CORRECCIÓN DE DATOS IMPERFECTOS EN BASES DE DATOS MÚLTIPLES, ANÓMALAS, ABIERTAS Y PÚBLICAS MEDIANTE APRENDIZAJE AUTOMÁTICO

Ing. Antonio SOTTILE BORDALLO
Profesor Titular Efectivo Computación
Lic. Daniel CAVALLER RIVA
Profesor Asociado Efectivo Computación
Lic. Héctor Nicolás SOSA
Adscripto Computación
Lic. Diego Vidal SILVA
Adscripto Computación
Cont. Cristian Darío ORTEGA YUBRO
Adscripto Computación
Ing. Norma Lidia AZCURRA
Especialista Externa

INTRODUCCIÓN

Los datos abiertos, públicos, personales y los informatizados (Pascual et al. 2000), como cualquier otro tipo de dato, ejemplo los datos no tradicionales (Álvarez et al. 2016), están expuestos a anomalías, y estas anomalías responden a distintas causales. Un dato es anómalo o inconsistente, cuando ese dato encuadra en los conceptos de imprecisión, incertidumbre y vaguedad (Álvarez et al. 2016). La imprecisión es la carencia de exactitud en la expresión, ocurre cuando el dato es desconocido, o su atributo es impreciso por su naturaleza. La incertidumbre manifiesta una situación determinada en donde no hay seguridad acerca de la veracidad. Por último, la vaguedad sucede cuando la información es afectada por imprecisión, por incertidumbre o ambos. A veces es difícil establecer los límites entre la imprecisión y la incertidumbre. La anomalía de los datos también procede de otras deficiencias como datos erróneos o inconsistentes conocidos como outliers, por su denominación en inglés. La detección de las anomalías en los datos y el correcto análisis de sus causales aportarán elementos en la búsqueda de indicios reveladores de ilícitos y hechos de corrupción, independientemente del tipo de organización, y si la organización se encuentra en el Sector Público o en el Sector Privado. Las anomalías en los datos coexisten en las múltiples bases de datos a las que accede el Estado, es decir en aquellas bases de datos y en aquellos datos no tradicionales que utiliza y produce la Administración Pública, sean estos datos de origen interno como los de origen externo. En el caso puntual de un organismo de la Administración Pública, como lo son aquellos organismos que administran tributos, los datos conceptualizados como no tradicionales proceden de diversas fuentes y orígenes, por ejemplo, como los que se detallan:

- El sitio web de las Administradoras Tributarias,
- Otros sitios y portales de la Administración Pública Provincial
- Los canales de acceso y de relación con los contribuyentes como la oficina virtual
- Las redes sociales
- Los correos electrónicos de los contribuyentes
- Las imágenes
- Los datos geográficos, y otras fuentes de datos.

Si las bases de datos y los datos no tradicionales que tienen a disposición las Administraciones Tributarias, contienen información acerca de todas las actividades sujetas a gravamen 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 una causa, pudiendo evidenciar estas causas ciertos ilícitos y hechos de corrupción cuando existe la connivencia del contribuyente con el empleado público o funcionario público. Los ilícitos fiscales como la evasión fiscal, y los hechos de corrupción que se relacionan con los aspectos impositivos son uno de los tantos problemas a resolver. Esos problemas son independientes de los sistemas tributarios preestablecidos y del nivel de desarrollo del país, y de cada una de sus provincias, ya que en gran parte tienen que ver con el grado de nivel de Cultura Tributaria que se posea, la cual se debilita con tales hechos de corrupción.

El propósito de la presente investigación, es el desarrollo teórico del análisis causal de las anomalías de los datos, teniendo en cuenta el gran volumen de datos que existen en la Administración Pública, especialmente en las administraciones tributarias, demostrando que los procesos metodológicos de la Minería de Datos y Aprendizaje Automático contribuyen a evidenciar indicios de ilícitos y hechos de corrupción, a través de la aplicación de algoritmos.

Objetivos

- Detectar anomalías en los datos provenientes de bases de datos múltiples, y las inconsistencias de los datos no tradicionales, sean de origen interno y de origen externo a la organización, en este caso, una Administración Tributaria.
- Analizar las causales de las anomalías y clasificarlas, segregando aquellas causales que manifiesten indicios de ilícitos y hechos de corrupción, del resto de las causales de las imperfecciones en los datos de las bases de datos múltiples y en los datos no tradicionales.
- Corregir solamente las anomalías en los datos que puedan aseverarse que su causal de inconsistencia no responde a indicios de ilícitos y hechos de corrupción, para lograr la mayor calidad de los datos bajo análisis con el propósito de garantizar optimizar los modelos, en virtud de los objetivos planteados para la Minería de Datos.
- Utilizar el conocimiento adquirido de la aplicación de los modelos, para inducirlo en los procesos metodológicos del Aprendizaje Automático, y redefinir los procesos y procedimientos de la organización de una Administración Tributaria, con los procesos metodológicos de la Minería de Procesos, como segunda instancia.

Originalidad e importancia

La utilización de la Minería de Datos, la Minería de Procesos, y el Aprendizaje Automático en la Administración Pública, puntualmente en administraciones tributarias, para detectar patrones ocultos en los datos que habilitan la descripción de los comportamientos, con el objetivo que estos comportamientos evidencien ilícitos y posibles hechos de corrupción, impactarán en la construcción del bienestar de la ciudadanía, para que estos ilícitos y hechos de corrupción no lleguen a ser condicionantes de la calidad de vida que merecen los ciudadanos, y de la calidad de los servicios públicos a los que ellos acceden, y contribuyan a fortalecer la cultura tributaria y el contrato social.

El correcto uso de la Minería de Datos, la Minería de Procesos y el Aprendizaje Automático en la Administración Pública, como en el Sector Privado, aportarán calidad a los datos, tanto a los de origen interno, como a los de origen externo de la organización, a través de su depuración llevada a cabo en el Proceso de Extracción del Conocimiento. Por tal motivo, es muy importante el estudio de las causales de las anomalías en los datos que se presenten, y las razones para

---

[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.

vincular estas anomalías de los datos a los posibles ilícitos y hechos de corrupción en los distintos organismos de la Administración Pública, especialmente en las Administraciones Tributarias, por la gran cantidad de datos que estos organismos acceden y producen en su quehacer diario. En la Provincia de Mendoza como en el resto del país, la Administración Tributaria además de ser un consultor de datos de distintas fuentes y orígenes, es un productor de datos a través de las distintas aplicaciones y sistemas informáticos que utiliza y pone a disposición del contribuyente, siendo esas herramientas informáticas adquiridas a terceros o producidas por sus recursos humanos. La importancia en la calidad de los datos que se manipulan en las Administraciones Tributarias, teniendo siempre presente la sensibilidad del mismo, está íntimamente relacionada con los patrones de comportamientos de los contribuyentes que contienen esos datos, abarcando información muy valiosa, como por ejemplo las actividades gravadas que desarrollan, y otros datos implícitos en los metadatos. Por eso es clave la integración del Big Data a la Administración Tributaria, porque la comparación de diferentes métodos estadísticos y diferentes bases de datos impositivas permite detectar irregularidades (Stankevicius y Leonas 2015). La creación de modelos para evidenciar ilícitos y hechos de corrupción se basa en el análisis de los datos del contribuyente, las anomalías de los datos, y los procesos y procedimientos definidos por las Administraciones Tributarias, a través de la Minería de Datos, la Minería de Procesos y el Aprendiza Automático. El conocimiento resultante del Proceso de Extracción del Conocimiento también debe ser un pilar para el diseño de aplicaciones informáticas inteligentes que sean destinadas a optimizar la interacción y orientación del contribuyente con la Administración Tributaria a través de un servicio que se oriente mucho más que satisfacer la experiencia del usuario, desburocratizando así largos procedimientos y tortuosos trámites, que impactarán primordialmente en el búsqueda del fortalecimiento de la Cultura Tributaria.

Ejes de las propuestas

Ejes Centrales

- Clasificar las anomalías detectadas en los datos de fuente interna, de bases de datos múltiples, y las inconsistencias en los datos no tradicionales, que permitan orientar el proceso del análisis de las causales de esas anomalías.
- Agrupar las causales de las anomalías de los datos en dos grupos:
    - las que manifiesten indicios de ilícitos y hechos de corrupción y
    - las que no manifiestan indicios de ilícitos y hechos de corrupción,
- Orientar esas causales al Proceso de Extracción del Conocimiento, utilizando algoritmos que sean adecuados a los objetivos planteados.
- Inducir el conocimiento al Aprendizaje Automático, modelar y aplicar Minería de Procesos.

Ejes Periféricos

- Poner a disposición de la Facultad de Ciencias Económicas – Sede Central y Delegación San Rafael – de la Universidad Nacional de Cuyo, los resultados y las conclusiones obtenidas en la presente investigación, para fomentar el estudio continuo y profundizado de la Minería de Datos, la Minería de Procesos, y el Aprendizaje Automático, revelando su importancia para el profesional de Ciencias Económicas, y desarrollando una metodología estandarizada.
- Publicar los resultados obtenidos y las conclusiones arribadas de la presente investigación en las Jornadas Provinciales de Ciencias Económicas, en el Congreso Nacional de Ciencias Económicas, en las Jornadas Argentinas de Informática e Investigación Operativa, y en el repositorio digital del Sistema Integrado de Documentación de la Universidad Nacional de Cuyo.

MARCO TEÓRICO

Los procesos metodológicos de la Minería de Datos, la Minería de Procesos y del Aprendizaje Automático 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 digitales, siendo este conocimiento previamente desconocido. El conocimiento adquirido cuando es inducido, ajustará los procedimientos de la organización. Este ajuste puede realizarse a través de la Minería de Procesos (IEEE Task Force on Process Mining s. f.). Por ello, se infiere el siguiente axioma: la calidad resultante del conocimiento no trivial que se extraiga dependerá en gran medida, de la calidad que posean los datos que se tenga al alcance. Por el contrario, la baja calidad de los datos concluirá en una extracción de conocimiento no trivial deficiente, y esta baja calidad de los datos, puede provenir de las anomalías que tienen los datos, como por ejemplo el ruido de los datos, los valores perdidos y los datos atípicos entre otras inconsistencias. Los modelos que se generan en la Minería de Datos pueden ser descriptivos o predictivos (Kuna 2014), y dependerá de ello la aplicación de los algoritmos que se utilicen para minar los datos y también en el Aprendizaje Automático, como aprendizaje supervisado y aprendizaje no supervisado, ya que los algoritmos son especificos. El algoritmo de agrupamiento se orienta al aprendizaje no supervisado, donde la agrupación de los datos está relacionada con las características comunes que los datos poseen. Este algoritmo es muy utilizado cuando se quiere descubrir conocimiento oculto, patrones de comportamiento y valores extremos de los datos (Kuna 2014). Por eso cuando se analiza la distancia entre los datos de un conjunto de datos, el criterio general del análisis es que cuanto mayor es la distancia entre un dato de una base de datos y el resto de la los datos del conjunto de datos, mayor es la posibilidad de considerar al dato como anómalo. El Aprendizaje Automático en definitiva es un proceso de inducción del conocimiento adquirido a través de distintos lenguajes de programación que interpreten los algoritmos como por ejemplo el lenguaje Python® o por medio de programas que incluyen los algoritmos específicos de Aprendizaje Automático como por ejemplo RapidMiner Studio®.

Estado del Arte

Las metodologías que se aplican en la Minería de Datos, la Minería de Procesos, y el Aprendizaje Automático no son utilizadas por la Administración Pública Provincial, ni se las considera como válidas para analizar anomalías de datos digitales, por consiguiente, no se orientan a detectar y evidenciar ilícitos y hechos de corrupción, tal vez, por desconocimiento, por falta de capacitación o falta de recursos, tanto humanos como materiales, o por subestimar los beneficios del Big Data. Además esas metodologías pueden generar aportes muy valiosos a la Auditoría Gubernamental, cuando se quiera evidenciar patrones en la gran cantidad de datos digitales que la Administración Pública produce, como también puede aportar a las distintas áreas de Control de Gestión de la Administración Pública, o a los Sistemas de Calidad implementados. Actualmente el Estado Provincial no utiliza Big Data. Más aún, no existe oferta ni demanda de profesionales capacitados dentro de la esfera de la Administración Pública que certifiquen conocimientos en Big Data, a través de la certificación como Científico de los Datos, Analista de Datos o similar, expedida por una autoridad certificante especializada en la materia. No obstante ello, actualmente existe una innumerable cantidad de información acerca de estas metodologías analíticas de los datos orientadas a la detección de fraude en el Sector Privado con el objeto de prevenir corrupción contable y financiera, evidenciándose con su implementación los beneficios de su utilización, lo que lleva a considerar las potencialidades que su correcta aplicación pueda significar en la concreción de la modernización del Estado y

en la lucha contra la corrupción, aportando transparencia en la Administración Pública de manera eficiente. Puntualmente en las Administraciones Tributarias su utilización sería muy beneficiosa y oportuna, siendo este tipo de organización gubernamental por la tarea que realiza, las que manejan y producen la mayor cantidad de datos, los que no solo son de materia tributaria, sino que abarcan aspectos relevantes del análisis económico tanto nacional como regional, ya que contienen por ejemplo, información relevante de las actividades económicas que se llevan a cabo, con lo que podrían desarrollarse modelos descriptivos como modelos predictivos de la economía.

Hipótesis de las Propuestas

1. Los datos tienen anomalías. No puede concebirse que en la gran cantidad de datos a los que se accede y genera la Administración Pública no existan errores. Existen, y tienen una razón de ser.
2. La experimentación, aplicación y uso de técnicas, procesos y algoritmos de corrección de datos anómalos en bases de datos múltiples, independientemente de su condición, es decir datos abiertos, públicos, e informatizados, constituirán procedimientos metodológicos para evidenciar indicios de comportamientos ilícitos y hechos de corrupción.
3. La evaluación de la calidad de los datos y la corrección de las anomalías de los mismos que no se ajustan a los estándares de calidad, examinando sus causales, es factible y muy valioso y oportuno en las Administraciones Tributarias, ya que fortalecen el sistema de control interno, y el sistema de gestión de la calidad.

El Proceso de Extracción de la Información o del Conocimiento

Existen diversas metodologías para la implementación de un Proceso de Extracción del Conocimiento. En el Anexo I – Metodología Propuesta, de la presente investigación, se desarrolla una metodología que se considera oportuna y estandarizada a los objetivos planteados, basándose en los principios rectores del CRISP-DM[2].

Minería de Procesos

De acuerdo al manifiesto del IEEE[3], la Minería de Procesos es una disciplina que se ubica entre la Inteligencia Computacional y la Minería de Datos, por un lado, y la Modelización y el Análisis de los Procesos por otro, y tiene por objeto descubrir, monitorear y mejorar los procesos reales, a través del Proceso de Extracción del Conocimiento de los registros de los eventos disponibles en los actuales sistemas. La Minería de Procesos incluye el descubrimiento automático de procesos, la verificación, la minería de redes organizacionales, la construcción de modelos automatizados de simulación, la extensión de los modelos y otros. Con técnicas de Minería de Procesos se puede verificar el cumplimiento de normativas y establecer la validez y confianza de la información en los procesos críticos de la organización. Estas técnicas asumen que se pueden registrar eventos secuencialmente y que cada evento se refiere a una actividad,

---

[2] *CRISP-DM (Cross Industry Standard Process for Data Mining) fue concebido en el año 1996. En 1997 se puso en marcha como un proyecto de la Unión Europea dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR Corporation y Ohra, una compañía de seguros. La primera versión de la metodología se presentó en Bruselas en marzo de 1999, y fue publicada como una guía paso a paso, de la minería de datos. Actualmente IBM es la principal empresa que promueve la metodología CRISP-DM haciendo disponibles algunos de los documentos originales para su descarga, y ha incorporando esta metodología a su producto SPSS Modeler.*

[3] *http://www.win.tue.nl/ieeetfpm/doku.php?id=start*

y se relaciona a un caso particular. Por lo tanto, los registros de eventos pueden almacenar información adicional, como por ejemplo, la marca del tiempo, la cual puede encontrarse en los metadatos. Uno de los principios rectores de la Minería de Procesos, justamente hace hincapié en los eventos registrados, sin tener en cuenta en donde se registran. Los eventos pueden estar almacenados en múltiples bases de datos, correos electrónicos, registros de transacciones, y otras fuentes internas o externas de datos, o en datos no tradicionales. Estos eventos deben ser confiables, es decir, la registración de esos datos debe ser de calidad, logrando que cualquier evento registrado debe poseer una semántica bien definida, y los datos de los eventos deben ser seguros. De esta manera y con esta concepción, pueden ajustarse los procesos definidos por la organización en virtud de los eventos.

Anomalías de los Datos

Las razones por las cuales existen datos inconsistentes y bases de datos anómalas se deben a distintas causales, como pueden ser:

- La carga incorrecta de los datos.
- Errores en el software utilizado o incompatibilidades entre distintos softwares.
- El dato proviene de una población distinta.
- Algún tipo de ilícito.
- Un hecho de corrupción.

Por eso, cuando se realizan tareas trabajando con datos categóricos o cuando no se trabaja con una distribución estándar de los datos, la identificación de estos datos anómalos es muy difícil. La presencia de anomalías en los datos en una base de datos pueden instaurar distorsiones en los resultados al realizar cualquier tipo de análisis sobre esa base de datos, por ejemplo, en los análisis o modelización de tipo predictiva. Por lo tanto, la distorsión que produce este tipo de datos puede tener consecuencias graves dependiendo del campo de aplicación en los cuales son utilizados. Buscar datos anómalos realizando consultas manuales o formalizar un análisis de tipo secuencial sobre todos los datos de una organización, requiere conocer previamente los posibles valores anómalos de esos datos y conocer las inconsistencias que puedan aparecer. Teniendo en cuenta que el tamaño y el volumen de las bases de datos día a día se incrementa, en muchos casos, esta búsqueda de datos anómalos y de inconsistencias con consultas manuales o análisis secuenciales sobre los datos, se torna una tarea impracticable y tediosa, con lo que llevarla a cabo de esa manera, puede generar que no solamente se cometan nuevos errores provocando otras anomalías en los datos o generación de ruido, sino que se omita el conocimiento implícito que tienen esos datos, por concentrar la atención en esa tarea específica y no poseer la capacidad de trabajar con grandes volúmenes de datos. Por eso es acertado para este tipo de búsquedas el uso de algoritmos específicos. La mayoría de los algoritmos utilizados para la detección de datos anómalos resuelven un tipo específico de problema, y la solución dependerá directamente del dominio del problema que se quiera resolver, en nuestro caso, del dominio de las políticas tributarias, su concepción y conocimiento de la Administración Tributaria, adoptando para ello, conceptos de distintas disciplinas, como la Estadística Analítica, la Minería de Datos, la Minería de Procesos y el Aprendizaje Automático. Hoy en día, es inestimable poseer y desarrollar mecanismos que permitan automatizar la búsqueda de los datos anómalos, entre los cuales, la aplicación metodológica del Proceso de Extracción de Conocimiento resulta fascinante debido a su capacidad para detectar patrones y relaciones entre los datos que no son evidentes a simple vista.

METODOLOGÍA Y RESULTADOS OBTENIDOS

Etimológicamente metodología en griego “metodólogos” significa tratado del método; es decir, a primera vista, una metodología trata del método, entendiendo por método el camino para

alcanzar una meta (del Moral et al. 2008). Cualquier metodología para ser tenida por tal, debe cumplir una serie de condiciones que pueden agruparse en dos grandes clases. La primera clase, formada por las así llamadas condiciones formales de adecuación, independientes del dominio donde se emplee la metodología y, en consecuencia, todas y cada una de las metodologías deben cumplirlas. Para la presente investigación el dominio está comprendido por la explotación de datos digitales de una Administración Tributaria. La segunda clase, constituida por las condiciones materiales de adecuación, las que son específicas de cada dominio en donde la metodología concretamente se use. No se vuelca en la presente investigación el desarrollo en su totalidad de la metodología descripta en el Anexo I – Metodología Propuesta, por una cuestión de tiempo, solamente se puntualizan aquellos pasos necesarios que responden a los objetivos buscados en la presente investigación.

Proceso de Extracción del Conocimiento

El proceso de extracción de conocimiento se aplica a datos que son los resultantes de inspecciones fiscales a los contribuyentes, determinando de oficio el impuesto que debían pagar en un determinado intervalo de tiempo, acotado este tiempo por los plazos de prescripción. Cuando el contribuyente conforma una inspección fiscal, es decir cuando da su consentimiento y opta por no recurrir la determinación, lo que finalmente queda es que regularice ese ajuste impositivo determinado en la inspección fiscal. Esa regularización, en la Provincia de Mendoza puede llevarse a cabo con la generación de una Liquidación de Deuda (LD). Ese proceso de generación de la Liquidación de Deuda (LD) concluye con la emisión de un Boleto de Deuda que permite pagar. Además, se debe cargar en la cuenta corriente del contribuyente, aquellos periodos a regularizar. Las Liquidaciones de Deudas (LD) son generadas por los operadores del sistema tributario, sistema que registra las diferentes operaciones transaccionales fiscales, y esa liquidación generada la valida el contribuyente, quien recibe el Boleto de Deuda. La cantidad de Liquidaciones de Deudas generadas en el ejercicio 2017 es aproximadamente de 100.000 registros. Del total de los registros se conforma un subconjunto que está compuesto por liquidaciones de deuda que poseen características que evidencian modificaciones de las declaraciones juradas de los contribuyentes, y además que los valores consignados en los campos Impuesto y Retención son coincidentes. Todos los datos personales (Pascual et al. 2000) que contiene la planilla de cálculos utilizada para modelar, han sido alterados para proteger la identidad del contribuyente, y los datos bajo análisis se utilizan exclusivamente para evaluar el comportamiento del algoritmo matemático de Aprendizaje Automático estudiado. Los datos al estar contenidos en una planilla de cálculos no poseen integridad referencial. El término integridad referencial se refiere a la exactitud o corrección de los datos en las bases de datos. (Date 2001). La planilla de cálculos posee diecisiete (17) columnas y sesenta (60) filas. Las denominaciones de las columnas son las siguientes:

01) SPO_CUIT: Clave Única de Identificación Tributaria del contribuyente.
02) SPO_DENOMINACION: Nombre o razón social del contribuyente.
03) NRO_INSCRIPCION: Número de inscripción del Impuesto.
04) ROL: Identificación del Impuesto.
05) TIPO_LD: Código que identifica el tipo de Liquidación de Deuda.
06) NUMERO_LD: Número que se le asigna a la Liquidación de Deuda.
07) EJERCICIO: Año Fiscal.
08) ANTICIPO: Mes correspondiente al Año Fiscal.
09) 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 en calidad de incentivo.
17) FEC_ALTA: Fecha de alta de la Liquidación de Deuda.

El análisis se concentra en dos columnas: 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, que en ciertos períodos fiscales correspondientes a declaraciones juradas mensuales, 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 de datos digitales en cero. 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 estudiado, ni a los procedimientos establecidos resultantes de su sistema de calidad.***
Por tal motivo la muestra debería agruparse en principio por las características descriptas, dos conjuntos de datos, resultantes en dos clusters:

- un conjunto donde IMPUESTO y RETENCIÓN tienen los mismos valores, y
- otro conjunto donde no los tienen.

Esa coincidencia de los valores entre la columna IMPUESTO y la columna RETENCIÓN se repite quince (15) veces en una muestra de cincuenta y nueve (59) registros, lo que representa aproximadamente, el veinticinco por ciento (25%) de la muestra seleccionada bajo análisis. Para continuar el proceso con un operador de Aprendizaje Automático en RapidMiner Studio® debería limpiarse del Conjunto de Ejemplos (ExampleSet) los datos anómalos, con lo cual, concebida correcta y documentadamente las causales de las anomalías, podrían depurarse esos datos inconsistentes de las bases de datos, aportando mayor calidad a la muestra de datos, para el diseño de los modelos a validar. Por consiguiente, cuando la causal de la anomalía no tiene una explicación lógica, ni documentada como por ejemplo podrían ser datos perdidos, errores de los datos, incoherencias en los datos, o metadatos ausentes o erróneos, entonces estamos en presencia de algún indicio de ilícito y/o hecho de corrupción. Una hipótesis que puede considerarse en el análisis de las causales, es justamente la exploración de los metadatos contenidos en los datos de la muestra, como por ejemplo su coincidencia en la estampa del tiempo, y modelizar paralelamente con algoritmos de análisis de metadatos.

Imagen 1 Planilla de Cálculos, producida por los autores

	C	D	E	F	G	H	I	J	K	L	Q
1	NRO_INSCR	ROL	TIPO_LD	NUMERO_LD	EJERCICIO	ANTICIPO	IMPUESTO	SALDO_CUENTA	RETENCION	BOLETO	FEC_ALTA
2	123456	IB	LI1	199900972170	2.003	1	7.296,03	0,00	0,00	2003000650306	23/10/2006 13:49:26
3	123456	IB	LI1	199900972170	2.003	2	9.194,67	0,00	0,00	2003000650307	23/10/2006 13:49:26
4	123456	IB	LI1	199900972170	2.003	3	8.333,33	0,00	0,00	2003000650308	23/10/2006 13:49:26
5	123456	IB	LI1	199900972170	2.003	4	8.556,51	0,00	0,00	2003000650309	23/10/2006 13:49:26
6	123456	IB	LI1	199900972170	2.003	5	11.710,26	0,00	0,00	2003000650310	23/10/2006 13:49:26
7	123456	IB	LI1	199900972170	2.003	6	12.594,14	0,00	0,00	2003000650311	23/10/2006 13:49:26
8	123456	IB	LI1	199900972170	2.003	7	14.641,75	0,00	0,00	2003000650312	23/10/2006 13:49:26
9	123456	IB	LI1	199902782076	2.003	7	9.194,67	0,00	9.194,67	2003000650312	01/12/2014 11:25:24
10	123456	IB	LI1	199900972170	2.003	8	10.268,43	0,00	0,00	2003000650313	23/10/2006 13:49:26
11	123456	IB	LI1	199902782076	2.003	8	9.856,39	0,00	9.856,39	2003000650313	01/12/2014 11:25:24
12	123456	IB	LI1	199903397744	2.003	8	10.200,00	0,00	10.200,00	2003000650313	22/06/2017 09:59:17
13	123456	IB	LI1	199903404040	2.003	8	10.268,43	3.404,08	0,00	2003000650313	07/07/2017 10:00:56
14	123456	IB	LI1	199900972170	2.003	9	12.759,08	0,00	0,00	2003000650314	23/10/2006 13:49:26
15	123456	IB	LI1	199902782076	2.003	9	12.759,08	0,00	12.759,08	2003000650314	01/12/2014 11:25:24
16	123456	IB	LI1	199903397744	2.003	9	12.700,00	0,00	12.700,00	2003000650314	22/06/2017 09:59:17
17	123456	IB	LI1	199903404040	2.003	9	12.759,08	0,00	0,00	2003000650314	07/07/2017 10:00:56
18	123456	IB	LI1	199900972170	2.003	10	9.455,48	0,00	0,00	2003000650315	23/10/2006 13:49:26
19	123456	IB	LI1	199900972170	2.003	11	11.516,05	0,00	0,00	2003000650316	23/10/2006 13:49:26
20	123456	IB	LI1	199903397744	2.003	11	11.500,00	0,00	11.500,00	2003000650316	22/06/2017 09:59:17
21	123456	IB	LI1	199903404040	2.003	11	11.516,05	0,00	0,00	2003000650316	07/07/2017 10:00:56
22	123456	IB	LI1	199900972170	2.003	12	9.279,83	0,00	0,00	2003000650317	23/10/2006 13:49:26
23	123456	IB	LI1	199903397744	2.003	12	9.200,00	0,00	9.200,00	2003000650317	22/06/2017 09:59:17
24	123456	IB	LI1	199903404040	2.003	12	9.279,83	0,00	0,00	2003000650317	07/07/2017 10:00:56

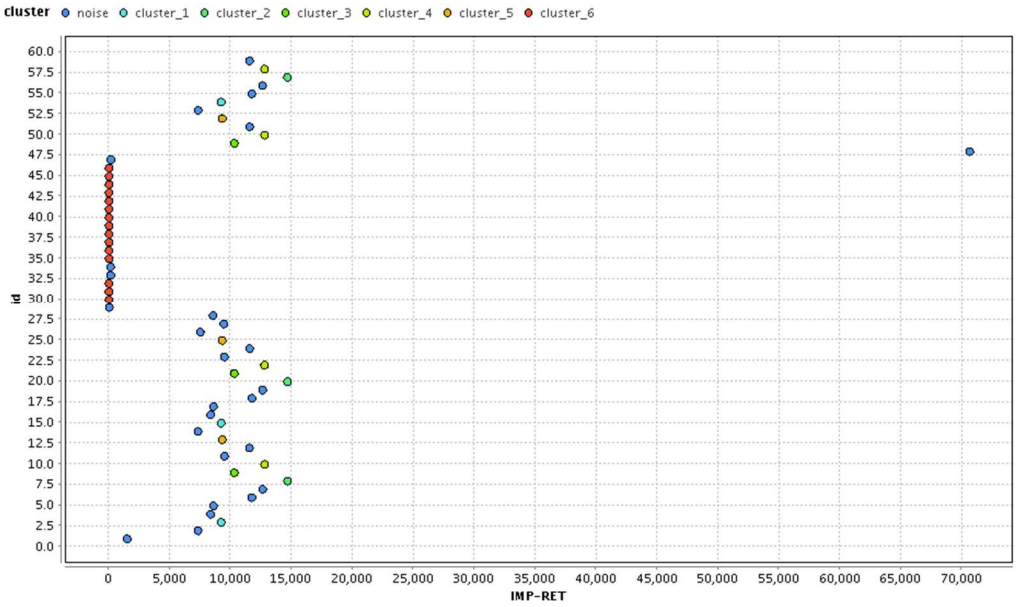
Resultados Obtenidos en Modelo I

La modelización utiliza un Algoritmo de Segmentación de la aplicación RapidMiner Studio® versión 8.2.[4]. Este Algoritmo de Segmentación se denomina Agrupamiento de Soporte Vectorial (Support Vector Clustering, SVC), y el mismo se desarrolla matemáticamente en el Anexo III – Algoritmia Utilizada. El proceso que se ejecuta para el desarrollo de lo que se denomina Modelo I, es el que se describe en el Anexo II – RapidMiner Studio Versión 8.2.

Para preparar los datos y poder aplicar el algoritmo SVC en la Etapa Modelización se ejecuta un proceso previo para crear una nueva columna en la muestra de los datos, y esta columna tendrá los valores resultantes de restar de la columna IMPUESTO los valores de la columna RETENCIÓN, de esta forma, los valores que sean 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, ya que esa carga en la generación de la Liquidación de Deuda, se hace en presencia del contribuyente, quien debe regularizar sus obligaciones impositivas determinadas por el inspector fiscal. Creada la columna impuesto menos retención denominada IMP-RET, se ejecuta el algoritmo con sus parámetros por defecto, y el resultado de la ejecución arroja siete (7) clusters que agrupan los siguientes valores:

- Cluster 0, 28 items de distintos valores distinto a cero.
- Cluster 1, 03 items con el valor 9194.670
- Cluster 2, 03 items con el valor 14641.750
- Cluster 3, 03 items con el valor 10268.430
- Cluster 4, 04 items con el valor 12759.080
- Cluster 5, 03 items con el valor 9279.830
- Cluster 6, 15 items con el valor 0

Imagen 2 Gráfico Agrupamiento de Soporte Vectorial Modelo I, producida por los autores



La cantidad de items, cincuenta y nueve (59), coincidentes con registros de la muestra.

Los valores del cluster cero (0) representan a los registros de veintocho (28) periodos mensuales del Impuesto sobre los Ingresos Brutos, en los cuales el campo IMPUESTO no es igual al

[4] https://rapidminer.com

campo RETENCIÓN y tampoco esos valores son iguales a los valores que contienen los clusters uno (1), dos (2), tres (3), cuatro (4) y cinco (5). A estos valores, los del cluster cero (0), el algoritmo los define como ruido del conjunto de ejemplos (ExampleSet). Los valores del cluster seis (6) forman parte del grupo que representan a los registros de quince (15) periodos mensuales en los cuales el campo IMPUESTO es igual al campo RETENCIÓN, ya que el valor de aplicar la función matemática establecida para el atributo IMP-RET, es cero (0).
Y los valores de los cluster uno (1), dos (2), tres (3), cuatro (4) y cinco (5), corresponden a periodos mensuales con identico valor distinto pero distinto de cero. A los datos contenidos en los distintos cluster se les puede adicionar información para arribar a ciertas conclusiones, casi como si se tratara de la generación de un modelo de aprendizaje automático supervisado:

- Las valores establecidos por el operador del sistema tributario en la columna RETENCION no se pueden constatar con respaldo documental que avale esa registración, siendo esos valores en definitiva, datos que no se ajustan al modelo de datos ni a los procedimientos establecidos por la Administración Tributaria.
- Puede identificarse el operador de sistema que hizo la carga de los datos.
- Y del análisis del año de cada uno de los registros puede encontrarse una lógica en la anomalía del dato.

De la muestra total de cincuenta y nueve registros (59), veintiocho (28) de ellos, que representan el cuarenta y ocho por ciento (48%) 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 algoritmos seleccionado como ruido. Dieciséis (16) 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 veintisiete porciento (27%) 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 porciento (25%) de la muestra. Como el objetivo planteado es el descubrimiento de patrones en los items de la muestra, que evidencien posibles ilícitos por parte del contribuyente, y/o hechos de corrupción, y los items fueron preparados a tal fín, 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 6 representan el cincuenta y dos porciento (52%) de la muestra, siendo estos valores reveladores de ciertos patrones y puede establecerse que:

- Los valores del Cluster 1 al 5 pueden evidenciar un posible ilícito, como por ejemplo, evasión fiscal, siempre y cuando en el análisis de las causas de 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
- Los valores del Cluster 6, 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 generacion de la Liquidación de Deuda (LD).

Resultados Obtenidos en Modelo II

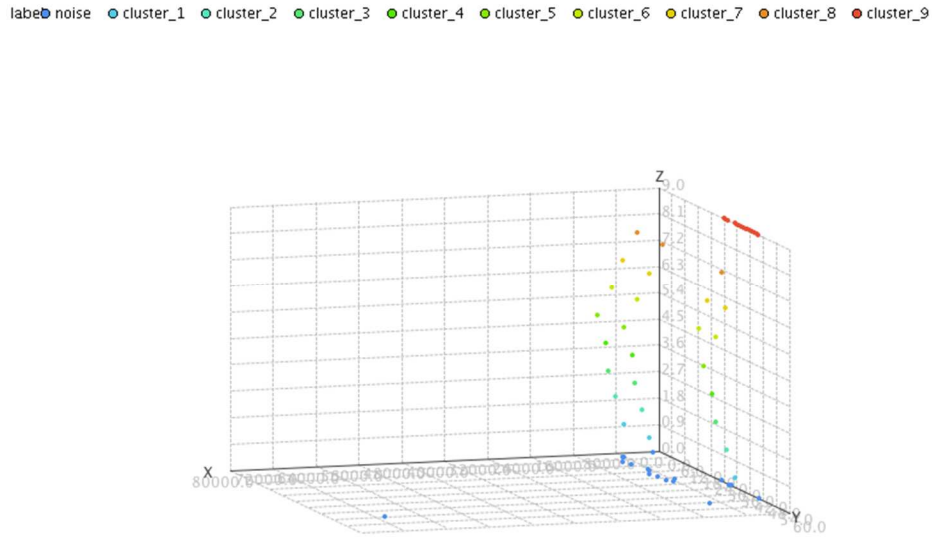
Ajustado los parámetros por defecto del algoritmo de Segmentación en el cual se cambia el valor de convergencia el cual especifica la precisión de las condiciones de los clusters, el algoritmo Agrupamiento de Soporte Vectorial (Support Vector Clustering, SVC) arroja por resultado diez (10) clusters que agrupan los siguientes ítems:

- En el Cluster 0, 18 items, conjunto de valores distintos.
- En el Cluster 1, 03 items, con el valor 7296.030
- En el Cluster 2, 03 items, con el valor 9194.670

- En el Cluster 3, 03 items, con el valor 11710.260
- En el Cluster 4, 03 items, con el valor 12594.140
- En el Cluster 5, 03 items, con el valor 14641.750
- En el Cluster 6, 04 items, con el valor 12759.080
- En el Cluster 7, 04 items, con el valor 11516.050
- En el Cluster 8, 03 items, con el valor 9279.830
- En el Cluster 9, 15 items, con el valor 0

La agrupación es más precisa, corrigiendose así el conjunto de datos del Cluster 0, pasando de 28 items (Modelo I) a 18 items (Modelo II).

Imagen 3 Gráfico Agrupamiento de Soporte Vectorial Modelo II, producida por los autores



CONCLUSIONES

Hasta ahora no hay una aplicación específica de algoritmos personalizados para Minería de Datos impositivos (Liu et al. 2012) Tampoco existen algoritmos específicos en la Minería de Procesos, ni en el Aprendizaje Automático en la Provincia de Mendoza, a través de la ejecución de algoritmos matemáticos predictivos o descriptivos. No obstante, ello, el análisis de datos con los algoritmos de segmentación contribuyen 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 de análisis, en virtud de los objetivos planteados oportunamente y de la preparación de los datos.

Las observaciones de los agrupamientos de los datos y sus características comunes pueden revelar datos con 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, o datos inconsistentes, conduce al descubrimiento de pequeños conjuntos de datos que serán significativamente muy diferentes al resto de los datos bajo análisis, y justamente el análisis de estos datos anómalos 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 evidenciar ilícitos y hechos de corrupción, sin perder de vista que la premisa es que exista calidad en los datos en las bases de datos tributarias, con lo cual habría poco lugar para la existencia de inconsistencias de este tipo, hecho aún más llamativo, cuando estas anomalías responden a un patrón de conducta de un mismo contribuyente, de un conjunto de contribuyentes, de un ejercicio específico o de un operador

determinado del sistema tributario, sin una causal asertiva. El algoritmo propuesto de Agrupamiento de Soporte Vectorial aplicado a un gran volumen de datos, una vez que el modelo ha sido validado, puede descubrir de las inconsistencias de los datos, anomalías y ruidos, las fuentes y orígenes de estas anomalías, y dependerá de cómo se planteen los objetivos, que son la base de la preparación de los datos, detectar ilícitos y/o hechos de corrupción.
Posteriormente, aquellas causales que expliciten las imperfecciones de los datos, y no sean ilícitos ni hechos de corrupción, permitirán segregar y limpiar estos datos inconsistentes de la base, para ir depurando la base de datos, corrigiendo las cuentas corrientes de los contribuyentes, optimizando la calidad del dato, y contribuyendo a los procesos de aprendizaje automático en la generación de las instrucciones necesarias en el lenguaje de programación seleccionado.

PROYECCIONES

- Validar algoritmos de agrupamiento de forma jurisprudencial para que exista como prueba suficiente en denuncias contra hechos de corrupción y fraude, tanto para el Sector Público como para el Sector Privado.
- Validar los algoritmos de modelización predictiva y descriptiva para otros procesos, como:
    - Árbol de Decisiones.
    - Árbol de Reglas.
    - ID3.
    - Reglas de Inducción.
    - Aprendizaje Profundo.
    - Red Neural.
    - Máquinas de Soporte Vectorial.
    - Agrupamiento Aglomerativo.
    - Agrupamiento Aplanado.
    - Matriz de Correlación.
    - Matriz Anova.
- Minimizar las cantidades de fraude e ilícitos.
- Proyectar la metodología y los conceptos de Minería de Datos propuesta por medio del Aprendizaje Automático para el Sector Público y el Sector Privado.
- Proponer ante la Oficina Nacional de Tecnología de la Información (ONTI), u organismo que corresponda,. la metodología resultante y la parametrización para su regulación normativa.

Anexo I – Metodología Propuesta

Se considera al Proceso de Extracción del Conocimiento necesariamente como un Proyecto Profesional Interdisciplinario. De esa manera se puede establecer un contexto integral que impactará en la elaboración y desarrollo de los modelos, la extracción del conocimiento y el procedimiento de inducción del mismo. Ese Proyecto Profesional Interdisciplinario no concluye cuando se generan los modelos necesarios, sino que el mismo estará relacionado con otros proyectos que iniciaran, como por ejemplo la inducción del conocimiento adquirido en el Aprendizaje Automático y la Minería de Procesos en tiempo real. Debe documentarse en forma analítica cada una de las etapas del Proceso de Extracción del Conocimiento, generando posteriormente un informe global que contenga una descripción de los puntos críticos de cada una de ellas, con el objetivo de que otros equipos interdisciplinarios de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él. La metodología propuesta se adapta a los objetivos del análisis de las causas de los datos anómalos, y que este análisis pueda contribuir en detectar ilícitos y posibles hechos de corrupción. Por tal motivo debe analizarse con mayor

profundidad los datos que evidencien patrones sospechosos, y los que tengan ruidos, y no eliminarlos de la muestra bajo análisis hasta que se deduzcan sus causales. El ciclo de vida del Proceso de Extracción del Conocimiento propuesto consiste de seis etapas:

Cuadro 1 Resumen de puntos críticos de los etapas, producida por los autores

ETAPAS	TIEMPO	RECURSOS	RIESGOS
I – Comprensión	a definir de acuerdo a la magnitud de la organización.	Analista de datos y Científicos de datos	Cambios externos a la organización.
II – Exploración, Preparación y Selección	a definir de acuerdo a la cantidad de datos a preparar.	Analista de datos y Científicos de datos	Problemas con los datos seleccionados y con la tecnología
III – Limpieza y Transformación	a definir de acuerdo a la cantidad de datos para explorar, limpiar y transformar	Científico de datos	Problemas con los datos explorados y con la tecnología
IV – Modelización	a definir de acuerdo a la cantidad de modelos definidos	Científico de datos	Problemas con modelos no adecuados
V – Evaluación	a definir de acuerdo a la cantidad de modelos para evaluar	Analista de datos y Científicos de datos	Cambios externos a la organización e incapacidad de implementación
VI – Implementación y Difusión del Conocimiento	a definir de acuerdo al alcance de la implementación propuesta	Científico de datos	Cambios externos a la organización e incapacidad de implementación

Etapa I – Comprensión

En la presente etapa se determinan todos los objetivos para el Proceso de Extracción del Conocimiento. Para lograr definir esos objetivos, se tendrá en cuenta las siguientes sub etapas, las cuales pueden ser ejecutadas indistintamente, es decir, sin un orden predeterminado, y estas sub etapas son las que se detallan a continuación:

a) Relevar los pilares de la Organización bajo estudio. Estos pilares son:
    a. La Visión.
    b. La Misión.
    c. Los Objetivos Estratégicos de la Organización.
b) Evaluar el Sistema de Control Interno de la Organización bajo estudio, de acuerdo a lo indicado en el informe COSO[5], en virtud de sus elementos:
    a. Ambiente de Control,
    b. Evaluación de los Riesgos,
    c. Actividades de Control,
    d. Información, Comunicación y
    e. Actividades de Supervisión.

[5] *COSO (Committee of Sponsoring Organizations of the Treadway) es una Comisión voluntaria constituida por representantes de cinco organizaciones del sector privado en Estados Unidos, para proporcionar liderazgo intelectual frente a tres temas interrelacionados: la gestión del riesgo empresarial (ERM), el control interno, y la disuasión del fraude.*

c) Determinar los objetivos del Proyecto.
    a. Establecer el tipo de Proyecto, predictivo o descriptivo.
    b. Identificar las áreas organizativas involucradas.
    c. Evidenciar las motivaciones del Proyecto.
d) Elaborar los Planes del Proyecto.
    a. Diseñar el Plan Estratégico del Proyecto.
    b. Diseñar el Plan Individual del Proyecto.

Las sub etapas a), b) y c) aportan los elementos que permiten se elabore el Informe Final. Este Informe Final, con los Planes del Proyecto constituirán el resumen de la Etapa I denominada Comprensión. El tiempo será una variable dependiente del tamaño de la organización objeto del Proyecto, y el intervalo razonable a establecer es de una (1) semana a cinco (5) semanas.

Etapa II – Exploración, Preparación y Selección de los Datos

La Etapa de Exploración, Preparación y Selección de los Datos comienza con la recolección de los datos iniciales necesarios para llevar a cabo el Proyecto de acuerdo a la Planificación individual, y continúa con las actividades necesarias que permiten familiarizarse con las fuentes de información de la organización, la estructura de los datos existentes, y la ponderación de la calidad de esos datos. El análisis de la calidad de los datos deja de manifiesto las posibles anomalías existentes, anomalías que suelen incrementarse cuando se trata de bases de datos múltiples y datos no tradicionales, sean de origen interno, como de origen externo.

Las sub etapas son:

a) Recoger los datos iniciales.
    a. Identificar el origen y la fuente de los datos.
    b. Generar informe de recopilación de los datos y sus métodos de captura.
b) Describir los datos.
    a. Detallar la cantidad de datos disponibles.
    b. Describir su interrelación.
c) Detectar anomalías de los datos.
    a. Identificar las posibles causales de anomalías de los datos.
    b. Agrupar las causales.
d) Definir la calidad de los datos.
    a. Visualizar valores perdidos.
    b. Evidenciar errores de los datos.
    c. Evidenciar errores de los metadatos.

Exploración de los Datos

Los datos pueden provenir de distintas fuentes, por lo tanto resulta necesario el análisis exploratorio de los datos, para hallar la distribución de los datos, la simetría y normalidad, y por consiguiente todas las correlaciones con la información. Para ello, se debe graficar los datos, lo que permitirá visualizar la estructura de los datos. Las herramientas de visualización pueden ser, por ejemplo, el histograma de frecuencias, el gráfico de caja y bigotes, el gráfico de simetría y el gráfico de dispersión.

Selección de los Datos.

Cuando los datos han sido preparados, se procede a seleccionar aquellos datos relevantes a los objetivos propuestos en la Planificación Individual del Proyecto.

Existen dos formas posibles de seleccionar los datos: la selección de elementos o la selección de atributos, o una combinación de los dos, es decir elementos y atributos. Cuando exista una exclusión de ciertos datos del conjunto de datos bajo análisis, los cuales no formarán parte de la modelización, debe dejarse constancia documentada de cuales fueron los motivos por los que se tomó la decisión de su exclusión, teniendo en cuenta para ello:

a) La Identificación de los atributos que están íntimamente relacionados con los objetivos propuestos del Proyecto. Los atributos que no estén relacionados, podrían en principio suprimirse.
b) La evaluación de si la calidad de los atributos de los datos en concreto, impactan en la validez de los resultados, si no lo hacen, en principio podrían suprimirse.

Etapa III – Limpieza y Transformación de los Datos

La etapa de Limpieza y Transformación de los Datos abarca todas aquellas actividades necesarias para construir un conjunto final de datos que se utilizarán con el algoritmo seleccionado en la modelización de acuerdo al tipo de modelo a construir, a partir de los datos que ya han sido explorados, preparados y seleccionados. Las actividades que se incluyen en esta etapa, son las siguientes:

- selección de tablas,
- selección de elementos y
- selección de atributos,

En esta etapa no deben eliminarse las anomalías de los datos, sino que se debe analizar y descubrir cuáles fueron las causas que generaron las inconsistencias, para poder de esta forma aumentar la calidad de los datos, porque el algoritmo que se aplique en la siguiente etapa deberá recibir los datos con un formato especifico, como por ejemplo el formato que es requerido por el algoritmo Agrupamiento de Soporte Vectorial (Ben-Hur et al., s. f.). Las sub etapas son:

a) Limpiar los datos,
b) Integrar los datos, y
c) Otorgarles un formato específico.

De manera intuitiva se afirma que la calidad de una observación dentro de un conjunto de datos (DataSet), se refleja por la relación que los mismos tienen con otras observaciones del mismo conjunto de datos que se obtuvieron bajo similares condiciones. Es común encontrar datos que parecen ser distintos que el resto de los datos, por tener valores más pequeños o más grandes, o por tener características distintas que el resto de la muestra, lo que podemos analizar, por ejemplo, con la interpretación de los clusters resultantes de aplicar cualquiera de los algoritmos de Segmentación, es decir cualquiera de los siguientes algoritmos:

- K – Means,
- X – Means,
- K – Medoids,
- DBScan,
- Agrupamiento Aleatorio, y
- Agrupamiento Aglomerativo.

Limpieza de los Datos

La limpieza implica observar más de cerca los problemas existentes en los datos que se han seleccionado para incluir en el Proyecto. Estos problemas pueden ser, los que se detallan en el cuadro siguiente:

Cuadro 2 Problemas de los Datos, producida por los autores

PROBLEMAS DEL DATO	SOLUCION
Datos perdidos	Excluir las filas, o imputarles un valor estimado.
Errores de los datos	Utilizar recursos lógicos para descubrir los errores manuales,
Incoherencias	Aplicar un esquema de codificación simple, convirtiendo y/o sustituyendo los valores.

Metadatos ausentes o erróneos	Examinar manualmente los campos sospechosos, y comprobar el significado correcto.

El informe que se genera describiendo la limpieza de los datos, detalla documentadamente las modificaciones que se llevaron a cabo en los datos del Conjunto de Datos (DataSet), puntualizando el impacto posible que la limpieza genera en cualquier algoritmo de Segmentación. El informe debe incluir en su confección, aquellos puntos que den respuesta a las siguientes interrogantes:

a) ¿Qué tipos de ruidos en los datos se han producido?
b) ¿Qué métodos se han utilizado para eliminar el ruido de los datos?
c) ¿Qué técnicas han demostrado ser eficaces y cuáles no?
d) ¿Qué atributos no han podido ser recuperados?
e) ¿Qué datos han sido excluidos por el ruido?
f) ¿Qué impacto generan los datos excluidos?

Los Ruidos en los Datos

Los datos que contienen ruidos van a interferir en la modelización, y esa interferencia puede llegar a ser significativa.

Es Ruido en los Datos 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. Ciertos métodos de análisis y procesamiento de datos pueden introducir ruidos al Conjunto de Datos (DataSet), y ello podría darse en cualquiera de las etapas previa a la obtención del conocimiento.

Etapa IV – Modelización

En la modelización se aplican los algoritmos, como por ejemplo el Agrupamiento de Soporte Vectorial, el cual es el pertinente a los objetivos planteados en la Planificación Individual del Proyecto, para generar varios modelos (cuantos más modelos se formen, mejor), y se calibran sus parámetros a los valores óptimos, de acuerdos a los resultados de su ejecución. Obviamente existen varios algoritmos para el mismo proyecto. En esta etapa se debe:

a) Seleccionar las técnicas y herramientas apropiadas
b) Diseñar los casos de modelado, en virtud de los objetivos
c) Construir los modelos, y
d) Dar valor a cada uno de los modelos.

Se debe registrar los ajustes y datos utilizados en cada uno de los modelos elaborados, y al finalizar la etapa de modelización, se podrá disponer de tres tipos de informaciones que son relevantes y de soporte vital para el Proyecto:

1) La configuración de los parámetros utilizados.

La mayoría de los modelos tienen diferentes parámetros que deben ajustarse para controlar su proceso de modelado y ejecución del mismo.

Generalmente se modela con los parámetros preestablecidos, y luego se ajustan los valores por defecto. Debe tomarse nota cada uno de los ajustes realizados, para poder automatizar el proceso de parametrización llevado a cabo, con nuevos Conjunto de Datos (DataSet).

2) Los modelos producidos, y
3) La descripción de los resultados de estos modelos.

Al examinar los resultados que arroja la ejecución de un modelo determinado, se debe tomar nota de esos resultados, lo que permitirá compararlos con los resultados de otros modelos desarrollados, y de la comparación se obtendrá aquel que satisfaga mejor los objetivos planteados en la Planificación Individual del Proyecto.

Luego, se registra esa comparación de modelos en el Informe Final de la presente etapa, informe que contará con la siguiente información adicional, la cual responde a los interrogantes que se detallan a continuación:

a) ¿se puede llegar a conclusiones significativas a partir de este modelo?
b) ¿revela este modelo otros patrones?
c) ¿el modelo presenta problemas de ejecución?
d) ¿el tiempo de procesamiento del modelo es el indicado?
e) ¿el modelo presenta problema de calidad de datos?
f) ¿hay incoherencias de cálculos en el modelo?

Etapa V – Evaluación

El llegar a la etapa de Evaluación significa que anteriormente se han generado uno o varios modelos, los cuales parecieran alcanzar la calidad suficiente y esperada desde la perspectiva del análisis del Conjunto de Datos inicial (DataSet).

Por tal motivo, antes de concebir al proceso como definitivo, debe evaluarse a fondo y en profundidad los modelos que se han generado, y revisar todos los pasos críticos que se ejecutaron para crearlo. Posteriormente debe compararse el modelo evaluado con los objetivos de la Organización, objetivos que fueron conocidos y relevados en la Etapa I denominada Comprensión, y los que fueron plasmados y procesados en la Planificación Individual del Proyecto. Un objetivo clave de la evaluación de los modelos, es determinar si hay algún punto específico relevante que concierne a la organización, que no haya sido considerado suficientemente, o que haya sido subestimado, y que impacta en el Proceso de Extracción del Conocimiento. Al final de presente etapa, se obtendrá una decisión sobre la aplicación de los resultados. Las sub etapas de la Evaluación son:

a) Evaluar los resultados obtenidos,
b) Revisar, y
c) Llevar acciones a cabo.

En esta etapa se determinan cuales de los modelos evaluados son los modelos más precisos para considerárselos finales, es decir, aquellos modelos definitivos que estén listos para ser aplicados o que evidencian patrones interesantes, o patrones que no se tuvieron en cuenta en la Planificación Individual del Proyecto. Es oportuno en esta etapa crear un método de valoración de los modelos.

Los modelos deben ser clasificados en virtud de la precisión que arrojan sus resultados, y de la facilidad e interpretación de los mismos. Una vez clasificados se los evalúa con mayor profundad, para utilizar los conocimientos extraídos de los datos, y volcarlos en la ejecución de los procesos y operaciones tendientes al Aprendizaje Automático. Luego, de la observación de estos resultados, las conclusiones podrán aportar material para los procesos de los métodos utilizados en la Minería de Procesos (IEEE Task Force on Process Mining s. f.). Las sub etapas de la Evaluación son:

a) Comprender los resultados de los modelos evaluados y encontrar su sentido lógico.
b) Establecer si existe incoherencias que necesitan una mayor exploración.
c) Explorar más de un modelo y comparar cada uno de los resultados.

Etapa VI – Implementación y Difusión del Conocimiento

La obtención del modelo o de los modelos de extracción de conocimiento, no es la conclusión del Proyecto. Incluso, si uno de los objetivos del modelo, es el de extraer conocimiento proactivo o analítico de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que se lo pueda usar a través de la inducción. Dependiendo de los requisitos preestablecidos en los objetivos de la etapa conocimiento, la presente etapa puede ser tan simple como la generación de un informe final de calidad, o tan compleja como la realización periódica

y automatizada de un proceso de análisis de datos en la organización, lo que requerirá las siguientes sub etapas:

a) Planificar la difusión del conocimiento,
b) Monitorear el plan, mantenerlo, y
c) Generar informes y revisiones periódicas

Por lo tanto, en esta instancia deben reunirse los resultados obtenidos, los modelos evaluados y los descubrimientos y patrones logrados que evidencien posibles ilícitos y hechos de corrupción, por tal motivo todo el proceso tiene que ser continuado.

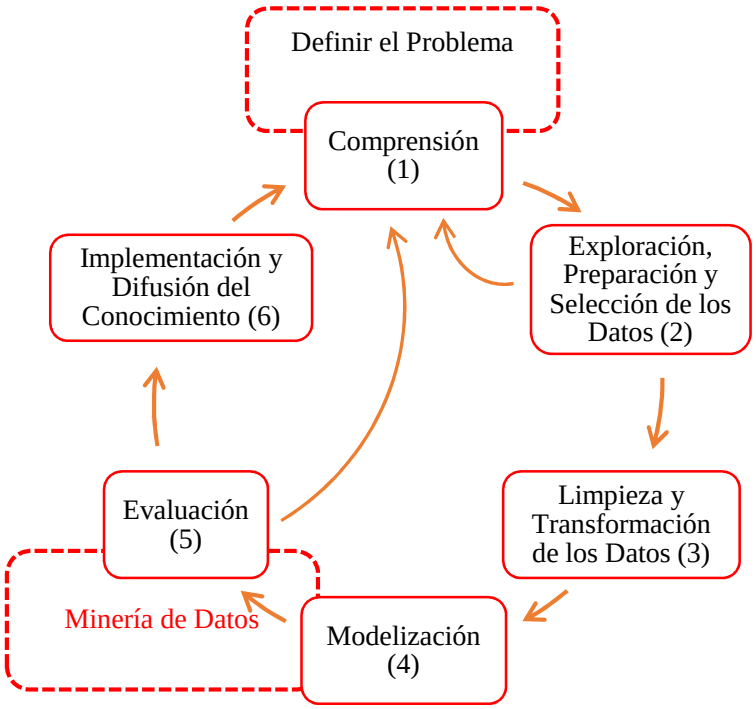
El modelo debe permitir que sea modificado y reutilizado, dentro de un enfoque sistémico con retroalimentación, lo que se logra con el Autoaprendizaje Automático. El Informe Final de esta etapa contiene las siguientes pautas que permiten contestar las siguientes interrogantes:

a) ¿qué factores se necesitan controlar?
b) ¿cómo debería determinarse que un modelo ha expirado?
c) ¿puede reconstruirse el modelo con nuevos datos?
d) ¿puede utilizarse este modelo para otros fraudes o ilícitos?

Entonces, contestadas las interrogantes anteriores se debe volcar las respuestas en las observaciones, recomendaciones y conclusiones del Informe final, el que contendrá los siguientes puntos:

a) Descripción detallada del problema original.
b) Detalle de los pasos utilizados de la Metodología.
c) Costos.
d) Desviaciones de la Planificación Individual del Proyecto.
e) Resumen de los resultados, incluyendo los modelos en los descubrimientos.
f) Plan propuesto para la difusión del conocimiento.
g) Recomendaciones para futuros Proyectos.

Esquema 1 Interacción de las Etapas de la Metodología, producida por los autores



Anexo II – RapidMiner Studio® Versión 8.2

Para la presente investigación se utiliza la herramienta RapidMiner Studio®, versión 8.2. Al ejecutar la herramienta RapidMiner Studio®, como primera medida se procederá a la

importación de los datos consideraros para análisis. Todos los modelos que se describan a continuación, utilizan algoritmos de Segmentación. La diferencia entre los modelos, es como se preparan los datos del Conjunto de Datos para la entrada de datos del algoritmo (input), y las distintas parametrizaciones de los algoritmos utilizados, lo que impactará en el resultado de la ejecución del proceso.

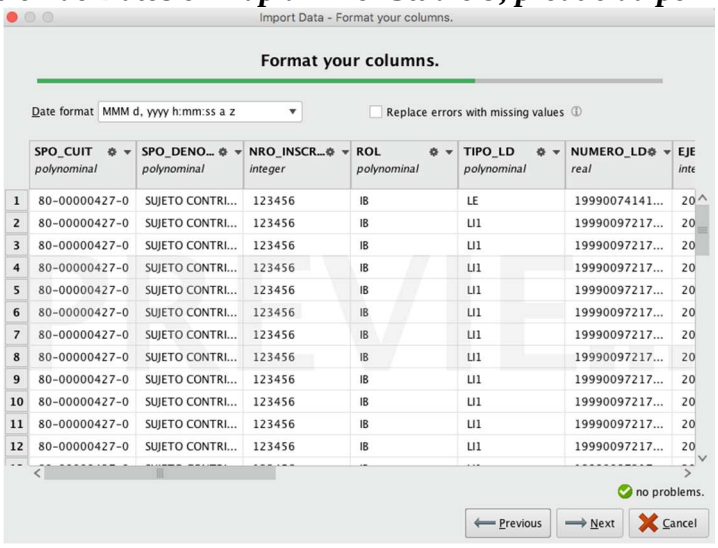
Modelo I

En la generación del Modelo I, el algoritmo Agrupamiento de Soporte Vectorial es aplicado solamente a un atributo, el que se denomina IMP-RET (Impuesto menos Retención) y que fue creado para poder agrupar con mayor claridad los datos. Ese procedimiento se describe a continuación. El aplicar el algoritmo indicado, presupone no solo conocer cómo se comporta el algoritmo con el Conjunto de Ejemplos (ExampleSet), sino también comprender que son los datos contenidos en la planilla de cálculos bajo análisis, y sus interrelaciones.

Etapa II – Exploración, Preparación y Selección de los Datos

Cuando se procede con la importación de los datos, RapidMiner Studio® preguntará ¿de donde son los datos?, otorgando dos opciones para indicar la procedencia de los datos para analizar, una opción será “en la computadora” y la otra opción “en una base de datos”. Seleccionado el archivo o la base de datos para proceder a la importación, RapidMiner Studio® otorga un formato estándar a las columnas de los datos resultantes del proceso de importación, las que no las llama columnas, sino que las denomina atributos, indicando en ese preciso momento si hubo algún problema en la importación, y durante el proceso muestra una vista previa.

Imagen 4 Importación de Datos en RapidMiner Studio®, producida por los autores



También otorga la opción de reemplazar los errores que se detecten correspondientes a valores perdidos, y de cambiar el formato de los datos, como por ejemplo, el formato fecha, suministrando varias opciones de formatos, los cuales serán reemplazados en los valores del Conjunto de Datos (DataSet). Cuando la importación concluye, se visualiza el resultado en el área de trabajo Resultados (Results). Luego, para empezar a trabajar en el Proyecto de Extracción del Conocimiento, se pasa al área de trabajo denominada Diseño (Design), en la cual, en esa área de trabajo, se pueden visualizar los objetos que forman parte del proyecto, conectandolos entre sí, y ejecutando cada operador. Los datos importados se guardan en un repositorio de datos, lo que facilitará su lectura en el diseño de múltiples modelos. Ese repositorio puede ser un:

- Repositorio local: es el que aloja el archivo en la computadora o dispositivo desde el que se está ejecutando la herramienta.

- Repositorio en la nube: espacio que asigna RapidMiner Studio®, previa registración en línea.

El Conjunto de Datos (DataSet) es denominado por RapidMiner Studio® Conjunto de Ejemplos (ExampleSet), siendo llamados ejemplos (Example), a lo que se conoce como fila en una planilla de cálculos y Atributos Regulares (Regular Attribute) lo que corresponde a las columnas en una planilla de cálculos. Por ese motivo el proceso de importación indica Conjunto de Ejemplos cincuenta y nueve (59) Ejemplos y diecisiete (17) Atributos regulares (ExampleSet 59 Examples). Puede haber Atributos Especiales (Special Attribute) los que aparecen por ejemplo, cuando se aplican procesos determinados o algoritmos específicos como el Atributo Especial Cluster que aparece luego de aplicar los algoritmos de Segmentación, siempre y cuando se seleccione la casilla de verificación de los parámetros de ajustes del algoritmo que se esté utilizando.

Etapa III – Limpieza y Transformación de los Datos

Solamente se desarrollan algunos puntos prácticos que pertenecen a esta etapa de acuerdo a lo establecido en el Anexo I – Metodología Propuesta, sin desarrollarla en su totalidad, por una cuestión de practicidad y simplificación del ejemplo. RapidMiner Studio® denomina proceso a las etapas que se definen para el proyecto, y estas etapas las llama operadores, los cuales van conectados entre sí, por el punto de entrada de datos, y el punto de salida de los mismos. Lo primero que se realiza, es la lectura los datos de la planilla de datos importada previamente, a través del operador retrieve.

Recuperar (Retrieve)

Este operador puede acceder a la información almacenada en el repositorio y cargarlos en el proceso que se está desarrollando. El operador Recuperar (Retrieve) carga un objeto en el proceso de RapidMiner Studio®. El objeto cargado es habitualmente un Conjunto de Ejemplos (ExampleSet), pero también puede ser una Colección de Datos o un Modelo. Recuperar de los datos almacenados dentro de un repositorio, otorga la ventaja de que también se recuperen las propiedades de metadatos, esta característica es muy importante, ya que los metadatos brindan información adicional sobre el objeto que se recupera, como, por ejemplo, los nombres y tipos de los atributos, su rango y cuántos valores faltantes hay. Los metadatos le permiten configurar fácilmente los parámetros de otros operadores, por ejemplo, puede seleccionar atributos de una lista de atributos disponibles. A continuación, se seleccionan los atributos con los que se avanzará en el proceso…

Selección de Atributos (Select Attributes)

Se utiliza el proceso Selección de Atributos (Select Attributes), con la opción Subconjunto (Subset), que permite elegir las columnas deseadas de la planilla de cálculos para el análisis de los datos, y la aplicación de los algoritmos. En el presente modelo se seleccionan dos columnas, la columna IMPUESTO, y la columna RETENCIÓN, ya que las mismas son las que arrojan valores coincidentes. El operador Selección de Atributos (Selected Attributes) proporciona diferentes tipos de filtros para facilitar la selección de atributos. Las distintas alternativas son: selección directa de atributos, selección de un subconjunto de atributos, selección mediante una expresión regular o selección de atributos sin valores perdidos, y la selección de una de ellas dependerá en gran medida, del algoritmo que se utilice, y de los valores de entrada que requiera tal algoritmo.

Generar Atributos (Generate Attributes)

El operador Generar Atributos (Generate Attributes) construye nuevos atributos, a partir de los atributos del Conjunto de Ejemplos (ExampleSet), y constantes arbitrarias usando expresiones

matemáticas las que son definidas por el Analista de Datos. Los nombres que poseen los Atributos del Conjunto de Ejemplos de entrada, se pueden usar como variables de las expresiones matemáticas que se definan para la generación y el comportamiento de los nuevos atributos. Durante la ejecución de este operador, las expresiones matemáticas consignadas por el Analista de Datos se evalúan para cada Ejemplo (Example).
Por lo tanto, este operador no solo crea nuevas columnas (nuevos atributos), sino que también rellena esas columnas con los valores resultantes de la expresión matemática definida. En el desarrollo del presente Modelo, se genera el Atributo con el nombre IMP-RET (Impuesto menos Retención), que es el consiguiente de restar a los valores existentes en la columna IMPUESTO, los valores que representan las retenciones, los que están en la columna RETENCION, defiendo la expresión matemática IMPUESTO – RETENCION. RapidMiner Studio® automáticamente interpreta en la función matemática, los nombres de los atributos, tal como se lo puede visualizar en la Ilustración e Imagen siguiente Generar Atributos en Rapidminer Studio®

Imagen 5 Generar Atributos en RapidMiner Studio®, producida por los autores

La ejecución de este operador arrojará como resultado 59 Ejemplos de un Atributo regular, denominados IMP-RET, y cada Ejemplo contendrá un valor positivo, es decir, no existen en la muestra retenciones sin impuestos, o retenciones mayores a los valores del impuesto, por eso el intervalo de los posibles resultados parte desde cero, a un valor positivo. Cuando ese valor es cero, significa que el impuesto de un periodo mensual determinado es igual a la retención del impuesto, con lo cual, en principio no hay obligación impositiva para el contribuyente en ese periodo mensual ya que no tendría impuesto a pagar en ese periodo. Eso es lo que indica un valor cero. La causal de esta anomalía de datos será determinante.

Etapa IV – Modelización

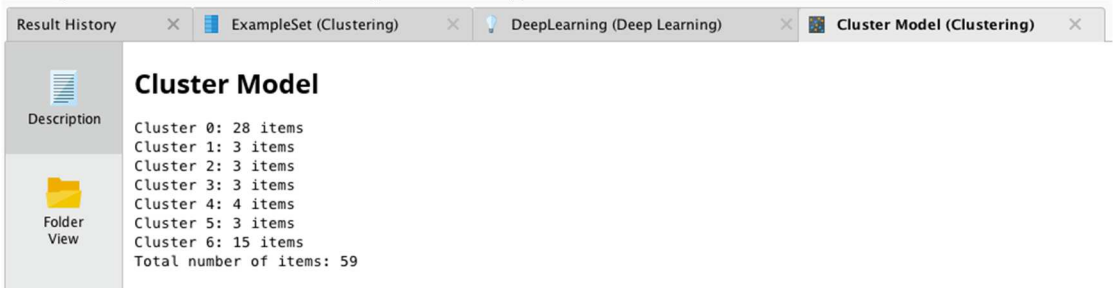
Agrupamiento de Soporte Vectorial (Support Vector Clustering)

El Agrupamiento de Soporte Vectorial (Support Vector Clustering) es la agrupación de objetos que son similares entre sí y diferentes a los objetos que pertenecen a otros grupos, sin especificación previa de los grupos. La agrupación es una técnica para extraer información de datos no etiquetados, y puede ser muy útil en muchos escenarios diferentes, para encontrar comportamientos similares, y en el caso bajo estudio, comportamientos similares del mismo contribuyente, en diferentes declaraciones juradas mensuales declaradas en la Liquidación de Deuda, resultante de una fiscalización de los inspectores de la Administración Tributaria Mendoza. En la parametrización de las opciones que permite la ejecución de este algoritmo, se puede definir que se agregue al Conjunto de Ejemplos (ExampleSet) un Atributo especial, como lo es el atributo Cluster, que tiene como valor, los números de los distintos clusters, o si los datos se consideran ruidosos. Luego de esta instancia el ExampleSet será ExampleSet Clustering con dos Atributos especiales, uno es el Atributo Cluster, y el otro el Atributo Id, que es el número del Example. El resultado de la ejecución del proceso resulta en 7 Clusters, distribuyendo los items de los valores de cada Ejemplo del Atributo IMP-RET (Impuesto menos retención) de la siguiente manera:

- En el Cluster 0, 28 items, conjunto de valores distintos.
- En el Cluster 1, 03 items, con el valor 9194.670

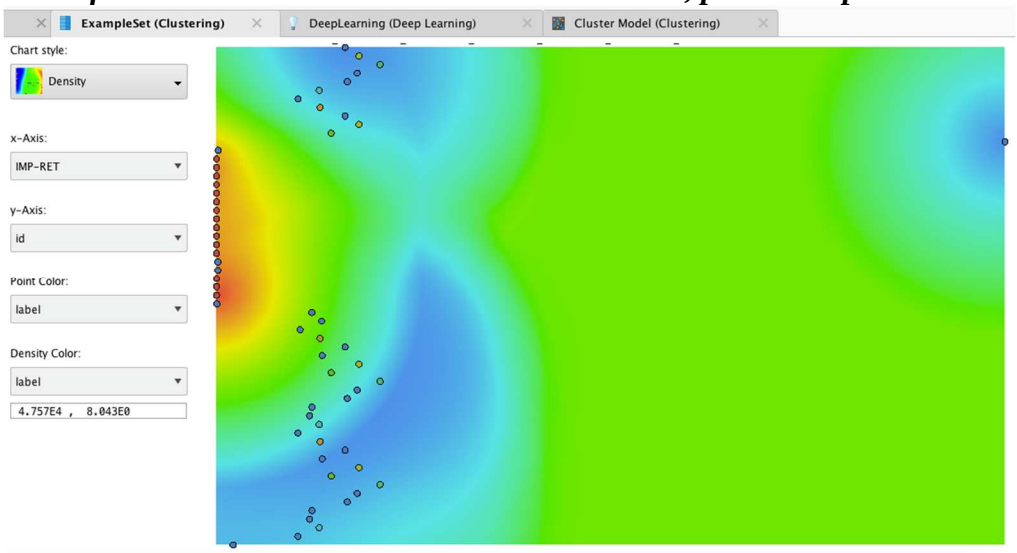
- En el Cluster 2, 03 items, con el valor 14641.750
- En el Cluster 3, 03 items, con el valor 10268.430
- En el Cluster 4, 04 items, con el valor 12759.080
- En el Cluster 5, 03 items, con el valor 9279.830
- En el Cluster 6, 15 items, con el valor 0

Imagen 6 Modelo I Clúster, producida por los autores



Los valores atípicos deforman las distancias, y producen clusters unitarios.

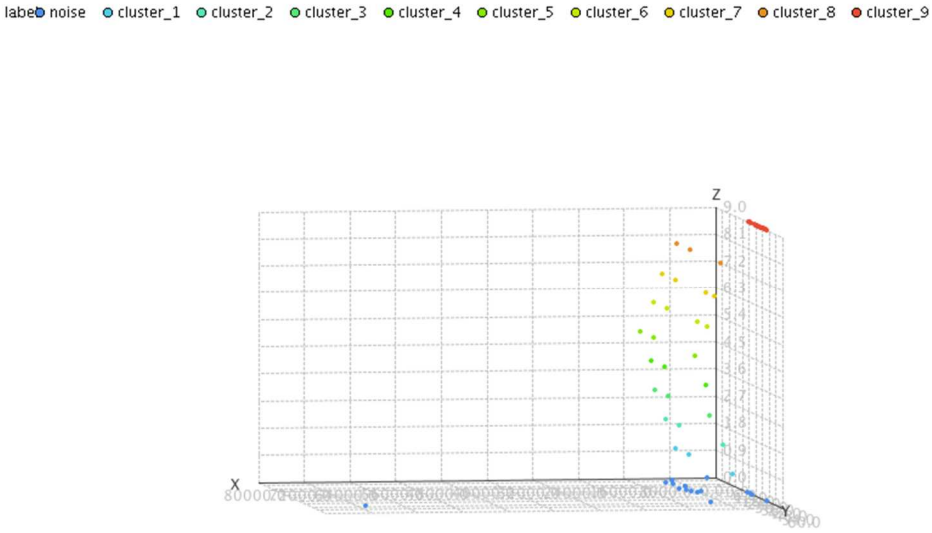
Imagen 7 Gráfico de Densidad de los Clusters del Modelo I, producida por los autores



Modelo II

En la generación del Modelo II, se modifica el valor opcional predeterminado que se denomina épsilon de convergencia correspondiente al algoritmo Agrupamiento de Soporte Vectorial, del valor 0,001 al valor propuesto 0,05. Ese parámetro optimiza la precisión y la cantidad de clusters.

Imagen 8 Gráfico avanzado del Modelo II, producida por los autores



Etapa IV – Modelización

Agrupamiento de Soporte Vectorial (Support Vector Clustering)

El resultado que arroja el uso del algoritmo, con la parametrización indicada anteriormente es de 10 Clusters, distribuyendo entre ellos, los items de los valores de cada Ejemplo del Atributo IMP-RET (Impuesto menos retención) de la siguiente manera: el Cluster 9 posee 15 items con el valor cero, igual que en el Modelo I, por lo que se aprecia la mayor precisión, conteniendo cada cluster los siguientes items:

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

Anexo III – Algoritmia Utilizada

Un algoritmo es un procedimiento lógico que se utiliza para resolver un problema. La elección del algoritmo a utilizar, depende del tipo de datos a analizar, los objetivos propuestos en el Proceso de Extracción del Conocimiento, la estructura de los datos, las anomalías de los datos, la cantidad de atributos de los datos, y otras. Los algoritmos de la Minería de Datos pueden implementarse en cualquier lenguaje de programación, lo que facilita el proceso de Aprendizaje Automático. Las herramientas de Minería de Datos, como por ejemplo RapidMiner Studio®, permiten implementar los algoritmos de manera sencilla, en una interfaz gráfica.

Agrupamiento de Soporte Vectorial (Support Vector Clustering)

Se selecciona el algoritmo Agrupamiento de Soporte Vectorial (Support Vector Clustering, SVC) en contraste con la mayoría de los algoritmos de agrupamiento (Cluster), porque el resto no tienen ningún mecanismo para tratar el ruido de los datos o los valores atípicos. (Ben-Hur et al., s. f.). Los puntos de datos se mapean desde el espacio de datos a un espacio de

características utilizando un núcleo kernel gaussiano. En el espacio de características, se busca la esfera más pequeña que encierra los datos digitales. Esta esfera se asigna al espacio de datos, donde forma un conjunto de contornos que encierran los puntos de datos. Esos contornos se interpretan como los límites del clúster. Los puntos encerrados por cada contorno separado están asociados con el mismo clúster. El algoritmo Agrupamiento de Soporte Vectorial (SVC) puede tratar con 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. Todos los puntos que no estén en ningún Cluster (Cluster 0), se los considera ruido.

Algoritmo SVC

Usando la transformación no lineal Φ de x a un espacio, se busca la esfera del radio más pequeña $\mathcal{R}$, lo que se describe 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:

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

con $\xi_j \geq 0$.

Para resolver este problema, se utiliza la mecánica de Lagrange[6], es decir, el langrangiano:

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

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 lleva a:

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

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

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

Las condiciones complementarias de Roger Fletcher[7] (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 $u_i = 0$, $\beta_i = C$. Eso de determinara Vector de Soporte Limitado (BSV).

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 (SV).

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

---

[6] *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.*

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

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:

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

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

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

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

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

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

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

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

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

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

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

Y el radio de la esfera es:

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

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

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

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

Asignación de Conglomerados (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 componentes (clusters) cualquier camino que los conecte debe salir de la esfera en el espacio de caracteristicas.
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

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

Los clusters se definen entonces, como los componentes conectados del gráfico inducido por A

Método de Agrupación

El método de agrupación no tiene un sesgo explícito ni del número, ni de la forma de los clusters. Tiene 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,"$\boldsymbol{p}$", 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 clústeres superpuestos.

Bibliografía

- Álvarez, Kity, Betzaida Romero, José Cadenas, David Coronado, y Rosseline Rodríguez. 2016. «Arquitectura para la Gestión de Datos Imperfectos en la Era de Big Data». *Revista Venezolana de Computación* 3 (2): 47-56. http://saber.ucv.ve/ojs/index.php/rev_vcomp/article/view/11729.
- Ben-Hur, Asa, David Horn, Hava T. Siegelmann, y Vladimir Vapnik. 2001. «Journal of Machine Learning Research». *Support Vector Clustering* 2001.
- Date, C.J. 2001. *Introducción a los Sistemas de Bases de Datos*. Séptima. México: Pearson Educación.
- Fletcher, Roger. 2000. *Practical Methods of Optimization*. 2nd ed.
- IEEE Task Force on Process Mining. s. f. «Manifiesto sobre Minería de Procesos». Manifiesto sobre Minería de Procesos. Accedido 22 de diciembre de 2017. http://www.win.tue.nl/ieeetfpm/lib/exe/fetch.php?media=shared:pmm-spanish-v1.pdf.
- 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.
- 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.
- Moral, Anselmo del, Juan Pazos, Esteban Rodríguez, Alfonso Rodríguez - Patón, y Sonia Suárez. 2008. *Gestión del Conocimiento*. Madrid, España: Thomson Editores Spain.
- Pascual, Rafael, José Genoud, Guillermo Aramburu, y Mario Pontaquarto. 2000. «Ley Nº 25326». InfoLEG. 4 de octubre de 2000. http://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm.
- 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.
- Stankevicius, Evaldas, y Linas Leonas. 2015. «Hybrid Approach Model for Prevention of Tax Evasion and Fraud». *20th International Scientific Conference «Economics and Management 2015 (ICEM-2015)»* 213 (Supplement C): 383-89. https://doi.org/10.1016/j.sbspro.2015.11.555.