Análisis de Causales en Datos Tributarios con Anomalías

Ing. Antonio Sottile Bordallo, Lic. y Prof. Daniel Guillermo Cavaller Riva,
Cdor. Cristian Darío Ortega Yubro

Prof. Titular, Prof. Asociado y Prof. Adscripto
Cátedra Computación
Universidad Nacional de Cuyo, Facultad de Ciencias Económicas
{antonio.sottile, daniel.cavaller,
cristian.ortega}@fce.uncu.edu.ar
http://fce.uncuyo.edu.ar

Abstract. Actualmente en las administraciones tributarias existe un gran volumen de datos. Estos datos contienen implícito un conocimiento que puede ser extraído, este conocimiento dependerá de la calidad de los datos, y en esa cantidad de datos no puede concebirse que no posean anomalías. Los datos tienen anomalías, y las mismas responden a distintas causales. El análisis de las causales en las anomalías de los datos permitirá deducir si ellas responden a ilícitos o hechos de corrupción.

Keywords: Minería de Datos, Datos Anómalos, Algoritmos, Aprendizaje Automático, Big Data, Ruido.

Abstract. Currently in the tax administrations there is a large volume of data. These data implicitly contain knowledge that can be extracted, this knowledge will depend on the quality of the data, and in that amount of data it cann't be conceived that they do not have anomalies. The data have anomalies, and they respond to different causes. The analysis of the causes in the anomalies of the data will allow to deduce if they respond to illicit or acts of corruption.

Keywords: Data Mining, Anomalous Data, Algorithms, Machine Learning, Big Data, Noise.

1 Introducción

Las anomalías en los datos coexisten en las bases de datos y en los datos no tradicionales a los que puede acceder y los que produce una administración tributaria, sean estos datos tanto de origen interno, como de origen externo.
Si las bases de datos y los datos no tradicionales que tienen a disposición estas administraciones tributarias contienen información acerca de todas las actividades objeto

adfa, p. 1, 2011.

de gravámenes de los sujetos pasivos[1] de los tributos, el análisis de ciertas anomalías en los datos podría dar lugar al descubrimiento de patrones que responden a distintas causas, pudiendo evidenciar estas causas ciertos ilícitos por parte de los contribuyentes o hechos de corrupción cuando existe la connivencia del contribuyente con el empleado público o el funcionario público. El propósito de la presente investigación, es el desarrollo teórico del análisis causal de ciertas anomalías de datos tributarios, demostrando que la metodología de Minería de Datos contribuye a evidenciar indicios de ilícitos y hechos de corrupción, a través de la aplicación de ciertos algoritmos.

2 Marco Teórico

Los procesos metodológicos de la Minería de Datos se utilizan para extraer de los datos, conocimiento proactivo o analítico (SAS® Institute Inc 2015), optimizando todo el potencial que despliega el Proceso de Extracción de Conocimiento no trivial (Kuna 2014), conocimiento que se encuentra de manera implícito en los datos, datos que son aportados, en gran medida, por los contribuyentes.
Por ello, se puede inferir el siguiente axioma: la calidad resultante del conocimiento no trivial que se extraiga dependerá en gran medida, de la calidad que posean los datos.
La Etapa Limpieza y Transformación de los Datos, del Proceso de Extracción del Conocimiento, de acuerdo a la Metodología CRISP-DM o basada en ella, analiza la influencia y causal de las inconsistencias y/o anomalías de los datos.
Los algoritmos de agrupamiento se orientan al aprendizaje no supervisado, donde la agrupación de los datos está relacionada con las características comunes que ellos poseen, en la presente investigación, estos datos corresponden a determinaciones de oficio de los inspectores que concluyeron con en el proceso de determinación, en la generación de Liquidaciones de Deuda (LD). Esos algoritmos mencionados son muy utilizados cuando se quiere descubrir conocimiento oculto, patrones de comportamiento y valores extremos de los datos (Kuna 2014). Al analizar la distancia entre los datos de un conjunto de datos, el criterio general de análisis en principio es que cuanto mayor es la distancia entre un dato de una base de datos y el resto de la los datos, mayor es la posibilidad de considerar al dato como anómalo, incosistente o con ruido.

3 Anomalía en los Datos

Las razones por las cuales pueden existir datos anómalos:

1. La incorrecta carga de datos.
2. Errores de los programas utilizados (software) o incompatibilidades entre distintos programas.

[1] Sujetos pasivos: son sujetos pasivos de los tributos aquellas personas (humanas, jurídicas, etc.) sometidas al cumplimiento de las obligaciones tributarias, tanto en su carácter de contribuyentes o meramente como responsables de los tributos.

3. El dato es de una población distinta.
4. Algún tipo de ilícito, como evasión fiscal.
5. Algún posible hecho de corrupción.

Por eso cuando no se trabaja con una distribución estándar de los datos, la identificación de esas incosistencias, anomalías o detección de ruido en los datos es muy difícil. Buscar datos anómalos realizando consultas manuales o formalizar un análisis de tipo secuencial sobre todos los datos de una administración tributaria, aunque sea en uno de sus procesos, como lo es la generación de las Liquidaciones de Deuda (LD) requiere conocer previamente las incosistencias y/o anomalías de los datos que podrían aparecer.

4 Software Aplicado

Para el desarrollo de la presente investigación se utilizó RapidMiner Studio®[2] versión 8.2.1, herramienta de fácil uso, con una amplia gama de algoritmos y variadas opciones de visualización. El software elegido se puede integrar con otros programas y lenguajes como Phyton y R.

5 Agrupamiento de Soporte Vectorial

Se selecciona el algoritmo Agrupamiento de Soporte Vectorial (Support Vector Clustering, SVC) en contraste con otros algoritmos de agrupamiento porque el resto de los algoritmos no tienen ningún mecanismo para tratar el ruido de los datos, o los valores atípicos (Ben-Hur et al. 2001).
Agrupamiento de Soporte Vectorial trata con los valores atípicos y el ruido de los datos mediante el empleo de una constante de margen suave que permite que la esfera en el espacio de características no encierre todos los puntos.
Por consiguiente aquellos puntos que no estén en ningún "cluster", se los considera ruido, siendo ruido para la presente investigación todo aquello que no es de interés o es irrelevante, lo que degrada o distorsiona los datos, los contamina y/o impide o limita el estudio o uso de la información en el análisis de las causales en las anomalías de los datos. Es decir, será ruido en los datos bajo análisis, aquello que queda fuera de los límites de los objetivos que se plantean en la Minería de Datos.

5.1 Desarrollo del Algoritmo Agrupamiento de Soporte Vectorial

Usando la transformación no lineal Φ de x a un espacio, se busca una esfera del radio mas pequeña $\mathcal{R}$, lo que puede describirse con las siguientes restricciones:

$$\| \Phi(x_j) - a \|^2 \leq \mathcal{R}^2 \quad \forall j ,$$

donde $\| . \|$ es la norma euclidiana, y a el centro de la esfera. Las restricciones se van incorporando al agregar valor ξ_j:

[2] https://rapidminer.com

$$\| \Phi(x_j) - a \|^2 \leq \mathcal{R}^2 + \xi_j \ , \qquad (1)$$

con $\xi_j \geq 0$.

Para resolver este problema, se utiliza Lagrange[3], es decir, el langrangiano:

$$L = \mathcal{R}^2 - \sum_j (\mathcal{R}^2 + \xi_j - \| \Phi(x_j) - a \|^2)\xi_j - \sum \xi_j u_j + C \sum \xi_j \ , \qquad (2)$$

donde $\xi_j \geq 0$ y $u_j \geq 0$ son los operadores de Lagrange. C es una constante y $C\beta_j$ es una penalización de $L = \mathcal{R} - término$.
Poniendo a 0 la derivada de L con respecto a $\mathcal{R}$, a y ξ_j respectivamente entonces:

$$\sum_j \beta_j = 1 \quad (3)$$

$$a = \sum_j \beta_j \Phi(x_j) \quad (4)$$

$$\beta_j = C - u_j \quad (5)$$

Las condiciones complementarias de Roger Fletcher[4] (Fletcher, Roger 2000) resultan en:

$$\xi_j u_j = 0 \quad (6)$$

$$(\mathcal{R}^2 + \xi_j - \| \Phi(x_j) - a \|^2)\beta_j = 0 \quad (7)$$

Entonces, un punto x_i donde $\xi_i \geq 0$ y $u_i \geq 0$ se encuentra fuera de la esfera del espacio de características. Si $u_i = 0$, $\beta_i = C$ determinará un Vector de Soporte Limitado.
Un punto x_i con $\xi_i = 0$ se asigna al interior o la superficie de la esfera del espacio caracteristico. Si es $0 < \beta_i < C$ entonces implica que $\Phi(x_i)$ se encuentra en la superficie de la esfera del espacio característico. Ese punto se lo denomina Vector de Soporte. Los puntos Vector de Soporte se encuentran en los límites del clúster, los puntos Vector de Soporte Limitado se encuentran fuera de los límites, y todos los otros puntos se encuentran dentro de ellos, por lo tanto, cuando $C \geq 1$ no existen Vectores de Soporte Limitados por la restricción de la ecuación (3).
Con esas relaciones, se eliminan las variables $\mathcal{R}$, a y u_j convirtiendo el lagrangiano en la forma dual de Wolfe, que es una función de las variables β_j:

[3] Joseph Louis de Lagrange, astrónomo y matemático ítalo – francés, desarrollo una función escalar por la cual se puede obtener la evolución temporal, las leyes de conservación y otras propiedades importantes de un sistema dinámico, considerándose este operador el más fundamental que describe un sistema físico. Con un langragiano se puede explorar la mecánica en sistemas alternativos de coordenadas cartesianas, como coordenadas polares, cilíndricas y esféricas.

[4] Roger Fletcher fue galardonado en 1997 con el Premio Dantzig por sus contribuciones fundamentales a los algoritmos de optimización no lineal.

$$W = \sum_j \Phi(x_j)^2 \beta_j - \sum_{i,j} \beta_i \beta_j \Phi(x_i).\Phi(x_j). \quad (8)$$

Como las variables u_j no aparecen en el lagrangiano, entonces se reemplaza por las restricciones:

$$0 \leq \beta_j \leq C, \qquad j = 1, \ldots, N \quad (9)$$

Siguiendo el método Vector de Soporte se representan los productos de puntos $\Phi(x_j).\Phi(x_j)$ mediante un Kernel $K(x_i, x_j)$. Usando el núcleo gausiano:

$$K(x_i, x_j) = e^{-q\|x_i - x_j\|^2}, \quad (10)$$

con el parámetro de ancho q. Los núcleos polinomiales no producen representaciones de contornos ajustados de clusters. El wolfe – langragiano resulta ahora:

$$W = \sum_j K(x_j, x_j)\beta_j - \sum_{i,j} \beta_i \beta_j K(x_i, x_j) \quad (11)$$

Para cada punto x se define su distancia en el espacio de características desde el centro de la esfera:

$$\mathcal{R}^2(x) = \| \Phi(x) - a \|^2 \quad (12)$$

De acuerdo a la ecuación (4), y la definición del kernel, entonces:

$$\mathcal{R}^2(x) = K(x,x) - 2\sum_j \beta_j K(x_j, x) + \sum_{i,j} \beta_i \beta_j K(x_i, x_j) \quad (13)$$

Y el radio de la esfera es:

$$\boldsymbol{\mathcal{R} = \{\mathcal{R}(x_i) \mid x_i \; un \; Vector \; de \; Soporte\}} \quad (\mathbf{14})$$

Los contornos que encierran los puntos en el espacio de datos, están definidos por el conjunto:

$$\{x \mid \mathcal{R}(x) = \mathcal{R}\} \quad (15)$$

De acuerdo a la ecuación (14) los puntos Vectores de Soporte se encuentran en los límites del clúster, mientras que los puntos Vectores de Soporte Limitado están fuera, y todos los demás puntos se encuentran dentro de los clusters.

5.2 Asignación de Clusters

Se realiza la asignación de los puntos con un enfoque geométrico $\mathcal{R}(x)$ basado en la siguiente observación: dado un par de puntos de datos que pertenecen a diferentes clusters cualquier camino que los conecte debe salir de la esfera en el espacio de características.

Tal camino contiene un segmento de puntos. Esto conduce a la definición de la matriz de adyacencia A_{ij} entre los pares de los puntos x_i y x_j

$$A_{ij} = \begin{cases} 1 \; si \; para \; todo \; Y \; en \; el \; segmento \; de \; línea \; que \; conecta \; x_i \; y \; x_j, \mathcal{R}(Y) \leq \mathcal{R} \\ 0 \; de \; otra \; manera \end{cases} \quad (16)$$

5.3 Método de Agrupación

El método de agrupación no tiene un sesgo explícito ni del número, ni de la forma de los clusters. Tienen dos parámetros permitiendo obtener varias soluciones de agrupamiento.

El parámetro q del núcleo gaussiano determina la escala a la que se sondean los datos y, a medida que aumenta, los grupos comienzan a dividirse.
El otro parámetro, p, es la constante de margen suave que controla el número de valores atípicos. Este parámetro permite analizar puntos de datos con ruidos y separarlos entre clusters superpuestos.

6 Desarrollo y Resultados

El análisis de las anomalías en los datos tributarios se aplicará sobre aquellos datos que son resultantes de inspecciones fiscales realizadas a contribuyentes determinándole de oficio el impuesto que debían pagar en un determinado intervalo de tiempo, acotado este tiempo por plazos de prescripción, lo que incide en su exigibilidad.
Cuando un contribuyente otorga su conformidad en una inspección realizada, es decir da su consentimiento y opta por no recurrir la determinación de oficio, finalmente entonces queda la regularización de ese ajuste impositivo determinado. Esa regularización se lleva a cabo con la generación de una Liquidación de Deuda (LD). Ese procedimiento de generación de la Liquidación de Deuda (LD) concluye con la emisión de un Boleto de Deuda que le permite pagar al contribuyente, imputando tales ajustes y pagos en la cuenta corriente del contribuyente en los períodos determinados por la inspección.
Las Liquidaciones de Deudas (LD) son generadas por los operadores del sistema tributario de la administración tributaria, sistema que registra diferentes operaciones transaccionales fiscales (planes de pagos, declaraciones juradas, etc) y esa Liquidación de Deuda (LD) generada es validada por el contribuyente al recibir del operador del sistema, el Boleto de Deuda para proceder a la cancelación de su obligación tributaria.
La cantidad de registros correspondientes a Liquidaciones de Deudas (LD) generados en el ejercicio 2017 es aproximadamente de 100.000 (de distintos contribuyentes y actividades). Del total de los registros se conforma un subconjunto de datos compuesto por liquidaciones que poseen características en los datos que evidencian modificaciones de las declaraciones juradas de los contribuyentes y que los valores consignados en los campos Impuesto y Retención son coincidentes. La muestra seleccionada al azar para analizar las causales de las imperfecciones es la correspondiente a uno de esos contribuyentes, y esa muestra posee 59 registros.

6.1 Datos de Planilla de cálculos

Los datos están contenidos en una planilla de cálculos, por consiguiente no poseen integridad referencial. Las denominaciones de las columnas son las siguientes:

1. SPO_CUIT: Clave Única de Identificación Tributaria del contribuyente.
2. SPO_DENOMINACION: Nombre o razón social del contribuyente.
3. NRO_INSCRIPCION: Número de inscripción del Impuesto.
4. ROL: Identificación del Impuesto.
5. TIPO_LD: Código que identifica el tipo de Liquidación de Deuda.

6. NUMERO_LD: Número que se le asigna a la Liquidación de Deuda.
7. EJERCICIO: Año Fiscal.
8. ANTICIPO: Mes correspondiente al Año Fiscal.
9. IMPUESTO: Impuesto declarado por el contribuyente.
10. SALDO_CUENTA: pago del mes anterior a favor del contribuyente.
11. RETENCION: Retención del impuesto practicada al contribuyente.
12. BOLETO: Número del boleto de pago.
13. ACTIVIDAD: Código de actividad del contribuyente, nomenclador Ley Impositiva.
14. IMPORTE_MULTA: Multa generada al contribuyente.
15. MULTA_TERMINO: Multa con intereses resarcitorios.
16. MONTO_INSPECTOR: Porcentaje de la multa para el inspector (incentivo).
17. FEC_ALTA: Fecha de alta de la Liquidación de Deuda.

El análisis se concentra en dos columnas de la planilla de cálculos: la columna IMPUESTO y la columna RETENCIÓN, donde se observa que existen valores coincidentes, lo que implicaría sin ejercer un análisis integral de las causales de esta anomalía en los datos, que en ciertos períodos fiscales correspondientes a declaraciones juradas mensuales de acuerdo al impuesto a que se refieren, el contribuyente no tendría que pagar tributo alguno ya que el saldo del impuesto a pagar será cero, porque ese valor ha sido retenido.
En todas las coincidencias bajo análisis, nunca el valor de la retención supera al valor del impuesto, lo que conlleva a que no hay un saldo negativo, iniciando el intervalo en cero, hacia valores positivos.
Lo que se sabe, en virtud del análisis previo en el cual se identifica las relaciones de los datos entre sí, es que los valores que han sido consignados en la columna RETENCIÓN que son iguales a la columna IMPUESTO son datos que no se ajustan ni al modelo de datos, ni a los procedimientos establecidos resultantes de su sistema de calidad.

6.2 Análisis de los Datos

Se aplica el algoritmo Agrupamiento de Soporte Vectorial a un atributo del conjunto de ejemplos (ExampleSet), el que se denomina IMP-RET (que significa Impuesto menos Retención) y que fue creado para poder entender con mayor claridad el comportamiento de los datos bajo análisis. Ese atributo entonces tendrá los valores resultantes de restar de la columna IMPUESTO de la planilla de cálculos, los valores de la columna RETENCIÓN. Los valores que sean iguales a cero (0) serán representativos de los registros en los que el valor determinado del Impuesto sobre los Ingresos Brutos es igual a la retención cargada manualmente por el operador y validada por el contribuyente en la generación del Boleto de Deuda.

6.3 Ejecución del Algoritmo

Ajustado los parámetros por defecto del algoritmo de Segmentación cambiando el valor de convergencia que especifica la precisión de las condiciones de los clusters, el algoritmo Agrupamiento de Soporte Vectorial (Support Vector Clustering, SVC) arroja por resultado diez (10) clusters que agrupan los siguientes ítems, los que se pueden visualizar en la Figura 1 – Dispersión de clusters en tres dimensiones – y en la Figura 2 – Dispersión de clusters en dos dimensiones – :

1. En el Cluster 0, 18 items, conjunto de valores distintos.
2. En el Cluster 1, 03 items, con el valor 7296.030
3. En el Cluster 2, 03 items, con el valor 9194.670
4. En el Cluster 3, 03 items, con el valor 11710.260
5. En el Cluster 4, 03 items, con el valor 12594.140
6. En el Cluster 5, 03 items, con el valor 14641.750
7. En el Cluster 6, 04 items, con el valor 12759.080
8. En el Cluster 7, 04 items, con el valor 11516.050
9. En el Cluster 8, 03 items, con el valor 9279.830
10. En el Cluster 9, 15 items, con el valor 0

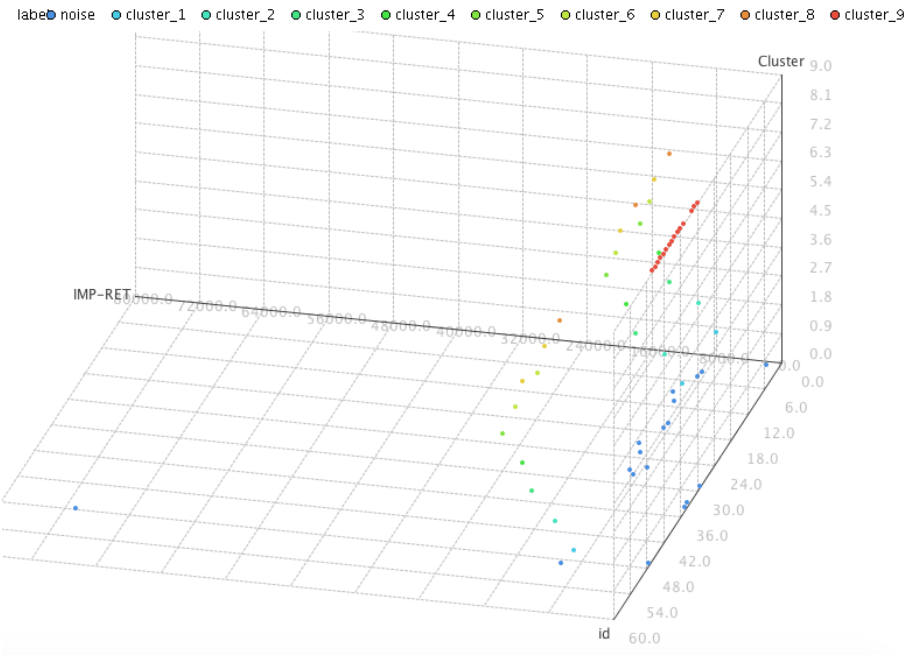

Fig. 1. Dispersión de clusters en tres dimensiones.

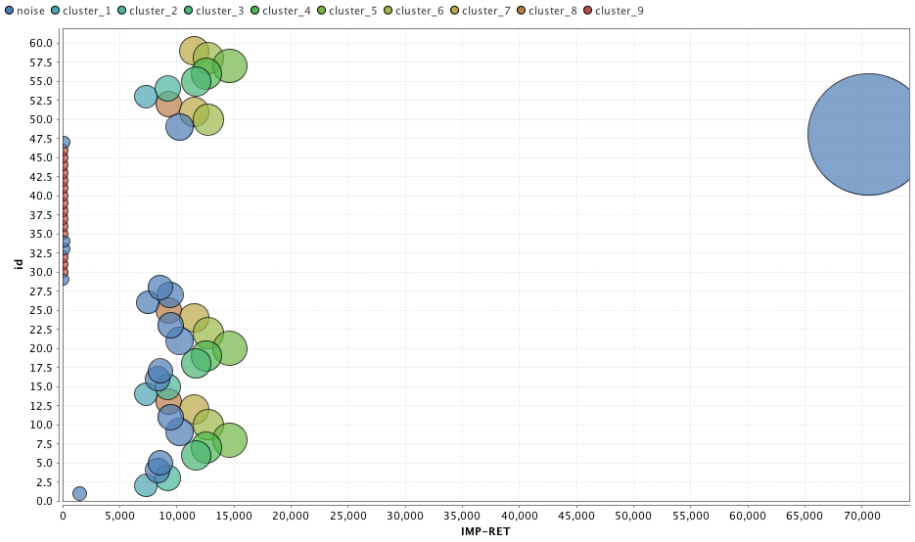

Fig. 2. Dispersión de clusters en dos dimensiones.

6.4 Primeras deducciones del análisis

A los datos agrupados en los distintos clusters se les puede adicionar información para arribar a ciertas deducciones, casi como si se tratara de la generación de un modelo de aprendizaje supervisado:

1. Las valores imputados por el operador del sistema tributario, que se encuentran en la columna RETENCION no pueden constatarse con respaldo documental que los avale, siendo esos valores en definitiva, datos que no se ajustan al modelo de datos ni a los procedimientos establecidos por el sistema de calidad de la administración tributaria.
2. Puede identificarse el operador de sistema que hizo la carga de los datos.
3. Y del análisis del año de cada uno de los registros puede inferirse una lógica de la incosistencia del dato.

De la muestra total de cincuenta y nueve registros (59), dieciocho (18) de ellos, que representan el treinta y un por ciento (31%) de la muestra, en principio no demandarían un análisis extra a los objetivos planteados, y son los datos agrupados en el Cluster 0, definido por el algoritmo seleccionado como ruido (Noise).
Veintiseis (26) registros quedaron agrupados en distintos clusters, porque el valor del impuesto a pagar es el mismo en distintos ejercicios, lo que puede visualizarse en el gráfico de tres dimensiones, representando este agrupamiento el cuarenta y cuatro por ciento (44%) del conjunto de los datos.

Y quince (15) registros en donde el valor del impuesto declarado es igual al de la retención, representando este cluster el veinticinco por ciento (25%) de la muestra.
Como el objetivo planteado es el descubrimiento de patrones en los ítems de la muestra que expliciten la causal de la anomalía en los datos y que evidencien posibles ilícitos por parte del contribuyente, y/o hechos de corrupción, y los ítems fueron preparados a tal fin, el cluster 0 es definido como ruido, porque en principio, esos datos no evidencian con claridad posible ilícitos y/o hechos de corrupción, pero ello no significa que no puedan existir.
Los ítems del cluster 1 al cluster 8 representan el cuarenta y cuatro por ciento (44%) de la muestra, siendo estos valores reveladores de ciertos patrones y puede establecerse que:

4. Los valores del cluster 1 al cluster 8 pueden evidenciar un posible ilícito, como por ejemplo, evasión fiscal, siempre y cuando en el análisis de las causas de la imperfección de los datos no pueda aseverarse con certividad que esos datos corresponden por ejemplo a errores de carga de las Liquidaciones de Deuda por parte del operador o a otras causas, y
5. Los valores del cluster 9, verificándose que la imputación de los valores al campo RETENCIÓN se hizo de forma manual por el operador del sistema tributario, podría evidenciar un posible hecho de corrupción, porque podría haber connivencia entre el contribuyente y el operador, ya que la carga se efectúa en presencia del contribuyente, y es este el que valida la carga y generación de la Liquidación de Deuda (LD), recibiendo el Boleto de deuda, con el que cancela los ajustes resultantes de la determinación impositiva.

7 Conclusiones

Si la causal de la anomalía tiene una explicación lógica o documentada como por ejemplo datos perdidos, errores de los datos, incoherencias en los datos, o metadatos ausentes o erróneos, entonces no estamos en presencia de algún indicio de ilícito y/o hecho de corrupción. El aplicar el algoritmo indicado, presupone no solo conocer cómo se comporta el algoritmo con el conjunto de ejemplos (ExampleSet), sino también comprender los datos que se están evaluando, y sus interrelaciones.
Hasta ahora no hay una aplicación específica con algoritmos para Minería de Datos Impositivos (Liu et al. 2012). Tampoco existen algoritmos de Aprendizaje Automático Impositivo, correspondientes a modelos predictivos o descriptivos.
El análisis de datos impositivos con algoritmos de segmentación contribuye en la detección del comportamiento de un contribuyente, o de un grupo de contribuyentes, ya que las posibilidades de parametrización y la creación de modelos responden a distintas alternativas en virtud de los objetivos planteados oportunamente y de la preparación de los datos, de acuerdo a la metodología CRISP-DM o de alguna metodología basada en ella (Lopez-Pablos 2013).
Las observaciones de los agrupamientos de datos impositivos y sus características comunes pueden revelar anomalías, las cuales deben advertirse porque ellas pueden estar exteriorizando ilícitos y hechos de corrupción, en virtud de sus causales. Por eso,

la detección de datos atípicos tributarios, o datos tributarios anómalos, conduce al descubrimiento de pequeños conjuntos de datos que serán significativamente muy diferentes al resto de los datos tributarios bajo análisis, y justamente el análisis de estos datos incosistentes y sus causales será más valioso que el análisis general de todos los datos de la muestra, basándose ello en que justamente los objetivos del análisis de los datos se concentra en determinar las causales de la anomalía, sin perder de vista que la premisa es que exista calidad en los datos de las bases de datos tributarias, con lo cual no habría lugar para la existencia de anomalías de este tipo, hecho aún más llamativo, cuando estas incosistencias responden a un patrón de conducta de un mismo contribuyente, de un conjunto de contribuyentes, de una actividad determinada, de un ejercicio específico o de un operador del sistema tributario, sin una causal asertiva.

El correcto análisis de las anomalías en los datos tributarios permitirá explicitar las causales de esas anomalías, y si se puede determinar que no son ilícitos por parte de los contribuyentes, ni hechos de corrupción, entonces permitirá segregar y limpiar estas anomalías de las bases tributarias, para ir depurandola, corrigiendo las cuentas corrientes de los contribuyentes, optimizando la calidad del dato tribuario, para contribuir a procesos de aprendizaje automático, y una correcta y valiosa extracción de conocimiento no trivial.

Referencias

1. Ben-Hur, Asa, David Horn, Hava T. Siegelmann, y Vladimir Vapnik. 2001. «Journal of Machine Learning Research». *Support Vector Clustering* 2001.
2. Fletcher, Roger. 2000. *Practical Methods of Optimization*. 2nd ed.
3. Kuna, Horacio Daniel. 2014. «Procedimientos de explotación de información para la identificación de datos faltantes con ruido e inconsistentes». Universidad de Málaga. http://sistemas.unla.edu.ar/sistemas/gisi/tesis/UM-TD-Horacio-KUNA.pdf.
4. Liu, Bin, Guang Xu, Qian Xu, y Nan Zhang. 2012. «Outlier Detection Data Mining of Tax Based on Cluster». *2012 International Conference on Medical Physics and Biomedical Engineering (ICMPBE2012)* 33 (Supplement C): 1689-94. https://doi.org/10.1016/j.phpro.2012.05.272.
5. Lopez-Pablos, Rodrigo. 2013. «Elementos de Ingeniería de Explotación de la Información aplicados a la Investigación Tributaria Fiscal». *Anales Asociación Argentina de Economía Política*.
6. SAS® Institute Inc. 2015. «La Minería de Datos de la A a la Z: Como Descubrir Conocimientos y Crear Mejores Oportunidades». SAS® The Power to Know. 2015. https://www.sas.com/content/dam/SAS/es_mx/doc/assets/26-mineria-datos-a-z.pdf.