TESIS DE MAESTRÍA

MAESTRÍA EN INGENIERÍA EN SISTEMAS DE LA INFORMACIÓN

Título:

“Procedimiento de detección de datos anómalos y ruido en administración pública para un caso de contralor civil”

Autor: Rodrigo López-Pablos

Director de Tesis: Horacio Daniel Kuna

Buenos Aires - 2018

A la familia, por estar siempre en todo momento

A los viejos Pedro y Rodolfo, por sus ejemplos de honradez, honestidad y trabajo

Agradecimientos

A la Universidad Tecnológica Nacional, por ofrecernos los medios humanos y tecnológicos de frontera necesarios para la contribución al conocimiento científico, institucionalidad desde la cual se forja esta tesis de posgrado.

A la Biblioteca de Congreso de la Nación, al proporcionarnos un espacio casi familiar donde fue escrita la mayor parte de esta tesis de posgrado, la BCN funcionó casi de segundo hogar dada la cantidad de horas trabajo allí realizadas.

Esta tesis no hubiera sido posible sin el apoyo de nuestro director y maestro Horacio Daniel Kuna, a quién admiramos y guardamos el mas profundo de los respetos tanto en lo académico como en lo humano.

A Ma. Florencia Pollo por su entusiasmo, recomendaciones y encontrarla siempre con la mejor onda para animarnos a finalizar nuestras investigaciones, así como su permanente predisposición y ayuda para finalizar esta tesis, por todo eso y mucho más: gracias Flor.

A Ramón García Martínez, de quién aprendimos a amar las ciencias de la computación, mediante su palabra y pasión por la docencia e investigación computacional.

A los profesores e investigadores de la UTN que participaron con útiles comentarios y recomendaciones sobre la elaboración de este documento.

Al personal docente y administrativo de la Escuela de Posgrado de la Facultad Regional Buenos Aires, entre ellos Anahí, Jimena, Paola y quienes acompañaron con diligencia y apego en cada trámite y consulta efectuada en la sede de posgrado Castro Barros.

A los compañeros de clase magistral, en la etapa de realización de los cursos, a quienes les debo su solidaridad *uteniana* sincera, desinteresada y paciente con los ilusos que no venimos del palo informático.

ÍNDICE

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

RESUMEN

Las técnicas y procesos de detección de campos anómalos y con ruido en datos públicos abiertos han sido escasamente empleadas con propósitos cívicos en la lucha contra la corrupción, aún así estas pueden ser de suma utilidad en la evaluación de la calidad de bases de datos así como en el descubrimiento de indicios de comportamiento corrupto. En esta tesis se desarrolla y articulan procedimientos híbridos de detección de datos anómalos y ruido para investigar, experimentar y validar su aplicación en sistemas de declaraciones juradas públicas disponibles actualmente en datos públicos abiertos en Argentina.

Palabras clave: contralor cívico, datos públicos, declaraciones juradas, bases de datos, datos anómalos y ruido.

ABSTRACT

The techniques and processes for detecting anomalous and noisy fields in open public data have been scarcely used for civic purposes in the fight against corruption, although they could be useful in the evaluation of the quality of databases as well as in the discovery of signs of corrupt behavior. This thesis develops and articulates hybrid anomalous data and noise detection procedures to research, experiment and validate its application in public official's affidavit systems currently available through open public data in Argentina.

Keywords: civic control, open public data, public official's affidavit, anomalous data and noise detection.

LISTA DE ABREVIACIONES

ACIJ	Asociación Civil por la Igualdad y la Justicia.
BD/BBDD	Base de datos/Bases de datos.
CAATs	Computer Assisted Audit Techniques. Técnicas de Auditoria Asistidas por Computador.
C4.5	Algoritmo de inducción desarrollado por Quinlan (1993).
DBSCAN	Algoritmo de detección de datos anómalos basado en densidad.
DJ/DDJJ	Declaración Jurada/Declaraciones juradas patrimoniales.
DLND	Sitio on-line interactivo de DDJJ del Diario La Nación "Data".
EI	Explotación de la información.
K-Means	Algoritmo que permite clasificar un conjunto de objetos en un número K de clusters.
LOF	Local Outlier Factor. Factor de anomalía local.
LRD	Local Reachability Density. Densidad de accesibilidad local.
OA	Oficina anticorrupción.
OGP	Open Government Partnership. Alianza para el gobierno abierto.
PRISM	Algoritmo de aprendizaje basado en reglas.
PUEC	Políticas Universales de Empoderamiento Digital Ciudadano.
RB	Redes bayesianas, naive bayes o redes de creencias.
RN	Redes neuronales.
RM	Rapid Miner, software de minería de datos.
SOM	Self-Organizing Map, Mapas auto-organizados.
SPV	Support Vector Machines. Maquinas de vector de soporte.
TDIDT	Top Down Induction Decision Trees. Árboles de decisición de arriba hacia abajo.
TI	Teoría de la información.

1. CAPÍTULO INTRODUCTORIO

En este Capítulo Introductorio se despliega la presentación y motivación inicial que hace al esfuerzo realizado en esta investigación, desde la introducción primera a nuestro tema de tesis, detallando su contenido en toda su extensión, la descripción del problema que aquí nos reúne poniendo de relieve la importancia en la actualidad que representa el acceso a la información pública para el contralor cívico, las hipótesis planteadas y los objetivos de investigación propuestos, su alcance.

1.1 INTRODUCCIÓN AL TEMA DE LA TESIS

La ingeniería de explotación de la información y la minería de datos han sido aplicadas generalmente a cuestiones dentro del ámbito privado, de los negocios y las corporaciones en general, no así pensado como un conjunto de técnicas al servicio de los organismos del Estado ni como herramienta destinada a fortalecer el contralor Ciudadano, al menos en el caso Argentino, no se observa un desarrollo sustantivo de estas tecnologías en administración pública.

Sus campos de aplicación interdisciplinar abordaron generalmente la utilización de estas técnicas como herramientas vinculadas a las problemáticas frecuentes ligados a la empresa o dirigidas a la gran corporación lo que cuadró la morfología semántica del campo del arte; consecuencia de ello, se observa en su uso recursivo de términos en donde se suele referirse a “negocios” y “clientes” al sujeto aplicativo en lugar de “ciudadanos” o “contribuyentes”, así como a “consumidores” en lugar de usuarios de servicio público, servicio comunitarios, etc. donde tanto en uno u otro caso la información explotada resulta igualmente útil.

En esta tesis se propone explorar los indicios que las herramientas que ofrecen las técnicas de EI en un área donde han sido sub-explotadas tanto por la administración pública como por los organismos de contralor, por ende presentando un vacío epistémico. Estas técnicas tampoco han sido consideradas en casos específicos de casos para el control cívico Ciudadano a escala individual, transparencia administrativa o resorte de proyección de organizaciones civiles encargadas del control y la auditoría cívico-ciudadana de las personas.

Sin una sociedad civil robusta peligran los derechos de los individuos y de los grupos que no adhieren incondicionalmente al sistema, el fortalecimiento de las sociedades civiles requiere como condición de posibilidad, el fortalecimiento de una ética compartida por todos los miembros de esa misma sociedad (Cortina, 1994). Al tiempo presente, tal fortalecimiento vuelve necesaria el uso de procesos informáticos como herramental mínimo para asegurar su fortalecimiento.

Siendo que la información es la columna vertebral y epicentro organizativo para la toma de decisiones organizacional de cualquier entidad tanto pública como privada, la calidad de los datos, y el conocimiento que de allí se explote guarda una importancia sin precedentes para el tratamiento de los datos en las instituciones.

Los estudios e investigaciones científicas del fenómeno de la corrupción, a través del uso de técnicas de minería de datos, estuvieron enfocados generalmente a la apreciación de la corrupción desde niveles de percepción social, así como de relacionar supuestas conectividades entre variables macroeconómicas y percepción de la corrupción pero no directamente sobre los sujetos de poder que de caer en actos corruptivos más perjuicio desencadenaría en el tejido social, i. e. sobre los representantes sociales que deben garantizar el bienestar general y el cumplimiento de la constitución, leyes y normas de convivencia.

La *transparencia* depende conceptualmente de la calidad de la información organizacional que se audita, ésta misma esta definida como la *"cualidad de un gobierno, empresa, organización o persona de abierta a la divulgación de información, normas, planes, procesos y acciones"* (International Transparency, 2009). En tiempos digitales, la continuación natural de políticas de transparencia se traduce en accesos a la información pública más efectivos y eficientes para la ciudadanía, así como la capacidad de brindar herramientas efectivas y la capacitación necesaria al auditor ciudadano para el control de sus propios representantes.

En términos republicanos, la democracia indirecta representativa depende, cada vez más e inevitablemente, de ciudadanos informados capaces de ejercer sus derechos civiles, y tener representantes probos que como paradigma ciudadano obren con el ejemplo pues encarnan durante un tiempo el poder delegado de los habitantes. Dada la complejidad de inherente a toda realidad social las herramientas de EI se vuelven una herramienta más de la ciudadanía y los Estados para su auto control y regulación.

Mediante esta propuesta se introduce la utilización de técnicas de minería de datos en el ámbito público en general y contribuir en los procesos cívicos y de control en particular. En este sentido estas nuevas técnicas contribuirán a hacer operativo el derecho a la información pública lo cual implica en términos de prevención de la corrupción, dar a los ciudadanos las herramientas para ejercer sus derechos y controlar el accionar del Estado (Naciones Unidas, 2004; Raigordsky y Geler, 2007) brindando las herramientas computacionales para crear conocimiento a partir de esa información se estará contribuyendo al empoderamiento de esa misma ciudadanía.

La carencia en el haber de caja de herramientas para la utilización de datos públicos desde instituciones de auditoría y control del propio Estado, a través de los mecanismos institucionales constituidos para esa tarea, en explotar información y generar conocimiento a través del uso de estas técnicas, eleva la oportunidad de aplicar procesos de sistemas de información en este sentido.

Asimismo, la aplicación de técnicas de minería de datos en general y de detección de outliers y ruido en particular han sido esquivas a la resolución de problemas ligados al contralor civil Ciudadano, actividades avocadas al control de los dineros públicos por parte de funcionarios, la efectividad y eficiencia de sus gestiones, etc. En Argentina, la existencia de procedimientos para la detección de outliers y ruido desarrollado por Kuna (2014) constituyen una oportunidad de experimentación que este trabajo de posgrado intenta explotar extrayendo información a partir DDJJ patrimoniales.

Por otra parte también es la primera vez que se propone la utilización de técnicas especificas de detección de outliers para ayudar a acometer dichos objetivos cívico institucionales.

1.1.1 CONTENIDO DE LA TESIS

El contenido, partes y estructura de la tesis comprende seis (6) Capítulos, incluyendo Capítulo Introductorio, Anexos y Glosario de atributos. Como podrá apreciarse, su contenido y estructura respetan el lineamiento para la presentación de tesis de maestría y doctorado adoptado por la Circular N°1- 2017 de la Escuela de Posgrado UTN FRBA.

Capítulo Introductorio:

En este Capítulo Introductorio hace a la presentación inicial al tema de la tesis que hace a este proyecto de investigación y su contenido, el cual consta de las siguientes partes:

Introducción al tema de la tesis: Presentación inicial y motivacional al tema de investigación que hace a esta tesis.

Descripción del problema: Se describe el problema complejo en donde se situa, direcciona y responde el proyecto de investigación de esta tesis. Se plantea el problema de la corrupción pública y las hipótesis de aplicación de procesos de explotación de la información para ayudar al combate de su flagelo.

Hipótesis: Se plantean las hipótesis del proyecto de investigación.

Objetivos de la investigación: Se presentan los objetivos de investigación así como su alcance.

Capítulos Centrales:

Capitulo 2: Se describe el Estado del arte sobre la auditoría de sistemas en organismos públicos y privados, así como los antecedentes en la utilización de técnicas de minería de datos para la detección de fraude.

Capítulo 3: Se desarrolla la solución propuesta para la problemática planteada, la descripción de bases de datos y la metodología a ser empleada.

Capítulo 4: Se aproxima a un caso empírico con bases de datos reales para buscar validar la metodología propuesta.

Capítulo 5: En un Capítulo final posterior, una vez obtenidos los resultados, se expondrán las conclusiones finales y aportes realizados en esta tesis, futuras líneas de investigación y breve discusión final.

Bibliografía y publicaciones derivadas: Posteriormente al despliegue de las referencias bibliográficas se presentan las publicaciones derivadas a la que dió lugar esta tesis, también se incluye un Anexo, el cual se describe a continuación.

Anexos

A manera de apéndice técnico, teórico y práctico, estos se subdividen en dos partes de la siguiente manera:

Anexo A: En el primer Anexo se presenta pormenorizadamente la metodología para la configuración y programación del procedimiento IV para la detección de outliers, la cual contempla las reglas de determinación de outliers para la unión de los resultados de los algoritmos de detección de datos anómalos y ruido (LOF y DBSCAN) y las mismas reglas para los algoritmos de clasificación (RB, C4.5, PRISM)

Anexo B: El segundo Anexo describe operativamente en RM la aplicación procedimental pormenorizada de los procedimientos III y IV para la detección de outliers y ruido.

Anexo C: Contempla un Glosario con los atributos de la BD empleada, para comprender el contenido de las DDJJ utilizada para el descubrimiento de datos anómalos en datos abiertos de FFPP.

1.2 DESCRIPCIÓN DEL PROBLEMA

En esta Sección se presenta la problemática general en la que se enfoca e impulsa este trabajo de tesis: la corrupción en las sociedad, su naturaleza, taxonomía y las políticas que posibilitan su lucha, su detección.

1.2.1. LA CORRUPCIÓN, SU NATURALEZA COMPLEJA, Y SU IMPACTO EN LA SOCIEDAD

La corrupción es un flagelo que vulnera los derechos humanos de las personas, acentúa la desigualdad social y afecta el desarrollo de la población. El sistema democrático requiere, necesariamente, del aporte de la ciudadanía en el fortalecimiento de las instituciones, como un ejercicio indirecto de la soberanía del pueblo representado.

El 9 de diciembre de 2013, el secretario general de la ONU Ban Ki-Moon, se dirigía a la asamblea general con las siguientes palabras[1]:

> *"La corrupción impide el crecimiento económico al elevar los costos y socava la gestión sostenible del medio ambiente y los recursos naturales. Así mismo, quebranta los derechos humanos fundamentales, agrava la pobreza e incrementa la desigualdad al desviar fondos de la atención de la salud, la educación y otros servicios esenciales. Los efectos perniciosos de la corrupción los sienten miles de millones de personas en todo el mundo"*
>
> (UNODC, 2013)

[1] Un año después a este pronunciamiento, el 9 de diciembre de 2014 se declaraba esa misma fecha como día mundial contra la corrupción.

En sentido amplio la corrupción se entiende como "el abuso del poder para beneficio propio" (Tranparency International, 2009) el abuso autoritario del poder, el cual algunos autores puede comprender:

- El mal uso del poder político.
- Un poder encomendado que puede estar en el sector privado tanto como en el público.
- Un beneficio particular, referido a beneficios personales para la persona que hace mal uso del poder, incluyendo también a miembros de su familia inmediata y de sus amigos.

La definición por otra parte de corrupción política es entendida también como la manipulación de políticas, instituciones y normas de procedimiento en la asignación de recursos y financiamiento por parte de los responsables de las decisiones políticas, quienes se abusan de su posición para conservar su poder, estatus y patrimonio (Transparency International, 2009).

Por otra parte el hecho corruptivo puede tasarse de diferentes maneras: (i), como *corrupción negra*, a aquellas acciones ocultas como el soborno y la extorsión; (ii), *corrupción blanca*, cuando el acto corruptivo es aceptado ampliamente por la comunidad y la corrupción gris en donde existe un debate abierto en la sociedad y medios sobre si esta se considera correcta o no (Cisneros, 2011; Heidenheimer et. al., 1989).

Desde la complejidad del fenómeno de la corrupción, la percepción y aceptación de los hechos de corrupción en las sociedades puede asociarse a contextos y escenarios particulares; por ejemplo, una sociedad puede encontrarse en algún punto medio entre dos escenarios o contextos corruptivos: una sociedad completamente corrupta donde la misma es estructural o una donde se ve mínimamente afectada donde la corrupción se presenta de forma marginal (Gómez & Bello, 2009). Los contextos y escenarios donde el fenómeno corruptivo marginal o estructural puede variar como presenta la Figura 1.

Figura 1: Contextos y escenarios de corrupción (Gómez y Bello, 2009).

Los mismos autores describen las consecuencias de una sociedad que sufre un escenario de corrupción estructural la cual se caracteriza por las siguientes:

- Pérdida de legitimidad del sistema político.
- Asignación de recursos ineficiente e ineficaz.
- Destrucción del profesionalismo y las capacidades de las personas.
- Segregación, marginación y desánimo de las personas honestas.
- Pérdida de previsibilidad sobre el futuro de las organizaciones del sistema.

Por otra parte, el impacto político y social de la corrupción en la sociedad es múltiple y por su particularidad compleja es difícil de mensurar directamente. Para comprender el verdadero alcance del mismo es necesario entender el fenómeno de manera integral, en donde el perjuicio social del acto corruptivo se encuentra mediado por la concurrencia de otros hechos y situaciones que impacta colateralmente en múltiples víctimas indirectas, impactando de lleno en lo público y acrecentando las diferencias sociales, como se observa en la Figura 2.

Figura 2: Impacto de la corrupción (Gómez y Bello, 2009).

Como se desprende de la Figura 2, la corrupción no solo afecta la estabilidad política y mina la confianza societaria -la cual se percibe por ejemplo al ver un noticiero- esta también tiene consecuencias indirectas en la desigualdad y el empobrecimiento poblacional, y consecuentemente sobre el capital social. Sin embargo, cuando el impacto del incremento y presencia de la corrupción se ve amplificado en el empeoramiento del escenario en la corruptela social, la percepción de su perjuicio en la misma -si la sociedad cuenta con los mecanismos cívicos e institucionales para canalizar su curso- se traduce en un mayor participación y control cívico de la propia población como auto-reflejo de esa percepción, cuando esta percepción es genuina.

De lo contrario, si la propia sociedad víctima y culpable al mismo tiempo de un escenario corruptivo no cuenta con las herramientas institucionales, técnicas y ciudadanas provistas por el propio Estado, las sociedades civiles y la ciudadanía se encontraran con mayores inconvenientes y una disipación de sus energías para lograr un Estado más efectivo en el logro del bienestar.

1.2.2. EL PROBLEMA DEL ACCESO CIUDADANO A LA INFORMACIÓN PÚBLICA

El acceso a la información pública es uno de los cuatro pilares fundamentales necesarios, junto a la transparencia fiscal, la revelación patrimonial de los funcionarios oficiales y el compromiso ciudadano, para la existencia de un gobierno abierto multiparticipativo (OGP, 2012).

En Argentina aún no existe ni se ha planteado una estrategia en como sería una mejor participación digital ciudadana (Vercelli, 2013), si bien existe un proyecto de ley de acceso a la información pública este aún no ha sido promulgado como tal lo que hace de factor negativo adicional, además de la necesaria formación civil del ciudadano, la no disponibilidad de la información pública de libre acceso al ciudadano, coarta las posibilidades de una vida democrática plena a la ciudadanía para una participación activa y responsable de todos los actores de la sociedad en la construcción de los consensos necesarios para el fortalecimiento de todo sistema social organizado.

En este sentido el acceso de la ciudadanía a la información pública guarda un rol central en la acción cívica de la misma cuando se afirma que:

> "*Un ciudadano que enfrenta vacíos y lagunas de información sobre los asuntos públicos no tendrá la oportunidad de expresar su opinión en temas inherentes a la administración gubernamental, por lo que mucho menos podrá juzgar los actos de sus gobernantes y como consecuencia a la hora de ejercer su derecho de elegir a estos poca certeza tendrá de que su elección sea correcta; quebrantándose de esa forma uno de los pilares fundamentales, la participación ciudadana, que sustenta a todo el Estado de derecho."* (Cisneros, 2011).

El acceso a la información pública es fundamental para que desde la propia base poblacional ciudadana permita el auto-control y auditoría cívica de los gobernados, y que, naturalmente para la naturaleza digital de las transacciones informacionales, es ineludible que los datos a los que aquellos accedan sean de la mejor calidad posible.

Si bien el recabo de información se encuentra oficiado por la Oficina Anticorrupción (OA), el acceso a las mismas se hace por pedido particular o colectivo lo que dificulta el acceso rápido y efectivo a los datos. Además de problemas burocráticos de modernización en el dominio público, como lo son el necesario y obligado traspaso de archivos desde el formato papel al digital; la OA, se enfrenta a dificultades de excesiva centralización, lo que resulta inconveniente para realizar controles formales y sustantivos, obstáculos en el acceso público a las declaraciones juradas de funcionarios. (Ferro, Giapponi & Gómez, 2007). (Ferro, Giapponi & Gómez, 2007).

Tales problemas prácticos de acceso se solucionan parcialmente a través del nuevo rol del periodismo de BD en la publicación de las declaraciones juradas on-line de funcionarios públicos por parte de organizaciones civiles representa un paso en ese sentido, un esfuerzo de digitalización y difución de avanzada de fácil acceso para el público en general lo representa el desarrollado por Directorio Legislativo, CIJC y DLND (DD.JJ. Abiertas, 2013). A pesar de que tales DDJJ presentan algunos errores la falta de difusión de los mismos descubre la carencia de una estrategia Estatal frente sistemas cívico-institucionales e informacionales frente a una sociedad digitalmente más demandante.

La disposición pública de esa información pone de relieve, además de la impostergable formación digital de la ciudadanía, la necesidad de contar con herramientas para el análisis y explotación de la información. Si bien se han hecho intentos desde las ciencias de la computación para encontrar soluciones tecnológicas para la mejora de los procesos de participación ciudadana directa (Colombo et al., 2013); pocos han sido los casos que encaren la participación indirecta que ejerce el individuo como auditor y contralor cívico de sus propios representantes así como de sociedades civiles abocadas a esa tarea. Por esos motivos, en el siguiente Capítulo, se repasan los antecedentes en el campo para conocer los verdaderos alcances de una aplicación herramental.

1.3 HIPÓTESIS

En esta Sección se presentan las hipótesis que son objeto de este proyecto de investigación, en línea con el problema de la corrupción que engloba a toda sociedad, la posibilidad de contar BBDD públicos y acceso abierto abre nuevos horizontes para el uso original y novedose de técnicas de minería de datos y explotación de la información, ello nos permite plantear las siguientes hipótesis de investigación:

1. Las técnicas de minería de datos conforman una herramienta fundamental para descubrir conocimiento útil en la lucha contra la corrupción y el comportamiento corrupto en las sociedades modernas.

2. Las técnicas de minerías de datos en general y las técnicas de detección de datos anómalos y con ruido en particular resultan útiles y efectivas para la lectura, análisis y tratamiento de datos públicos abiertos.

3. Las BBDD de datos públicos abiertos disponibles actualmente en Argentina pueden ser objeto de análisis experimental para la aplicación y utilización de técnicas novedosas de minería de datos.

De acuerdo a las hipótesis propuestas se articulan los objetivos de la tesis como se describe inmediatamente a continuación.

1.4 OBJETIVOS DE LA INVESTIGACIÓN

En línea con las hipótesis planteadas en la Sección anterior, esta tesis tiene como objetivo la construcción y validación de un modelo de procesos para la utilización de procedimientos de explotación de la información, en base a procesos desarrollados por Kuna (2014); que en su aplicación combinada en forma inédita y sin precedentes para el contralor civil, permita detectar valores anómalos en BBDD en organismos públicos como lo son las declaraciones juradas obligatorias de funcionarios públicos en Argentina (DDJJ). A manera de validación empírica, el trabajo se aproximará en un caso para la detección de anomalías en DDJJ.

Estos pueden enumerarse de la siguiente manera:

1. Evaluar los procedimientos de detección de datos anómalos y ruidos existentes en DDJJ de funcionarios públicos.

2. Cuantificar y caracterizar la presencia de datos anómalos y ruido en DDJJ de funcionarios públicos.

3. Determinar la calidad de los datos en bases de DDJJ de funcionarios públicos.

4. Adaptar, mejorar y optimizar los procedimientos de detección de outliers y ruido a BBDD de DDJJ.

5. Validar el desarrollo de procedimientos que permitan detectar valores anómalos en DDJJ de funcionarios públicos.

6. Promover la utilización de técnicas y herramientas de detección de datos anómalos y explotación de la información para el descubrimiento y revelación de comportamiento corrupto por parte de la función cívica ciudadana así como contribuir a la formación del ciudadano auditor que haga al control ciudadano de la información Estatal.

1.4.1 ALCANCE

El alcance extendido de la tesis comprende el establecer un paradigma para la aplicación de procesos de explotación de la información en el ámbito público, ya sea para los auditores dentro de la administración pública como para la ejercicio cívico y de contralor de la propia ciudadanía representada; *i.e.*, lo que hace a las funciones de las organizaciones civiles.

En base a su aplicación efectiva en datos públicos reales se justifica la necesidad de acceso a la ciudadanía de datos públicos de manera de contribuir a la sociedad en su capacidad de auto-gestión y auto-auditoría a través de organizaciones civiles y del propio Estado contralor.

2. ESTADO DE LA CUESTIÓN

En este primer Capítulo central, se expone el estado de la cuestión en el campo de la minería de datos y los procesos de explotación de la información para la detección de datos anómalos, en la sección subsiguiente se repasa el estado del arte sobre minería de datos así como se describen los antecedentes correspondientes a la utilización de dichas técnicas avocadas al descubrimiento de fraude y corrupción pública y privada. Posteriormente también se abordan los métodos y enfoques para la detección de datos anómalos.

2.1 LA MINERÍA DE DATOS, SU ESTADO DEL ARTE

La minería de datos puede definirse como un proceso de extracción no trivial disponible de manera explícita en BBDD (Schiefer et al., 2004; Clark, 2000); siendo que tal conocimiento en cuestión previamente desconocido, resulta posteriormente adquirido de alguna utilidad inherente para la sociedad en general o en particular.

Por esos motivos la minería de datos puede ser enmarcada como un elemento dentro de un proceso más general dirigido al descubrimiento de conocimiento dentro de grandes sets de datos *"Knowledge Discovery in Databases"* (KDD) (Fayyad et al., 1996); figurativamente, aquel representa la tercera fase en el proceso de descubrimiento de conocimiento como se presenta en el siguiente cuadro.

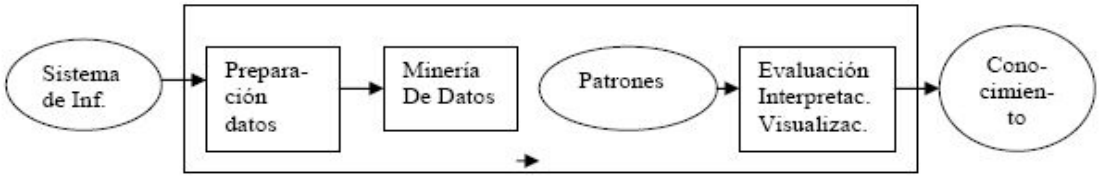
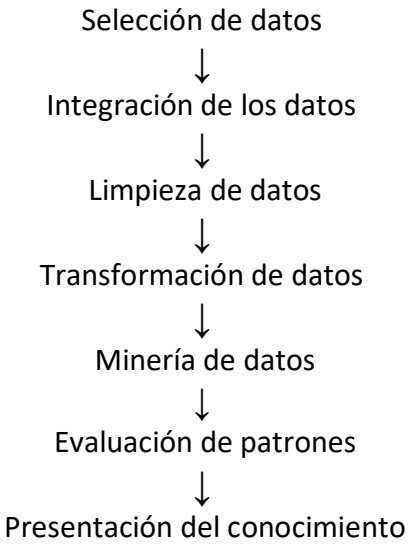


Figura 3: Proceso KDD (Kuna, 2014).

Como se desprende de la Figura 3, la minería de datos es un componente más de un proceso más general de descubrimiento de conocimiento en BBDD, precedido por la etapa de preparación de los datos y proseguido por la de obtención de patrones, producto de la aplicación de técnicas de minería de datos y materia prima para la obtención de conocimiento *a posteriori* de la fase de evaluación e interpretación de patrones.

Análogamente, la minería de datos también puede ser considerada parte de un proceso de EI (Larose, 2005), dentro del cual la minería de datos viene a conformar la quinta fase del proceso (Britos et al., 2005): *a priori* de la evaluación de patrones y *a posteriori* de la transformación de datos como se describe a continuación.



Modelos y técnicas de la minería de datos

A su vez los modelos de minería de datos pueden ser tipificados de dos maneras básicas:

- Modelos predictivos: Se utiliza fundamentalmente la clasificación donde según una BD, donde se busca predecir a qué clase pertenece una nueva instancia, asignándose previamente un valor en cada tupla correspondiente a una clase predeterminada. Para análisis predictivos numéricos se utiliza la regresión.

- Modelos descriptivos: Buscando generar etiquetas y/o agrupaciones estos modelos exploran las propiedades de los datos bajo análisis, donde se utiliza fundamentalmente el clustering así como la correlación y factorizaciones para evaluar el grado de similitud entre atributos numéricos; reglas de asociación, para la búsqueda de asociaciones inter-atributos no explicitas.

Las técnicas de minería de datos por otra parte pueden clasificarse entre técnicas basadas en análisis estadístico y técnicas basadas en sistemas inteligentes.

Por un lado algunas de las tecnologías basadas en análisis estadístico más usadas son las siguientes.

- Análisis de agrupamiento
- Análisis de varianza
- Análisis discriminante
- Prueba Chi-cuadrado
- Regresión
- Series de tiempo

Por otra lado, algunas de las tecnologías basadas en sistemas inteligentes pueden enumerarse de la siguiente manera (García Martinez et al., 2003).

- Algoritmos de inducción
- Perceptrón multicapa
- Maquina de vector soporte
- Algoritmos *a priori*

- Redes neuronales SOM
- Algoritmos genéticos
- Redes bayesianas
- Algoritmo del vecino más próximo
- Variedad de técnicas estocásticas, relacionales, declarativas, difusas, así como múltiples tecnologías híbridas.

Resumidamente, puede explicitarse relacionalmente los vínculos entre técnicas y modelos de minería de datos en la Tabla 1 a continuación.

Técnicas	Predictivo		Descriptivo		
	Clasificación	Regresión	Agrupamiento	Regla de Asociación	Correlaciones/ Factorizaciones
Redes Neuronales	■	■	■		
Arboles de decisión ID3, C4.5, C5.0	■				
Arboles de decisión CART	■	■			
Otros árboles de decisión	■	■	■	■	
Redes de Kohonen			■		
Regresión Lineal		■			■
Regresión Logística	■			■	
K-Means			■		
A priori				■	
Naive Bayes	■				
Vecinos más próximos	■	■	■		
Algoritmos genéticos y evolutivos	■	■	■	■	■
Máquinas de vectores de soporte	■	■	■		
Análisis discriminante multivalente	■				

Tabla 1: Relación entre técnicas y métodos de minería de datos.

2.2. MINERÍA DE DATOS PARA LA DETECCIÓN DE FRAUDE Y CORRUPCIÓN

Como consecuencia de la realidad actual que enfrentan los sistemas y procesos deben enfrentarse a crecientes volúmenes de información, la minería de datos posee por lo tanto una ventaja teórica sobre las técnicas manuales en la búsqueda de evidencias al evitar la subjetividad de los análisis y optimizar sustancialmente los tiempos requeridos para realizar (Kuna, 2014).

Un claro antecedente en el uso de herramientas computacionales para la detección de fraude (Abbot et al., 1998), donde se testearon un conjunto de modelos para la detección de falsos positivos en transacciones fraudulentas estableciendo la superioridad de árboles de decisión y redes neuronales para acometer esa tarea.

En otro caso de aplicación (Spatis, 2002) se desarrollaron dos modelos orientados hacia la detección de una gerencia fraudulenta en empresas privadas, a través del empleo de regresiones lógicas, representa un caso aplicado en este sentido. Otros estudios que utilizaron minería de datos en la búsqueda de formas de corrupción en la actividad privada han sido relacionadas con el fraude bancario, más especificamente a los vinculados con tarjetas de crédito y aseguradoras de automotores (Foster & Stine, 2004).

A diferencia del caso de la corrupción pública, los casos de corrupción privada compleja, representada por las actividades fraudulenta persistentes en la alta gerencia corporativa, si supone un paradigma en la utilización de técnicas y herramientas de EI. Estos incluyen toda falsificación intencional de estados contables y financieros con la intención de obtener beneficios ilegales (Wang et al., 2006; Wang, 2010), lo cual se hace no solo para evadir o eludir impuestos sino también con el objetivo de confundir a inversores y acreedores.

En este sentido los delitos de corrupción corporativa vinculados generalmente con la alta gerencia, no solo han venido han aumentado considerablemente (Koskivaara, 2004) sino que se

han profundizado a partir de la crisis financiera de 2008 y por consiguiente los requerimientos para detectar, definir, y reportar fraude financiero y contable han aumentado (Yue X. et al., 2007). Antes del agravamiento de la corrupción corporativa global varios autores efectuaron contribuciones que contribuyan a la lucha de esta problemática compleja de corrupción privada.

Otros autores (Green & Choi, 1997) realizaron una clasificación de fraudes en la alta dirección corporativa a partir del uso de redes neuronales; así como se analizó el trabajo de auditores en la utilización de CAATs a través de sistemas de expertos para una mejor discriminación de riesgo de fraude corporativo en distintos niveles de gestión (Eining et al., 1997).

Respecto de corrupción corporativa interna también se llegaron a construir modelos de detección de fraude interno (Fannin & Cogger, 1998) por medio de la utilización de redes neuronales junto a ratios financieros y variables cualitativas aseverando que su modelo era más eficaz que métodos estadísticos tradicionales.

Otros autores como Kirkos y otros (Kirkos et al., 2007) fueron incluso más allá elaborando una comparación de técnicas de minería de datos en la auditoría corporativa de sistemas; en el cual a partir de los estados contables y financieros de 76 firmas griegas se evaluaron el rendimiento de tres técnicas particulares de EI: árboles de decisión, redes neuronales y redes bayesianas en dos fases, una de entrenamiento y otra de validación, encontrando que las RN eran más eficaces a la hora de determinar las firmas que sufrían fraude con el modelo sin entrenar mientras que las RB mostraron un mejor desempeño con los modelos entrenados en etapa de validación.

2.2.1 CLASIFICACIÓN DE TÉCNICAS DE MINERÍA PARA LA DETECCIÓN DE FRAUDE

Diversos autores han especificado una clasificación de la corrupción privada la cual se manifiesta en la forma de diversos fraudes financieros, en (Ngai et al., 2011) puede apreciarse una clasificación exhaustiva de estos últimos los cuales pueden comprender fraudes bancarios, tarjetas de crédito, blanqueo o lavado de dinero y fraude hipotecario. Los fraudes llevados acabo por aseguradores comprenden un amplio universo y comprenden las estafas por cobro fraudulento de seguros por cosechas, seguros de salud, automóviles etc. Finalmente, quizás la categoría más compleja de corrupción privada sea la corporativa, fraudes financieros

internacionales más complejos que involucran la falsificación de información corporativa, la especulación financiera, etc.

Figura 4: Clasificación de técnicas de explotación de la información para la detección de fraude financiero (Ngai et al., 2011).

La Figura 4 presenta una clasificación de fraudes y técnicas de minería para combatirlo. Similarmente otros autores (Sowjanya y Jyotsna, 2013), hicieron también lo propio en pos de confeccionar una taxonomía abarcativa de la aplicación de minería de datos en la detección de fraude, como se presenta en la Figura 5.

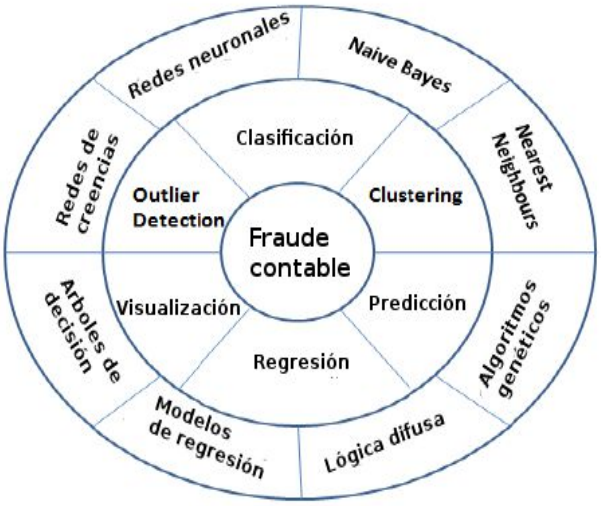


Figura 5: Algoritmos y técnicas de minería de datos para el descubrimiento de fraude fiscal (Sowjanya y Jyotsna, 2013).

En las clasificaciones exhaustivas desde la Figura 4 como la Figura 5 se vislumbra tácitamente la poca importancia que se le ha dado desde el uso de la minería de datos a los problemas cívicos y de administración pública.

Ahora considerando propiamente al fenómeno corruptivo a nivel público y sistémico, un esfuerzo reciente se despliega en el trabajo de Ransom (2013) en uno de los pocos trabajos en el campo abordando el fenómeno de la corrupción como caso de estudio, midieron los niveles de discusión sobre corrupción usando técnicas de minería de texto (*Text Mining*) y extracción de la información (*Information Extraction*) de un cuerpo de datos extraído de noticias y reportes sobre corrupción en 215 países en un lapso de cinco años, lo que les permitió discutir que tan prominente era el debate público sobre corrupción en cada nación. En el estudio los autores afirman de la utilidad de la minería de datos para la diseminación de ideas y conceptos en pequeñas organizaciones civiles que luchen contra la corrupción, reivindicando la utilidad de la minería de datos como herramienta en ciencias sociales, el monitoreo y la investigación científica.

Por otra parte, otros autores (Huysmans et al., 2006; Huysmans, et al. 2008) aplican técnicas de minería en un estudio transversal en cross entre países buscando relacionando variables

macroeconómicas a niveles percibidos de corrupción en dos pasos: usando mapas auto-organizados de Kohonen (SOM) *a priori* y maquinas de vector soporte -*Support Vector Machines*- (SVM) supervisadas *a posteriori* son entrenadas para pronosticar niveles de corrupción futura para cada país, y volcando *a fortiori* tales pronósticos nuevamente en mapas auto-organizados para comparar distintos modelos de comportamiento.

En los casos anteriores, aunque ambos representan esfuerzos rigurosos en cuanto a la búsqueda de la verdad, los mismos representan un enfoque informativo para la sociedad, pero de alcance netamente unidimensional y acotado, sin considerar la actividad del ciudadano cívico y sus organizaciones civiles que buscan una herramienta para "auditar" a sus conciudadanos representantes.

Sin desmedro de las restantes técnicas de minería de datos, las técnicas de detección de outliers o valores anómalos ha sido escasamente implementada para la solución escenarios de fraude, y por supuesto mucho menos aún, dentro del estudio de la corrupción en el ámbito público. Un caso de aplicación en este sentido para el descubrimiento de corrupción privada en el uso específico de técnicas de detección de datos anómalos fue materializado sobre transacciones de tarjeta de crédito (Aleskerov et al., 1997) de manera de buscar algún tipo de patrón corrupto relacionado con la existencia de fraude.

Figura 6: Técnicas de EI usados para la detección de corrupción pública.
(Elaboración propia)

En la Figura 6, podemos plantearnos la cuestión de si es posible la aplicación de técnicas de detección de outliers ahora para la detección de fraude en el ámbito público, aspectos los cuales poseen un alcance de trascendencia social e institucional particular con efectos profundos sobre la sociedad representada.

Los campos anómalos u outliers -siguiendo a Hawkins (1980)- se definen como un dato que por ser muy diferente a los demás pertenecientes a un mismo conjunto de datos, i.e.: una base de datos contenedora de tales campos, puede considerarse que fue creada por un mecanismo diferente; lo que, en el descubrimiento de tales mecanismos, radica el conocimiento latente en cada base analizada.

Dada la carencia de procesos y algoritmos de detección de datos anómalos dirigidos al problema de la corrupción pública en la sociedad civil, como se expone en la Figura 6, se detecta una ventana de oportunidad para la propuesta de búsqueda y experimentación tanto novedosa como oportuna de procesos desarrollados recientemente para la detección de campos anómalos y ruido, útiles para la detección patrones en BBDD civiles y públicas; estas, harto sensibles al descubrimiento de nueva información y conocimiento de posibles procesos corruptivos que afectan el bienestar y el tejido social como se expone a continuación en los siguientes enfoques y métodos.

2.3. ENFOQUES Y MÉTODOS PARA LA DETECCIÓN DE OUTLIERS

La búsqueda de valores anómalos en datos se remonta a más de dos siglos atrás donde ante la presencia de datos discordantes se procedía a descartarlos del resto de la muestra (Boscovich, 1757). A lo largo del tiempo fueron evolucionando en técnicas más complejas y eminentemente luego del desarrollo de técnicas de cómputo y la diseminación del computador.

Como estudiaron varios autores (Chandola et al., 2009), las técnicas de detección de datos anómalos poseen componentes principales asociados con cualquier técnica o método de detección de outliers como se describe a continuación en la siguiente Figura 7.

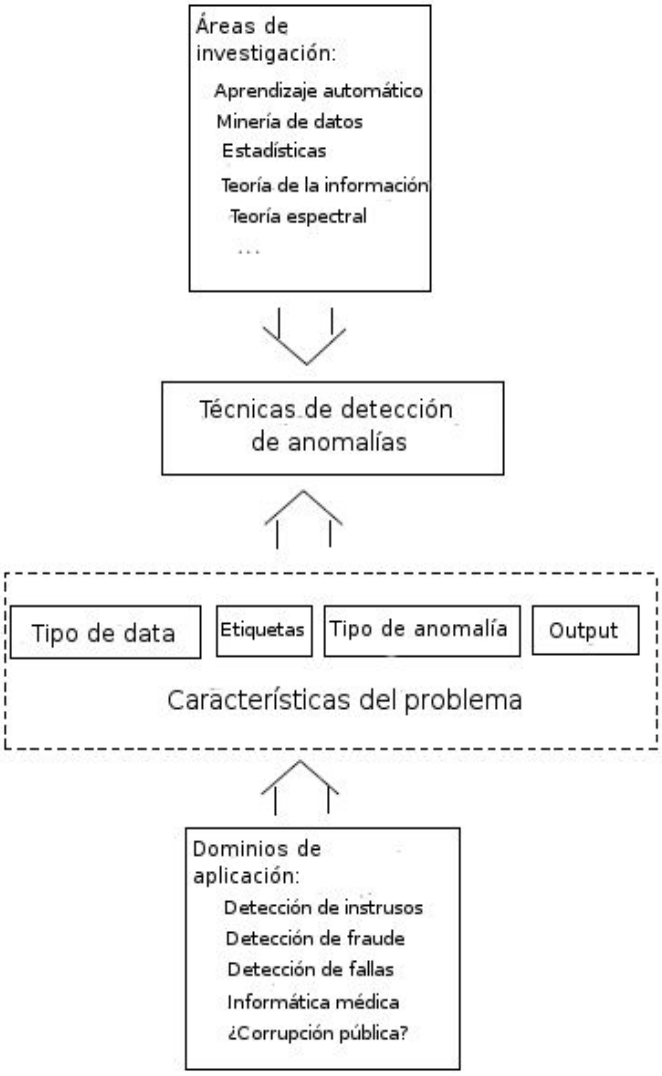


Figura 7: Componentes de las técnicas de detección de anomalías
(Elaboración propia en base a Chandola et al., 2009).

Como se desprende de la Figura 7, la minería de datos viene a ser simplemente un área más potencial más para el descubrimiento de conocimiento científico, puesto que al mismo tiempo considera las estadísticas, aprendizaje automático, teoría de la información y teoría espectral como áreas de investigación útiles para su uso.

Estas técnicas a su vez, para abordar la detección de outliers, presentan tres enfoques básicos para su resolución (Hodge & Austin, 2004; Chandola et al., 2009):

Enfoque tipo 1:	Detección de outliers con aprendizaje no supervisado.
Enfoque tipo 2:	Detección de outliers con aprendizaje supervisado.
Enfoque tipo 3:	Detección de outliers con aprendizaje semi-supervisado.

Tabla 2: Enfoques para el problema de la detección de outliers (Kuna, 2014).

En la Tabla 2, se describen los tres enfoques posibles que pueden adoptar técnicas de detección de datos anómalos y ruido. Retornando a los objetivos que propone este esfuerzo de tesis, es necesario volver a la incógnita de si el ámbito de aplicación de estas técnicas -que ya han sido probadas para la detección de fraude, de intrusos, de fallas e incluso de informática médica- pudieran aplicarse análogamente esta vez para la detección y/o instrumento en la lucha de la corrupción pública.

Dada la carencia de procesos y algoritmos dirigidos a la detección de datos anómalos en resolución de la problemática de ética pública en las sociedades, como se expone en la Figura 6, se detecta una ventana de oportunidad para la propuesta de búsqueda y experimentación. En este sentido, y a sabiendas que las DDJJ se conforman habitualmente por datos alfanuméricos, se descubre con gran potencialidad de aplicación hipotética la utilización de los procedimientos híbridos III y IV de detección de datos anómalos desarrollado recientemente por Kuna (2014), los cuales identificamos idóneos por las siguientes razones:

- Son procedimientos desarrollados recientemente y representan el estado del arte en lo que hace a la detección de campos anómalos y ruido.

- Son procedimientos óptimos para la detección de ruido en BBDD alfanuméricas como generalmente se encuentran caracterizados los datos públicos.

- Son de fácil aplicabilidad y ejecución con los programas de minería de datos disponibles actualmente.

La idoneidad de estos procedimientos son propuestas como solución hipotética para un caso de contralor para la mejora de los datos públicos y cívicos de una población.

3. SOLUCIÓN PROPUESTA

En este Capítulo se propone la solución propuesta a la problemática planteada, la metodología y los procesos para llegar a su resolución. En la primera sección se abre el camino a la explicación de los mecanismos y sistemas político sociales que promueven el desglose de escenarios corruptivos complejos para luego ahondar en técnicas y procesos de minería de datos que soporten los primeros.

Hasta ahora, el abordaje de los trabajos científicos sobre corrupción que emplearon minería de datos lo hicieron propugnando una unidireccionalidad del conocimiento sin considerar la realidad digital contemporánea en la que se encuentra situada la ciudadanía actualmente. Sin considerar tal situacionalidad y potencial digital de la ciudadanía, como tantos otros trabajos (Ransom, 2013; Huysman et al., 2008, 2006); desde las bases usadas para realizar las técnicas de minería de textos en uno, como los índices de percepción de corrupción en el otro, puede llegar a ser discutidas aspectos de acuerdo a las fuentes utilizadas así como su inherente subjetividad, los análisis intra-países también pueden resultar dificultosos de interpretar puesto que al ser la corrupción un fenómeno complejo, este puede adquirir distintos aspectos culturales endógenos para cada país.

Este trabajo pretende abordar el problema de la corrupción desde una perspectiva sistémica diferente, de manera de no solo ofrecer una herramienta para el mayor control a la ciudadanía y organizaciones sociales, sino también poner de relieve la situación patrimonial del alto funcionario publico y así contribuir a poner en valor el ejemplo ético y político como herramienta para la prevención del comportamiento corrupto.

3.1. MEDIDAS PARA LA PREVENCIÓN DE LA CORRUPCIÓN

Prevenir la corrupción y promover la transparencia en la gestión pública, implica desarrollar una mirada compleja que integre políticas y prácticas punitivas con políticas y prácticas preventivas (Gómez y Bello, 2009), un paradigma reciente que se aleja del enfoque tradicional que se apoyaba fundamentalmente en la penalidad de agente corrupto.

La respuesta penal punitiva al problema de la corrupción como única solución demostró ser limitada y poco sustentable en el largo plazo (Canavese 2005, Gómez y Bello 2009) por lo que es necesario desarrollar una mirada compleja que combine políticas y prácticas punitivas con prácticas preventivas de los hechos corruptivos.

La propuesta economicista era presentada generalmente en términos probabilísticos y de utilidad. En este enfoque pueden destacarse la necesidad de reducción de los incentivos a que el funcionario actúe deshonestamente. Siguiendo a Becker (1968) este sugiere aumentando la probabilidad de que el funcionario sea descubierto. Por otra parte el mismo autor sugiere la necesidad de incrementar la satisfacción moral de comportarse honestamente y acrecentar el salario de los funcionarios. Otros autores de este enfoque (Canavese, 2005) ya advirtieron de la inutilidad de la penalidad como solución a este problema social complejo. La principal deficiencia de este enfoque es que se encuentra restringido a una visión asignativa de agentes económicos maximizadores de utilidad rentística sin un capital social preexistente, poniendo en duda su aplicación real concreta.

Pero la lucha contra la corrupción es también una obligación para los Estados nacionales, la Convención de las Naciones Unidas Contra la Corrupción (Naciones Unidas, 2004; Raigorodsky y Geler, 2007), de la cual Argentina suscribió el acta que afirma que cada Estado procurará [...] *establecer y fomentar prácticas eficaces encaminadas a prevenir la corrupción*. Descartado los enfoques anteriores, las políticas anti-corrupción deben recaer en la prevención.

Según el manual de medidas prácticas contra la corrupción de las Naciones Unidas (Naciones Unidas, 1993; Gómez y Bello, 2009) se consideran centrales los tres siguientes aspectos para acometer dicha tarea:

- El fortalecimiento y la construcción institucional.
- La prevención.
- Toma de conciencia de la ciudadanía.

La primera referente al **fortalecimiento y construcción institucional**, comprende la creación de agencias anti-corrupción; un ombudsman o defensor del pueblo; el fortalecimiento del Poder Judicial, la rendición de cuentas precisa y oportuna; códigos de conducta; comités de integridad nacional o comisiones anti-corrupción; encuentros de integridad nacional para desarrollar planes de acción; fortalecimiento de gobiernos locales.

En segundo lugar, **la prevención**, hace a las declaraciones de activos y pasivos de funcionarios públicos como un instrumento para ello así como la creación de una autoridad de monitoreo internacional para la transparencia de los contratos del sector público en las transacciones comerciales internacionales; islas de integridad y pactos de integridad; grupos de coordinación de información: cooperación anti-corrupción del sector privado, reducción de la complejidad en los procedimientos y de la discrecionalidad.

Finalmente **la toma de conciencia de la ciudadanía**, la que se logra a través del acceso a la información pública; movilización de la sociedad civil a través de la educación pública; planes de acción anti-corrupción; capacitación de la prensa y periodismo de investigación.

En este sentido especialistas en el campo ya habían manifestado también de la importancia de los dos últimos puntos dado que hacer operativo el derecho a la información pública en términos de prevención de la corrupción es proporcionar a los ciudadanos las herramientas para ejercer sus derechos y controlar la acción del Estado (Raigorodsky y Geler, 2007; Naciones Unidas, 2004).

Visión integral que hasta puede entenderse como análogamente compatible a la propuesta a la de OGP, el cual estipula los siguientes cuatro aspectos clave que aseguran un gobierno abierto (OGP, 2012), como se desprende del siguiente cuadro.

Condiciones para un gobierno abierto y participativo			
Transparencia fiscal	Acceso a la información	Divulgación patrimonial de funcionarios	Compromiso ciudadano

Tabla 3: Aspectos de los gobiernos abiertos (OGP, 2012).

En la Tabla 3 se presentan los criterios básicos de gobierno abierto los cuales son necesarios para incrementar la responsabilidad de la administración pública, fortalecer el compromiso ciudadano y luchar contra la corrupción (OGP, 2012).

3.1.2 LOS SISTEMAS DE DDJJ PATRIMONIALES COMO MECANISMO DE PREVENCIÓN DE LA CORRUPCIÓN

La revelación y divulgación del patrimonio de los funcionarios oficiales es uno de los cuatro pilares para la elegibilidad de un gobierno abierto (OGP, 2012). Estos sistemas de información se encuentran representados en nuestro país por los sistemas o regímenes de declaraciones juradas; los cuales, poseen tres funciones básicas para lo cual fueron implementados (Gómez y Bello, 2009):

- Controlar la evolución patrimonial de los funcionarios de la función pública para prevenir enriquecimiento ilícito y otros delitos de corrupción.
- Detectar y prevenir conflictos de intereses e incompatibilidades de la función pública.
- Como mecanismo de transparencia y prevención de la corrupción pública.

Por lo tanto, si a esas políticas preventivas en los cuales se incluyen los regímenes y sistemas de declaraciones juradas patrimoniales como herramienta fundamental; se combinan con la

utilización de procedimientos y técnicas de minería de datos, se podrá fortalecer aún más el poder preventivo de la herramienta institucional como factor de la herramienta computacional.

En el sentido preventivo todo régimen de DDJJ de funcionarios públicos representan una herramienta que posibilita:

- El control del adecuado cumplimiento de las funciones públicas que desempeñan los funcionarios.
- Prevenir el desvío de sus deberes éticos.
- Corregir incumplimientos detectados.

Los mismos son documentos donde se expone, la variación patrimonial de oficiales públicos durante el desempeño de sus funciones. Sus antecedentes laborales -en especial aquellas relaciones contractuales mantenidas en forma simultanea en el desempeño del cargo público-; así como relaciones laborales que hayan cesado en un tiempo relativamente breve anterior a la toma de funciones. Un sistema de DDJJ de funcionarios también es un sistema de información comprende los siguientes elementos:

Figura 8: Sistema de DDJJ patrimoniales (Gómez y Bello, 2009).

Como se expone en la Figura 8, los sistemas de DDJJ se componen de un universo de funcionarios públicos con sus atributos correspondientes, las DDJJ *per se*, y la cantidad y cualidad de la información como output. Estos sistemas constituyen una de las herramientas centrales para la prevención y el combate de la corrupción, ya que representan instrumentos que

permiten mejorar significativamente los niveles de control sobre los funcionarios así como las vías para establecer su responsabilidad (Raigorodsky y Geler, 2007). Un sistema de declaraciones juradas, por otra parte, además que aumenta el costo-oportunidad de cometerse conductas desviadas así como mecanismo de transparencia y prevención de la corrupción (Ferro G., Giupponi L., Gómez, N. 2007).

Siguiendo uno de autores más relevante sobre el fenómeno de la corrupción (Klitgaard, 1992) el mantenimiento de los estándares de comportamiento ético por parte de los funcionarios públicos dependerá de la interacción de tres factores:

- Honestidad personal
- Nivel de ambición
- Sentimientos de pertenencia e integración al grupo

Un sistema o régimen de declaraciones juradas tiene una incidencia profunda por sobre estos factores personales puesto que, además de aumentar el costo-oportunidad de acometer conductas desviadas, permite la satisfacción moral del funcionario de cumplir con el deber de declarar su patrimonio y así rendir cuentas a la ciudadanía (Ferro, Giapponi y Gómez; *ut supra*; Becker, 1968).

Un sistema de DDJJ abierto y digitalizado disponible para la ciudadanía en acceso real, contribuirá a la reducción de la corrupción, en cuanto mayor transparencia (mediante el incremento del acceso a la información pública), reducción de la brecha digital cívica (mediante la (auto)participación ciudadana en el contralor civil), y una red isonómicamente neutral[2]

[2] De isonomía o isonómico: igualdad ante la ley, proveniente del marco civil de internet brasilero que busca la no discriminación en la regulación de internet.

(Vercelli, 2014) al proporcionar al Ciudadano la posibilidad de usar su tiempo conectivo cívico y digitalmente para la construcción de su propio proyecto fenoménico (López-Pablos, 2015a).

En Argentina según la Ley Nacional de Ética en la función pública (Ley Nacional nro. 25.188), están obligados a realizar declaraciones juradas patrimoniales, los funcionarios con categoría de Director o equivalente, los diputados y senadores del Poder Legislativo, los Jueces y Secretarios del Poder Judicial, el Presidente de la Nación, los Ministros, Secretarios y Subsecretarios del Poder Ejecutivo Nacional entre otros; considerándose falta grave el incumplimiento de la presentación de la declaración jurada; *i. e.*, que un funcionario puede recibir una condena penal si falsea u omite datos en sus DDJJ.

3.2 PROCEDIMIENTOS PARA LA DETECCIÓN DE OUTLIERS Y RUIDO EN BBDD

Generalmente almacenados en bases de datos relacionales, los datos algunas veces son considerados anómalos; *i. e.* diferentes al resto de los datos, ya sea por un error o por negligencia o malintención. La existencia de estas anomalías coarta definitivamente la capacidad de la información para generar predicciones y patrones en los datos, menguando la calidad de los mismos y por consiguiente el conocimiento y las decisiones resultantes de aquellos. Por otra parte al considerar algunos atributos económicos relacionados con la riqueza la existencia de posibles datos anómalo podría encontrarse relacionado con concentraciones de riqueza, ingreso y concentración extraordinarias de poder económico. Tales concentraciones de riqueza no representan un crimen al ser fruto del trabajo honrado, muchas veces, consecuencia de una o varias generaciones de familias de empresarios y trabajadores, pero que, al alcanzar ciertos grados de concentración a la vez que de poder político y social; respecto a la responsabilidad cívica y social del FFPP relativo a la sociedad en que se encuentra inserto, exigen de aquel una moralidad solidaria superior, doblemente ejemplar y proba en relación a aquellas clases más relegadas y excluidas que no han tenido la planificación suficiente, así como de la sociedad toda que espera esfuerzo y ejemplaridad de sus representantes en la función pública.

Dado que no es posible detectar campos de datos anómalos utilizando un solo algoritmo para todos los posibles entornos de posibles anomalías que puedan presentarse, se plantea la

necesidad de abordar la implementación de procedimientos híbridos para la detección integral de los mismos (Kuna, 2014), de manera de posibilitar la detección de campos anómalos teniendo en cuenta los dominios específicos de su implementación, las características de la bases de datos, la tipicidad su dimensionalidad, etc. Por otra parte, además del problema que plantea la detección de outliers se plantea la existencia de inliers o datos expuestos como atípicos cuando en realidad no presentan un comportamiento de un verdadero outlier.

Siguiendo procedimientos de explotación de la información para la identificación de datos faltantes, ruido e inconsistencias (Kuna, 2014), se definen reglas para la combinación híbrida de algoritmos en un proceso formal para dicha tarea estipulado en 4 procedimientos. El mismo plantea la utilización de combinaciones de algoritmos para cada uno de los procedimientos, sucintamente se contempla cada uno en la Tabla 4 como sigue.

Procedimiento I:	Set normal, LOF; con enfoque no supervisado y semisupervisado.
Procedimiento II:	Set normal, LOF y K-Means; con enfoque no supervisado.
Procedimiento III:	Set alfanumérico con atributo objetivo, algoritmos de inducción C4.5, LOF y TI; con enfoque no supervisado y supervisado.
Procedimiento IV:	Set alfanumérico sin atributo objetivo, LOF, DBSCAN C4.5, PRISM y K-Means; con enfoque no supervisado y supervisado.

Tabla 4: Algoritmos por procedimiento híbrido para la detección de outliers y ruido (Kuna, 2014).

La utilización de métodos híbridos con enfoques supervisado, no-supervisado y semisupervisado para la detección de outliers y ruido ya se habían constituido como la mejor alternativa al lograrse la mayor ganancia de información, reducción del espacio de búsqueda y optimización de los procesos (Kuna et al., 2009; Kuna et al., 2010a; Kuna et al., 2011). Investigaciones recientes

han determinado que es posible afirmar que la combinación de algoritmos de distinta naturaleza y también la combinación de procedimientos permite detectar outliers con un nivel de confianza mayor al 60%, entendiendo (Kuna, 2014).

En los procedimientos híbridos I y II expuestos se combinan ambos para detectar campos de outliers en bases de datos numéricas (Kuna et al., 2012; Kuna et al., 2013a); mientras que el procedimiento III detecta campos de outliers en bases de datos alfanuméricas conteniendo un atributo clase (Kuna et al., 2012c; Kuna et al., 2013b; Kuna, 2014), igualmente al procedimiento IV, el cual también detecta campos en bases alfanumérica solo que sin un atributo target (Kuna et al., 2014; Kuna, 2014).

3.3 METODOLOGÍA HÍBRIDA PARA LA DETECCIÓN OUTLIERS

Metodológicamente se propone la utilización de una metodología híbrida para la detección de datos anómalos y ruido. Este proveerá de un procedimiento que sirva para garantizar el control y la calidad de los datos e información que se piense auditar en la esfera pública.

La aplicación de la metodología híbrida para la detección de outliers en bases de datos y ruido seleccionado (Kuna, 2014) puede estandarizarse de acuerdo según dependa de la utilidad de cada procedimiento, un entorno de determinado de BBDD, la algoritmia utilizada, sus técnicas y enfoques como se despliega a continuación en el siguiente cuadro.

Procedimiento	Entorno	Algoritmos y técnicas	Enfoques
I	BBDD numéricas con o sin atributo objetivo	LOF; metadatos	Tipo 1 y 3
II	BBDD numéricas con o sin atributo objetivo	LOF; K-Means	Tipo 1
III	BBDD alfanuméricas con un atributo objetivo	C4.5, Teoría de la información; LOF	Tipo 1 y 2
IV	BBDD alfanuméricas que no contienen un atributo objetivo	LOF; DBSCAN; C4.5; RB; PRISM; K-Means	Tipo 1 y 2

Tabla 5: Procedimientos según entorno, algoritmos y enfoque (Kuna, 2014).

Como se desprende de la Tabla 5, los procedimientos I y II permiten detectar outliers en bases de datos numéricas basándose en metadatos de registros considerados normales y en el algoritmo LOF que proporciona un valor local de outlier de donde se define la medida de anomalía para una tupla considerada; sin embargo, para las DDJJ de funcionarios públicos poseen un entorno alfanuméricos, lo que intuye a la utilización de los procedimientos III y IV para la búsqueda de datos anómalos.

3.3.1 PROCEDIMIENTO PARA LA DETECCIÓN DE OUTLIERS BBDD ALFANUMÉRICOS CON ATRIBUTO CLASE

Advirtiendo que una gran parte de las BBDD son de naturaleza alfanumérica lo que exige una mayor complejidad procedimental para llegar a la detección y localización de tuplas candidatas a anómalas.

El procedimiento III (Kuna et al., 2012c; Kuna et al., 2013b; Kuna, 2014) es capaz de detectar campos de outliers en bases de datos alfanuméricas con un atributo target u objetivo, combinando la aplicación de un algoritmo de inducción (C4.5), con elementos de teoría de la información de Shannon y LOF, como se describe en el siguiente cuadro de relaciones.

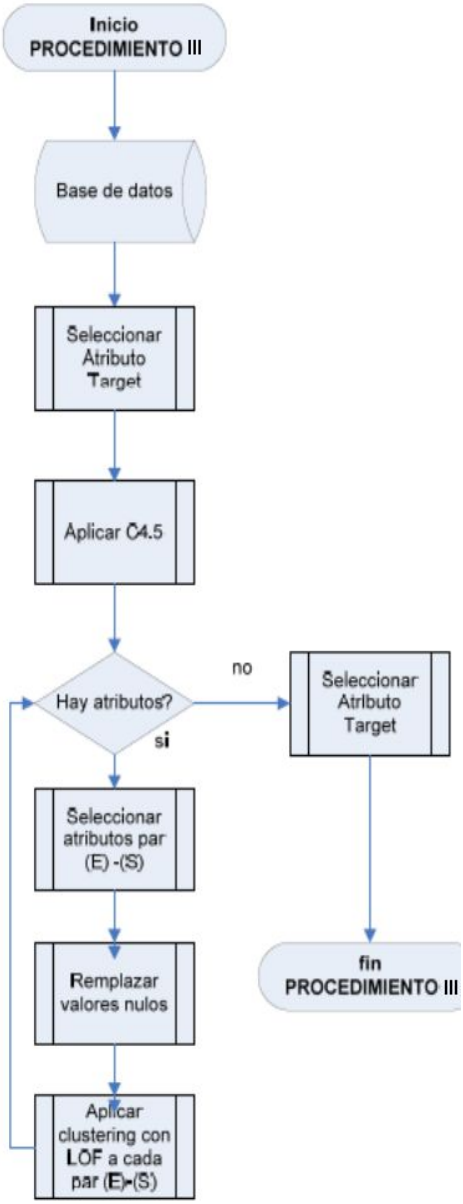


Figura 9: Procedimiento (III) de detección de outliers, BBDD alfanuméricas y con atributo clase (Kuna, 2014).

El procedimiento descripto anteriormente en el diagrama relacional de la Figura 9 superior anterior, este también puede describirse en seudo-código como sigue inmediatamente.

❍ Entrada BD

- Determinar atributo clase
- Aplicar algoritmo **C4.5**
- Y para cada atributo significativo:
  - Con (E): atributo seleccionado (entrada) y (S) atributo clase se arma el bin-conjunto (E)+(S).
  - Los datos de entrada (E) son analizados reemplazando los valores nulos por etiquetas.
  - Se aplica **LOF** al bin-conjunto (E)+(S) generando un atributo outlier.
- Se filtran los pares de datos del bin (E)-(S) con valores de outliers ≠ 0, el cual indica la presencia de dato/s anómalo/s al no aportar información y producir ruido en relación al atributo clase.

❍ Salida BD con campos anómalos detectados.

Según la experimentación empírica con BBDD reales y artificiales, este procedimiento demostró una efectividad promedio que suele rondar el 90% (Kuna, 2014).

3.3.2 PROCEDIMIENTO PARA LA DETECCIÓN DE OUTLIERS CON BBDD ALFANUMÉRICOS SIN ATRIBUTO CLASE

Pero sin embargo no en todos los abordajes con BBDD alfanuméricas se contempla la asumsión de un atributo clase determinado para el descubrimiento de datos anómalos y conocimiento. Igualmente que el anterior procedimiento anterior, el procedimiento IV también fue concebido para la detección de campos anómalos en BBDD alfanuméricas sin un atributo clase, donde se combinan primariamente algoritmos diseñados específicamente para la detección de outliers (LOF y DBSCAN)[3] para posteriormente aplicar algoritmos de clasificación (C4.5, PRISM y RB)[4] para finalizar con un algoritmo de clusterización de gran simpleza como lo es el K-Means.

[3] Revisar los Anexos A.1 para obtener una descripción de la union de resultados de la aplicación de LOF y DBSCAN y el Anexo A.1.1 para conocer la parametrización correspondiente a sus reglas necesarias para la determinación de outliers.

[4] Revisar los Anexos A.2 para obtener una descripción de la union de resultados de la aplicación de LOF y DBSCAN y el Anexo A.2.1 para conocer la parametrización correspondiente a sus reglas necesarias para la determinación de outliers.

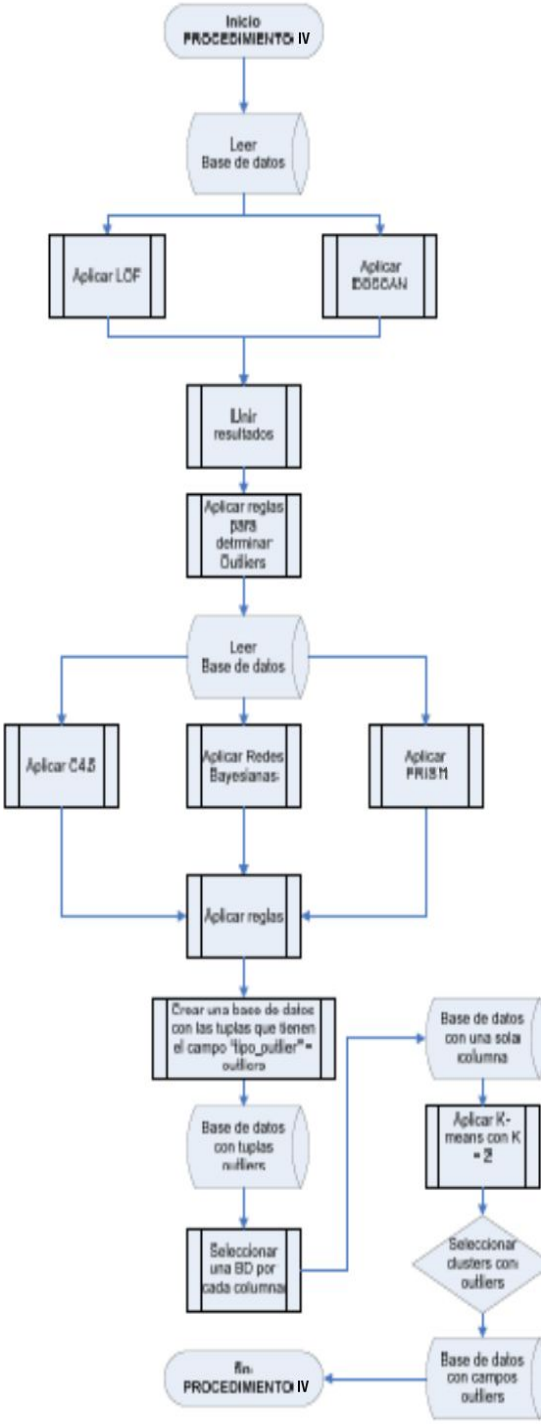


Figura 10: Procedimiento (IV) de detección de outliers, BBDD alfanuméricas y sin atributo clase (Kuna, 2014).

Como lo muestra la Figura 10 *up supra* el procedimiento propuesto tiene las siguientes etapas, las que pueden ser expresadas en seudo-código de la siguiente manera:

❍ **Entrada** BD(a)

- Leer BD(a)
  - ■ Aplicar algoritmo **LOF**; agregar atributo `'valor_LOF'` a cada tupla y gravar resultados según **reglas** de determinación de outliers **LOF-DBSCAN**.
  - ■ Aplicar algoritmo **DBSCAN**; agregar atributo `'valor_DBSCAN'` a cada tupla y gravar resultados según **reglas** de determinación de outliers **LOF-DBSCAN**.
- Unir resultados; agregar atributos `'tipo_outlier'` a cada tupla; gravar resultados según unión de resultados de algoritmos **LOF- DBSCAN**.
- Leer BD(a) con atributo `'tipo_outlier'`
  - ■ Aplicar algoritmo **C4.5**; determinar valor de atributo clase `'tipo_outlier'` para **C4.5**; gravar resultados.
  - ■ Aplicar algoritmo **RB**; determinar valor de atributo clase `'tipo_outlier'` para **RB**; gravar resultados.
  - ■ Aplicar algoritmo **PRISM**; determinar valor de atributo clase `'tipo_outlier'` para **PRISM**; gravar resultados.
- Unir resultados
  - ■ Aplicar reglas para la detección de outliers de algoritmos de clasificación.
  - ■ Gravar resultados en cada tupla de atributo clase `'tipo_outlier'` con valor `'limpio'` o `'outlier'`
  - .■ Crear BD(b), t.q. por cada valor cuyo valor sea `'tipo_outlier'` = `'outlier'`.
  - ■ Clusterizar la primera columna de BD(b), t.q. `'tipo_outlier'` = `'outlier'` con **K-Means** asumiendo **K** = 2
  - ■ Calcular diferencia entre los centroides de los clusters conformados; el cluster más alejado del centroide contiene los campos considerados anómalos.
  - ■ Repetir el procedimiento para cada columna de BD(b) que contenga un valor de `'tipo_outlier'` = `'outlier'`.

❍ **Salida**, BD(a) con campos anómalos detectados.

3.4 METODOLÓGICA COMBINATORIA DE LOS PROCEDIMIENTOS HÍBRIDOS PROPUESTOS

A partir de los procedimientos descriptos en la Sección 4.3, en la presente se despliega la combinación propuesta de ambos procedimientos híbridos para la detección de datos anómalos en BBDD alfanuméricas; en la cual, partiendo de una BD de DDJJ inicial se ejecutan los mismos. En la siguiente figura emerge el meta procedimiento final como compuesto de ambos procedimientos descriptos en la Sección 4.3, a partir de una BBDD de DDJJ como entrada.

Figura 11: Combinación de procedimientos híbridos propuesto para la detección de campos anómalos en BBDD de DDJJ.

En seudo código, a partir de la Figura 11 se desprenden las siguientes etapas para la combinación de los procedimientos híbridos propuestos.

- **Entrada** BD(ddjj)
  - Leer BD(ddjj)
    - Aplicar **procedimiento alfanumérico III**.
      - En caso que se encuentren outliers se evalúa la calidad de la BBDD y se exploran indicios de corrupción.
      - En caso contrario se lee la BD(ddjj) nuevamente y se aplica el **procedimiento alfanumérico IV**.
  - Leer BD(ddjj)
    - Aplicar **procedimiento alfanumérico IV**.
      - En caso que se encuentren outliers se evalúa la calidad de la BBDD y se exploran indicios de corrupción.
      - En caso contrario se lee la BD(ddjj) nuevamente y se aplica el **procedimiento alfanumérico III**.
- **Salida**, calidad de la BD(ddjj) evaluada, con o sin indicios de corrupción detectados.

El meta procedimiento propuesto en la Figura 11, supone aplicar los procedimientos híbridos alfanuméricos de detección de anomalías descripto *ut supra* sobre las BBDD preparadas de DDJJ para la detección de outliers; de esta manera, la posibilidad de detección de falsos positivos propone la retroalimentación tanto de uno como otro procedimiento alfanumérico propuesto. De esta forma, como se pretende hechar luz en el siguiente Capítulo aplicado, se busca evaluar la calidad de los datos públicos implicados en el análisis a primera luz; para en un análisis más profundo *a posteriori* en base a los campos y atributos detectados, soslayar indicios de comportamiento corrupto implícito, existentes como output de la información pública procesada.

4. APROXIMACIÓN A UN CASO DE CONTRALOR CIVIL

En este Capítulo se valida el modelo de solución y metodología propuesta mediante la evaluación de DDJJ desde el que podría ser el rol de un ciudadano o una organización civil buscando evaluar información disponible de la administración pública y/o de sus representantes políticos. En la sección subsiguiente se describe los materiales y datos a ser utilizados, la preparación de los mismos. La metodología a ser aplicada es la combinación de los procedimientos híbridos de detección de datos anómalos y ruido III y IV descriptos en Capítulo anterior[5]. A continuación se describen las características de los algoritmo injerentes en la validación.

4.1. MATERIALES Y DATOS

Para la validación metodológica se emplea las declaraciones juradas de funcionario públicos argentinos disponibles a través de la oficina anticorrupción así como también publicadas en los organismo civiles Directorio Legislativo, Poder Ciudadano y la Asociación Civil por la Igualdad y la Justicia (ACIJ) y el DLND con su plataforma de acceso Declaraciones Juradas Abiertas (DD.JJ. Abiertas, 2015), con última actualización en el sistema al 14 de abril de 2015.

De un total de 1550 DDJJ totales del sitio interactivo, 539 DDJJ son correspondientes a 99 funcionarios del poder ejecutivo, 843 DDJJ correspondientes a 313 funcionarios del poder legislativo, y 168 DDJJ correspondientes a 87 funcionarios del poder judicial; disponibles en el sitio interactivo actualizados a su última versión del vez al 14 de abril de 2015. Estas comprenden la exposición de DDJJ de funcionarios públicos Argentinos de mayores cargos en los tres poderes de la República, como se desprende del siguiente cuadro de la Tabla 6.

[5] Se destaca que la metodología planteada no corresponde a la aplicación de una metodología formal de minería, como podría ser CRISP-DM -ver Sección 5.2-, sino a la combinación de procesos híbridos para la detección de campos anómalos y ruido.

Poder Ejecutivo	Poder Legislativo	Poder Judicial
539 DDJJ 99 FFPP	843 DDJJ 313 FFPP	168 DDJJ 87 FFPP

Tabla 6: BD de DDJJ por poder republicano.

Las BBDD de DDJJ abiertas presentan la siguiente estructura en cuanto a sus atributos para cada declaración jurada de funcionario público.

```
dj.funcionario =  (ddjj_id, ano, tipo_ddjj,
                   url_document_cloud, poder,
                   persona_dni, persona_id, nombre,
                   nacimento, egreso, ingreso, cargo,
                   jurisdiccion, barrio, cant_acciones,
                   descripcion_del_bien, destino,
                   entidad, fecha_desde_cargo,
                   fecha_hasta_cargo, localidad, modelo,
                   nombre_bien_s, origen, pais, periodo,
                   porcentaje, provincia, ramo,
                   superficie, unidad_medida_id,
                   tipo_bien_s, titular_dominio,
                   moneda_mejoras, mejoras,
                   moneda_valor_adq, valor_adq,
                   moneda_valor_fiscal, valor_fiscal,
                   u_medida, vinculo)
```

Presentando un total de 41 atributos de los cuales se detectaron atributos que presentaban campos vacíos (nacimiento, egreso, periodo) t. q. `dj.funcionario(Ø) = (nacimiento, egreso, periodo)` mientras que por razones de redundancia -como se explica más adelante en la preparación de datos- se descartan los siguientes atributos de `dj.funcionario(E)=(url_document_cloud, u-medida)` al estar este ya representado por la

variables, descartándose los datos del sitio en la red, por ser irrelevante para el análisis pertinente así como uno de los dos atributos se encontraba repetido y por ende redundante.

Asimismo, la incompletitud de varios atributos se distinguen por su alta proporción de valores missings en la forma de gran cantidad de tuplas con valores no declarados donde, en todos los casos estos superaban el 70% de las observaciones totales juradas -como se aprecia en algunos de los atributos en negrita en el último cuadro-, los atributos x y z son un reflejo de ello donde más de la mitad de los items no presentan datos de ningún tipo. `dj.funcionario (E)=(barrio, localidad, provincia, ramo, fecha_desde_cargo, fecha_hasta_cargo)`. El resto de los atributos resaltados en negrita fueron preservados para generar los nuevos valores patrimoniales totales tanto en superficie como en valor monetario homogéneo los que luego del cálculo fueron descartados por poseer información redundante, altos niveles de incompletitud e inconsistencia como se explaya inmediatamente a continuación en la siguiente subsección.

Por otra parte, es necesario destacar que si bien se trata de información de funcionarios públicos de primera línea, como buenas prácticas el atributo `'dni_persona'` es descartado por principios de integridad y privacidad de los datos públicos.

4.1.1 PREPARACIÓN Y TRATAMIENTO DE LOS DATOS

La enorme variedad de bienes, servicios y sus características en una BD tan extensa sin una codificación y estandarización de tantos items, nos obligo a enfocar nuestro análisis solamente en los bienes inmuebles, con lo cual preparamos los datos de la siguiente manera.

Uno de los problemas a los que se enfrentó es la de la preparación de las bases de datos disponible en el sitio on-line de DLND. El sitio DNLD ofrece opciones para la descargar de DDJJ disponibles ya sea individualmente por funcionario público individualizado o para cada poder republicano separadamente[6], ofrece una opción que a la fecha, este ofrecía la descarga de las DDJJ por poder republicano o por funcionario público individualizado, aún así, la opción de

[6] Hay que mencionar, que en las primeras fases del proyecto, al momento de las primeras recolecciones de datos a lo largo de todo 2014, la opción de descarga completa de DDJJ del poder legislativo se encontraba inoperativo lo que obligaba a armar la data agregada individualmente lo cual resultaba muy engorroso.

descarga de datos para el poder legislativo -el poder más profuso de los tres- aunque ofrecida por DLND no parecía funcionar correctamente.

Puesto que al momento de la elaboración de este trabajo no se encontraba disponible la totalidad de las bases legislativas, confeccionar una base para todas las DDJJ de ese poder, a partir de bases individuales, implicó una ardua tarea. Por otra parte eran prolíficos los errores de tipeo en los atributos tanto en las versiones anteriores como a la presente de abril de 2014, *e.g.*: celdas vacías o incompletas, el atributo '`poder`' en todas las DDJJ de individuos del poder legislativo estaban cargadas como `'ejecutivo'`.

Otro inconveniente a la hora de trabajar con las bases de DDJJ de DLND fue que los encargados de realizar la transcripción de datos desde las DDJJ crudas de la OA a formato digital incorporaron sin necesidad de carácteres y glifos especiales en apellidos, nombres y entidades lo cual presenta inconvenientes con la decodificación de caracteres a texto plano al usar ASCII. En la misma línea, la falta de codificación de los atributos -al haber inconsistencias en la caracterización y denominación de algunos atributos- dificultó el tratamiento de los mismos al utilizar procesadores de texto plano y su consecuente perdida de tiempo de procesamiento humano.

Atributos generados

La multiplicidad de valoración monetaria -expresada en Pesos Ley, Australes, Pesos y Monedas extranjeras- y de superficie -hectáreas y metros cuadrados- con que se encuentran desplegados los datos en DLND presentaron otro problema a la hora de utilizar una misma escala valorativa para la búsqueda de valores outliers. La necesidad de una homogeneización de tales valores exigió la generación de nuevos atributos que aglutine en uno solo los valores monetarios y de superficie. Siendo así se procedió a generar los siguientes atributos:

```
dj.patrimoniales (Gen)=(superficiem2, valor_patrim, val_decl)
```

De esta manera, con los dos primeros atributos generados se resume todos los valores de patrimoniales expresados en superficie homogeneizados en metros cuadrados -`superficie2`- así como el valor patrimonial total actualizado expresado solamente en pesos argentinos -`valor_patrim`[7]-. Una vez ya contando con un valor patrimonial homogéneo se procede a la generación del atributo polinómico 'val_decl' el cual cataloga el valor declarado de los bienes inmuebles del oficial político según su valor comparativo entre su valor fiscal y el precio declarado del bien al momento de adquisición si es que lo tiene. Este atributo que resulta central para la validación de uno de los procedimientos.

Como se verá más adelante el atributo '`val_decl`' contempla las siguientes categorías descriptas a continuación en la Tabla 7:

Atributo generado Valor declarado (`val_decl`)
Valor de Mercado
Valor Fiscal
Valor Subfiscal
No Declara
Sin Datos

Tabla 7: Categorías del atributo generado: '`val_ decl`'.

De acuerdo al valor declarado de un bien determinado este atributo se elabora, a partir de la diferencia de valor existente entre el valor de adquisición de un bien determinado y su valor fiscal. Siendo así el atributo permite distinguir al oficial político que en su DJ opta por declarar solamente el valor fiscal del mismo solamente y no el de mercado representado por el valor de

[7] Para la conformación de este atributo se actualizaron valores monetarios teniendo en cuenta la inflación acumulada desde la finalización de los programas monetarios de Peso Ley, y Austral (Rapoport, 2010) así como se convirtieron aquellos valores en moneda extranjera correspondientes a los de cotización al 14-4-2015 fecha de actualización de la BD en DLND.

adquisición del bien generalmente superior al valor fiscal. El atributo también incluye los casos en cuanto el valor de adquisición es inferior al valor fiscal así como el de no declaración de valor patrimonial alguno.

La BD de DDJJ de inmuebles de preparada finalmente para su tratamiento en la búsqueda de valores outliers finalmente quedara configurada de la siguiente manera de acuerdo a los atributos pertinentes.

```
dj.funcionario =  (ddjj_id, ano, tipo_ddjj, poder,
                   persona_id, nombre, ingreso, cargo,
                   jurisdiccion, cant_acciones,
                   descripcion_del_bien, destino, localidad,
                   nombre_bien_s, origen, pais, porcentaje,
                   provincia, tipo_bien_s, titular_dominio,
                   vinculo, superficiem2, val_decl,
                   valor_patrim)
```

Consecuentemente como se desprende del cuadro anterior, la BD preparada para su tratamiento en la validación de los procedimientos cuenta ahora con 24 atributos y 6627 tuplas, como se describe en la Tabla 8.

Características	Cantidad
Cantidad de tuplas	6627
Cantidad de atributos	24

Tabla 8: Características generales de la BD de DDJJ preparada.

En donde cada tupla corresponde a un item, bien o servicio determinado en poseción patrimonial declarada por el oficial político.

4.1.2 ENTORNO TECNOLÓGICO DE LA EXPERIMENTACIÓN

El entorno tecnológico en el cual se desarrolló la experimentación comprendió la utilización de RM en su distribución 5.3.015 para Windows, una planilla de cálculo en formato `csv` contenedora de las DDJJ de inmuebles preparada con un tamaño en disco de 1916KB.

El equipo de experimentación se trató de un computador personal Lenovo, con micro AMD C-60 APU Radeon(tm) de 1.00GHz. y 3.60 GB de RAM utilizable.

4.2. ALGORITMOS UTILIZADOS

Los algoritmos de aplicación utilizados corresponden a las técnicas de minería pertenecientes a los procesos de detección de outliers y ruido en BBDD expuestos superficialmente y de forma general anteriormente en la Sección 4.2.

Aunque existen una gran variedad de algoritmos que utilizan a la minería de datos para la obtención de conocimiento sumamente útil para las bases de datos (Kuna et al., 2010a), en los procesos sugeridos como solución para la detección de campos anómalos se utilizarán los siguientes algoritmos de clasificación, agrupamiento y teoría informativa.

4.2.1 ALGORITMOS DE CLASIFICACIÓN

Los algoritmos de clasificación a ser utilizados en ambos procedimientos serán los algoritmos C4.5 (Procedimiento III, Procedimiento IV) , RB (Procedimiento IV) y PRISM (Procedimiento IV), como se detalla a continuación.

C4.5

Algoritmo de inducción para la detección de campos significativos en forma recursiva a través de particiones del tipo *dephfirst* (primero en profundidad) de manera de obtener la mayor ganancia de información posible (Quinlan, 1993), de la familia de algoritmos TDIDT capaces de realizar tareas de clasificación. El algoritmo actúa consecutivamente buscando los ejemplos de una BD dada que posea la mayor ganancia de información posible en cada partición de la misma.

Al considerar atributos discretos el algoritmo toma el número de valores posibles que puede tomar el atributo. La aplicación del algoritmo C4.5 permite reducir el espacio de búsqueda para la detección de datos anómalos dentro de la BD a solo aquellos campos que son relevantes en el set de datos considerado *i.e.* aquellos atributos que en la BD aportan más información para clasificar al atributo clase u objetivo y ergo optimizar la performance.

El seudo-código del algoritmo puede describirse de la siguiente manera:

Entrada

Si **S** = vacío

Se devuelve un único nodo con valor falla:

Si todos los registros de **S** tienen el mismo valor para **C**

Devolver un único valor con el valor más frecuente de **C** en los registros de **S**:

Si **R** = Vacío

D ← atributo con mayor proporción de ganancias informativa (**D**;**S**) entre los atributos de **R**;
Siendo $\{d_j | j=1,2, \ldots, m\}$ los valores del atributo **D**;
Siendo $\{d_j | j=1,2, \ldots, m\}$ los subconjuntos de **S** correspondientes a los valores de d_j respectivamente;

Devolver árbol con la raíz nombrada como **D** y con los arcos nombrados que van respectivamente a los árboles. `C4.5 (R – {D}, C, SI), C4.5 (R-{D}, C, S2), C4.5 (R-{D}, C, Sm)`

Salida

Para dividir los datos el sistema debe decidir entre tres pruebas posibles a ejecutar en cada nodo:

[i] Prueba estándar: para variables discretas, se ejecuta con un resultado y una rama para cada valor posible de la variable.

[ii] Prueba discreta: también para variables discretas donde los valores posibles se asignan a un número variable de grupos con un resultado posible para cada uno de ellos, en lugar de asignarlos para cada valor.

[iii] Prueba binaria: para valores continuos se realiza una prueba binaria suponiendo una variable A determinada, se defina *a priori* el valor límite de un umbral Z, t. q. A <= Z y A > Z.

RB

Grafo acíclico dirigido, basado en un clasificador bayesiano *naive*, el cual proporciona una técnica que a través de un modelo probabilístico permite la representación de las relaciones existentes dentro de los diferentes campos y sus ponderaciones respectivas en una BD determinada (Jensen, 1996).

En todo grafo acíclico dirigido posee nodos los cuales representan a variables aleatorias y sus aristas las influencias causales entre las variables que pueden ser continuas o discretas. Si un nodo es padre de otro nodo significa que es causa directa del segundo. El modelo de grafos posee un funcionamiento conveniente para representar conocimiento en contextos donde existen contextos con altos grado de incertidumbre, así como la representación gráfica de las dependencias e interdependencias de las variables que conforman el dominio que permite abstraerse a un modelo casual cualitativo (Pearl, 1988).

Formalmente, una red bayesiana es una tupla $B = (G, \Theta)$, donde G representa el grafo acíclico dirigido y Θ al conjunto de distribución de probabilidades $P(X_i | Pa(X_i))$ para cada una de las variables i t.q. $i = (1, \ldots, n)$, dado un gafo G, $Pa(X_i)$ son todos los padres para cada la variable X_i.

Hay dos maneras posibles de representar una red bayesiana:

[i]: Como una base de reglas donde cada arista representa un conjunto de reglas que permite asociar a variables o atributos involucrados en el análisis, donde las probabilidades hacen de cuantificadores a dichas reglas.

[ii]: Como una distribución de probabilidad conjunta de las variables representadas en la red bayesiana.

Una distribución de probabilidad puede representarse de la siguiente manera.

$$I = (X, Y|Z) \Leftrightarrow P(X|YZ) = P(X|Z)$$

La ecuación anterior abstrae un modelo de dependencias y relaciones condicionadas donde `X`, `Y`, `Z` son subconjuntos de variables, `I(X,Y|Z)` plasma una relación de independencia condicional; por otra parte, la probabilidad conjunta de n variables se encuentra especificada por el producto de las probabilidades dado a cada padre, como se expresa de la siguiente manera.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} (PX_i|Pa(X_i))$$

Buscando ahora obtener la forma característica de árbol que más se aproxime a la distribución real, lo que para lograrlo se utiliza una medida de la diferencia de información `I(•)` entre la

distribución real (P) y la distribución aproximada (P*), formalmente representada de la siguiente manera.

$$I(P, P^*) = \sum_X P(x) \log \left(\frac{P(X)}{P^*(X)} \right)$$

Resumiendo el problema en uno de optimización de la diferencia de información, se busca minimizar el valor de `I(•)` considerando la información mutua entre pares de variables tal que la minimización de `I(•)` pueda expresarse como sigue.

$$I(X_i, X_j) = \sum_X P(X_i, X_j) \log \left(\frac{P(X_i, X_j)}{P(X_i)P(X_j)} \right)$$

Siendo que la diferencia de información es una función negativa de la suma de las ponderaciones de todos los pares de variables que constituyen un árbol (Chow & Liu, 1968); ergo, encontrar el árbol más próximo equivale a encontrar el árbol con mayor peso.

A partir de los postulados teóricos para la conformación de una RB el algoritmo original para la determinación la red bayesiana óptima tendrá el siguiente pseudo-código.

I. Calcular la información mutua entre todos los pares de variables t.q. `n(n-1)/(2-pares)`

II. Ordenar la informaciones mutuas de mayor a menor.

III. Seleccionar la rama de mayor valor como árbol inicial

IV. Agregar la siguiente rama mientras no forme ciclo, si lo forma, desecharlo.

V. Repetir consecutivamente el paso 4. hasta que se cubran todos los atributos del set de datos `(n-1)` ramas

PRISM

Comprende un algoritmo de aprendizaje de extracción de reglas, el cual asume como dado que el set de datos a ser tratado no posee ruido, lo que resulta en una ventaja al crear reglas que cubren una gran parte de los elementos de las BB.DD. aislando instancias para un análisis discriminado (Cendrowska, 1987).

El algoritmo tiene el siguiente de seudo-código.

Para cada clase **C**

Sea **E** = ejemplo de entrenamiento
Mientras **E** tenga ejemplos de la clase **C**
Crea una regla **R** con **LHS** vacío y clase **C**
Until **R** es perfecta do
Para cada atributo **A** no incluido en **R** y cada valor v,
Considera añadir la condición **A** = **v** al **LHS** de **R**
Selecciona el par **A** = **v** que maximice p/t
Se agrega **A** = **v** a **R**
Elimina de **E** los ejemplos cubiertos por **R**

Con esta estructura, el algoritmo funciona de la siguiente manera:

[i] Siendo "*t*" el numero de ejemplos que están cubiertos por la regla y sea 'p' el número de ejemplos positivos que cubre la regla.

[ii] El algoritmo PRISM agrega condiciones a reglas que maximicen la relación 'p/t'.

[iii] Como se van eliminando los ejemplos que va cubriendo cada regla, las reglas que se construyen deben interpretarse en orden.

[iv] Las reglas que dependen del orden de su interpretación se conocen como listas de decisión.

[v] Con varias clasificaciones es posible seleccionar la regla que cubra más ejemplos, y en el caso que no se tenga una clasificación se escoge la clase mayoritaria.

4.2.2 ALGORITMOS ESPECIALIZADOS

Los algoritmos especializados en la detección de datos anómalos a ser utilizados en serán el algoritmos LOF (Procedimiento III y Procedimiento IV) y el algoritmo DBSCAN (Procedimiento IV), como se detalla a continuación.

LOF

Constituye un algoritmo especialmente diseñado para la detección de outliers el cual considera la densidad de los datos para determinar un factor local de outliers determinando en que medida cada tupla es anómala sin necesidad de contar con un atributo clase u objetivo (Breuning et al., 2000).

> **Definición 1** Un objeto p en una BD 'D' es un outlier si el cardinal del conjunto t.q. {q ∈ D|d(p,q) ≤ dmin} ≤ {100-pct}% del set de datos 'D', siendo d la distancia y p, q los elementos de la BD 'D' se encuentran a una distancia menor que dmin de p. Lo anterior es equivalente a la siguiente expresión: DB(pct, dmin)-Outlier).

Definición 2 Para todo entero positivo k, la k-distancia del objeto p, denominada k-distancia(p) se define como la distancia d(p,o) entre p y un objeto $o \in D$ t.q.:

[i] Para al menos k objetos $o' \in D|\{p\}$ se cumple $d(p,o') \leq d(p,o)$

[ii] Para a lo sumo $k-1$ objetos $o' \in D|\{p\}$ se cumple $d(p,o') \leq d(p,o)$

Definición 3 Dada una k-distancia de p *i.e.* la k-distancia del vecindario de p contiene todos los elementos y objetos cuya distancia de p no es mayor que la k-distancia. e.*g.* Nk-distancia(p)(p) = $\{q \in D|\{p\}$ t.q. $d(p,q) \leq$ k-distancia(p)$\}$, estos objetos q se denominan como k-vecinos de p, definiendose así una noción del grado de vecindad.

Definición 4 La distancia de accesibilidad de un objeto p respecto de un objeto o, siendo k un número natural, la distancia de accesibilidad de respecto al objeto o se define como $\text{reach-dist}_k(p,o) = \max\{$k-distancia(o), $d(p,o)\}$.

Dos parámetros definen la noción de densidad:

[i] el parámetro MinPts que especifica el número de objetos;

[ii] mientras que el parámetro $\text{reach-dist}_{\text{MinPts}}(p,o)$ se considera una medida del volumen para la determinación de la densidad en la vecindad de un objeto p.

Para su operación debe mantenerse MinPts como único parámetro y se usan los valores t. q. $\text{reach-dist}_{\text{MinPts}}(p,o)$ t. q. $o \in N_{\text{minPts}}(p)$. Los dos, determinan el umbral de densidad utilizados por los algoritmos de agrupamiento, respecto este umbral, los elementos de la BD se encuentran conectados -o disociados- del resto si sus elementos superan -o no- un umbral de densidad dado. En cada caso

para detectar los valores atípicos basados en densidad es necesario comparar las densidades de los distintos conjuntos de objetos, lo que conlleva a la determinación dinámica del conjunto de objetos dentro de la BD.

Definición 5 El concepto de densidad de accesibilidad local de p, LRD (p) se define de acuerdo a la siguiente ecuación.

$$\mathrm{LRD}_{\mathrm{MinPts}}(p) = 1/\left(\frac{\sum_{o \in N_{\mathrm{MinPts}}(p)} \text{reach-dist}_{\mathrm{MinPts}}(p,o)}{|N_{\mathrm{MinPts}(p)}|}\right)$$

La ecuación anterior representa el valor de la densidad local del objeto p, LRD(p) $\in [0,\infty]$ obtenida a partir de la inversa de la distancia media del parámetro $\text{reach-dist}_{\mathrm{MinPts}}(p,o)$ sobre la base de los vecinos más cercanos MinPts-p por otra parte. Esta a su vez juega un central para la determinación del factor local de outlier como se despliega a continuación.

Definición 6 El factor local de outlier de p, LOF(p), se define ecuacionalmente de la siguiente forma.

$$\mathrm{LOF}_{\mathrm{MinPts}}(p) = \frac{\sum_{o \in N_{\mathrm{MinPts}}(p)} \mathrm{LRD}_{\mathrm{MinPts}}(o)/\mathrm{LRD}_{\mathrm{MinPts}}(p)}{|N_{\mathrm{MinPts}(p)}|}$$

Donde el valor de LOF de un objeto p, LOF(p), representa el grado en que se considera al elemento p como uno atípico, es la medida de la relación de p con los p vecinos más cercanos del parámetro MinPts. Como se desprende de la ecuación cuando menor (mayor) sea LRD(p) y mayor (menor) sea LRD(o), la accesibilidad local de los vecinos más cercanos, de los valores de p de MinPts mayor (menor) será el valor del LOF(p).

DBSCAN

Sin necesidad de contar con un atributo clase u objetivo, el algoritmo DBSCAN aplica principios análogos a los utilizados por LOF (Saad & Hewahi, 2009), agrupando filas definidas como outliers dentro de un cluster discriminándolas del resto de las filas de la BD.

Este constituye el primer algoritmo especialmente concebido para la detección de outliers (Ester et al., 1996). Los algoritmos basados en densidad como DBSCAN localizan regiones de alta concentración de puntos, los cuales se encuentran separados entre si por zonas de menor densidad. Es un algoritmo relativamente fácil de implementar donde la densidad de los puntos depende del área del radio de vecindad asumido.

La noción principal del algoritmo es la de encontrar todos los puntos centrales posibles, siendo que los puntos centrales de un grupo son aquellos que poseen un área de vecindad que contienen un número mínimo de puntos en un radio determinado. Consecuentemente para la detección de campos anómalos en este algoritmo es necesario establecer el concepto de punto central, como se describe el primero a continuación:

> **Definición punto central**: Un punto `'p'` se define como central si y solo sí este posee un área de vecindad `'N'` correspondiente a un radio `'Eps'` que excede un umbral determinado `'MinPts'`, t.q. se cumpla con la siguiente condición:
>
> $$|\mathrm{N}_{\mathrm{Eps}}(\mathrm{p})| \geq \mathrm{MinPts}$$

Como se desprende del anterior los puntos centrales poseen un área de vecindad para un determinado radio que tiene un número mínimo de puntos (`MinPts`). Tal área de vecindad se describe en la siguiente definición.

> **Definición área de vecindad**: El área de vecindad `N` de un punto `p` que pertenece a un set de datos `D` esta dado por un radio `Eps` t.q. cumpla con la siguiente expresión:
>
> $$\mathrm{N}_{\mathrm{Eps}}(p) = \{\mathrm{q} \in \mathrm{D} | \mathrm{dist}(\mathrm{p},\mathrm{q}) \leq \mathrm{Eps}\}$$

La ecuación anterior estipulan que la densidad de los puntos depende del radio del área de la vecindad especificado como limite del mismo. La morfología del área de vecindad por otro lado estará determinada por la elección de la medida de distancia entre dos puntos, *e.g.*: si se decide utilizar la distancia rectilínea entre dos puntos DBSCAN tenderá a crear grupos de forma rectangular.

Los puntos que no son centrales, *i.e.*: los que quedan fuera de los grupos formados a partir de los puntos centrales, se llaman puntos ruido, en cambio los puntos que no son ni ruido ni centrales se dice que son puntos borde.

Operacionalmente, el algoritmo comienza seleccionando un punto `p` arbitrario, si es que `p` cumple con la definición y es un punto central, se inicia la construcción de un grupo y se incorporan en dicho grupo todos los objetos denso-alcanzables a partir del centro `p`. De otra manera si `p` no es central, *i.e.*: si fuesen puntos ruido y puntos borde, se lee otro elemento del set de datos y así consecutivamente hasta que la totalidad de los objetos de la BD sean procesados.

Cada grupo creado poseerá por ende puntos centrales o de borde, los puntos de ruido quedan fuera de los grupos generados; además, los grupos creados por el algoritmo pueden tener a su vez más de un punto central y compartir puntos de borde en varios grupos.

4.2.3 OTROS ALGORITMOS: TEORÍA DE LA INFORMACIÓN Y DE AGRUPAMIENTO

Adicionalmente se utilizará un algoritmo basado en la teoría de la información (Procedimiento III) y el algoritmo de agrupamiento K-Means (Procedimiento IV), como se detalla a continuación en los dos siguientes algoritmos.

Teoría de la información (TI)

La teoría de la información busca determinar la cantidad de información promedio que contienen los símbolos usados (Shannon, 2011), para su medición se utiliza mismo concepto de entropía

(H) aplicado en ciencias naturales para medir el desorden de un sistema, el mismo que adaptado a teoría de la información adquiere la siguiente forma.

$$\mathrm{H} = \sum_{k=1}^{m} \mathrm{p}_k \log \frac{1}{\mathrm{p}_k}$$

Entendida ahora la ecuación de la entropía con respecto a la teoría de la información, cuanto menor probabilidad de aparición de un símbolo haya mayor será la cantidad de información que este aportaría. Siendo la entropía el valor de una esperanza de probabilidad *a priori*, cuanto menor sea la probabilidad de un símbolo determinado en el sistema informativo mayor será la entropía que este genere perturbando el sistema al momento de su aparición.

Respecto la detección de valores anómalos, la teoría permite, por medio de la medición de la entropía, encontrar ruido presente en los mismos, puesto que la TI proporciona una medida del grado de las características extraídas de cada elemento en relación a la clase que pertenece.

Ferreyra, otro autor interesado en la aplicación de la TI (Ferreyra, 2007), propone la posibilidad de tratar los datos en binomios de entrada (E) y salida (S) de manera de detectar los outliers en cada atributo como ruido en el mensaje trasmitido, *i. e.* una mayor entropía del mensaje; siendo E el mensaje emitido y S el recibido, procedimentalmente si el atributo de entrada E presenta una baja densidad respecto del de atributo de clase de salida S, el primero presenta una alta probabilidad de que fuera considerado como elemento anómalo.

K-Means

El algoritmo K-Means es un algoritmo de clusterización ampliamente utilizado por su simpleza y su eficacia. Creado a partir de un trabajo de MacQueen (MacQueen, 1967) permite clasificar un conjunto de objetos en un número K de clusters, donde K es un número determinado *a priori*. Cada cluster es representado por una media ponderada la cual viene a ubicar a su centroide, el cual se encuentra efectivamente en el medio de los elementos que componen el cluster. Luego de la ubicación de todos los centroides, el algoritmo ubica al resto de los puntos en la clase de su centroide más cercana para *a posteriori* recalcular los centroides reubicando cada uno de los puntos en cada conglomerado, proceso de recálculo consecutivo que finaliza en la iteración que no produzca más cambios en la distribución de los puntos respecto su anterior inmediata.

La operatividad del algoritmo K-Means se resume en cuatro pasos, siendo O un conjunto de objetos t.q. $Dn = (x_1, x_2, ...x_n)$ para todo i, son los elementos de un cluster y con $x_i \in R$ y k, v1, los centro de los clusters, se tiene que:

Paso I: Determinación del valor de **K** y una forma aleatoria inicial de **K** objetos para los primeros clusters. Para cada cluster **K** el valor inicial del centroide será x_i los cuales son los únicos objetos de Dn que pertenecen al cluster.

Paso II: Reasignar los objetos del cluster según una medida de distancia para cada elemento x asignándosele el que está más próximo al objeto conglomerado.

Paso III: Recálculo de los centroides de cada cluster una vez que todos los objetos asignados.

Paso IV: Repetición de los pasos II y III hasta que no se produzcan más reasignaciones.

El algoritmo K-Means presenta por otra parte las siguientes tres desventajas:

[i] Solo puede aplicarse con atributos numéricos ya que es necesario calcular el punto medio.

[ii] No conocer *a priori* el valor de K puede hacerlo poco eficaz, aunque existen varias métricas capaces de validar el valor de K.

[iii] El algoritmo es sensible a los valores anómalos, por lo tanto los considera como valores inliers.

Descriptos todos los pasos técnicos para la usanza de los procedimientos alfanuméricos para la detección de datos anómalos se procede inmediatamente a continuación la aplicación empírica contando con un atributo clase u objetivo.

4.3 EXPERIMENTACIÓN CON EL PROCEDIMIENTO DE DETECCIÓN DE OUTLIERS CON ATRIBUTO TARGET (PROCEDIMIENTO III)

La búsqueda de datos anómalos asumiendo una aproximación heurística que contemple la utilización de un atributo objetivo target constituye la aplicación empírica descripta en el Capítulo 3.3.1., -así como también describimos en el Anexo su aplicación práctica del flujo de minería en RM[8]-.

El procedimiento III al ser híbrido toma múltiples enfoques de minería para la detección de campos anómalos aborda en principio un enfoque de Tipo 2 al usar el algoritmo C4.5, un algoritmo de clasificación con aprendizaje supervisado no automático, siendo que el C4.5 débese configurarse manualmente respecto un atributo target y calibrarse el algoritmo para obtener la mayor ganancia informativa posible.

El procedimiento III tiene un enfoque de Tipo 1 de aprendizaje no supervisado al apoyar sus flujos de minería para la detección de outliers en TI y un algoritmo LOF de agrupamiento; resumidamente el enfoque híbrido del procedimiento III posee la siguiente configuración de tipos:

Tipo II → Tipo I

La implementación y los resultados de la sub-fase no-automática puede describirse a continuación donde se selecciona el atributo '`val_decl`' como atributo clase, a partir del cual se obtiene el siguiente árbol de inducción en la Figura 12, resultado de la configuración operativa en RM de la Figura 16, de la Sección B.1 del Anexo.

[8] Para estudiar el flujo de minería en RM para el Procedimiento III con mayor detalle, se ofrece al lector revisar el apartado Anexo B.1, donde de describe los flujos de minería con pormenor especificación operativa.

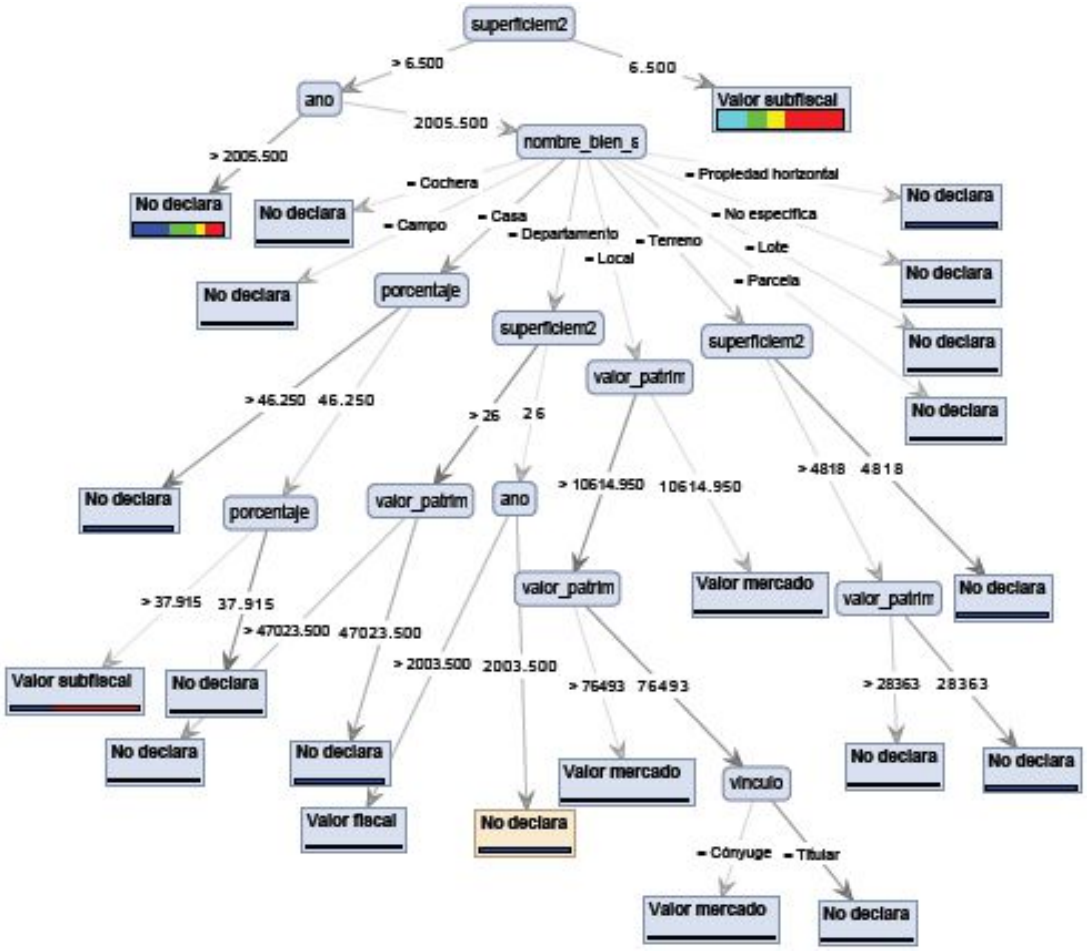


Figura 12: Árbol de inducción sobre la BD preparada con atributo target (`val_decl`).

Del presente árbol de inducción de la Figura 12, se identifican los atributos `'superficiem2'`, `'ano'`, `'nombre_bien_s'`, `'porcentaje'`, `'valor_patrim'` y `'vinculo'` como significativas en la ganancia informativa para la explicación del valor declarado en DDJJ.

A primera vista, una de las primeras observaciones que se hacen es que al declararse propiedades superiores a los 6500m^2 estas tienden a corresponder mayormente a un valor declarado subfiscal mientras que cuando los bienes declarados son menores a esa superficie ($\leq$ 6500m^2) se tiende

directamente a no declara el valor del inmueble si la DJ corresponde a partir de mediados de 2005.

De las DDJJ juradas anteriores o correspondientes a la mitad de 2005 la cualidad del valor declarado de los bienes en DDJJ de funcionarios oficiales dependerá de su tipo nominativo ('nombre bien s'), donde se tendió a no declarar su valor al tratarse de valores cochera, campo, propiedad horizontal, lote, parcela y en casos no especificados.

En este mismo rango, la superficie en metros cuadrados sigue teniendo significancia para el caso de departamentos y terrenos, el valor patrimonial declarado para locales y la proporción porcentual propietaria del bien en cuestión (porcentaje) en el caso de casas. En estas se tiende a declarar un valor subfiscal de los inmuebles cuando el porcentaje de co-propiedad del mismo fluctúa entre el 46,25% y el 37,92% del bien, no declarándose valor en los casos restantes.

Entre la mitad de 2003 y mitad de 2005 el valor de los departamentos tiende a ser el fiscal cuando estos poseen una superficie igual o inferior a los $26m^2$. En cuanto a los locales comerciales estos tienden a ser declarado en cuanto a su valor de mercado siempre y cuando los titulares poseen un valor patrimonial declarado superior a los 76493 o el local en cuestión se encuentre inscripto a nombre del conjugue del oficial político, en caso contrario tiéndese a no ser declarado su valor. Del análisis de inducción se desprenden los siguientes atributos significativos para explicar la declaración patrimonial de bienes inmuebles de acuerdo a su valor relativo en la siguiente Tabla 9 desplegada como sigue.

Atributos significativos
superficiem2
ano
nombre_bien_s
porcentaje
val_patrim
vinculo

Tabla 9: Atributos significativos detectados (Procedimiento III).

Ahora *a posteriori* de la primera fase no automática se pone en marcha los presupuestos de TI con atributo de entrada (E)-salida (S), y sobre este, un algoritmo de agrupamiento LOF de aprendizaje automático.

En una primera instancia, con los atributos significativos detectados con el algoritmo de clasificación general C4.5; se elaboran los "bins" simulando un sistema de la información (Ferreyra, 2007; Kuna, 2014), con los atributos significativos detectados como entrada (E) y el atributo clase considerado como salida (S). Operativamente se replica la configuración de RM de la Figura 17 para cada "bin", como también se desprende su selección de la Tabla 14 (Ver Anexo B.1).

En una segunda instancia, se ejecutan los flujos de minería para cada bin, donde se confeccionan los 6 (seis) bins (E)-(S) correspondiente a cada atributo significativo encontrado, a la vez que reemplazando los valores nulos en atributos no numéricos por la etiqueta 'nulos' como dicta el procedimiento III (Kuna, 2014), arribando finalmente a los siguientes producto por bin considerado, como se describe en la Tabla 10 a continuación.

Bins Entrada-(Salida)	**Outliers detectados**	**Bins anómalos sospechosos Media o Moda(Moda)**
superficiem2-(val_decl)	968 (∞)	38733.15-(No declara)
ano-(val_decl)	122 (∞)	2001-(Mercado)
nombre_bien_s-(val_decl)	209 (∞)	Prop. Horizontal-(Fiscal)
porcentaje-(val_decl)	252 (∞)	29.18528-(Sin datos)
val_patrim-(val_decl)	12 (∞)	146528.3-(Subfiscal)
vinculo-(val_decl)	30 (∞)	Conviviente-(Subfiscal)

Tabla 10: Bins de Entrada-Salida con outliers detectados.

Como se desprende de la Tabla 10, el algoritmo LOF calcula como infinito para cada tupla de cada bin como valor anómalo, lo que en teoría de sistemas y su aplicación empírica como procedimiento, se interpreta que cuanto menor sea la probabilidad de aparición de una tupla (E)-(S), mayor es la posibilidad que corresponda a una inconsistencia, siendo el resultado LOF la relación existentes entre el elemento de E y el de S.

De los resultados se tiene que la superficie cuadrada declarada (`superficiem2`), seguida del porcentaje accionario (`porcentaje`), nombre del bien (`nombre_bien_s`), y año de la DJ (`ano`) son los atributos que presentaron mayor cantidad de valores anómalos de entrada (E) con atributo con valor declarado (`val_decl`) (S) como atributo target.

Los atributos pertenecientes a los "bins" de la Tabla 10, constituyen consecuentemente aquellos; que en ese orden, presentan mayores inconsistencias en relación a la (des)información provista respecto el valor declarado de tales bienes.

4.4 EXPERIMENTACIÓN CON EL PROCEDIMIENTO DE DETECCIÓN DE OUTLIERS SIN ATRIBUTO TARGET (PROCEDIMIENTO IV)

Esta sección tiene como objetivo la detección de outliers en la misma BD alfanumérica de DDJJ preparada para proceder con la validación del Procedimiento IV descripto en la Sección 4.3.2. para la misma BD preparada hacia tal fin en la Sección 5.1.1.

Al igual que el procedimiento III, aunque un poco más compleja, la aproximación híbrida que supone la ejecución del procedimiento IV le corresponde la siguiente configuración de tipos de enfoques de minería:

Tipo I → Tipo II → Tipo I

BD-DDJJ(a) → BD-DDJJ(b)

(Fase I) (Fase II)

Donde *a priori* para la detección de campos anómalos considera un enfoque de Tipo I, esto es, de detección de outliers con aprendizaje no supervisado el cual no precisa el conocimiento previo de lo datos. Seguido de un enfoque más cercano al Tipo II, *i.e.*: detección de outliers con aprendizaje supervisado (algoritmos PRISM, RB, y C4.5 los cuales deben configurarse manualmente[9] ya explicados con anterioridad en la Sección 5.2); para volver nueva y finalmente a un enfoque de Tipo I, esta vez con una BD(b) elaborada para la aplicación de algoritmos de agrupamiento donde se busca determinar las distancias de los centroides de aquellos atributos candidatos de contener datos anómalos, para lo que se crea una nueva BD(b), se calcula su LOF para finalmente agrupar las distancias obtenidas mediante el algoritmo K-means.

De acuerdo a la aplicación de determinación de reglas para la detección de outliers, en una aproximación convencional de acuerdo con el umbral LOF aceptado en 1.5 y el mismo cálculo paralelo aceptando mayores o menores niveles de confianza (con umbrales[10] [1575, 1675]) lo que permite descubrir valores que puedan representar posibles falsos positivos, así como en caso contrario, asegurar la ausencia absoluta de valores anómalos.

A partir del flujo de minería correspondiente a la primera fase del Procedimiento IV cumpliendo con la primer fase procedimental, en la programación del flujo para la unión de los algoritmos de clasificación en RM (C4.5-RB-PRISM)[11] en la detección de outliers se obtuvieron los siguientes resultados:

[9] En el Anexo B se detalla con mayor pormenorización la configuración manual de estos algoritmos para lograr la mayor ganancia informativa posible.

[10] En la Tabla 15 del Anexo se exponen todos los valores de umbrales de valor LOF utilizados.

[11] En el Anexo A.2 se describe en detalle la unión de algoritmos de clasificación así como sus reglas de determinación de outliers empleados en el Anexo A.2.1 para su programación en RM.

Algoritmo	Outliers detectados
UNIÓN RB-C4.5-PRISM	2531

Tabla 11: Outliers detectados por unión de algoritmos de clasificación.

En la Tabla 11 se presentan los resultados de la aplicación de los algoritmos de clasificación, donde fueron identificados 2531 tuplas como outliers sospechosos, de un total de 6627 tuplas de bienes inmuebles, tenemos que 38,19% de las tuplas de la BD de DDJJ son candidatas a ser sospechosas de contener campos anómalos. Una nota respecto los outliers detectados, donde observamos que no se presentaron outliers dobles dado que se obtuvo solo un agrupamiento único de un solo cluster en los subprocesos anteriores a la unión algorítmica.

En la segunda fase a partir de la utilización de BD(b), luego de la unión entre los algoritmos de clasificación y el naive bayes se procede a clusterizar la base de donde se obtuvieron e identificaron dos agrupamientos. Notar que los valores obtenidos operativamente en RM pueden apreciarse en la Figura 13 en RM *ut infra,* correspondiente al output de las operaciones descriptas en el Anexo B.2.

De los dos agrupamientos detectados, se tiene la siguiente Tabla 12, donde se presentan los correspondiente valores promedios de distancia de cada grupo como sigue.

Columna outlier transpuesta (BD(b))	Valor promedio (Cluster_0)	Valor promedio (Cluster_1)
outlier (BD(b))	1.797	1079

Tabla 12: Distancia promedio de cada tupla-centroide para cada cluster.

Donde se observa una mayor distancia promedio para aquellos atributos considerados portadores de valores anómalos (`Cluster_0`) de aquellos considerados normales (`Cluster_1`). Con más detalle para cada atributo, se observa ahora las distancia de cada atributo en particular respecto cada centroide, los siguientes resultados considerando solo los outliers con valores LOF superiores a los umbral, como se despliega a continuación en la siguiente Tabla 13.

Atributo(id)	Valor distancia Cluster 0	Valor distancia Cluster 1
ddjj_id(1)	**1.841**	
ano(2)	**1.518**	
tipo_ddjj(3)		1.302
poder(4)		1.232
persona_id(5)		1.136
nombre(6)		1.063
ingreso(7)		1.062
cargo(8)		1.035
jurisdiccion(9)		1.009
cant_acciones(10)		0.937
descripcion_del_bien(11)		0.926
destino(12)		0.942
localidad(13)		1.042
nombre_bien_s(14)		0.998
origen(15)		1.054
pais(16)		1.003
porcentaje(17)		1.071
provincia(18)		1.104
tipo_bien_s(19)		1.050
titular_dominio(20)		1.076
vinculo(21)		1.038
superficiem2(22)	**2.033**	
val_decl(23)		1.095
valor_patrim(24)		1.237
outlier(25) (Transpuesto BD(b))		1.326

Tabla 13: Distancia del centroide para cada atributo.

De la Tabla 13 se desprende que los atributos más sospechosos de poseer campos anómalos son el identificador de declaración jurada (`ddjj_id`), el año de presentación de DJ (`ano`) y la superficie cuadrada en propiedad del oficial político (`superficiem2`). Por otra parte, si bien el atributo valor patrimonial (`valor_patrim`) posee un valor alto, este aún así pertenece al grupo del centroide más cercano.

Como se aprecia de los resultados arribados de la siguiente Figura 13, obtenida de la aplicación empírica tales valores corresponden al output de los flujos de minería en RM como pantalla última del flujo de minería del procedimiento IV (Ver Figura 29 del Anexo B.2), el cual se aprecia a continuación[12].

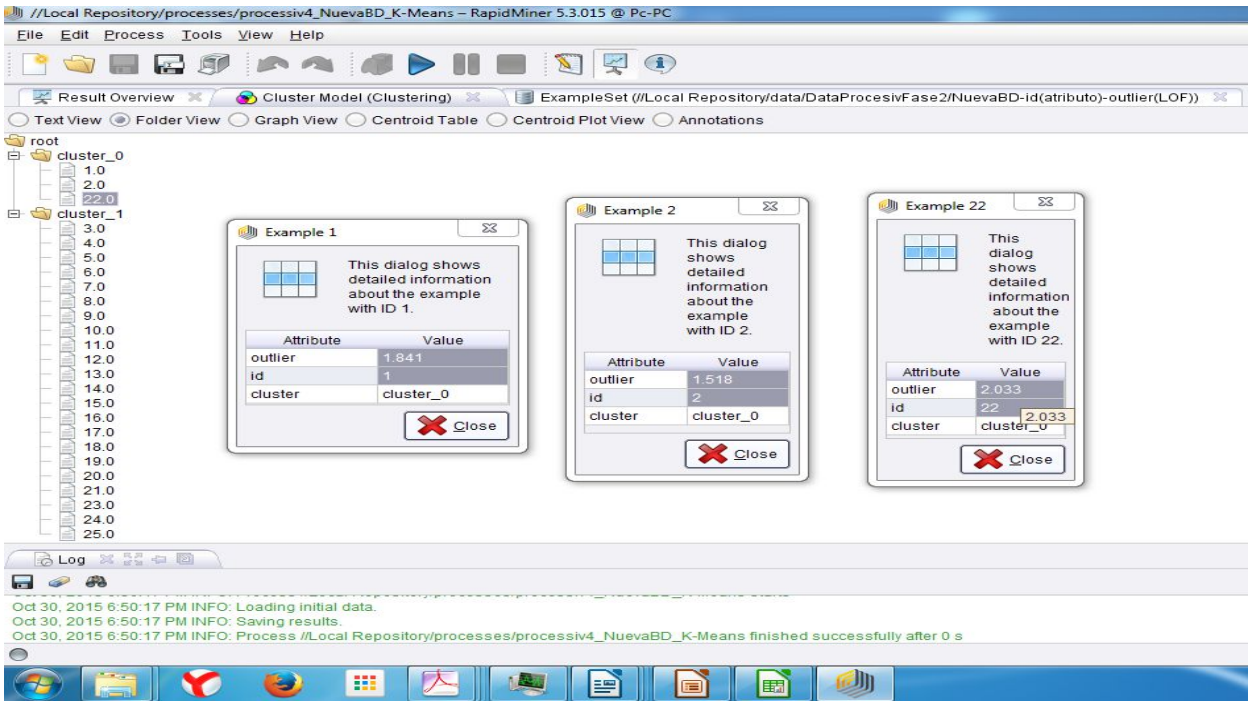


Figura 13: Clusterización de la columna transpuesta (RM)

De la Figura 13, y sus correspondientes distancias de la Tabla 12 y 13, representan el output final de minería del Procedimiento IV, de donde se desprende que los atributos '`superficiem2`',

[12] Para estudiar el flujo de minería en RM para el Procedimiento IV con mayor detalle, se ofrece al lector revisar el apartado Anexo B.2, donde se describe los flujos de minería con pormenor especificación operativa.

'ano' y 'ddjj_id' tuvieron las mayores distancias respectos su centroide, y por lo tanto, conteniendo la mayor cantidad de campos anómalos en las DDJJ de inmuebles.

4.5 RESULTADOS Y DISCUSIÓN DE LA EXPERIMENTACIÓN DE LOS PROCEDIMIENTOS III Y IV CON DDJJ

De la ejecución de los procedimientos III y IV para la detección de tuplas anómalas alfanuméricas en DDJJ, tanto en uno como en otro procedimiento, el atributo 'superficiem2' resultó significativamente sospechoso de contener tuplas anómalas. El atributo superficie cuadrada en metros cuadrados registró la mayor cantidad de valores outliers detectados totalizando 938 (Procedimiento III), así como la mayor distancia centroidal (Procedimiento IV) de todos los atributos.

En número tuplas sospechosas atributos correspondiente a los porcentaje accionario , con 252 tuplas sospechosas, y nombre de los bienes, con 209 tuplas sospechosas, escoltaron al atributo superficie cuadrada a la hora de presentar importancia en explicar el tipo de valor patrimonial declarado.

La tenencia inmobiliaria en metros cuadrados, indiferentemente de su tipo en superficie construida o cubierta, y el atributo año de presentación de la DJ sobre el bien inmueble parecen ser un fuerte factor de asimetría, y de fuente relacionante con la existencia de campos anómalos, tanto en su interpretación inductiva sobre el valor declarado de inmuebles respecto su valor de mercado o fiscal, así como atributo significante en la detección de campos anómalos en el procedimiento para BBDD alfanuméricas sin atributo objetivo.

Por otro lado, la experimentación sobre BBDD de datos públicas reales, hace difícil saber a ciencia cierta la totalidad de las tuplas outliers verdaderas respecto, mucho más tratándose de

economías emergentes inestable como la Argentina, pero sin dudas sirve de señalador para, por ejemplo, determinar posibles casos de enriquecimiento ilícito por parte de un funcionario, ya sea de cambios abruptos y repentinos en superficie cuadrada por la adquisición injustificada de inmuebles, variaciones en las proporciones de posesión accionaria, etc.

Más allá de la falta seria que hace a la no declaración, la cual es dominante en muchos atributos, en base al árbol de inducción según el tipo de valuación declarada de cada inmueble (Figura 12), pueden inducirse tramas ocultas que el los algoritmos en procedimiento desvelan, posibles tramas de evasión, elusión o incluso enriquecimiento indebido:

- Los inmuebles menores a 6500m^2 tienden a ser declarados a valor subfiscal.

- Los inmuebles mayores a 6500m^2, correspondientes a DDJJ de entre mediados de 2005 y 2012 tienden a no declararse.

- Las cocheras, campos, terrenos, parcelas, lotes, y propiedades horizontales sin especificación, cuando los inmuebles en general son mayores a 6500m^2, y las DDJJ son anteriores a mediados de 2005, tiende a no declararse su valor ante el fisco.

- Cuando el funcionario posee una casa, y una participación accionaria superior al 46,3% [100%; 37,9%) o inferior al 37,9 (37,9% ; 0%] prefiere no declarar su valuación, pero tiende a declararla subfiscalmente cuando posee un rango accionario entre el 46% y el 38% (46,3%; 37,9%).

- Cuando el funcionario posee un departamento, a partir de mediados de 2003 hasta mediados de 2005, con superficie menor o igual a 26m^2 tiende a ser declarados a su valuación fiscal, mientras que en DDJJ juradas anteriores a 2003 simplemente no lo hace. Lo que podría intuirse como una mejora cuantitativa la información vertida en los sistemas de DDJJ; no obstante, los departamentos mayores a 26m^2 simplemente tienden a no declararse.

- Respecto la posesión de locales comerciales la cuestión parece volverse interesante, puesto que aquellos locales comerciales que poseen una valuación mayor a $76493 [∞,

> $76493) o inferiores a $10615 ($10615, 0] tienden a ser declarados a su valor de mercado; a excepción de los locales con valuación con rango patrimonial entre [$76493, $10615], puesto que en este caso también se tenderá a declararlo a su valor de mercado, siempre y cuando, el inmueble se encuentra inscripto a nombre de su cónyuge, y no precisamente a nombre del oficial político, caso en el cual se tenderá a no declararlo.

El descubrimiento de estos patrones bien podría servir para la construcción de estrategias impositivas así como de contralor civil para la investigación de posibles tramas ocultas de enriquecimiento ilícito implícitos en los datos. En este sentido el uso de estos procedimientos para todo experto tanto de AFIP, AGN, OA así como para cualquier Ciudadano de a pié con acceso a datos públicos, podría ser de gran utilidad, en última instancia para el reforzamiento de los tejidos societarios.

Si bien las BBDD de DLND constituyen un esfuerzo mayor, de vanguardia, en acercar datos públicos a todo aquel que posea instrucción y acceso a internet, aquel aún no representan un canal oficial de contralor Ciudadano ni del Estado ni de cualquier otra asociación civil. Lamentablemente, la versión de BBDD trabajada aún posee fallas importantes en la presentación de sus datos, es menester que cualquier institución que se honre con proveer de información vital y útil para el bienestar cívico posea una estrategia de largo plazo en sus formas, procesos, y estabilidad temporal así como en su mantenimiento, para hacer así, un trabajo mejorable y perdurable en el tiempo.

Siendo que este trabajo abarcó DDJJ declaradas hasta el año 2012, año a partir del cual los FFPP lamentablemente no están obligados a declarar sus bienes conjuntamente con los de sus relativos más cercanos, es de esperar que futuras líneas de pesquisas con BBDD de DDJJ posteriores el atributo `'vinculo'` desaparezca junto a toda ganancia informativa social que para la retroalimentación cívico ciudadana, un hecho que coarta el verdadero sentido de transparencia informacional en la relación cibernética entre Estado y Sociedad.

5. CONCLUSIONES

En esta tesis se experimentó y se validó metodología de procedimientos híbridos (Kuna 2014) reutilizados de forma inédita para la detección de campos anómalos sobre DDJJ de inmuebles de oficiales políticos y argentinos de primera línea de función, siguiendo metodología reciente y pertinente en la detección de outliers y ruido. En este sentido la tesis logra obtener información útil para un caso de contralor civil mediante la detección de datos anómalos en DDJJ, algo hasta ahora inédito pues no existen antecedentes de detección de tuplas en DDJJ utilizando estos procedimientos.

Se validaron empíricamente procedimientos híbridos de detección de campos anómalos en BBDD reales, por primera vez en experimentación con datos públicos recientes. Estos de gran utilidad para la retroalimentación cívica societaria como lo son las DDJJ, donde se tomaron solamente bienes inmuebles, el valor patrimonial actualizado de las mismas y su valuación declarada relativa respecto el valor fiscal de los inmuebles. Tanto los atributos correspondiente a la superficie cuadrada en propiedad del declarante como el año de presentación parecen significativas en ambos flujos de minería de cada procedimientos.

5.1 BREVE DISCUSIÓN SOBRE GA Y CAPITAL SOCIAL

La importancia vital de los flujos informativos en la sociedad de información tanto en lo contemporáneo, como en lo porvenir inmediato, mediano y lejano, no puede quedar en disponibilidad de acceso abierto a DDJJ públicas solo a una única fuente para el contralor Ciudadano: la levedad e importancia de la existencia de las DDJJ en DLND debe ser apuntalada por sistemas de información análogos desde oenegés, OA, la AGN y otros organismos de contralor Civiles y Estatales. Toda institución que se honre de contribuir al acceso abierto debería hacerse de una estrategia de diseño de BBDD concisa y perdurable en el tiempo ya sea en la calidad de la información presentada como en la preservación de la misma.

Concretamente, los resultados arribados sugieren que más acuse de control debe dársele a la no declaración de inmuebles por parte de funcionarios públicos, así como el de una mayor y mejor

valuación monetaria real de su patrimonio; por otra parte, puede esperarse que contextos menos inflacionarios y mejores estadísticas públicas ayuden en esa tarea.

Sopesando la necesaria labor de la sociedad civil, los organismos de contralor del propio Estado como la es OA, AGN, etc. no solo deberían modernizarse adentrándose de lleno en el empleo de procesos como los aquí presentados, estos también deben asumir procedimientos de gestión de la información estandarizados, y acceso abierto en línea acorde a los tiempos de digitalización en que la sociedad se encuentra empoderándose[13]. *i.e.* adoptando plataformas de acceso abierto y un sistema de información ágil y estable en el tiempo para la población conectada, de manera de proponer una funcionalidad y modernización completa de los organismos de contralor del Estado. Esto exige la disponibilidad de datos públicos y un sistema de GA disponible en tiempo y forma para su acceso libre en sistemas e infraestructura disponible para cualquier Ciudadano cosa que en Argentina, a la fecha de la edición de este trabajo es aún inexistente.

Sin desmerecer la posesión de riqueza creada con el esfuerzo personal y familiar de varias generaciones de Ciudadanos empresarios y trabajadores argentinos, es indudable que Ciudadanos con recursos económicos asegurados poseen más posibilidades de acceder a la clase política, muchos de ellos mismos como el fruto del esfuerzo de varias generaciones de una familia. Como ejercicio cívico, cabe la posibilidad de cuestionarse a manera de ejemplificación cuantitativa, cuantas generaciones de una familia de clase media o media baja de hoy serían necesarios para acceder a una fortuna mínima para, por ejemplo, financiar la campaña electoral de alguno de sus vástagos; así como de la efectividad de los sistemas educativos para la generación de igualdad de

[13] La palabra *empoderamiento* proviene del término inglés *empowerment* originario de la escuela de psicología comunitaria norteamericana de principios de los 80s.

oportunidades, con la integración necesaria para que tales diferencias en riqueza monetaria no fueren causa ni escusa para la división y desmantelamiento del capital societario.

Con el objeto claro de lograr un tejido social más denso, donde la corrupción no amenace con tornarse estructural, la construcción de un GA de acceso efectivo; y capaz de incorporar en la cultura endógena de todo aglomerado Ciudadano, un mayor desenvolvimiento cívico en la participación digital que impulse y transforme al contralor civil societario en una actividad bien comunitaria más disponible a todo actor social que decida colaborar con su tiempo y energías en ese sentido. Aquella última, en obvia articulación con la formación cívica formal y tradicional, la cual ya no puede considerarse desligada de la realidad social informacional contemporánea.

Desarrollar una cultura generacional para la participación digital ciudadana se encontrará de hecho ligado a la formación cívica tradicional, pero incluso así se hace cada vez más necesaria una co-construcción [14] (Vercelli 2014) técnica normativa para una correcta ejecución que posibilite un empoderamiento cívico de la población de forma universal, sin ánimos clasistas, apartidario y responsable; como estrategia posible y mejorable en el largo plazo para combatir la corrupción.

No obstante el desarrollo de políticas perdurables de empoderamiento ciudadano debería incorporar toda estrategia posible para la lucha anticorruptiva, el empoderamiento digital que estas conllevan supone la reducción de las brechas digitales cívicas preexistentes en las poblaciones no conectadas, o que ya conectadas, no persiguen el empoderamiento cívico como objeto satisfactor, al inicio, pero capaces en potencia de hacer del Ciudadano en mejores electores, mejores proyectos políticos compatibles a proyectos generacionales individuales y colectivos reales, mediante los medios digitales disponibles, hacen a la satisfacción de su bienestar más allá de sus necesidades básicas sino la capacidad de su cultura originaria por trascender los tiempos a su descendencia, haciendo a la conservación de una infodiversidad (López-Pablos, 2015b).

[14] Relativo al desarrollo de normativa tecnológica para la regulación de Internet donde se advierte que las tecnología y regulación se *co-construyen* mutuamente al tener que emplear tecnicismos presentes en software y la red inserta en la conformación normativa propiamente dicha. Por ejemplo, los conflictos jurídicos entre Ubër y Taxis hacen a esa co-construcción normativa, ya que la aplicación de la normativa tradicional no es efectiva en los contextos actuales.

Las brechas cívico digitales preexistentes intrasocialmente, obran en contra de tales transferencias infodiversas, y podrían actual de contrapeso a las desigualdades preexistentes en los estratos societarios. Poniendo de relieve las carencias de una co-construcción (di)isonómica de una red isonómica de pares (re)regulada por un lado, y la falta de decisión y efectividad de políticas digitales inclusivas con responsabilidad y disciplina cívico-digital, en su necesidad de enriquecer la sociedad, así como del deber de protegerla y la capacidad de reconocer la interferencia sistemática de agentes externos que puedan amenazarla, tanto dentro como fuera de los Estados, por fuentes solapadas e intereses de poder concentrado muchas veces ajenos a los intereses comunitarios. Aún no existen procesos y herramientas para coadyuvar a la distinción entre tales poderes e intereses similares al que este trabajo aporta como proceso y artefacto de contralor civil; la formación cívica y la voluntad ciudadana en búsqueda de justicia siguen siendo absolutamente necesarios e irreemplazables para el despertar de las consciencias y voluntades colectivas socialmente responsables.

5.2 CONSIDERACIONES PARA EL FUTURO

En lo que hace a futuros esfuerzos en la detección de campos anómalos, no puede dejar de pensarse en la posibilidad de experimentar con variantes de los procedimientos kunianos III y IV, en los cuales podría probarse con reemplazar operadores tanto de clasificación como de clustering *e.g.*: utilizando K-Medoids en lugar de K-Means, o experimentar con otros algoritmos detectores de anomalías como COF, y alternarlos entre LOF y DBSCAN, etc., podrían ser otra opción para así conformar así sendos procedimientos IIIbis y IVbis podrían constituirse en una opción a ser explorada. Dado que aquí no se utilizó una metodología de minería formal es esperable y recomendable la utilización de CRISP-DM de manera formal en BBDD de DDJJ para explorar la posibilidad de búsqueda de nuevos resultados en la explotación de la información disponible con la que disponemos actualmente.

BIBLIOGRAFÍA

Abbott, D. W., Matkovsky, I. P., & Elder, J. F. (1998). An evaluation of high-end data mining tools for fraud detection. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, *3*, 2836 -2841.

Aleskerov, E., B. Freisleben, B., & Rao (1997). Cardwatch: A neural network based database mining system for credit card fraud detection. *Proceedings of the IEEE/IAFE, Conference on Computational Intelligence for Financial Engineering (CIFEr),* 220-226.

Becker, G. (1968) *Crime and Punishment: An Economic Approach.* Journal of Political Economy, 76(2), 169-217.

Bour, E. A. (2014, Marzo) "Corrupción, un punto de vista estratégico", *ANCE,* CABA. Recuperado de http://www.anceargentina.org/site/trabajos/Enrique%20Bour___%20Marzo%202014.pdf (válido al 1/3/2018)

Boscovich, R. J. (1757). De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientarum et Artum Instituto Atque Academia Commentarii, 4,* 353-396.

Breuning, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *ACM Sigmod Record, 29*(2), 93-104.

Canavese A. J. (2005). The Effects of Corruption Organization and Punishment on the Alloction of Resources. *Berkeley Program in Law & Economics. Latin American and Caribbean Law and Economics Association* (ALACDE) Annual Papers. Paper 19. Recuperado de http://repositories.cdlib.org/bple/alacde/19 (válido al 1/3/2018)

Cendrowska, J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, *27*(4), 349-370.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15.

Cisneros C. (2011). El libre acceso a la información pública: una necesidad para el desarrollo de la democracia representativa y sus instituciones. Tesis de maestría, FCJC-UNLP. La Plata. Recuperado de http://hdl.handle.net/10915/32206 (válido al 1/3/2018)

Colombo H. L., Antonini S. G., Chong Arias C. D., Istvan R. M., Peternoster F. M. (2013). Estudio de soluciones tecnológicas para el desarrollo de un modelo factible de participación ciudadana. *Proceedings del XV Workshop de Investigadores en Ciencias de la Computación, 954*-958. Recuperado de: http://hdl.handle.net/10915/27354 (válido al 1/3/2018)

Cortina A. (1994). *La ética de la sociedad civil*, Madrid: Ed. Anaya.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions* on, 14(3), 462-467.

DD.JJ. Abiertas (2015). LNData. Actualizado al 13/1/2014. Recuperado de http://interactivos.lanacion.com.ar/declaraciones-juradas/ (válido al 1/3/2018)

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD, 96*, 226-231.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. *MIT Press.*

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management, 7*(1), 21-41.

Ferreyra, M. (2007). Powerhouse: Data Mining usando Teoría de la información. Recuperado de http://web.austral.edu.ar/images/contenido/facultad-ingenieria/2- Data_Mining_basado_Teoria_Informacion_Marcelo_Ferreyra.pdf (válido al 1/3/2018).

Ferro G., Giupponi L., & Gómez, N. (2007, Enero). *Declaraciones Juradas de Funcionarios Públicos. Una herramienta para el control y prevención de la corrupción*. Tecnología informática y gestión pública 2° ed. – CABA: Oficina Anticorrupción. Ministerio de Justicia y Derechos Humanos, 88 p. Recuperado de http://www.anticorrupcion.gov.ar/documentos/libro%20ddjj%202ed.pdf (válido al 1/3/2018)

Foster, D. & Stine, R. (2004). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. *Journal of American Statistical Association, 99*, 303-313.

García Martínez, R., Servente, M., & Pasquini, D. (2003). *Sistemas Inteligentes*. *CABA, Argentina: Editorial Nueva Librería.*

Gómez N., & Bello M. A. (2009, Mayo). Ética, *transparencia y lucha contra la corrupcion en la administracion publica*, Manual para el ejercicio de la funcion publica, 1ra ed., CABA: Oficina Anticorrupción, Ministerio de Justicia y Derechos Humanos de la Nación. Recuperado de http://www.anticorrupcion.gov.ar/documentos/Libro%20SICEP%202da%20parte.pdf (válido al 1/3/2018)

Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing-A Journal Of Practice & Theory*,*16*(1), 14-28.

Heidenheimer, A., Johnston, M., Le Vine, V. (1989). *Political Corruption: A Handbook,* Transaction Publishers, New-Brunswick.

Hawkins, D. M. (1980). Identification of outliers. *London: Chapman and Hall.,* 11. Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126.

Huysmans J., Baesens B., & Vanthienen J. (2008). "A Data Miner's Approach to Country Corruption Analysis", *Intelligence and Security Informatics, Studies in Computational Intelligence,* 135, 227-247.

Huysmans J., Martens D., Baesens B., Vanthienen J., Van Gestel T. (2006). Country Corruption Analysis with Self Organizing Maps and Support Vector Machines, WISI, LNCS 3917, 103-114. Recuperado de https://lirias.kuleuven.be/bitstream/123456789/101626/1/ (válido al 1/3/2018)

Jensen, F. V. (1996*).* An introduction to Bayesian networks. *London: UCL press,* 210.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995-1003.

Koskivaara, E. (2004). Artificial neural networks in analytical review procedures. *Managerial Auditing Journal*, *19*(2), 191-223.

Klitgaard R. (1992). Controlando la corrupción, Buenos Aires: Ed. Sudamericana.

Kuna, H., García Martínez, R., & Villatoro, F. (2009). Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información. *Revista Iberoamericana de Tecnología en Educación y Educación en* Tecnología, *5*, 39-44.

Kuna, H., García-Martínez, R., & Villatoro, F. (2010a). Pattern Discovery in University Students Desertion Based on Data Mining. In *Advances and Applications in Statistical Sciences* Journal, *2*(2): 275-286.

Kuna, H., Caballero, S., Rambo, A., Meinl, E., Steinhilber, A., Pautch, G., García-Martínez, R., Villatoro, F. (2010b). Avances en procedimientos de la explotación de información para la identificación de datos faltantes, con ruido e inconsistentes. *Proceedings XII Workshop de Investigadores en Ciencias de la Computación,* 137-141.

Kuna, H., Caballero, S.. Rambo, A., Meinl, E., Steinhilber, A., Pautch, G., Rodríguez, D., García-Martínez, R., Villatoro, F. (2010c). Identification of Noisy Data in Databases by Means of a Clustering Process. *Ingeniería de Software e Ingeniería del Conocimiento: Tendencias de Investigación e Innovación Tecnológica en Iberoamérica,* 264-273.

Kuna, H., Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., Steinhilber, A., García-Martínez, R., Villatoro, F. (2011). Avances en procedimientos de la explotación de información con algoritmos basados en la densidad para la identificaciónde outliers en bases de datos. *Proceedings XIII Workshop de Investigadores en Ciencias de la Computación.* Artículo 3745.

Kuna, H., García-Martinez, R., &Villatoro, F. (2012a). Automatic Outliers Fields Detection in Databases. In *Journal of Modelling and Simulation of Systems*, *3*(1), 14-20.

Kuna, H., Rambo, A., Caballero, S., Pautsch, G., Rey, M., Cuba, C., García-Martínez, R., Villatoro, F. (2012b). Procedimientos para la identificación de datos anómalos en bases de datos. In *Proceedings of CONISOFT,* 184-193.

Kuna, H. , Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., García-Martínez, R., Villatoro, F. (2012c). Comparación de la efectividad de procedimientos de la explotación de información para la identificación de outliers en bases de datos. *Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación,* 296-300.

Kuna, H., Villatoro, F., & García-Martinez, R. (2013a). Development and Comparison of Procedures for Outlier Detection in Databases. *Computers & Security*. (en evaluación).

Kuna, H., Pautsch, G., Rambo, A., Rey, M., Cortes, J., Rolón, S. (2013b). Procedimiento de explotación de información para la identificación de campos anómalos en base de datos alfanuméricas. R*evista Latinoamericana de Ingeniería de Software*, *1*(3): 102-106.

Kuna, H., García-Martínez, R., & Villatoro, F. (2014). Outlier detection in audit logs for application systems. *Information Systems.*

Kuna H. (2014, Marzo). *Procedimientos de explotación de la información para la identificación de datos faltantes con ruido e incosistentes*, Tesis doctoral, Universidad de Málaga.

Larose, D. T. (2005). Discovering knowledge in data: an introduction to data mining. *Wiley. Com.*

Ley Nacional "Ética en el ejercicio de la función pública", Ley Nacional nro. 25.188, Infoleg. Recuperado de http://www.infoleg.gov.ar/infolegInternet/anexos/60000-64999/60847/norma.htm (válido al 1/3/2018)

López-Pablos, R (2015a, Marzo). "Nociones cibernéticas e informáticas para una actualización de la ecuación de Klitgaard", Documento de Trabajo MEISI, Escuela de Posgrado UTN. Recuperado de http://hdl.handle.net/10915/44663 (valido al 1/3/2018)

López-Pablos, R (2015b, Agosto). "Constructos teóricos en economía común informática", La Matanza: Documento de Trabajo Doctoral, DCE-UNLaM. Recuperado de http://hdl.handle.net/10915/48130 (valido al 1/3/2018)

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1*(281-297), 14.

Naciones Unidas (2004). *Convensión de las Naciones Unidas contra la corrupción, contra la droga y el delito* (CCNUC), Oficina de las NU, Viena. Recuperado de http://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/04-56163_S.pdf (válido al 1/3/2018)

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559-569.

Open Government Partnership. (2012, Junio [2014, Marzo]). *OGP: Articles of Government*. Recuperado de http://www.opengovpartnership.org/sites/default/files/attachments/OGP%20ArticlesGov%20March%2019%202014_1.pdf (válido al 1/3/2018)

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. *Morgan Kaufmann.*

Quinlan, J. R. (1993). C4.5: programs for machine learning. *Morgan Kaufmann*.

Raigorodsky, N., & Geler L. (2007, Agosto). Convención de las Naciones Unidas contra la corrupción: nuevos paradigmas para la prevención y combate de la corrupción en el escenario global, 1ra ed., CABA: Oficina Anticorrupción, Ministerio de Justicia y Derechos Humanos de la Nación. Recuperado de http://www.anticorrupcion.gov.ar/documentos/Libro%20CNUCC%202ed.pdf (válido al 1/3/2018)

Ransom J. (2013, Junio). *Replicating Data Mining Techniques for Development: A Case of Study of Corruption*, Lund University, Master Thesis, Master of Science in International Development and Management. Recuperado de http://lup.lub.lu.se/record/3798253/file/3910587.pdf (válido al 1/3/2018)

Rapoport M. (2010) "Una revisión histórica de la inflación Argentina y de sus causas", CABA: Aportes de Economía Política en el Bicentenario de la Revolución de Mayo.

Saad, M. K., & Hewahi, N. M. (2009). A comparative study of outlier mining and class outliermining. *COMPUTER SCIENCE LETTERS*, *1*(1).

Shannon, C. E. (2001 [1948]). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3-55.

Sowjanya, S., & Jyotsna G. (2013, Noviembre). Application of Data Mining Techniques for Financial Accounting Frad Detection Scheme. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(11), 717-724.

Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, *17*(4), 179-191.

Transparency International (2009, Julio). *The anti-corruption Plain Language Guide*, TI E-Book. Recuperado de http://media.transparency.org/fbooks/pubs/pl_guide/ (válido al 1/3/2018)

Vercelli, A. (2014). Repensando las regulaciones en la era digital: ¿llego la hora de (re)regular Internet?, CABA: FCE-UBA,*Voces en el Fenix*,5(40), 14-21. Recuperado de http://www.vocesenelfenix.com/sites/default/files/pdf/14_art2-n40fenix40baja1.pdf (valido al 1/3/2018)

Vercelli, A. (2013). La participación política ciudadana en la era digital, Monterrey: CIES, *Virtualis*, enero-julio 4(7), 115-129. Recuperado de http://micampus.ccm.itesm.mx/documents/14896/111549100/virtualis07.pdf (valido al 1/3/2018)

Yue, X., Wu, Y., Wang, Y. L., & Chu, C. (2007, Septiembre). A review of data mining-based financial fraud detection research, international conference on wireless communications, *Networking and Mobile Computing*, 5519–5522.

Wang, J., Liao, Y., Tsai, T. & Hung, G. (2006, Octubre). Technology-based financial frauds in Taiwan: issue and approaches, *IEEE Conference on: Systems, Man and Cyberspace,* 1120–1124.

Wang, S. (2010). A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research. *International Conference on Intelligent Computation Technology and Automation*, vol. 1, pp.50-53.

PUBLICACIONES A LAS QUE DIÓ LUGAR LA TESIS

Como producción científica derivada de esta tesis, se produjeron las siguientes publicaciones y trabajos científicos:

Comunicaciones a congresos:

López-Pablos, R. y Kuna H.D. (2016, Octubre). *Propuesta de detección de datos anómalos y ruido en declaraciones juradas públicas*, San Luis: Anales del XVIII Congreso Argentino de de Ciencias de la Computación. Pág. 674-682. ISBN 978-950-947449-9.

López-Pablos, R. y Kuna H.D. (2017, Septiembre). *Detección de datos anómalos y ruido en base de datos abiertas para el contralos público, utilizando técnicas de minería de datos.* Córdoba: Anales de las 46° Jornadas Argentinas de Informática e Investigación Operativa 46 JAIIO – STS. Pág. 105-127 ISSN: 2451-7631.

Revistas:

López-Pablos, R. y Kuna H.D. (2017). *A proposal for Outlier and Noise Detection in Public Officials' Affidavits.* Computer Science & Technology: XXII Argentine Congress of Computer Science Selecterd Papers (Eds. M. B. Piccoli *et. al.*) . Pág. 201-210 ISSN: 978-987-4127-28-0.

Documento de trabajo:

López-Pablos, R. (2015, Febrero) Nociones cibernéticas e informáticas para una actualización de la ecuación de Klitgaard, Documento de Trabajo MEISI, Escuela de Posgrado UTN.

Trabajo integrador de especialización:

López-Pablos, R (2016, Junio). Propuesta de detección de datos anómalos y ruido para un caso de contralor civil, TFI de Especialidad en Ingeniería en Sistemas de la Información, Escuela de Posgrado - FRBA, Universidad Tecnológica Nacional.

ANEXO

Como Sección complementaria y apéndice de esta tesis, este Anexo se divide en tres subsecciones esenciales para el entendimiento detallado y pormenorizado del funcionamiento y paramentrización del procedimiento IV –ver Sección 3.3.2- en el primer Anexo A, para luego en el Anexo B, ahondar en detalle en la programación operativa paplicada de ambos procedimiento III y IV mediante el uso de la herramienta de minería RM. Finalmente en el Anexo C, se despliega un Glosario con la descripción de los atributos de la BBDD de DDJJ utilizada en este estudio.

A. METODOLOGÍA HÍBRIDA PARA LA CONFIGURACIÓN PROCEDIMIENTO IV

Siguiendo literatura especializada en la detección de outliers en este anexo se despliegan las uniones híbridas de algoritmos así como las reglas de determinación de outliers siguiendo el procedimiento propuesto por (Kuna, 2014).

A.1 UNIÓN DE RESULTADOS POR APLICACIÓN DE ALGORITMOS LOF Y DBSCAN

Buscando optimizar los resultados de detección de datos anómalos, se explora un formato híbrido (Kuna, 2014) sugiriendo la unión en el uso de los algoritmos LOF y DBSCAN, como ilustra la siguiente Figura 14. Aproximación que consisten en la aplicación individual de cada uno de los algoritmos *a priori*, agregando un atributo binario capaz de indicar si la tupla es un outlier o no para el algoritmo utilizado, Operativamente es ilustrado en la Figura 17 del Anexo B.2 *ut infra*.

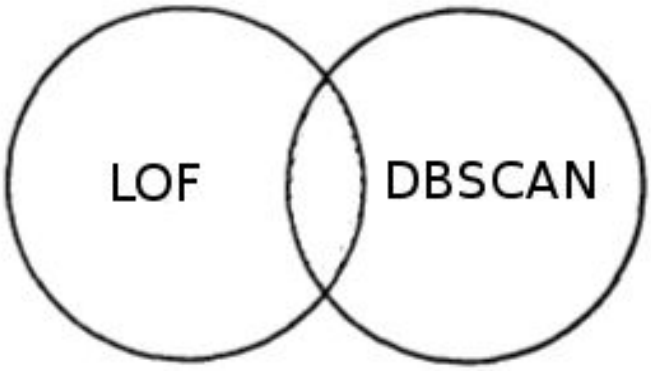


Figura 14: Unión de algoritmos LOF-DBSCAB.

A partir de la experimentación se estableció el siguiente criterio para la detección de datos anómalos en la aplicación del algoritmo especializado de LOF.

> **Proposición LOF**: De acuerdo a experimentación se determina que los mejores resultados de aplicación del algoritmo se tuvo con un **Limite de LOF** = `1.5` con **MinPtsMin** = `10`, **MinPtsMax** = `20`.

En el caso de la aplicación del algoritmo LOF cuando se obtiene un valor superior a 1.5 considérese un outlier de acuerdo a su proposición LOF (Kuna, 2014).

En el caso de aplicar DBSCAN los outliers son aplicados en el cluster "0". Débesele agregar cuatro columnas a la tabla de auditoría en cuestión añadiéndosele los siguientes cuatro atributos `'LOF'`,`'valor_LOF'`, `'valor_DBSCAN'`, y `'tipo_outlier'`. Para cada tupla se debe completar los valores de esos atributos de acuerdo al siguiente criterio:

Aplicar **LOF**

Gravar en `'LOF'` el valor que se obtuvo después de aplicar el algoritmo.

Si `'LOF'` ≤ `1.5` => `'valor_LOF' = '0'`
Si `'LOF'` ≤ `1.5` => `'valor_LOF' = '1'`

Aplicar **DBSCAN**

Si la tupla pertenece a un cluster ≠ `'0'`=> `'valor_DBSCAN' = '0'`
Si la tupla pertenece a un cluster = `'0'`=> `'valor_DBSCAN' = '1'`

A.1.1 REGLAS DE DETERMINACIÓN DE OUTLIERS PARA ALGORITMOS LOF Y DBSCAN

Ahora para la constitución de reglas de detección de datos anómalos se asignan los siguientes valores al atributo `'tipo_outlier' = 'outlier_doble'` cuando ambos algoritmos detectan la tupla como outlier, u `'outlier_simple'` cuando solo uno de los algoritmos considera a la tupla como outlier; para el caso en donde ninguno de los algoritmos detecta a la tupla como outlier, se asigna `'no_outlier'`. Las siguientes reglas sobre una BD aseguran la aplicación anterior como sigue.

Si `'valor_LOF' = '1'` y `'valor_DBSCAN' = '1'` =>
`'tipo_outlier' = 'outlier_doble'`

Si `'valor_LOF' = '0'` y `'valor_DBSCAN' = '1'` =>
`'tipo_outlier' = 'outlier_simple'`

Si `'valor_LOF' = '1'` y `'valor_DBSCAN' = '0'` =>
`'tipo_outlier' = 'outlier_simple'`

Si `'valor_LOF' = '0'` y `'valor_DBSCAN' = '0'` =>
`'tipo_outlier' = 'outlier_simple'`

Operativamente visible en la Figura 18 del mismo Anexo más abajo. Se busca ahora disminuir el número de falsos positivos y optimizar el número de outlier detectados en el procedimiento

desarrollado por los autores (Kuna et al., 2014), se propone realizar un ajuste de acuerdo al siguiente criterio para cada tupla.

```
Si 'tipo_outlier' = 'outlier_simple' y 'valor_LOF' = '1'    y
'LOF' > 1.575 => 'tipo_outlier' = 'outlier_simple'
```

```
Si 'tipo_outlier' = 'outlier_simple' y 'valor_LOF' = '1'    y
'LOF' ≤ 1.575 => 'tipo_outlier' = 'no_outlier'
```

Como podrá notarse se advierte un incremento en el valor límite del 5% del valor de acuerdo a la proposición de LOF se debe a una mayor exigencia para la detección de outliers para evitar falsos positivos.

```
Si 'tipo_outlier' = 'outlier_simple' y 'valor_DBSCAN'      = '1'
y 'LOF' > 1.425 => 'tipo_outlier' = 'outlier_simple'
```

```
Si 'tipo_outlier' = 'outlier_simple' y 'valor_DBSCAN' = '1' y
'LOF' ≤ 1.425 => 'tipo_outlier' = 'no_outlier'
```

Inversamente, ahora la disminución del valor límite de LOF se debe a un mayor grado de confianza en los resultados obtenidos del algoritmo DBSCAN a partir de la cantidad de errores detectados. Operativamente con RM, se hace lo propio con estos umbrales como se exponen las Figuras 20, 21 y la Tabla 15 desplegada *ut infra*.

A.2 UNIÓN DE ALGORITMOS DE CLASIFICACIÓN

Investigando las formas de volver el procedimiento aún más efectivo tanto en la detección de datos anómalos, en referencia a la propuesta procedimental de la sección 4.3.2, para la disminución de la cantidad de falsos positivos *a posteriori* de aplicar algoritmos especializados como LOF y DBSCAN, se propone la utilización de algoritmos de clasificación sobre la misma BD inicial ya tratada.

Estos algoritmos de clasificación son capaces de predecir el valor de atributos clase o target a partir de la consideración de otros atributos presentes en el mismo set de datos, los cuales permiten bajo determinadas reglas verificar si una tupla en cuestión es o no un outlier o un inlier.

Producto de la aplicación *a priori* de los algoritmos especializados LOF y DBSCAN, se cuenta con el atributo de clase único `'tipo_outlier'` indicador del resultado de cada tupla *a posteriori* de la aplicación de los algoritmos especializados, sin necesidad de crear un atributo clase. Este procedimiento contempla tanto la aplicación de un enfoque 1, con algoritmos especializados como de tipo 2 al usar algoritmos de aprendizaje como el C4.5 y el PRISM.

Por razones de efectividad procedimentales, la combinación de algoritmos de clasificación seleccionados consiste en la unión de los algoritmos C4.5 + RB + PRISM, como se despliega en la siguiente Figura 15, operativamente visible en la Figura 22 y 23 del Anexo B.2.

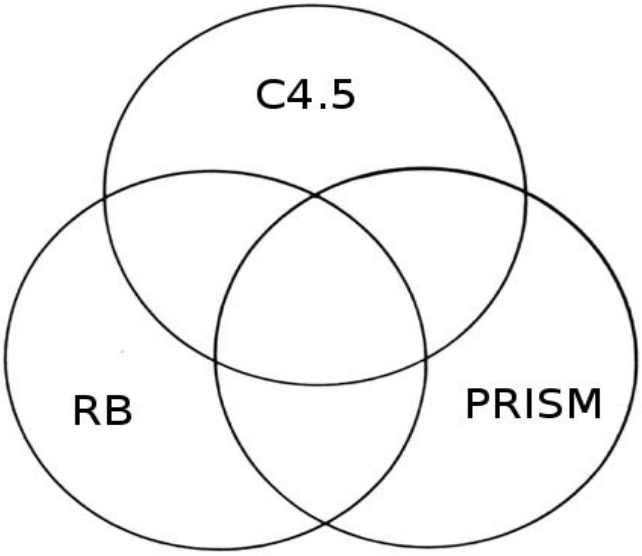


Figura 15: Unión de algoritmos C4.5-RB-PRISM.

La justificación en el uso de estos algoritmos radica en su relación de equilibrio entre las fortalezas y debilidades de cada uno en la detección de datos anómalos, puesto que mientras el algoritmo C4.5 manifestó buena efectividad para la detección de outliers no logró buenos resultados detectando falsos positivos; algo que si se pudo lograr aplicando RB, aunque la aplicación de este último disminuía la efectividad en la detección de outliers. Buscando equilibrar ambas estrategias el uso de PRISM termina por robustecer todo el procedimiento.

A.2.1 REGLAS DE DETERMINACIÓN DE OUTLIERS PARA ALGORITMOS DE CLASIFICACIÓN

Siempre dentro del procedimiento IV de detección de outliers de Tipo 1 para BB.DD. alfanuméricos desarrollado y siguiendo a los mismos autores (Kuna et al., 2014; Kuna, 2014), *a posteriori* de la aplicación de la combinación algoritmica LOF + DBSCAN especializada, se despliega un conjunto de reglas para la optimización de los resultados de la utilización simultánea de algoritmos de clasificación C4.5, PRISM y RB. A partir de la utilización de estos tres últimos se obtienen dos valores posibles derivados de su utilización: el valor `'outlier'` o el valor `'limpio'` para cada tupla del atributo `'tipo_outlier'` de los resultados obtenidos de aplicar LOF se los considera elementos de validación.

La aplicación algorítmica simultánea de los tres algoritmos de clasificación permite optimizar la detección de valores anómalos para el procedimiento de detección de sin un atributo target dado para entornos alfanuméricos con aprendizaje no supervisado.

Si `'tipo_outlier'` = `'outlier_simple'` => `'tipo_outlier'` = `'outlier_doble'`

De manera que de contar solo con dos valores en el atributo `'tipo_outlier'` se simplifica el proceso de clasificación.

Clasificar cada tupla con los algoritmos C4.5, RB y PRISM tal que:

Para cada tupla si se tiene para los 3 algoritmos de clasificación t. q.
`'tipo_outlier'` = `'outlier_doble'` => `'tipo_outlier'` = `'outlier'`

Si para el agoritmo de RB y PRISM `'tipo_outlier'` = `'outlier_doble'` => `'tipo_outlier'` = `'outlier'`

Si para el agoritmo de RB `'tipo_outlier'` = `'outlier_doble'` y para C4.5 y PRISM => `'tipo_outlier'` = `'no_outlier'` y `'valor_LOF'` > `1.575` => `'tipo_outlier'` = `'outlier'`

Si para el agoritmo C4.5 y PRISM `'tipo_outlier' = 'outlier_doble'` y para RB `'tipo_outlier' = 'no_outlier'` => `'tipo_outlier' = 'outlier'`

Si para el agoritmo C4.5 `'tipo_outlier' = 'outlier_doble'` y para PRISM `'tipo_outlier' = 'no_outlier'` y para RB `'tipo_outlier' = 'no_outlier'` y `'valor_LOF' > 1.65` => `'tipo_outlier' = 'outlier'`

Umbral que representa el 10% del valor límite establecido en la proposición LOF.

Si para el agoritmo PRISM `'tipo_outlier' = 'outlier_doble'` y para C4.5 `'tipo_outlier' = 'no_outlier'` y para RB `'tipo_outlier' = 'no_outlier'` y `'valor_LOF' > 1.65` => `'tipo_outlier' = 'outlier'`

En este caso también se asume como umbral que representa el 10% del valor límite establecido en la proposición LOF.

Para cualquiera de las tuplas que no cumpla estas condiciones => `'tipo_outlier' = 'outlier'`

De esta forma se consigue arribar a un atributo, que respondiente a la programación anterior, es capaz de determinar cuales tuplas o cuales no se encuentran dentro o fuera de la unión de los tres algoritmos generales de clasificación. Notar asimismo que estas reglas son visibles operativamente en RM como subprocesos en las Figuras 24, 25, 26 y 27 del siguiente Anexo.

B. APROXIMACIÓN OPERATIVA EMPÍRICA PROCEDIMENTAL EN RM

En el presente apartado del Anexo se describen los pasos procedimentales para la aplicación empírica de los procedimientos de detección de datos anómalos alfa-numéricas (Kuna, 2014), a través de la utilidad operacional que brinda el software especializado en minería de datos *Rapid Miner* (RM). En la Subsección B.1 se describe la anterior correspondiente al procedimiento III de la Sección 4.3.1 mientras que en la Subsección B.2 se hace lo suyo respecto el procedimiento IV ya explicado en la Sección 4.3.2 así como en el presente Anexo más arriba.

B.1 APLICACIÓN DEL PROCEDIMIENTO III EN "RM"

Siguiendo las fases propuestas por Kuna operativamente en la práctica en RM se selecciona para el proceso supervisado el atributo como '`label`' a través del operador **Set Role** el atributo '`val_decl`' y luego imputamos el algoritmo C4.5 (**Decision Tree**) para obtener el árbol de decisión correspondiente como describe el siguiente flujo de minería. Seteamos el algoritmo de inducción con la siguientes configuración de parámetros:

```
criterion: gain ratio
minimal size of split: 8
minimal leaf size 2
minimal gain: 0.1
maximal depth: 20
confidence: 0.25
no pruning
```

(Obteniendo como resultado experimental lo expuesto en la Figura 12, 5.3, *ut supra*)

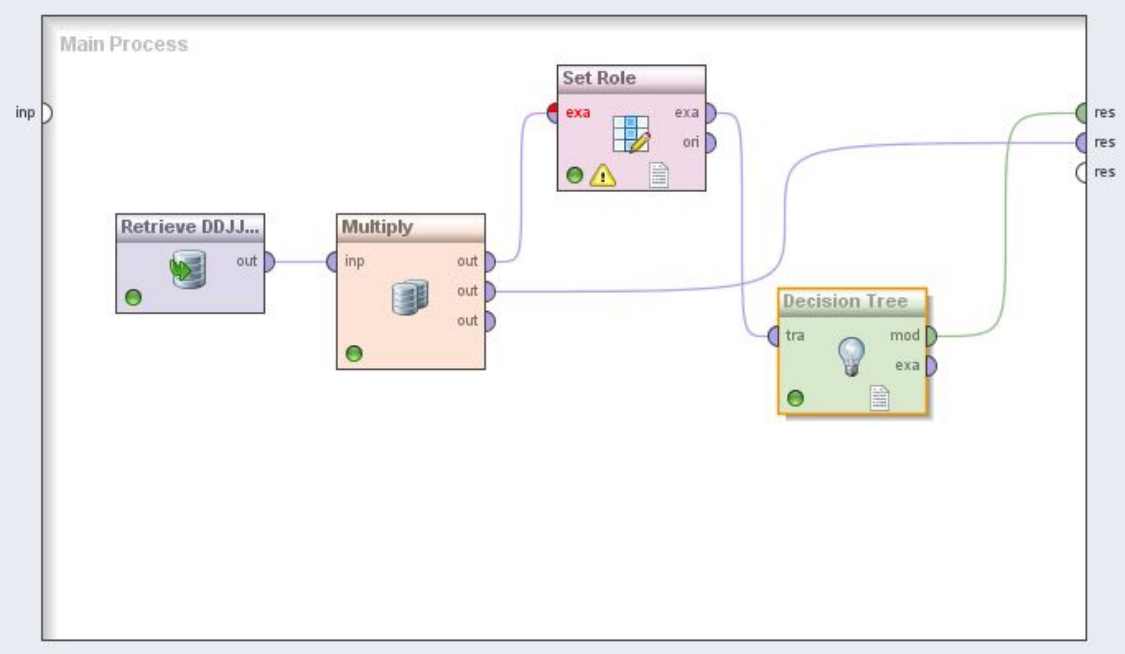


Figura 16: Flujo de minería en RM, (Procedimiento III[a]).

Desde el flujo anterior se obtuvieron los 6 (seis) atributos significativos que hacen de entrada a los bins correspondientes, donde el atributo supervisado hace de salida para cada bins analizado. Posteriormente, para cada bin E→S se hace correr el siguiente flujo de minería correspondiente a la figura superior:

BD_DDJJ → Multiply → Set Role (`'val_decl'`) → Decision Tree (C4.5)

De donde se obtiene el árbol de inducción correspondiente a la Figura 12 de la Sección 5.3.

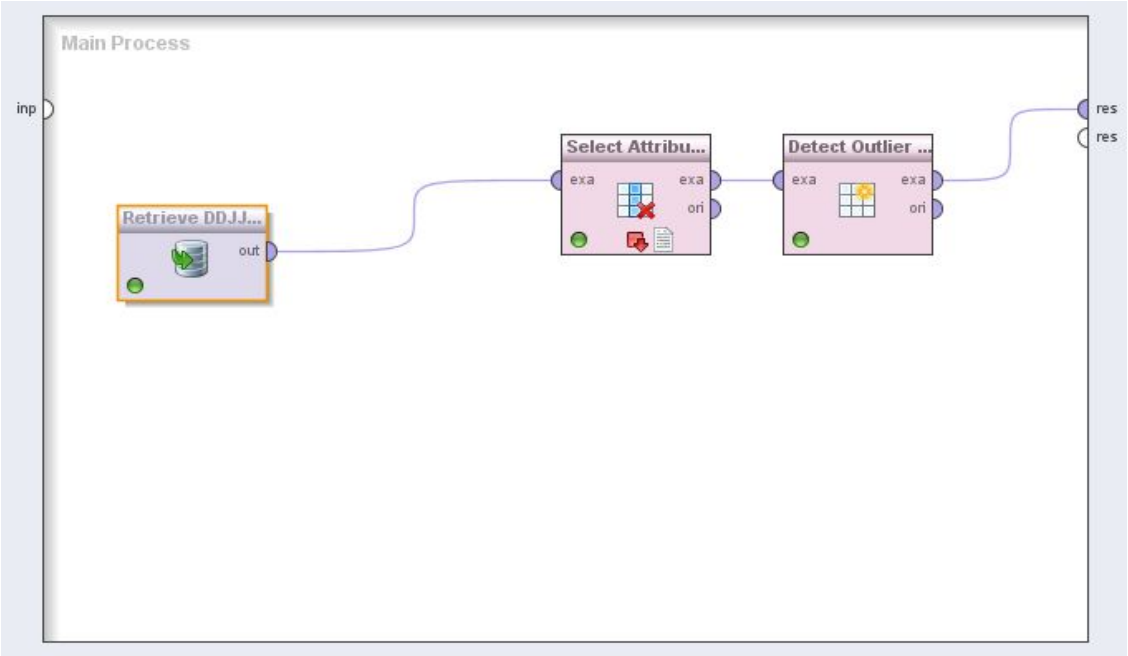


Figura 17: Flujo de minería en RM programable para cada bin Entrada-Salida, (Procedimiento III[b]).

De hacer correr el flujo de minería superior se obtuvieron los campos anómalos hipotéticos presentados en los resultados producto del output del algoritmo LOF con valor ∞ (infinito).

```
attribute filter type: subset
attibutes: iterion: Selects attributes
```

Repitiendo el flujo para las siguientes seis (6) bi-selecciones de atributos (E)-(S):

superficiem2	ano	nombre_bien_s	porcentaje	val_patrim	vinculo
val_decl	val_decl	val_decl	val_decl	val_decl	val_decl

Tabla 14: Selección de atributos en RM para seis (6) bins.

Donde se selecciona cada par de atributos correspondientes a los bins de Entrada-Salida correspondientes a los 6 (seis) bins de la Tabla 10. Para lo cual se programa el operador 'Select Attributes' donde en cada caso se baja una BD con los dos atributos y los resultados *a posteriori* de la apliacación de LOF.

B.2 APLICACIÓN DEL PROCEDIMIENTO IV EN "RM"

En la aplicación empírica del procedimiento IV en RM se procede a programar el flujo de minería para tal procedimiento donde se lee la BD de DDJJ preparada (**Retrieve**) donde se setea al atributo '`ddjj_id`' como atributo target identificatorio (**Set Role**), luego se multiplica el flujo para que los algoritmos de búsqueda **DBSCAN** y **LOF** puedan traccionar sobre la data. Ambos resultados son fusionados en uno solo posteriormente mediante el operador **Join** como se describe en la primera fase del procedimiento en la siguiente figura.

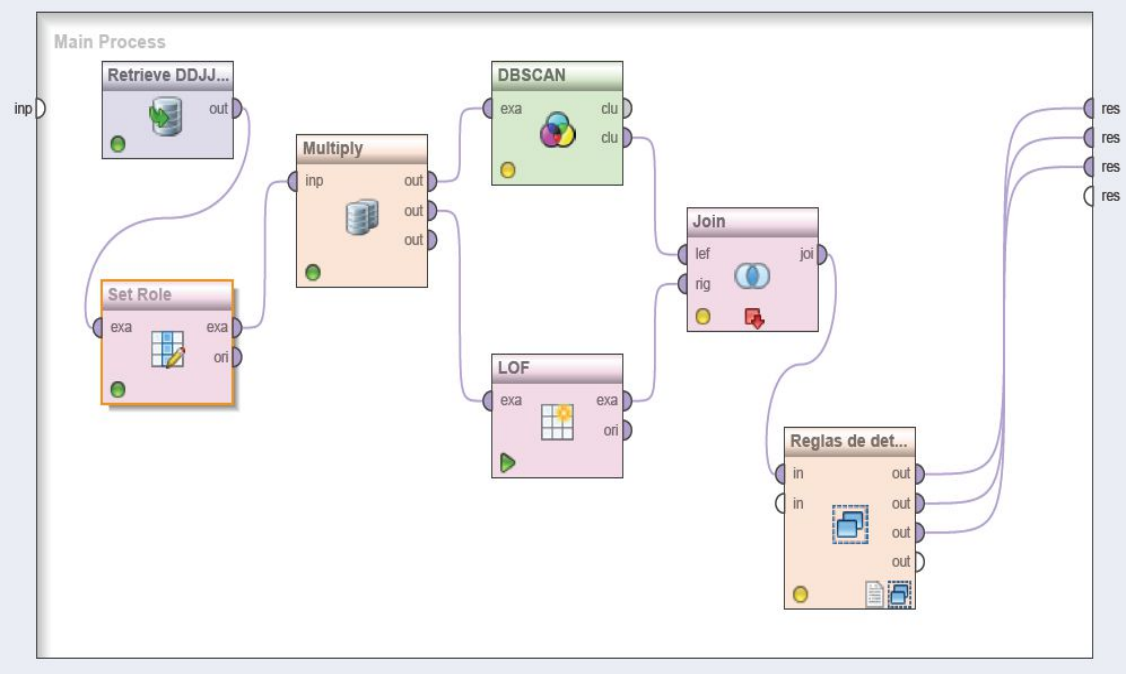


Figura 18: Flujo de minería en RM, (Procedimiento IV[a]).

Una vez fusionado los resultados de los escaneos en la búsqueda de campos anómalos se se programan las reglas de determinación de outliers como subproceso del anterior de acuerdo al límite estándar en el valor de LOF de 1.5. Para ello se generan los atributos `Valor_LOF` y `Valor_DBSCAN` (Operador **Generate Attribute**) con expresión de función '`outlier`' para el primero y '`cluster`' para el segundo, asignando *a posteriori* un rango '`false`' [0.0, 1.5]

mediante el operador **Numerical to Binominal** sobre el nuevo atributo '`Valor_LOF`'. Luego se reemplazan los valores '`false`' y '`true`' derivados por `0` y `1` en ambos atributos para luego fusionar ambos atributos en uno solo mediante el operador Generate Concatenation. Renombrando la concatenación en un solo atributo denominándolo '`tipo_outlier`' reemplazamos su denominación (**Replace**) siguiendo el procedimiento, según fuese este outlier simple, doble o no outlier.

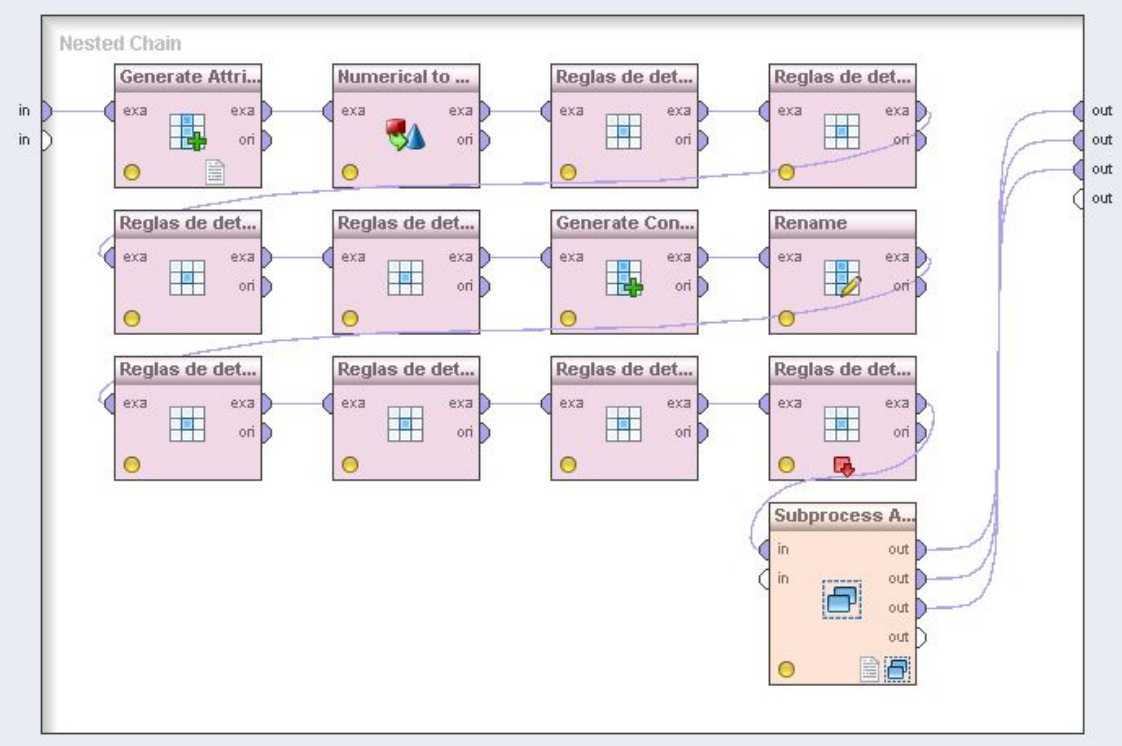


Figura 19: Flujo de minería en RM, (Procedimiento IV[b]).

Igualmente que en el subproceso anterior de la Figura 19 se hace lo propio para las otras variantes de umbrales, como se plasma en la siguiente figura, los atributos de outliers para las tres distintas variantes de outliers según su valor LOF (**Generate Attribute**), aún vacíos de contenido, se multiplica para cada uno de los tres casos (**Multiply**) los tres umbrales LOF correspondientes para cada caso quebrando en los umbrales 1.475 -valor LOF 5% inferior-, 1.575 -valor LOF 5% superior- y 1.65 -valor LOF 10% superior- mediante el operador '**Numerical to Binominal**', como se aprecia en la continuación del flujo en la Figura 20.

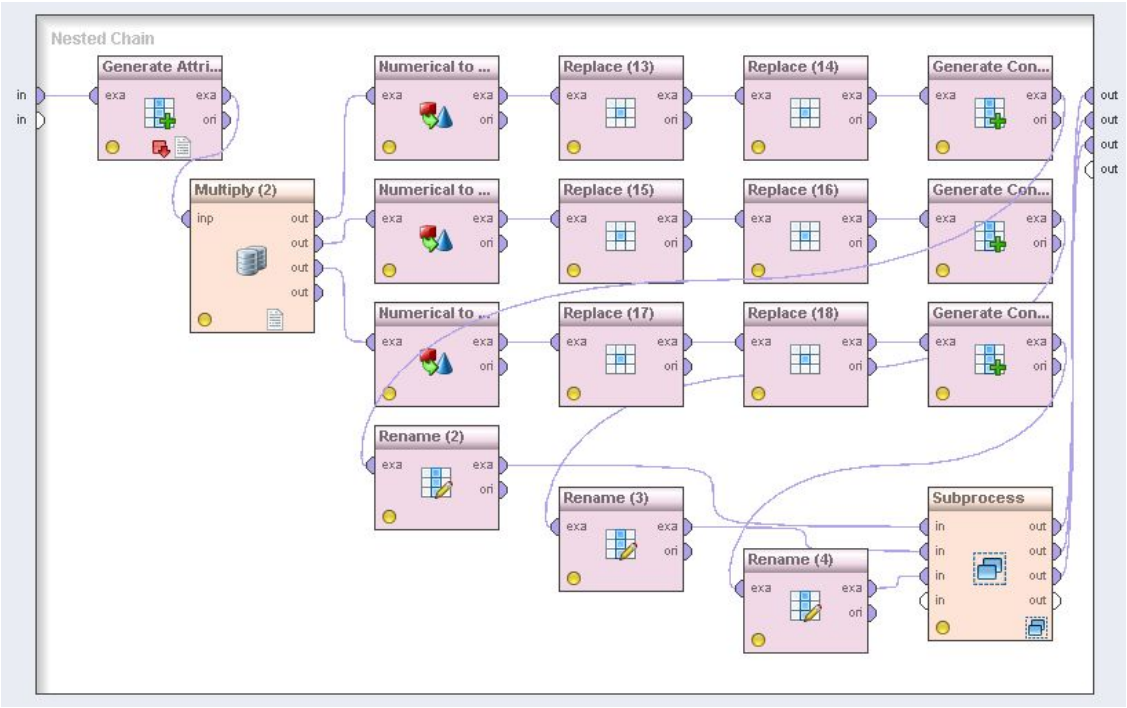


Figura 20: Flujo de minería en RM, (Procedimiento IV[c]).

Ut supra en la Figura 20, se aprecia la multiplicación del flujo de minería para cada caso de umbral de LOF y así crear un atributo de '`tipo_outlier`' para cada uno, para luego proceder con la nominalización de los valores binomiales de tales atributos, de acuerdo a los valores arribados; ergo, a partir de las combinaciones de 0 y 1, se reemplaza los valores binarios como '`no_outlier`' (`0_0`), '`outlier_simple`' (`0_1`, `1_0`) y '`outlier_doble`' (`1_1`) como describe la Figura 21.

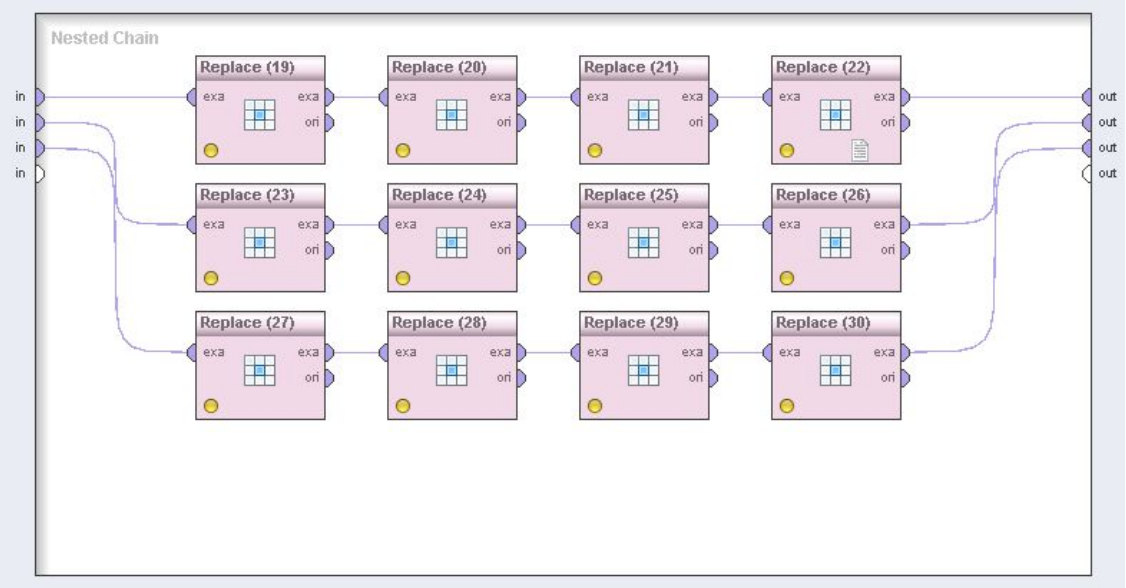


Figura 21: Flujo de minería en RM, (Procedimiento IV[d]).

De esta forma, hasta aquí, se plasman operativamente las reglas de determinación de outliers para LOF y DBSCAN en RM de la sección A.1.2. del Anexo.

Contando ahora con una BD de DDJJ con outliers clasificados para cada umbral correspondiente, se procede ahora en la siguiente etapa del procedimiento IV en su aproximación , el cual contempla el uso de algoritmos de clasificación sobre los outliers detectados *a priori*, la unión de los mismos en su utilidad de los resultados y la aplicación de reglas de determinación de outliers sobre tal unión, se procede de la siguiente manera.

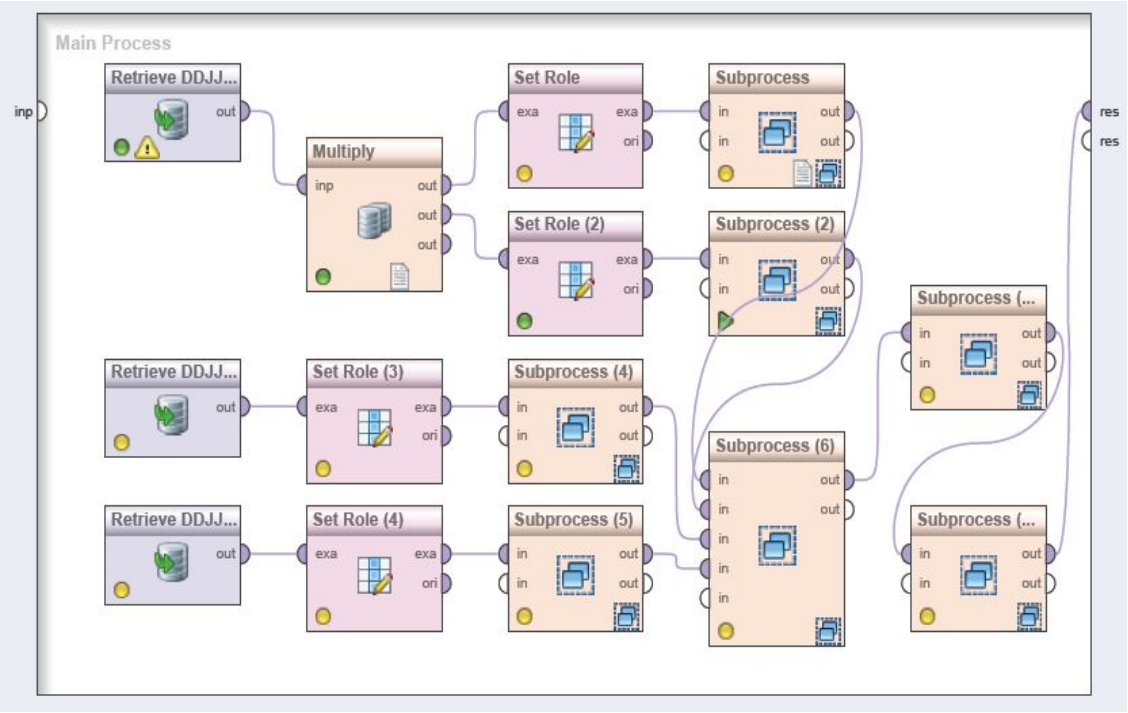


Figura 22: Flujo de minería en RM, (Procedimiento IV[e]).

De la Figura 13, donde hay tres BBDD con los cuatro (4) umbrales de LOF para la determinación de outliers; para cada caso, se setea un rol de atributo '`label`' para *a posteriori* usar algoritmos de clasificación con los outliers como label en los subprocesos (0-2-4-5). Si bien las tres BBDD tienen incorporado el antributo binomial valor_LOF en 1,5 como umbral, en la BD DDJJ-BD1 hacemos correr el flujo de minería tanto para el valor de LOF estándar como para su 5% superior, como se detalla a continuación.

BD	Umbral LOF	Nombre atributo
DDJJ-BD1	`1,5`	`valor_LOF`
DDJJ-BD1	`1,575`	`valor_LOFsup5%`
DDJJ-BD2	`1,65`	`valor_LOFsup10%`
DDJJ-BD3	`1,475`	`valor_LOFinf5%`

Tabla 15: Variantes de BBDD según umbral LOF(Procedimiento IV).

Previa a a la unión se aplica los algoritmos de clasificación C4.5 (operador **Decision Tree**), redes bayesianas (**Naive Bayes**) y PRISM (**Rule Induction**) de donde se estiman los valores de ocurrencia de outliers mediante el operador **Apply Model** con la misma BD de DDJJ del subproceso superior correspondiente a un umbral de LOF determinado para cada algoritmo, q.e.:

tres Retrieve de BD DDJJ repetidas para cada algoritmo de clasificación como se describe inmediatamente a continuación en la Figura 23.

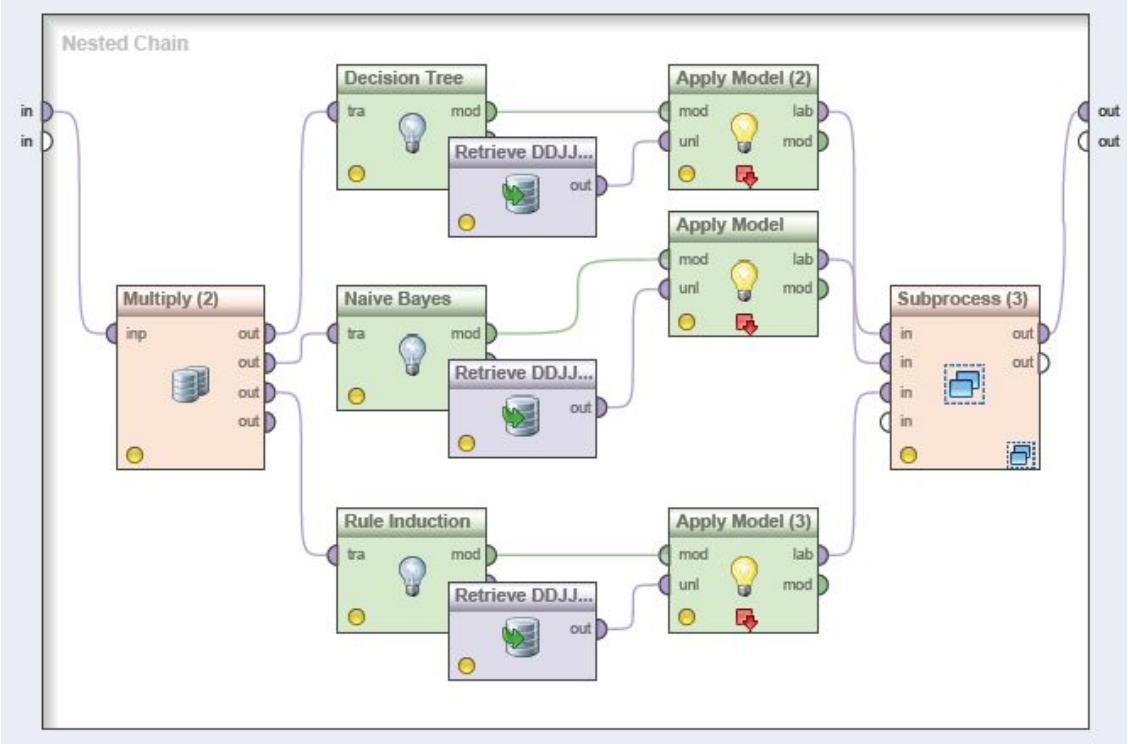


Figura 23: Flujo de minería en RM, (Procedimiento IV[f]).

En los procesos posteriores se procede a unificar, mediante los comandos **Join** con configuración de tipo '`inner`', primero todos los valores de predicción luego aplicar los modelos como se describe a continuación a partir de los pasos inmersos todos en el subproceso (Subprocess (3-7-8-9)), todos estos, análogos al siguiente proceso *ut infra*.

En este, se genera un atributo identificador para cada Subproceso 3-7-8-9, los cuales gestionan los flujos resultantes de aplicar los modelos predictivos de los algoritmos de clasificación; donde: se genera un atributo ID (**Generate ID**), se configura un rol para el nuevo atributo predictivo

como atributo regular (**Set Role**), se lo renombra (**Rename**) y se lo une (**Join**), como se ilustra a continuación.

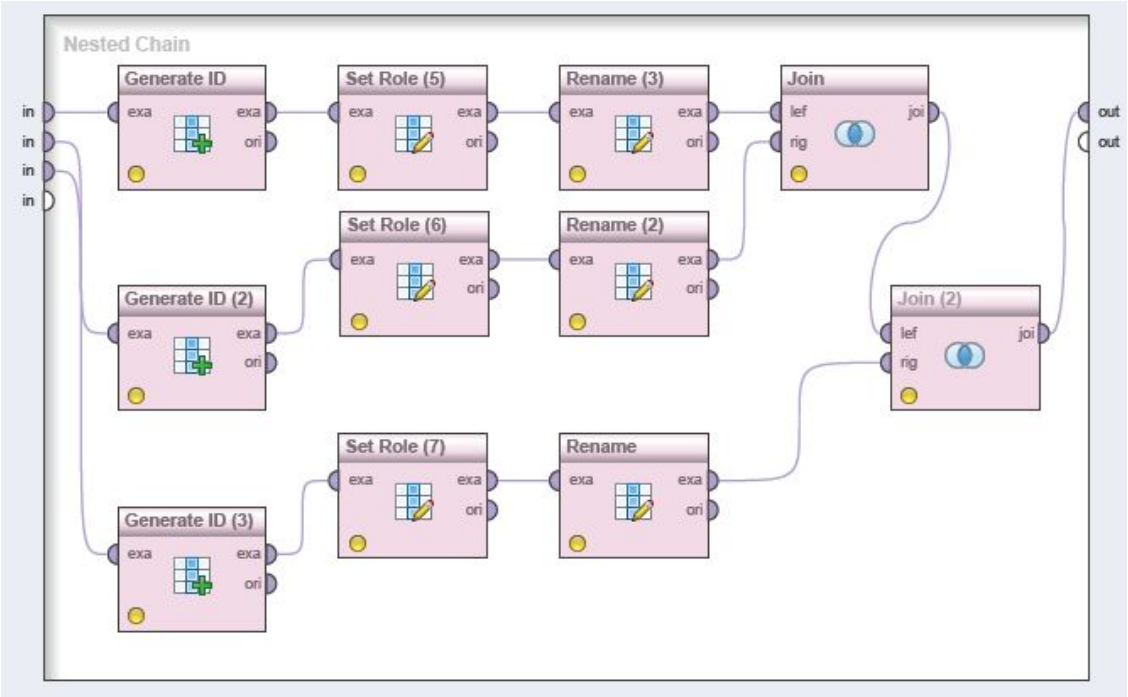


Figura 24: Flujo de minería en RM, (Procedimiento IV[g]).

En la Figura 24 se mergen todas las BBDD en una sola, con Join de tipo `'inner'`, de manera de proceder y concretar la unión todos los flujos de minería confluyen en el suproceso (6) donde la unión se hace efectiva a partir de la siguiente configuración y aplicación de operadores **Join** sucesivos.

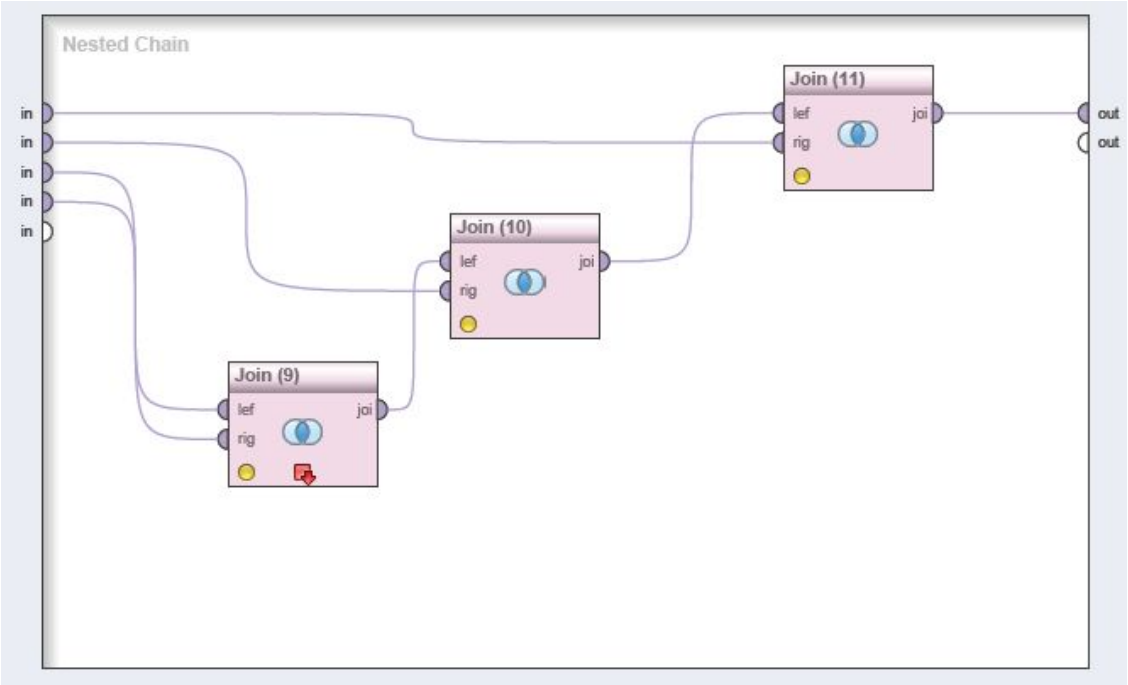


Figura 25: Flujo de minería en RM, (Procedimiento IV[h]).

Si bien ahora se tienen los valores estimados de los algoritmos de clasificación en una sola BD, aún no se los ha unido en un solo atributo para su tratamiento posterior acorde al procedimiento.

Para lograr una unión de los resultados precedidos por los algoritmos de clasificación se procede a una fusión concatenada de sus valores mediante el operador '**Generate Concatenation**' como se vislumbra a continuación.

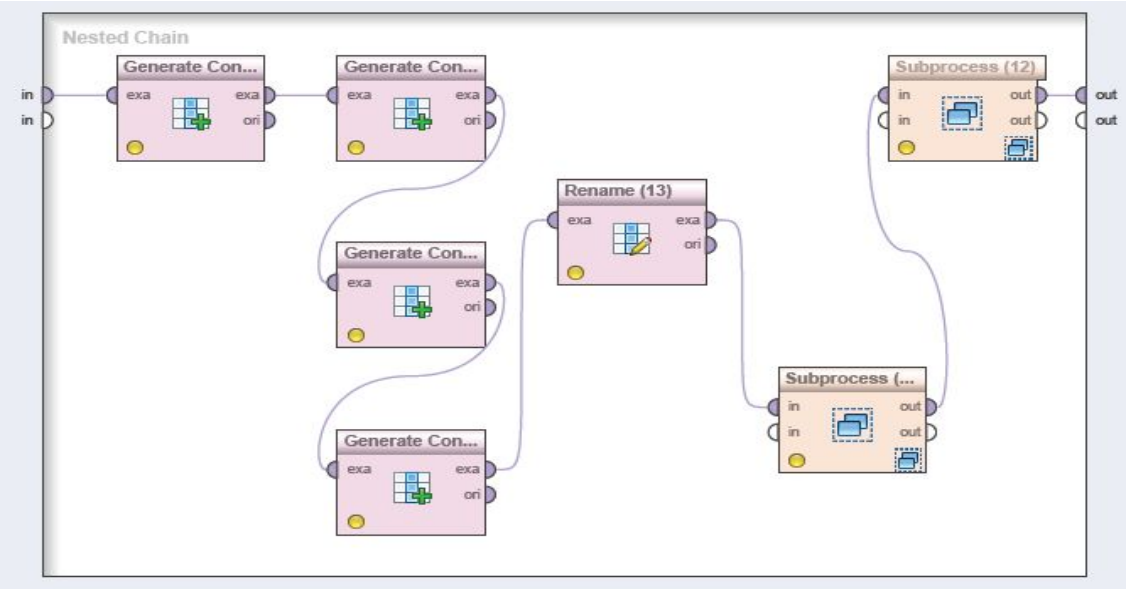


Figura 26: Flujo de minería en RM, (Procedimiento IV[i]).

Luego de la concatenación se renombra (**Rename** (13)) el atributo fusionado como `tipo_outlier_union` para *a posteriori*, en el subproceso (**Subprocess** (11)) se procede al reemplazo de la clasificación descripta en la sección A.2.1 según las mencionadas reglas en el binomio de tuplas '`outlier`' y '`limpio`' como describe la figura *ut infra* en la Figura 27.

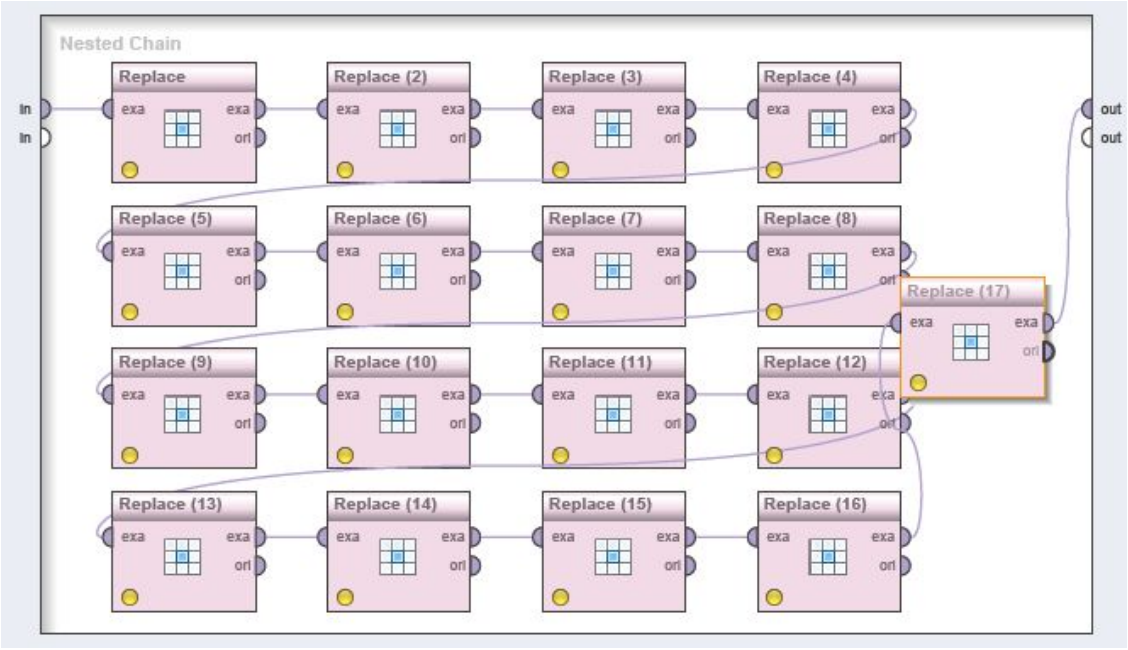


Figura 27: Flujo de minería en RM, (Procedimiento IV[j]).

(→Replace(x17)), y luego, mediante el subproceso (12), tomar solo los atributos originales. Ya reemplazando y renombrado las tuplas los de manera de cumplir con las reglas de determinación de outliers del Anexo A.2.1.

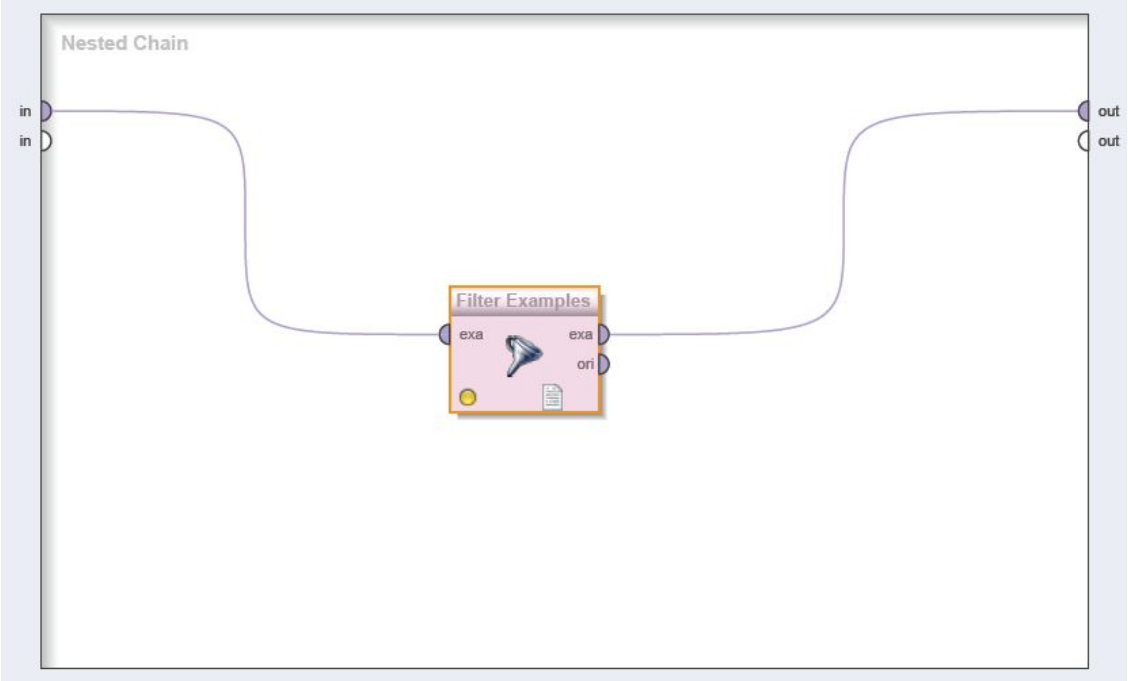


Figura 28: Flujo de minería en RM, (Procedimiento IV[k]).

(→ Filter Examples)

Una vez ya que la BD quedó reducida solamente a las tuplas outlier producto de los algoritmos de clasificación y el filtrado de las reglas de determinación de anomalías, pernoctando de 6627 tuplas a 2531; puesto que ahora, se filtraron las tuplas 'limpias' quedando la BD conformada solo con las tuplas 'outliers' (Operador **Filter Examples**) como se ilustra en la Figura 28. Siguiendo el procedimiento final de la Sección 4.3.2 -la aplicación última de LOF y K-Means para clusterizar los atributos sospechosos de contener campos anómalos-, se reduce la BD seleccionando los atributos originales para luego transponer la BD y creando un nueva (BD(b)), como se presenta a continuación en la performance de los dos primeros operadores.

Ya transpuesta la nueva BD(b), se aplica LOF, obteniéndose una nueva columna con valores LOF correspondiente a cada atributo. Notar que después de la transposición los atributos originarios de las DDJJ (**Transpose**) se presentan ahora en la forma de tuplas del atributo '`id`', el cual luego se renombra como '`atributos_ddjj`' para no crear confusión y purificando seleccionando y filtrando la nueva base de solo dos atributos. Se genera un atributo '`id`' para filtrar tuplas/atributos sobrantes no correspondientes a la BD original de DDJJ preparada (Operador **Generate ID**) y luego se filtran según el nuevo rango id que así lo posibilita (**Filter Example Range**).

Equivalente a la siguiente secuencia: (Select Attribute → Traspose → Detect Outlier (LOF) → Select Attribute → Rename → Generate ID →Filter Example Range).

Finalmente, se prosigue con la clusterización con algoritmo K-Means (**Clustering**) , programándose el operador K-Means de la siguiente forma en la Figura 29:

```
K = 2
max run = 10
```

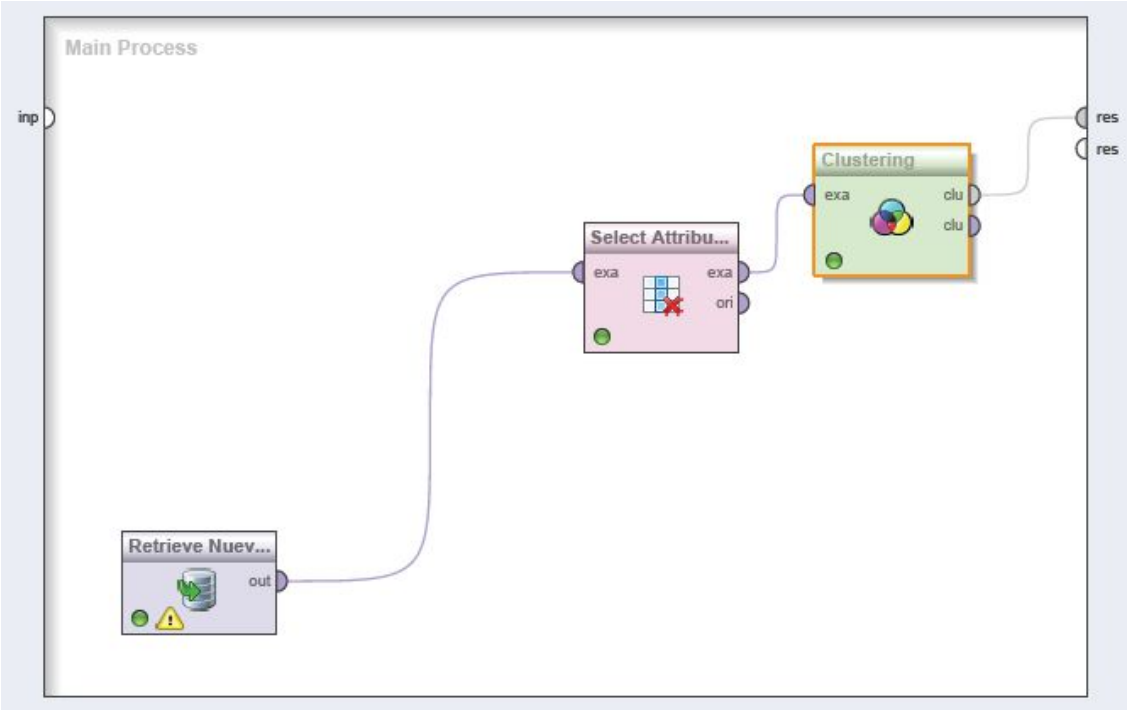


Figura 29: Flujo de minería en RM, (Procedimiento IV[l]).

La imagen de arriba correspondiente a la Figura 29, secuencia Retrieve (BD) → Select Attribute (outlier) → Clustering (K-Means)

De donde se obtuvieron los resultados de la Sección 5.4, identificándose los atributos sospechosos de cargar anomalías resultado del bi-agrupamiento programado[15].

[15] Notar que aquí se calcula la distancia del centroide para el LOF de cada atributo transpuesto; dado que al intentar hacerlo columna por columna; nos encontramos con valores infinito (infinity), los cuales impiden el calculo posterior de agrupamiento con K-means. Como propuesta de procedimiento adicional tales valores *infinity* podrían reemplazarse por un valor numérico extremo o por algún otro valor en máximo, o valor tope predeterminado, que permita el cálculo de la distancia columna por columna.

C. GLOSARIO DE ATRIBUTOS EN BBDD DE DDJJ

ano:	Período anual al que refiere la d.j. del funcionario público
tipo-ddjj:	Refiere al tipo de DDJJ del funcionario público
url:	URL a la imagen de la d.j. original
poder	Poder republicano al que pertenece el funcionario público.
apellido	Apellido/s del funcionario público.
nombre:	Nombre/s del funcionario público.
nacimiento:	Fecha de nacimiento del funcionario público.
egreso:	Fecha de egreso del funcionario del cargo público.
ingreso:	Fecha de ingreso del funcionario del cargo público.
cargo:	Cargo público que ocupa el funcionario.
jurisdicción:	Jurisdicción al que se encuentra circunscrito el bien del funcionario.
barrio:	Barrio al que se encuentra circunscrito el bien del funcionario.
cant-acciones:	Cantidad de acciones por firma declaradas por el funcionario público.
descripción:	Descripción del bien declarado.
destino:	Fin al que esta destinado el bien del funcionario declarante.
entidad:	Entidad tipo del bien declarado.
fecha-desde:	Fecha desde la cual se cuenta con el bien declarado.
fecha-hasta:	Fecha desde la cual no se cuenta con el bien declarado.
localidad:	Localidad al que se encuentra circunscrito el bien del funcionario.
modelo:	Modelo de diseño del bien declarado por el funcionario.
nombre-bien-s:	Nombre del bien declarado.
origen:	Origen patrimonial de los recursos que justifican la posesión o disposición del bien.
país:	País de origen del bien declarado por el funcionario.
periodo:	Período anual de la DJ del funcionario público.
porcentaje:	Proporción porcentual del bien correspondiente a la pertenencia del mismo por parte del declarante.
provincia:	Provincia en la que se encuentra situado el bien declarado.
ramo:	Ramo al que pertenece el bien declarado por el funcionario.
superficie:	Superficie en metros cuadrados del bien declarado, en caso de ser inmueble.

unidad-medida-id:	Categorización dicotómica de la unidad de medida física en que se encuentra mensurado el bien declarado por el funcionario: metros cuadrados (0) o en hectáreas (1).
tipo-bien-s:	Tipo del bien declarado según su cualidad.
Titular-dominio:	Titular del dominio del bien declarado.
moneda-mejoras:	Tipo de moneda en el cual se efectuó alguna mejora de un bien declarado.
mejoras:	Valor monetario destinado a las mejoras del bien declarado.
moneda-valor-adq:	Tipo de moneda en el cual se efectuó la adquisición del bien declarado.
valor-adq:	Valor monetario del bien adquirido declarado.
Moneda-valor-fiscal:	Tipo de moneda en la cual se efectuó el valor fiscal del bien declarado.
valor-fiscal:	Valor fiscal monetario del bien declarado.
u-medida:	Unidad de medida física mensurado el bien declarado por el funcionario: metros cuadrados (m2) o hectáreas (ha.).
vinculo:	Tipo de vinculo filiar con el co-propietario del bien declarado.