,00	0,72	0,68	0,65	0,48	1,00
TIE_IRC_15sinatip	0,04	-0,07	-0,09	-0,05	-0,05	-0,16	-0,02	-0,01	-0,01	0,69	0,74	0,69	0,54	0,79
TIE_IRC_16sinatip	0,02	-0,08	-0,09	0,01	-0,04	-0,14	-0,01	0,00	0,00	0,65	0,68	0,71	0,83	0,72
TIEpromedio_sector14	0,05	-0,03	0,01	0,01	-0,10	-0,13	-0,01	0,00	0,00	0,51	0,51	0,50	-0,03	0,55
TIEpromedio_sector15	0,01	-0,03	-0,04	-0,02	-0,09	-0,13	-0,01	0,00	0,00	0,52	0,53	0,52	-0,03	0,55
TIEpromedio_sector16	0,02	-0,09	-0,08	-0,03	-0,08	-0,13	-0,01	0,00	0,00	0,52	0,52	0,52	-0,04	0,55
Brecha_TIE14	-0,10	-0,02	-0,08	-0,10	0,00	-0,11	-0,01	-0,01	-0,01	0,49	0,44	0,42	0,60	0,80
Brecha_TIE15	0,03	-0,06	-0,08	-0,05	0,00	-0,10	-0,01	-0,01	-0,01	0,46	0,51	0,46	0,67	0,56
Brecha_TIE16	0,02	-0,03	-0,06	0,03	0,00	-0,08	-0,01	-0,01	-0,01	0,42	0,45	0,50	1,00	0,48
Compras_2014	-0,02	0,00	-0,03	-0,01	0,69	0,12	0,01	0,00	0,00	-0,02	-0,02	-0,02	-0,01	-0,10
Compras_2015	0,01	0,01	0,01	0,01	0,62	0,12	0,01	0,00	0,00	-0,03	-0,04	-0,03	-0,01	-0,10
Compras_2016	0,01	0,05	0,03	0,03	0,54	0,13	0,01	0,00	0,00	-0,04	-0,04	-0,04	-0,02	-0,10
Margen_utilidad14	-0,02	-0,01	-0,02	-0,02	0,00	-0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Margen_utilidad15	0,02	-0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	-0,01	0,01	0,00
Margen_utilidad16	0,00	0,02	0,02	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01

Variables	TIE_IRC_15 sinatip	TIE_IRC_16 sinatip	TIEpromedio_sector14	TIEpromedio_sector15	TIEpromedio_sector16	Brecha_TIE14	Brecha_TIE15	Brecha_TIE16	Compras_2014	Compras_2015	Compras_2016	Margen_utilidad14	Margen_utilidad15	Margen_utilidad16
Ventas_2014	-0,08	-0,07	-0,15	-0,13	-0,12	-0,01	-0,01	-0,01	0,97	0,93	0,89	0,00	0,00	0,00
Ventas_2015	-0,09	-0,08	-0,15	-0,14	-0,13	-0,01	-0,01	-0,01	0,94	0,97	0,92	0,00	0,00	0,00
Ventas_2016	-0,10	-0,09	-0,15	-0,14	-0,14	-0,01	-0,01	-0,01	0,91	0,95	0,96	0,00	0,00	0,00
IVA_ventas_2014	-0,01	0,00	-0,06	-0,04	-0,03	0,02	0,02	0,02	0,75	0,68	0,61	0,00	0,00	0,00
IVA_ventas_2015	-0,01	0,00	-0,06	-0,04	-0,03	0,02	0,02	0,03	0,73	0,70	0,62	0,00	0,00	0,00
IVA_ventas_2016	-0,01	0,00	-0,06	-0,04	-0,04	0,02	0,02	0,02	0,71	0,70	0,68	0,00	0,00	0,00
Impuestogen_totalvtas14	0,70	0,65	0,51	0,52	0,52	0,49	0,47	0,42	-0,02	-0,03	-0,04	0,01	0,00	0,00
Impuesto_liquidar_totalvtas14	0,55	0,52	0,41	0,42	0,42	0,43	0,37	0,34	-0,02	-0,03	-0,04	0,01	0,00	0,00
Vtas0af_totalvtas14	-0,70	-0,65	-0,51	-0,53	-0,52	-0,49	-0,46	-0,42	0,02	0,04	0,05	0,00	0,01	0,00
Vtas0sinaf_totalvtas14	-0,69	-0,65	-0,51	-0,53	-0,52	-0,48	-0,46	-0,42	0,03	0,04	0,05	0,00	0,01	0,01
Compras_ventas14	0,00	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	-1,00	0,00	0,00
IVAreten_IVAcompras14	0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
IVApagar_IVAcompras14	0,02	0,02	0,01	0,01	0,01	0,02	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00
Impuestogen_totalvtas15	0,74	0,68	0,51	0,53	0,53	0,44	0,52	0,45	-0,03	-0,04	-0,05	0,00	0,01	0,00
Impuesto_liquidar_totalvtas15	0,76	0,66	0,50	0,52	0,52	0,43	0,55	0,43	-0,02	-0,03	-0,04	0,00	0,01	-0,01
Vtas0af_totalvtas15	-0,74	-0,68	-0,51	-0,53	-0,53	-0,44	-0,52	-0,45	0,03	0,04	0,05	0,00	0,01	0,00
Vtas0sinaf_totalvtas15	-0,73	-0,67	-0,51	-0,53	-0,53	-0,44	-0,50	-0,44	0,03	0,04	0,05	0,00	0,01	0,00
Compras_ventas15	-0,01	0,00	0,01	0,01	0,01	0,00	-0,02	-0,01	0,00	0,00	0,00	0,00	-1,00	-0,01
IVAreten_IVAcompras15	0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,00
IVApagar_IVAcompras15	0,05	0,03	0,02	0,02	0,02	0,03	0,04	0,02	0,00	-0,01	0,00	0,00	0,00	0,00
Impuestogen_totalvtas16	0,69	0,73	0,50	0,52	0,52	0,43	0,47	0,51	-0,02	-0,03	-0,04	0,00	0,00	0,01
Impuesto_liquidar_totalvtas16	0,66	0,76	0,49	0,50	0,50	0,41	0,44	0,56	-0,02	-0,03	-0,04	0,00	0,00	0,01
Vtas0af_totalvtas16	-0,69	-0,71	-0,50	-0,51	-0,51	-0,42	-0,46	-0,50	0,02	0,03	0,04	0,00	0,01	0,00
Vtas0sinaf_totalvtas16	-0,68	-0,70	-0,50	-0,51	-0,52	-0,41	-0,45	-0,49	0,03	0,04	0,05	0,00	0,01	0,00
Compras_ventas16	0,00	-0,01	0,00	0,00	0,01	-0,01	0,00	-0,01	0,00	0,00	0,00	0,00	-0,01	-1,00
IVAreten_IVAcompras16	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,00
IVApagar_IVAcompras16	0,05	0,06	0,03	0,03	0,03	0,05	0,04	0,05	-0,01	-0,01	-0,01	0,00	0,00	0,00
Variacion_sim_vtas15_14	0,04	0,02	0,05	0,01	0,02	-0,10	0,03	0,02	-0,02	0,01	0,01	-0,02	0,02	0,00
Variacion_sim_vtas16_15	-0,07	-0,08	-0,03	-0,03	-0,09	-0,02	-0,06	-0,03	0,00	0,01	0,05	-0,01	-0,02	0,02
Variacion_sim_vtas16_14	-0,09	-0,09	0,01	-0,04	-0,08	-0,08	-0,08	-0,06	-0,03	0,01	0,03	-0,02	0,00	0,02
Variacion_sim_imp16_14	-0,05	0,01	0,01	-0,02	-0,03	-0,10	-0,05	0,03	-0,01	0,01	0,03	-0,02	0,00	0,02
Varianza_ventas16_14	-0,05	-0,04	-0,10	-0,09	-0,08	0,00	0,00	0,00	0,69	0,62	0,54	0,00	0,00	0,00
Exportaciones_Vtastot14	-0,16	-0,14	-0,13	-0,13	-0,13	-0,11	-0,10	-0,08	0,12	0,12	0,13	-0,02	0,00	0,00
Exportaciones_Vtastot15	-0,02	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01	0,01	0,01	0,01	0,00	0,00	0,00
Exportaciones_Vtastot16	-0,01	0,00	0,00	0,00	0,00	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,00
Varianza_export16_14	-0,01	0,00	0,00	0,00	0,00	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,00
Vtas12_Vtastot14	0,69	0,65	0,51	0,52	0,52	0,49	0,46	0,42	-0,02	-0,03	-0,04	0,00	-0,01	0,00
Vtas12_Vtastot15	0,74	0,68	0,51	0,53	0,52	0,44	0,51	0,45	-0,02	-0,04	-0,04	0,00	-0,01	0,00
Vtas12_Vtastot16	0,69	0,71	0,50	0,52	0,52	0,42	0,46	0,50	-0,02	-0,03	-0,04	0,00	-0,01	0,00
Brecha_TIE	0,54	0,83	-0,03	-0,03	-0,04	0,60	0,67	1,00	-0,01	-0,01	-0,02	0,00	0,01	0,01
TIE_IRC_14sinatip	0,79	0,72	0,55	0,55	0,55	0,80	0,56	0,48	-0,10	-0,10	-0,10	0,01	0,00	0,01
TIE_IRC_15sinatip	1,00	0,77	0,55	0,56	0,55	0,55	0,81	0,54	-0,09	-0,10	-0,10	0,00	0,01	0,00
TIE_IRC_16sinatip	0,77	1,00	0,52	0,53	0,53	0,49	0,55	0,83	-0,08	-0,09	-0,10	0,00	0,00	0,01
TIEpromedio_sector14	0,55	0,52	1,00	0,97	0,96	-0,06	-0,03	-0,03	-0,15	-0,16	-0,15	0,00	-0,01	0,00
TIEpromedio_sector15	0,56	0,53	0,97	1,00	0,97	-0,04	-0,04	-0,03	-0,13	-0,14	-0,14	0,00	-0,01	0,00
TIEpromedio_sector16	0,55	0,53	0,96	0,97	1,00	-0,03	-0,03	-0,04	-0,13	-0,14	-0,14	0,00	-0,01	-0,01
Brecha_TIE14	0,55	0,49	-0,06	-0,04	-0,03	1,00	0,69	0,60	-0,01	-0,01	-0,01	0,01	0,00	0,01
Brecha_TIE15	0,81	0,55	-0,03	-0,04	-0,03	0,69	1,00	0,67	-0,01	-0,02	-0,02	0,00	0,02	0,00
Brecha_TIE16	0,54	0,83	-0,03	-0,03	-0,04	0,60	0,67	1,00	-0,01	-0,01	-0,02	0,00	0,01	0,01
Compras_2014	-0,09	-0,08	-0,15	-0,13	-0,13	-0,01	-0,01	-0,01	1,00	0,96	0,92	0,00	0,00	0,00
Compras_2015	-0,10	-0,09	-0,16	-0,14	-0,14	-0,01	-0,02	-0,01	0,96	1,00	0,96	0,00	0,00	0,00
Compras_2016	-0,10	-0,10	-0,15	-0,14	-0,14	-0,01	-0,02	-0,02	0,92	0,96	1,00	0,00	0,00	0,00
Margen_utilidad14	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00
Margen_utilidad15	0,01	0,00	-0,01	-0,01	-0,01	0,00	0,02	0,01	0,00	0,00	0,00	0,00	1,00	0,01
Margen_utilidad16	0,00	0,01	0,00	0,00	-0,01	0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,01	1,00

Variables	Ventas_totales_15	Impuestogen_totalvtas14	Compras_ventas14	Compras_ventas15	Impuestogen_totalvtas16	Compras_ventas16	Variacion_sim_vtas15_14	Variacion_sim_vtas16_15	Varianza_ventas16_14	Brecha_TIE	Brecha_TIE14	Brecha_TIE16	Margen_utilidad16	Variacion_sim_vtas16_14	Variacion_sim_imp16_14	Variacion_sim_V12_Vtot16_14	Variacion_sim_Margutilidad16_14	Variacion_sim_Compras_ventas16_14
Ventas_totales_15	1.00	-0.09	0.00	0.00	-0.08	0.00	0.02	-0.02	0.66	-0.01	-0.02	-0.01	0.00	0.00	0.01	0.01	0.00	-0.01
Impuestogen_totalvtas14	-0.09	1.00	-0.01	0.00	0.75	0.00	-0.08	-0.01	-0.08	0.30	0.44	0.30	0.00	-0.01	-0.11	-0.24	0.00	-0.04
Compras_ventas14	0.00	-0.01	1.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	-0.01
Compras_ventas15	0.00	0.00	0.01	1.00	0.00	0.02	-0.01	0.01	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.01
Impuestogen_totalvtas16	-0.08	0.75	0.00	0.00	1.00	-0.01	0.09	0.00	-0.07	0.45	0.30	0.45	0.01	-0.01	0.17	0.36	0.00	0.07
Compras_ventas16	0.00	0.00	0.00	0.02	-0.01	1.00	0.00	-0.01	0.00	-0.01	0.00	-0.01	-1.00	-0.01	-0.01	0.00	0.00	0.01
Variacion_sim_vtas15_14	0.02	-0.08	0.01	-0.01	0.09	0.00	1.00	-0.13	-0.01	0.02	-0.09	0.02	0.00	0.49	0.61	0.32	0.00	0.03
Variacion_sim_vtas16_15	-0.02	-0.01	0.00	0.01	0.00	-0.01	-0.13	1.00	-0.02	0.00	0.00	0.00	0.01	0.62	0.49	-0.01	0.00	-0.02
Varianza_ventas16_14	0.66	-0.08	0.00	0.00	-0.07	0.00	-0.01	-0.02	1.00	0.00	-0.01	0.00	0.00	-0.02	-0.01	0.00	0.00	-0.01
Brecha_TIE	-0.01	0.30	0.00	0.00	0.45	-0.01	0.02	0.00	0.00	1.00	0.58	1.00	0.01	-0.03	0.08	0.20	0.00	-0.19
Brecha_TIE14	-0.02	0.44	0.00	0.00	0.30	0.00	-0.09	0.00	-0.01	0.58	1.00	0.58	0.00	-0.04	-0.10	-0.15	0.00	0.21
Brecha_TIE16	-0.01	0.30	0.00	0.00	0.45	-0.01	0.02	0.00	0.00	1.00	0.58	1.00	0.01	-0.03	0.08	0.20	0.00	-0.19
Margen_utilidad16	0.00	0.00	0.00	-0.02	0.01	-1.00	0.00	0.01	0.00	0.01	0.00	0.01	1.00	0.01	0.01	0.00	0.00	-0.01
Variacion_sim_vtas16_14	0.00	-0.01	0.01	0.00	-0.01	-0.01	0.49	0.62	-0.02	-0.03	-0.04	-0.03	0.01	1.00	0.78	0.00	0.00	-0.03
Variacion_sim_imp16_14	0.01	-0.11	0.01	0.00	0.17	-0.01	0.61	0.49	-0.01	0.08	-0.10	0.08	0.01	0.78	1.00	0.49	0.00	0.04
Variacion_sim_V12_Vtot16_	0.01	-0.24	0.00	0.00	0.36	0.00	0.32	-0.01	0.00	0.20	-0.15	0.20	0.00	0.00	0.49	1.00	0.00	0.16
Variacion_sim_Margutilida	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Variacion_sim_Compras_ve	-0.01	-0.04	-0.01	0.01	0.07	0.01	0.03	-0.02	-0.01	-0.19	0.21	-0.19	-0.01	-0.03	0.04	0.16	0.00	1.00