UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE

"METODOLOGÍA PARA EL ANÁLISIS DE PROCESOS DE NEGOCIO BASADA EN MINERÍA DE PROCESOS Y DE DATOS"

ANÍBAL TOMÁS SILVA OSSES

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: **JOSÉ LUIS MARTÍ L.**

PROFESOR CORREFERENTE: **PEDRO FRANCISCO GODOY B.**

ENERO - 2017

Agradecimientos

Me gustaría darle las gracias a todas las personas que me han acompañado en mi camino por la universidad. A mi familia, especialmente a mis padres Rosa y Juan, además a mi novia Paulina, quienes han estado incondicionalmente a mi lado. También a mis compañeros: Álvaro H., Álvaro R., Carlos, Enzo, Fabián, Francisco, Guillermo, Joaquín, Nicolás, Rodrigo y Teodoro; con quienes fomentamos una ayuda mutua durante todos estos años. Por último, agradecer a los profesores, quienes han hecho posible mi formación profesional.

Resumen

En la actualidad son muchas las plataformas que proveen la creación de diferentes tipos de datos, lo que ha generado un fuerte interés por analizar y generar conocimiento a partir de ellos. Esto ha fomentado el nacimiento de diversas disciplinas especializadas en el estudio de los datos provenientes de todo tipo de fuentes (sociales, empresariales, políticas, por mencionar algunas), dentro de las que destacan la minería de datos y de procesos.

El presente documento pretende definir un marco de trabajo que permita estudiar procesos de negocio de cualquier tipo de organizaciones a partir de la minería de procesos, integrando técnicas de minería da datos de forma complementaria que permitan obtener mayor conocimiento sobre éste.

Palabras Clave: metodología, proceso de negocio, minería de procesos, minería de datos, modelo de proceso, registro de evento.

Abstract

Currently, are many platforms who test the creation of different types of data, which has generated a strong interest in analyzing and generating knowledge from them. This has fomented the birth of differents disciplines specialized in the study of data from all kinds of sources (social, business, political, to name a few), within are data mining and process.

The present document pretend establish a framework that allow to study the business processes of any type of organizations from the process mining, integrating techniques of data mining of a complementary form that allow to get the best knowledge of about it.

Keywords: methodology, bussiness process, process mining, data mining, process model, event log.

Tabla de Contenido

1 Definición del Problema **2**

1.1 Contexto 2

1.2 Problema 3

1.3 Objetivos 4

1.3.1 Objetivo General 4

1.3.2 Objetivos Específicos 4

1.4 Alcance y Limitaciones 4

2 Estado del Arte **6**

2.1 Minería de Datos 6

2.1.1 Técnicas 7

2.1.1.1 Descriptivas 7

2.1.1.2 Predictivas 8

2.1.2 Herramientas 9

2.1.3 Metodologías 9

2.2 Minería de Procesos 14

2.2.1 Ramas de Estudio 15

2.2.1.1 Descubrimiento de Procesos 15

2.2.1.2 Análisis de Conformidad 18

2.2.1.3 Enriquecimiento de Procesos 18

2.2.2 Herramientas 19

2.2.3 Metodologías 20

3 Metodología Propuesta **26**

3.1 Descripción de la Metodología 27
3.2 Etapas 28
3.2.1 Planificación 28
3.2.2 Extracción 29
3.2.3 Procesamiento 29
3.2.4 SEMMA 30
3.2.5 Filtrado 31
3.2.6 Minería de Procesos 32
3.2.7 Evaluación 32
3.2.8 Mejoramiento de Procesos 33

4 Validación del Método Propuesto 34

4.1 Planificación 34
4.2 Extracción 37
4.3 Procesamiento 39
4.4 SEMMA 41
4.4.1 Agrupamiento 41
4.4.2 Asociación 46
4.4.3 Clasificación Mediante Árboles de Decisión 51
4.4.4 Clasificación Mediante Reglas de Inducción 54
4.5 Filtrado 56
4.6 Minería de Procesos 61
4.7 Evaluación 66
4.8 Mejoramiento de Procesos 71

Conclusiones 72

Referencias Bibliográficas **76**

Anexos **81**

Índice de Figuras

Figura 1 Modelo de la metodología KDD. 10

Figura 2 Modelo de la metodología CRISP-DM. 12

Figura 3 Modelo de la metodología SEMMA. 13

Figura 4 Red de Petri compuesta por 4 actividades, en donde B y C se ejecutan en paralelo. 16

Figura 5 Modelo BPMN compuesto por 4 actividades, en donde B y C se ejecutan en paralelo. 16

Figura 6 Grafo dirigido compuesto por 4 nodos. 17

Figura 7 Árbol de proceso compuesto por 4 actividades, un elemento *AND* ($\wedge$) y uno de continuación de flujo ($\rightarrow$). 17

Figura 8 Resumen de los diferentes tipos de procesamiento de datos de la metodología PM^2. 22

Figura 9 Resumen de las actividades de la etapa de minería y análisis de la metodología PM^2. 22

Figura 10 Visión global de la metodología PM^2. 23

Figura 11 Marco de trabajo propuesto para analizar procesos de servicios financieros. 25

Figura 12 Metodología propuesta para proyectos de minería de procesos apoyado con minería de datos. 28

Figura 13 Etapa de análisis de datos de la metodología propuesta, en donde se aplica el marco de trabajo SEMMA. 31

Figura 14 Red de Petri que representa el proceso en estudio. 36

Figura 15 Gráfico de coordenadas de centroides obtenidos con el algoritmo *k-means*. 44

Figura 16 Gráfico de coordenadas de centroides obtenidos con el algoritmo *k-medoids* sobre atributos continuos. 46

Figura 17 Árboles de decisión obtenidos al aplicar al algoritmo *random forest*. 53

Figura 18 Curva ROC obtenida al aplicar validación cruzada sobre el resultado del algoritmo *random forest* con un área bajo la curva de 0,998. 54

Figura 19 Curva ROC obtenida al aplicar validación cruzada sobre el resultado del algoritmo de *reglas de inducción* con un área bajo la curva de 0,877. 56

Figura 20 Modelo de proceso para el grupo 4, el cual posee la mayoría de las anulaciones del registro de eventos. 59

Figura 21 Modelo de proceso para el grupo 5, el cual no posee anulaciones. 59

Figura 22 Modelo del proceso entregado por Disco con instancias que no tienen la actividad *rechazar contrato*. 60

Figura 23 Modelo del proceso entregado por Disco considerando solamente las instancias que tienen en alguna parte de su flujo la actividad *rechazar contrato*. 60

Figura 24 Modelo del proceso filtrado en base al patrón seleccionado del árbol de decisión generado a través del algoritmo *random forest*. 61

Figura 25 Modelo del proceso encontrado por el algoritmo *inductive miner* para el *event log* del *cluster* 4 entregado por *k-means*. 62

Figura 26 Modelo del proceso encontrado por el algoritmo *inductive miner* para el *event log* del *cluster* 5 entregado por *k-means*. 63

Figura 27 Modelo del proceso encontrado por el algoritmo *inductive miner* para el *event log* de casos que no tienen la actividad *rechazar contrato*. . 63

Figura 28 Modelo del proceso encontrado por el algoritmo *inductive miner* para el *event log* de casos que tienen en su flujo la actividad *rechazar contrato*. 64

Figura 29 Modelo del proceso encontrado por el algoritmo *inductive miner* para el *event log* de casos en donde los atributos toman valores que aumentan la probabilidad de anular el contrato, los que se detectaron con el algoritmo *random forest*. 65

Índice de Tablas

Tabla 1 Ejemplo de registro de eventos (*event log*) estándar utilizado en minería de procesos. 15

Tabla 2 Tipos de actividades presentes en cada una de las tres metodologías presentadas de minería de datos. 27

Tabla 3 Resultados de distancias al aplicar *k-means* con diferente valores de k. 43

Tabla 4 Valores de coordenadas de los centroides entregados por el algoritmo *k-means* con *k=6*. 44

Tabla 5 Resultados de distancias al aplicar *k-medoids* con diferente valores de k. 45

Tabla 6 Valores de coordenadas de los centroides entregados por el algoritmo *k-medoids* con *k=6*. 46

Tabla 7 Modelo entregado por el algoritmo *k-means*, indicando la cantidad de elementos de cada grupo y cuántos contratos fueron vendidos y anulados para cada uno. 47

Tabla 8 Matriz transpuesta de algunas transacciones (casos) del proceso en estudio, en donde cada columna indica qué actividades se realizaron (1) y cuales no (0) en cada instancia. 48

Tabla 9 Precedentes de las reglas de asociación detectadas por el algoritmo *apriori* que tienen como conclusión la actividad *anular contrato*, señalando sus valores de soporte y confianza. 49

Tabla 10 Precedentes de las reglas de asociación detectadas por el algoritmo *FP-growth* que tienen como conclusión la actividad *anular contrato*, señalando sus valores de soporte y confianza. 50

Tabla 11 Matriz de confusión obtenida al aplicar validación cruzada sobre el resultado del algoritmo *random forest*. 53

Tabla 12 Matriz de confusión obtenida al aplicar validación cruzada sobre el resultado del algoritmo de *reglas de inducción*. 56

Tabla 13 Indicadores obtenidos a partir del análisis de conformidad para los 3 resultados de minería de procesos. 64

Tabla 14 Cantidad de actividades por modelo que demoran más de 24 horas en realizarse. 66

Introducción

En la actualidad la mayoría de las organizaciones lleva adelante una producción de datos considerable, implicando que cada día las fuentes de datos son más, lo que no se ve reflejado en la utilización de estos datos, principalmente en la obtención de conocimiento a partir de ellos. Se han creado muchas técnicas y algoritmos para el análisis de datos en diversas disciplinas de estudio, mas existen pocos marcos metodológicos que señalen la utilización de éstas herramientas, sino más bien son generales y deben ser adaptados al caso de estudio. De esto se desprende también que no hay metodologías que mezclen técnicas de diferentes ámbitos del análisis de datos.

Este trabajo plantea un método de trabajo para el análisis de procesos de negocio que tiene como base la minería de procesos, y que incorpora una etapa previa de estudio a través de la minería de datos, para lo que se han considerado metodologías ya existentes de estas disciplinas. Esta iniciativa nace de la necesidad de analizar de una manera mucho más completa los procesos de negocio que las organizaciones ejecutan; un estudio netamente enfocado en datos logrará generar conocimiento no descubierto desde una perspectiva de procesos, por lo que se pretende unificar estos dos desarrollos (el de minería de procesos y de datos) con el fin de obtener un análisis mucho más completo que permita entender de mejor forma el proceso de negocio.

El contenido del documento está dividido en 4 capítulos. El primero corresponde a la definición del problema, destacando el conflicto identificado y el contexto en que se desarrolla este. El capítulo 2 detalla el estado del arte de la minería de datos y de procesos. En el tercer capítulo se expone la metodología propuesta, definiendo todos pasos necesarios. El capítulo 4 presenta la validación del método propuesto, detallando el desarrollo de todas las etapas que lo componen. Finalmente se presentan las conclusiones del desarrollo realizado, además de una sección de anexos.

1. Definición del Problema

1.1. Contexto

En la actualidad gran parte de las organizaciones generan y almacenan datos, ya sea de sus clientes, proveedores, procedimientos o fuentes externas a ellas. Esto lo logran a través de sistemas de información que les permiten trabajar y procesar dichos datos. Así, se ha vuelto una necesidad generar conocimiento con el fin de mejorar su negocio y aportar mayor valor a sus clientes para este tipo de organizaciones.

El concepto de "ciencia de los datos" ya se ha instaurado en el mundo académico y profesional, haciendo referencia a todas las disciplinas y técnicas que se dedican a trabajar con grandes volúmenes de datos y obtener conocimiento de ellos. Dos de éstas áreas son la minería de datos [32] y la minería de procesos [29], las que, respectivamente, se encargan de obtener conocimiento de un fenómeno específico que aqueja a una organización, y estudiar los procesos de negocio de ésta misma a partir de los datos que generan las actividades que componen su flujo.

La minería de datos está fuertemente relacionada con el aprendizaje de máquinas (o automático) [10], permitiendo crear modelos computacionales que ayuden a predecir qué ocurrirá con instancias futuras de los datos estudiados, describir ciertas cualidades o descubrir patrones del negocio que representan los datos. Por otra parte, la minería de procesos tiene un rol más específico, enfocándose en inferir información de los procesos que son relevantes para las organizaciones. Esta rama de la ciencia de los datos nace a partir de la gestión de procesos de negocio [33] (*Business Process Management*), metodología que tiene como objetivo mejorar el desempeño de los procedimientos de dichas organizaciones, lo que se complementa con análisis de datos a través de diferentes algoritmos dedicados a esta área.

1.2. Problema

La cantidad de datos que generan las organizaciones está en aumento gracias a los sistemas de información que utilizan para apoyar sus labores; pero estos datos no están siendo procesados como debería, por lo que dichas entidades están perdiendo oportunidades de mejora en lo que compete a sus negocios, además de no satisfacer de forma óptima las necesidades de sus clientes. Complicando aún más la situación, conseguir un análisis apropiado de los datos en general no es fácil, se requiere un amplio entendimiento del fenómeno que se quiere estudiar y el origen de los datos, además de conocer y comprender el funcionamiento de las herramientas disponibles para realizar el análisis que se necesita, de forma tal de que los resultados obtenidos sean de calidad.

En general, previo al desarrollo de cualquier proyecto de estudio de datos, es necesario hacer las siguientes preguntas para orientar el trabajo: ¿qué conocimientos se desean?, ¿qué tareas se requerirán para obtener dichos conocimientos?, y ¿qué criterios serán utilizados para evaluar los resultados? El entender qué información se desea obtener permite determinar cuáles son las técnicas que abordarán el problema, con lo que se conseguirán resultados medibles en base a criterios debidamente especificados a través de las técnicas ejecutadas.

En la literatura pueden encontrarse diversas metodologías que proponen cómo implementar un correcto análisis de datos, principalmente para minería de datos; algunas de éstas son: KDD, CRISP-DM, SEMMA y Catalyst (P3TQ) [15]. De igual forma, pero en menor cantidad, ocurre con la minería de procesos, en donde se destacan metodologías como *Process Mining Project* (PM^2) [42] o *Life-cycle Model for Mining Lasagna Processes* (L*) [30] para cualquier tipo de procesos, u otras más específicas como la propuesta planteada en [19], la cual está orientada a procedimientos relacionados con salud. Con lo anterior es posible afirmar que hay muchas formas de cómo abordar un problema relacionado con análisis de datos, pero no existe un método o pauta bien

definido en donde se mezclen las dos disciplinas anteriores. Esto es lo que motiva el desarrollo de este trabajo, poder integrar 2 disciplinas de análisis de datos en un solo marco metodológico.

1.3. Objetivos

1.3.1. Objetivo General

Definir una metodología que integre la minería de datos y de procesos que permita apoyar el desarrollo de la utilización de técnicas de análisis de datos, siempre apuntando al entendimiento y mejora del negocio basado en lo que la organización dueña del proceso considere necesario.

1.3.2. Objetivos Específicos

- Establecer etapas concretas que conformarán la metodología propuesta especificando en qué consiste cada una, para así tener una guía clara de cómo implementar el método en futuras instancias.
- Determinar qué tareas de minería de datos son las que generan un conocimiento que aporte de mayor forma al análisis del proceso en cuestión en etapas posteriores de la metodología.
- Aplicar la propuesta a un caso de estudio real, para así entender y mejorar el proceso principal de una organización.

1.4. Alcance y Limitaciones

El presente trabajo no hace hincapié en la optimización de los algoritmos utilizados o cuál de los software empleados es el más indicado. Además, se ha seleccionado un conjunto limitado de técnicas de minería de datos para el desarrollo, procurando las que

sean mayormente útiles para el análisis desde una perspectiva descriptiva y no predictiva.

Sobre la validación de la propuesta, sólo se utilizará un caso de estudio, el que puede concluir de forma positiva o negativa para el dueño del proceso. Por esto cabe mencionar que existen muchos otros escenarios en que la metodología puede ser aplicada, en donde podrán ser requerido variar ciertos aspectos o utilizar técnicas que en este trabajo no serán empleadas.

La última etapa del marco propuesto (explicado en el capítulo 3) no podrá ser ejecutada, ya que no se dispone del acceso a la organización relacionada al caso de estudio más allá del análisis de sus datos y hacer propuestas de mejora en base a éste.

2. Estado del Arte

En este capítulo se definen los principales avances y elementos más relevantes de las disciplinas minería de datos y de procesos. Se especifican algoritmos y técnicas, metodologías de aplicación y algunas herramientas que permiten su uso.

En la actualidad el análisis de datos está presente en diversas áreas de estudio, considerando como base la estadística tradicional [20], abarcando hasta disciplinas capaces de entender fenómenos que ocurren en cualquier tipo de conjunto de datos como la minería de datos [28], o predecir hechos a través del aprendizaje de máquina [41]. También engloba estudios más específicos, como el de comprender grandes volúmenes de datos conectados entre sí gracias a la rama de redes complejas [2], o el análisis de procedimientos de negocios de entidades de diferentes ámbitos a través de la minería de procesos [22].

En la literatura existen muchas otras áreas de interés en el estudio de datos, mas este documento se centrará en la minería de datos y de procesos, por lo que a continuación se detalla una breve descripción de cada una de estas disciplinas.

2.1. Minería de Datos

La minería de datos [43] es una disciplina que tiene como objetivo obtener nuevos conocimientos en conjuntos de datos de gran tamaño. Para lograr esto hace uso de otras ramas del estudio de datos, como la estadística, la inteligencia artificial y el aprendizaje de máquinas. La minería de datos está compuesta por muchas diferentes técnicas y metodologías, las que permiten un uso estandarizado a través de un esquema bien definido. Los diferentes marcos de trabajo presentan etapas similares, como entendimiento del problema y de los datos, procesamiento de datos, modelado, y evaluación. Sobre las técnicas de minería de datos, existen dos grandes grupos, las que describen hechos

y las que predicen acontecimientos.

El uso de la minería de datos es transversal; así, en entornos financieros se puede usar para describir patrones de compra de clientes en grandes almacenes, o predecir por parte de un banco si un cliente es apto para un crédito o no. En ciencia, sirve para analizar redes genéticas, para predecir el desarrollo de enfermedades, entre otros. A continuación se definen las principales técnicas de minería de datos que permiten estos usos, dividiéndolas en descriptivas y predictivas.

2.1.1. Técnicas

2.1.1.1 Descriptivas

Técnicas destinadas a describir características en un conjunto de datos o patrones que sean interesantes de analizar. También son denominadas métodos de aprendizaje no supervisado, ya que no necesitan modelos a entrenar (de los que se esperaría una predicción en registros futuros) no existiendo una variable objetivo para la cual pronosticar un valor. Dentro de estas técnicas destacan:

- Agrupamiento: se encarga de agrupar los registros en base a diversos criterios, siendo éstos generalmente distancias entre los atributos que conforman el conjunto de datos (para valores continuos) o algún tipo de similitud en los valores que pueden tomar estas variables (para valores discretos). Los elementos de un mismo grupo (*cluster*) comparten características, lo que permite analizar éstos como un conjunto y obtener conocimientos específicos de ellos mismos.

- Asociación: esta técnica entrega como modelo expresiones de la forma $X \Rightarrow Y$, en donde X corresponde a una premisa e Y a una conclusión (si se cumple X, se cumplirá Y), pudiendo determinar relaciones de diferentes elementos que pertenecen a un mismo conjunto de datos. Los principales algoritmos que generan reglas de asociación se basan en 2 métricas para su ejecución, el soporte y la

confianza; el soporte es la proporción de elementos del conjunto que cumplen cierta característica sobre Y y X dentro de todas las transacciones, y la confianza corresponde a la frecuencia con que aparece Y dentro de las que cumplen con X.

2.1.1.2 Predictivas

Técnicas destinadas a construir modelos que realicen predicciones sobre un hecho particular a través de un conjunto de datos de entrenamiento. También son denominadas métodos de aprendizaje supervisado, ya que a partir de datos históricos (conjunto de entrenamiento) es posible generar modelos capaces de tomar otros datos de entrada y predecir qué sucederá con esos registros. Dentro de estas técnicas destacan:

- Clasificación: asigna una etiqueta a un registro que se desea clasificar en base a un modelo obtenido previamente a partir de un conjunto de datos, al cual se le denomina conjunto de entrenamiento, ya que "entrena" a un modelo para que prediga la etiqueta que tendrá un nuevo registro; además el modelo debe ser validado con un conjunto de prueba, el que debe estar compuesto por datos que no fueron utilizados en el entrenamiento. Los algoritmos más comunes de clasificación involucran: árboles de decisión, que corresponden a modelos gráficos que permiten entender qué sucederá con el nuevo registro en base a diferentes valores que toman los atributos utilizados en el modelo; reglas de inducción, sentencias de la forma "*si* condición *luego* conclusión", lo que permite predecir el resultado en base a un precedente; redes neuronales artificiales, conjunto de neuronas (nodos) conectadas entre sí, en donde cada una tiene varias entradas y salidas, asignando pesos a estos parámetros que permiten generar un modelo matemático que predice algún resultado.

- Regresión: esta técnica genera un modelo matemático (polinomio) que permite conocer el valor que tendrá una variable a predecir (dependiente) en base a otras (independientes). Al igual que en clasificación, requiere de un conjunto de entrenamiento que permita generar este modelo y otro de prueba para poder validarlo.

Además de la representación matemática del modelo y, dependiendo de la cantidad de atributos de los registros, es posible graficar el modelo encontrado y poder realizar análisis de forma cualitativa.

2.1.2. Herramientas

La minería de datos ya se ha posicionado como una técnica que apoya diferentes ámbitos de las organizaciones, como la toma de decisiones, predicciones requeridas y entendimiento de los registros, entre otras. Es de esperar entonces que existan diversas aplicaciones donde poner en práctica esta disciplina, tanto comerciales como de código abierto. Dentro de las herramientas con licencias pagadas se mencionan: SAS [34] e IBM SPSS Modeler [35], entre muchas otras; y de software libre se deben mencionar: Rapidminer [40], WEKA [43], Orange [37] y KNIME [36], entre otras. Cabe destacar que todas las aplicaciones mencionadas poseen una interfaz visual que facilita el uso de las técnicas de minería de datos, pudiendo ser utilizadas no sólo por expertos. Además de lo anterior, existen otra opción para realizar análisis de datos, como lo son los lenguajes de programación R [39] y Python [38], los que se usan en gran medida para estudios estadísticos.

Tener que hacer la elección de alguna de las herramientas antes mencionadas podría ser tarea difícil, pero esta labor debe basarse en el objetivo del proyecto. Primero que todo se debe identificar qué conocimiento se desea, con lo que se podrá saber qué técnica aplicar, además de elegir una aplicación que soporte el algoritmo a utilizar.

2.1.3. Metodologías

Considerando las diversas formas que pueden existir de cómo aplicar minería de datos, es que a través de su existencia también han surgido diferentes marcos de trabajo que proponen una pauta para realizar proyectos de esta disciplina. Las metodologías más destacadas se definen a continuación.

***Knowledge Discovery in Databases* (KDD)**

Se describe como un proceso no trivial de descubrimiento de información y conocimiento que puede ser útil a partir de un conjunto de datos, como se ejemplifica en [31]. La figura 1 [1] muestra las etapas de esta metodología, las cuales son:

1. Selección: se determinan cuáles serán las fuentes de datos y cuáles de éstos serán utilizados. Para ello se debe comprender qué tipo de conocimiento se quiere obtener.

2. Preprocesamiento: se deben preparar los datos ya seleccionados, ya sea limpiarlos, corregir los inconsistentes, y determinar qué hacer con los faltantes, entre otras acciones; esto con el fin de tener una estructura estándar de registros.

3. Transformación: se realiza un tratamiento a los datos previa generación de modelos, lo que puede requerir generar nuevas variables, agregación de datos, y normalización de éstos, entre otras acciones.

4. Minería de Datos: aquí se deben aplicar las técnicas conocidas para obtener modelos o patrones que representen o generen conocimiento útil a partir de los datos.

5. Interpretación y Evaluación: se detectan qué conclusiones entregadas por los modelos son relevantes, a través de evaluaciones sobre los resultados obtenidos.

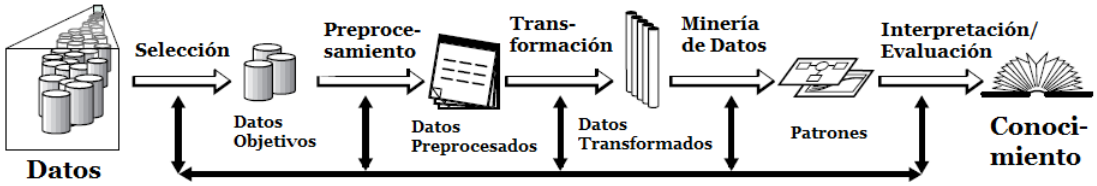

Figura 1: Modelo de la metodología KDD.

Como se ve en la figura 1, este marco de trabajo es iterativo, por lo que siempre es posible volver a una etapa anterior al concluir con alguna interpretación de los resulta-

[1] www.ceine.cl

dos y generar nuevos modelos en base a los creados anteriormente.

***Cross Industry Standard Process for Data Mining* (CRISP-DM)**

Siendo la metodología más utilizada en las organizaciones que requieren realizar minería de datos [27], CRISP-DM establece una pauta de cuáles pasos seguir en la realización de un proyecto de minería de datos. La figura 2 [2] muestra el modelo que describe este marco de trabajo, el que está compuesto por seis etapas, las que corresponden a:

1. Comprensión del Negocio: etapa en que se deben entender los objetivos del proyecto desde un punto de vista organizacional, para luego poder representar éstos en base a un problema de minería de datos.

2. Comprensión de los Datos: similar a la etapa de selección de KDD, se deben entender los datos en base al problema identificando cuáles serán utilizados.

3. Preparación de los Datos: corresponde a las etapas de preprocesamiento y transformación de KDD, en donde se deben limpiar los datos, agregarlos, tratar con datos faltantes, entre otras tareas (no existiendo un orden en estas actividades).

4. Modelado: se aplican las técnicas seleccionadas de minería de datos para obtener los modelos. Si es necesario se debe volver a la etapa anterior.

5. Evaluación: se analizan los modelos obtenidos en la etapa anterior en base a criterios establecidos.

6. Implantación: fase en que se implementan las acciones inferidas a partir de la información entregada por los modelos.

De igual forma que en KDD, CRISP-DM es un proceso iterativo, el que podrá realizarse tantas veces se crea necesario, ajustando los datos utilizados y los parámetros de las técnicas para obtener resultados diferentes.

[2] www.inteligenciamik.wikispaces.com

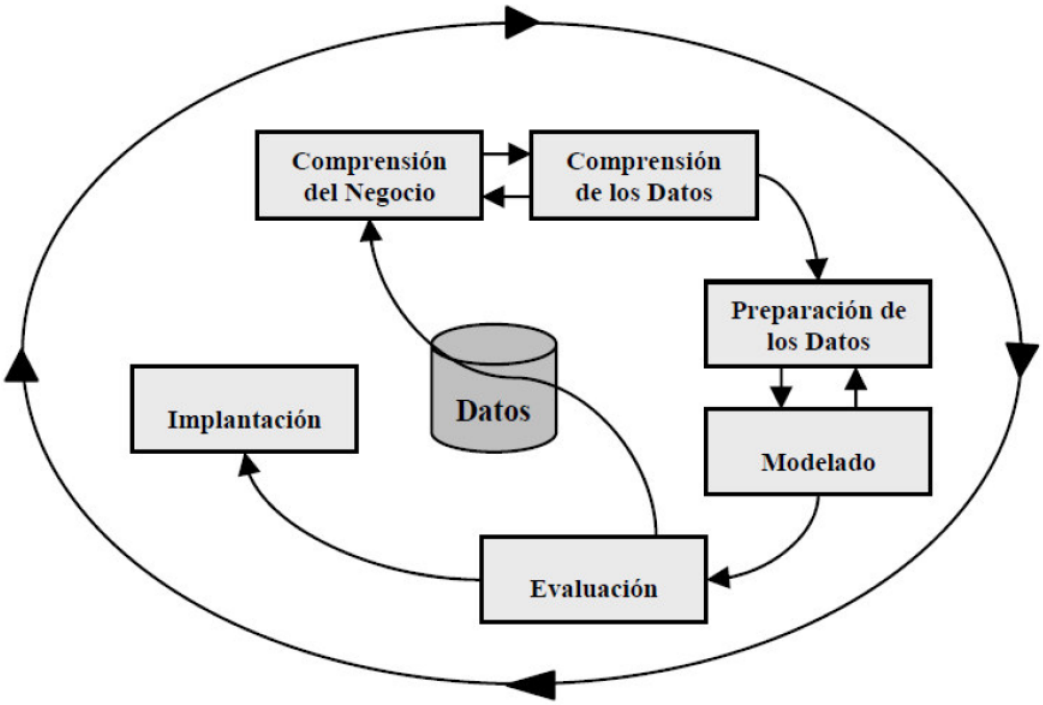

Figura 2: Modelo de la metodología CRISP-DM.

***Sample*, *Explore*, *Modify*, *Model*, *Assess* (SEMMA)**

Como se menciona en [26], la metodología SEMMA está enfocada en la aplicación de minería de datos y no en los objetivos organizacionales, a diferencia de CRISP-DM. Las etapas que componen esta metodología se muestran en la figura 3 [3] y son las que conforman su nombre.

1. Muestreo (*Sample*): fase opcional, en que se rescata una muestra de los datos que represente a la población de forma correcta o en base a alguna especificación.

2. Exploración (*Explore*): en esta etapa se deben inspeccionar los datos, detectando cuáles pueden ser anómalos (incompletos, erróneos, entre otros). Se puede realizar a través de medios visuales, apoyándose en algún *software*, con métodos estadísticos o cualquier técnica que permita la exploración de datos.

[3] www.actividadsenadisenocubosdedatos.blogspot.cl

3. Modificación (*Modify*): consiste en modificar o crear nuevos datos de ser pertinente, así poder ejecutar los modelos de los cuales se desea obtener resultados.

4. Modelado (*Model*): al igual que en las etapas de modelado y de minería de datos de las metodologías CRISP-DM y KDD, respectivamente, en esta fase se deben aplicar las técnicas que generarán modelos que permitan comprender las problemáticas planteadas al comienzo del proyecto.

5. Evaluación (*Assess*): etapa en que se deben evaluar los resultados obtenidos en el modelado, estimando la bondad de los resultados a través de diversas formas.

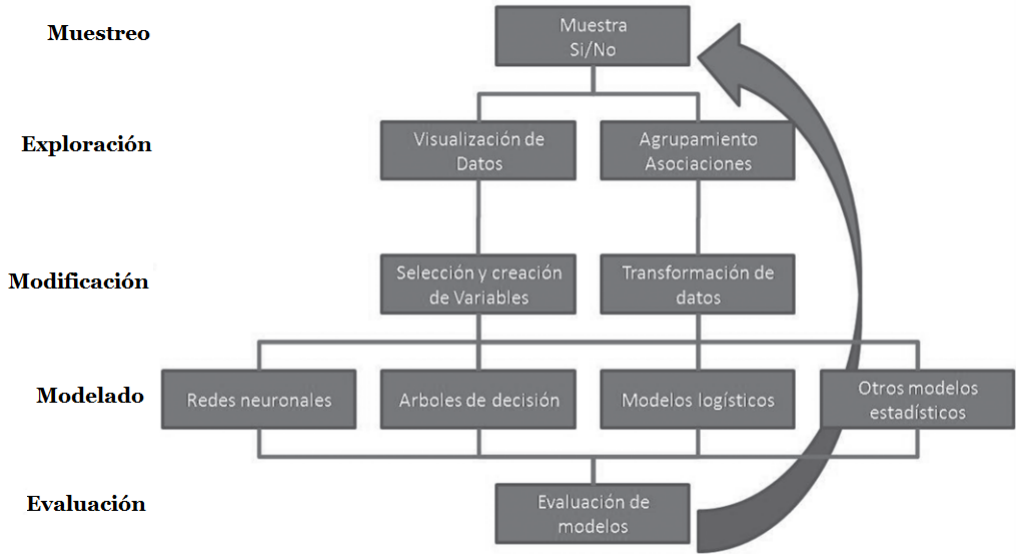

Figura 3: Modelo de la metodología SEMMA.

Con la presentación de estas tres propuestas de metodologías, se puede identificar que son similares; tienen etapas de entendimiento, procesamiento, modelado y análisis de resultados, además de ser procesos cíclicos que iteran mientras sea necesario. La diferencia más importante que puede señalarse, como se menciona en [30], es que CRISP-DM es la única que apunta a un entendimiento acabado del negocio previo trabajo o análisis sobre los datos; en contraste con KDD y SEMMA, en donde la primera etapa ya consiste en manipular los datos. Se destaca entonces un punto fuerte en CRISP-

DM, ya que el trabajo que conlleva aplicar este marco es enfocado en la problemática del negocio.

2.2. Minería de Procesos

El paradigma de desarrollo organizacional orientado a procesos ya se ha instaurado en entidades que impulsan una mejora continua en su negocio. Para lograr lo anterior ciertos elementos son requeridos, siendo fundamentales los sistemas de información. En base a los datos que estas aplicaciones generan nace la minería de procesos [2], una disciplina que intenta extraer conocimiento a partir de los registros de eventos que se generan en las diferentes actividades que componen los procesos, permitiendo el análisis a partir de los datos.

Como se menciona en [1], los procesos están conformados por una secuencia de actividades que se realizan de manera ordenada, las que deben comprender el funcionamiento de alguna organización que busca administrar los recursos involucrados en cada una de estas tareas. El elemento fundamental de minería de procesos son los registros de evento (*event log*), los que guardan características de un evento realizado en un momento determinado. Entre los atributos comúnmente utilizados están: identificador del caso, nombre de actividad, ejecutor que la realiza, y fechas de inicio y término (*timestamps*). Un conjunto ordenado de estos registros conforma un caso (traza), que corresponde a una instancia del proceso. La tabla 1 ejemplifica un registro de eventos sencillo, en donde cada identificador (*case id*) corresponde a una instancia o ejecución del proceso, las *actividades* a las diferentes tareas desarrolladas en él, los *recursos* a la persona que ha desarrollado dicha actividad, y los atributos *inicio* y *término* a las fechas en que se ejecutó la misma tarea. La creación de los *event log* debe ser pensada considerando el problema del negocio que se estudia; se debe comprender éste para determinar qué elementos aportarán información al estudio. Esta disciplina ha sido aplicada en diversas industrias, por ejemplo, en el área de la salud [7] y de la educación [8]; y en

diferentes tipos de sistemas, como son los sistemas legados de datos [18] o en plataformas colaborativas como SharePoint [1].

Tabla 1: Ejemplo de registro de eventos (event log) estándar utilizado en minería de procesos.

Case ID	Actividad	Recurso	Inicio	Término
1	*A*	a	01/01/2016	02/01/2016
1	*B*	a	02/01/2016	05/01/2016
2	*A*	b	01/02/2016	03/02/2016
1	*C*	a	07/01/2016	10/01/2016
2	*C*	c	06/02/2016	11/02/2016
3	*A*	c	18/02/2016	20/02/2016
1	*D*	a	15/01/2016	16/01/2016
3	*B*	d	21/02/2016	22/02/2016
2	*B*	c	06/02/2016	07/02/2016
3	*C*	c	25/02/2016	28/02/2016
3	*D*	d	01/03/2016	02/03/2016
2	*D*	b	13/02/2016	20/02/2016

La minería de procesos permite realizar tres tipos de análisis: el descubrimiento de procesos, la verificación de conformidad y el enriquecimiento de procesos, las que se explican en la siguiente sección.

2.2.1. Ramas de Estudio

2.2.1.1 Descubrimiento de Procesos

Consiste en encontrar modelos que describan los procedimientos, siendo las notaciones más utilizadas las que se listan a continuación, donde cada ejemplo representa

el registro de eventos de la tabla 1.

- Redes de Petri [24]: corresponde a una representación gráfica de un conjunto de eventos relacionados en un flujo que se describe a través de tres elementos: actividades (cuadrados), lugares o *places* (círculos), y fichas o *tokens* (círculo negro) dentro de los lugares (ver figura 4).

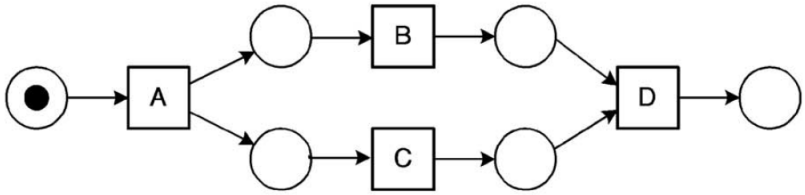

Figura 4: Red de Petri compuesta por 4 actividades, en donde B y C se ejecutan en paralelo.

- BPMN (*Business Process Model and Notation*) [12]: notación gráfica estandarizada que permite la creación de modelos de procesos de negocio. Posee una gran cantidad de elementos que permiten modelar diversas características de los procedimientos. La figura 5 ejemplifica esta notación.

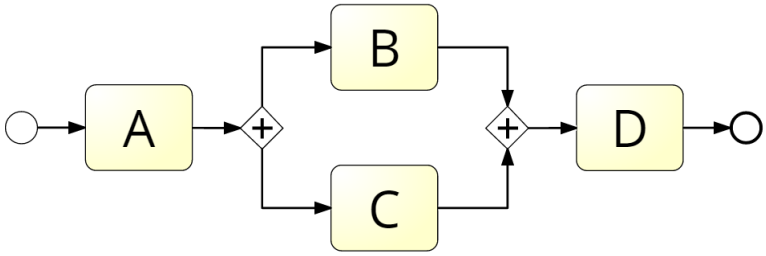

Figura 5: Modelo BPMN compuesto por 4 actividades, en donde B y C se ejecutan en paralelo.

- Grafos Dirigidos [13]: conjunto de nodos conectados entre sí a través de aristas, en donde éstas tienen sólo un sentido de flujo. La figura 6 consiste en un grafo dirigido de 4 nodos.

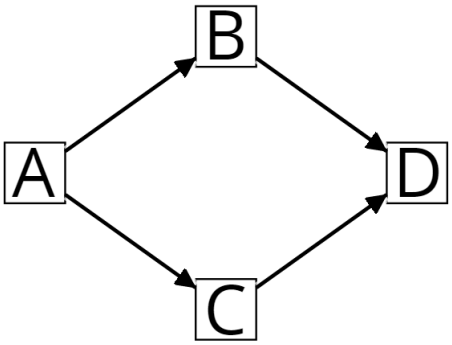

Figura 6: Grafo dirigido compuesto por 4 nodos.

- Árboles de Procesos [3]: al igual que un grafo, un árbol es un modelo en que sus nodos están conectados pero ahora con una jerarquía definida, en donde existe un único nodo raíz y varios nodos hoja. En el caso de la figura 7, el elemento de *continuación* (flecha superior) representa la raíz y las actividades las hojas.

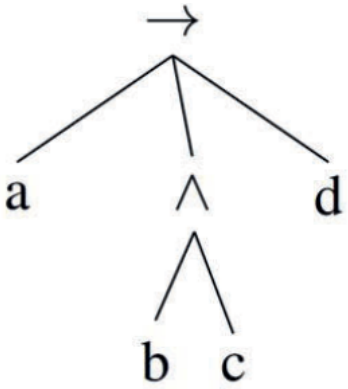

Figura 7: Árbol de proceso compuesto por 4 actividades, un elemento AND (∧) y uno de continuación de flujo (→).

Gran parte de los algoritmos empleados en el descubrimiento de procesos entregan una red de Petri como salida, la que en la mayoría de los casos es capaz de describir de buena forma el proceso. Pero esta notación no soporta todas las características posibles, como la relación *OR* entre las actividades (se realizar una, algunas o todas las tareas). Es por esto que se requería de notaciones más elaboradas que abarcaran todos los casos, como BPMN o árboles de procesos. Para conseguir los modelos anteriores

existen distintos algoritmos, siendo los más utilizados: *alpha miner* [16], *heuristic miner* [17], *genetic mining* [4], *fuzzy miner* [11] e *inductive miner* [14], entre otros. Cada uno de ellos posee características diferentes y generan modelos en las notaciones antes expuestas.

2.2.1.2 Análisis de Conformidad

El segundo tipo de estudio de procesos (complementario al descubrimiento) es el de verificación de conformidad, el que consiste en contrastar un modelo ya creado (puede ser a través de técnicas de descubrimiento) con un registro de eventos, con el objetivo de entender qué tan similar son los datos que el proceso genera contra alguna ejecución del proceso representada en el modelo utilizado. Esto se mide a través de dos principales indicadores, el ajuste (*fitness*) y la precisión, los que son obtenidos gracias al algoritmo *token replay* expuesto en [21].

- Ajuste: establece si es posible replicar el registro de eventos en el modelo dado, a través de un valor que varía entre 0 y 1; mientras más cercano a 1 se estará reproduciendo un mayor número de trazas de forma correcta.
- Precisión: mide la generalidad de un modelo en base a los datos. También utiliza un valor entre 0 y 1; mientras más cercano a 1 el modelo estará ajustado a los datos de buena forma, o en otras palabras, será menos general.

2.2.1.3 Enriquecimiento de Procesos

Permite ampliar el conocimiento de los procedimientos a partir del uso de información del registro de eventos que no es requerida obligatoriamente para realizar minería de procesos, específicamente, para el descubrimiento y en el análisis de conformidad.

Esta información puede corresponder a tiempos de ejecución de las actividades, relaciones de los ejecutores de las tareas y cálculo de probabilidad de flujo de las diferentes instancias de los procesos, entre muchas otras opciones sujetas a los datos que se tengan.

Otra forma de comprender la utilidad de la minería de procesos es destacar las aristas del proceso que esta disciplina estudia, las que pueden agruparse en cuatro fundamentales [22]:

- Perspectiva de Flujo: consiste en analizar el orden y relación de las actividades, lo que puede asociarse con el descubrimiento de modelos.
- Perspectiva de Caso: permite identificar las características de cada traza del proceso, así poder establecer similitudes o diferencias de éstas.
- Perspectiva Temporal: permite analizar los tiempos de ejecución de las distintas etapas del proceso, por ejemplo, saber cuánto demora una actividad en realizarse o conocer el tiempo de espera entre dos ramas del proceso que se deben efectuar en paralelo.
- Perspectiva Organizacional: analiza diferentes relaciones que tienen los ejecutores de las actividades que componen el proceso, como pueden ser el trabajo en conjunto, traspaso de tareas, subcontratación (entregar y recibir la realización de una actividad a un mismo ejecutor) y ejecución de tareas similares (ejecutores con mismo rol); pudiéndose determinar cada una de estas relaciones a través de los algoritmos [25]: *working together*, *handover of work*, *subcontracting* y *similar task*, respectivamente.

2.2.2. Herramientas

Para aplicar la minería de procesos, se puede mencionar que existen herramientas enfocadas en la utilización de técnicas que componen los distintos tipos de análisis ya

descritos. Las dos principales herramientas son ProM [23] y Disco [9]; la primera es un framework de código abierto que permite realizar el descubrimiento de modelos, el análisis de conformidad, y otros tipos de estudios siempre bajo la perspectiva de minería de procesos. En el caso de Disco, ésta es una aplicación comercial que permite generar un entendimiento inicial del proceso gracias a su simple interfaz y fácil uso; además, brinda otras funcionalidades para analizar más en detalle el proceso, permitiendo realizar el descubrimiento de éste, la aplicación de filtros o analizar el desempeño temporal, entre otras características. En el capítulo 4 del presente documento serán utilizados los dos software aquí descritos.

2.2.3. Metodologías

La minería de procesos es una rama del análisis de datos relativamente nueva, en donde lo presentado en [16] corresponde al comienzo de esta disciplina (año 2003). Considerando esto, y en contraste con la minería de datos, es de esperar que no existan muchos métodos formales que propongan una pauta de cómo realizar un proyecto de minería de procesos. En el artículo denominado PM^2: *A Process Mining Project Methodology* [3] se expone una metodología sencilla y estándar que permite un desarrollo completo de estudio de procedimientos a través de minería de procesos, la que se compone de las siguientes etapas:

1. Planificación: etapa en que es requerido estudiar el negocio y saber cuáles son los datos que éste genera, con el objetivo de identificar cuáles serán adecuados para crear el registro de eventos que permitirá realizar todo el análisis. Lo que se hace es seleccionar el proceso de negocio que interesa analizar y mejorar; luego, se deben identificar las preguntas de investigación entendiendo los objetivos del análisis; y por último, se debe crear el equipo adecuado de individuos, tanto analistas de procesos como expertos del negocio, que sea capaz de generar conocimiento que permita responder las interrogantes antes planteadas.

2. Extracción: aquí lo primero es determinar el alcance que los datos tendrán con-

siderando los objetivos establecidos en la etapa anterior; luego, se deben extraer los datos desde los diferentes sistemas de información. En paralelo, los expertos del proceso deben traspasar los conocimientos de éste a los analistas, para así estar pronto para comenzar el análisis. En este paso es necesario entender adecuadamente el objetivo del análisis a realizar, además de comprender la causa de un problema específico que la organización posee, para así no extraer datos equivocados o que puedan no aportar al análisis que se realizará.

3. Procesamiento de Datos: en esta etapa se deben crear vistas de los registros de eventos necesarias para responder las interrogantes, con lo que es posible generar agregaciones entre éstos con el fin de reducir la complejidad del análisis y mejorar los resultados de la minería de procesos. Posteriormente, se requiere enriquecer el registro de eventos con el objetivo de generar mayor información al momento de realizar el análisis. Por último, es necesario filtrar el *log* para reducir nuevamente la complejidad del análisis y que los resultados sean más claros. La figura 8 [4] muestra una mirada global de esta etapa.

4. Minería y Análisis: ésta es la etapa en donde se deben aplicar las técnicas de minería de procesos que permitan realizar los tres tipos de análisis antes descritos (descubrimiento de procesos, análisis de conformidad y enriquecimiento de procesos. Además es posible realizar cualquier otro tipo de análisis sobre los datos para comprender de mejor forma el proceso. La figura 9 [4] muestra una mirada global de esta etapa.

5. Evaluación: luego del análisis, se deben entender estos resultados. Para ello, en primera instancia es requerido diagnosticar el desarrollo, lo que implica interpretar correctamente los resultados, distinguir cuáles de éstos son interesantes o no, e identificar si las preguntas de investigación iniciales requieren ser iteradas para

[4] Van Eck, M. L., Lu, X., Leemans, S. J., & Van Der Aalst, W. M. (2015, Junio). *PM 2: A Process Mining Project Methodology, In International Conference on Advanced Information Systems Engineering (pp. 297-313). Springer International Publishing.*

un mejor entendimiento. Luego, se deben comparar los resultados con la información original y con los elementos claves del proceso o ejecutores de él. Esta etapa, además, permitirá intuir qué ideas son posibles para mejorar el proceso.

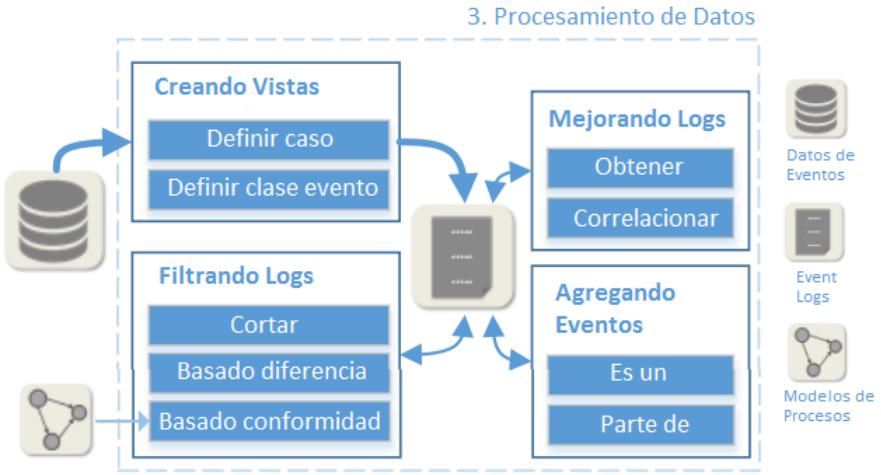

Figura 8: Resumen de los diferentes tipos de procesamiento de datos de la metodología PM^2.

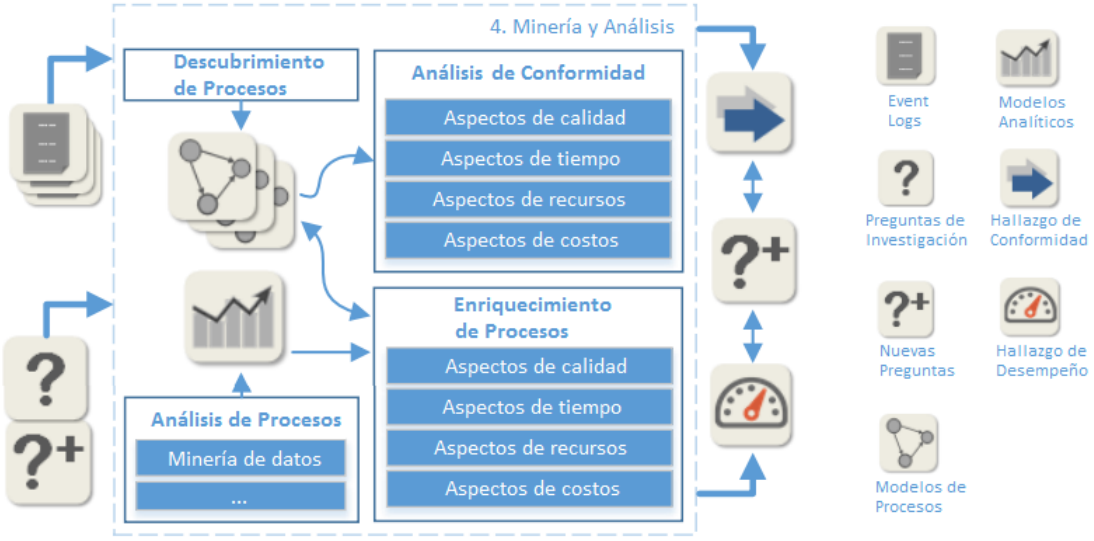

Figura 9: Resumen de las actividades de la etapa de minería y análisis de la metodología PM^2.

6. Mejora de Procesos y Soporte: consiste en implementar las mejoras detectadas para el proceso y soportar las operaciones de éste. Uno de los principales mo-

tivos de que exista la minería de procesos es lograr perfeccionamiento, por lo que esta última etapa es muy relevante. Además, es posible realizar mediciones de estas mejoras hechas al proceso a través de otros proyectos de análisis. Por último, apoyar las operaciones a través de minería de procesos es fundamental ya que permite detectar posibles problemas en la ejecución de los casos. Esto se logra a través de sistemas de información que soportan el proceso (midiéndolo y gestionándolo), mejorando la calidad de los resultados del procedimiento.

Como resumen de las etapas recién expuestas, en la figura 10 [4] se presenta el flujo de éstas y los principales elementos con los que interactúa.

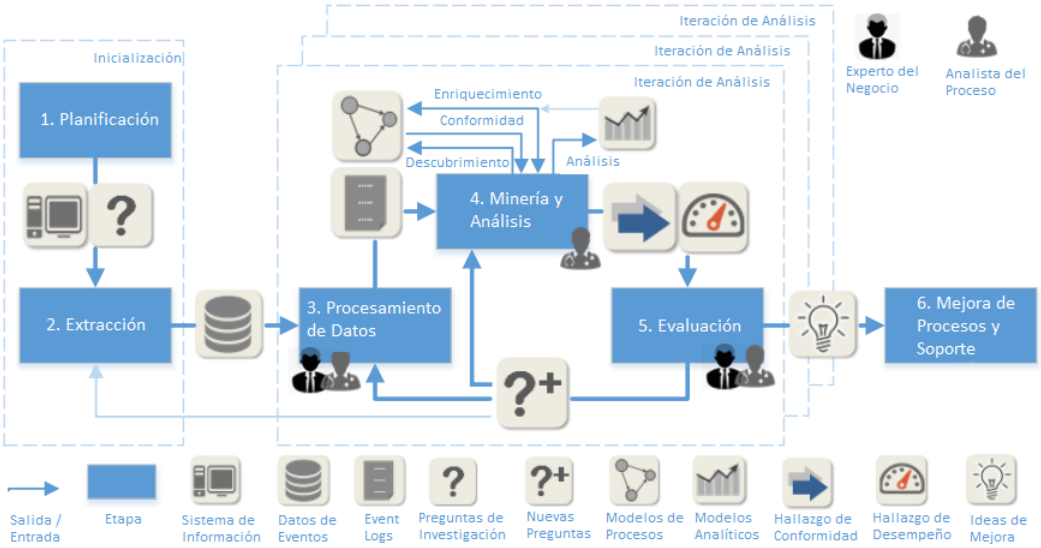

Figura 10: Visión global de la metodología PM^2.

Otra metodología bien definida es la que se propone y utiliza en [5], denominada *Process Mining Methodology Framework* y que está representada en la figura 11 [5]. Ésta fue validada en una organización que presta servicios financieros, aunque no se limita su uso a otro tipo de procesos de negocio. Las etapas que en este modelo se especifican son similares a las del anterior; se puede destacar que existe una fase iterativa de explo-

[5] De Weerdt, J., Schupp, A., Vanderloock, A., & Baesens, B. (2013). *Process Mining for the multi-faceted analysis of business processes—A case study in a financial services organization. Computers in Industry, 64(1), 57-67.*

ración de los datos, lo que es bastante útil ya que permite poner a prueba si el registro de eventos que se extrajo fue el indicado o no. Este marco divide el análisis en tres diferentes aristas del proceso: control de flujo, perspectiva de casos y organizacional, en donde todas pasan por un análisis de conformidad y de desempeño; y finaliza con una guía de posibles mejoras al proceso, análogo a la etapa 6 en PM^2. Las etapas de esta metodología se explicitan a continuación:

1. Preparación: fase en que se deben preparar los datos para el análisis. Está constituida por la extracción de los registros desde las diferentes fuentes, el preprocesamiento de los datos y la creación del registro de eventos.

2. Exploración: se indaga y conoce el conjunto de datos. De ser necesario, se utiliza alguna herramienta estadística que apoye la exploración.

3. Definición de perspectivas: se divide en tres enfoques que se pueden trabajar en paralelo, control de flujo (análisis relacionado con el descubrimiento de procesos), casos (entendimiento del procedimiento en base a sus casos) y análisis organizacional (relaciones entre los ejecutores del proceso). En esta etapa se pretende crear la base del análisis de dichas perspectivas.

4. Análisis: se realiza el análisis de descubrimiento de las perspectivas anteriores, para posteriormente profundizar en él a través de algoritmos de conformidad y desempeño sobre el proceso.

5. Resultado: en base a todo el análisis previo, se proponen pautas que permitan mejorar el proceso estudiado.

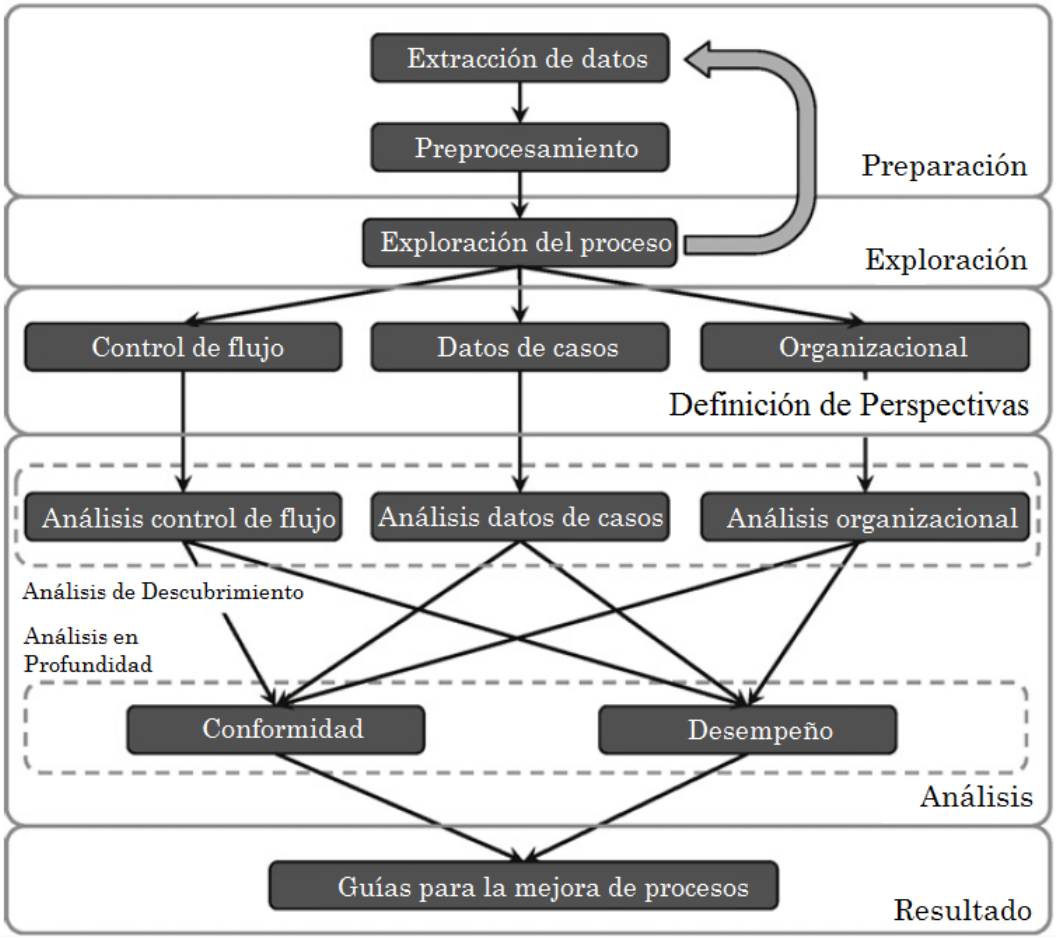

Figura 11: Marco de trabajo propuesto para analizar procesos de servicios financieros.

3. Metodología Propuesta

Como ya se mencionó en la introducción de este documento, uno de los objetivos del trabajo es plantear una metodología para proyectos de minería de procesos que esté apoyado con técnicas de minería de datos. Así, se han considerado los métodos estudiados en el capítulo anterior para generar una base al marco de trabajo que se propone.

Respecto a minería de procesos, si se comparan las dos metodologías vistas, se puede establecer que PM^2 es mucho más general y capaz de abarcar procesos con cualquier tipo de información, inclusive cuando no existan atributos como los ejecutores o los tiempos de ejecución de cada actividad, hecho que impediría hacer el análisis organizacional y de control de flujo con perspectiva temporal, respectivamente (los que se explicitan en el modelo propuesto en el caso de estudio de servicios financieros). Otro punto favorable de PM^2 es su fase iterativa, la que permite volver a plantear el *event log* luego de evaluar los resultados obtenidos de la etapa de análisis.

Sobre las metodologías de minería de datos, se podría pensar que las etapas que las componen son de diversas índoles, mas tienden a repetirse modificando pequeños aspectos, además de sus nombres. La tabla 2 muestra los tipos de actividades que KDD, CRISP-DM y SEMMA tienen, agrupadas por sus características en común.

En general las tres metodologías poseen fases similares: seleccionan datos, los que luego son procesados para obtener modelos, que posteriormente son evaluados. Aún así, es fácil notar que CRISP-DM es un marco de trabajo mucho más completo ya que incluye etapas de comprensión del negocio y de implementación; en cambio, KDD y SEMMA no le dan importancia a estas etapas.

Tabla 2: Tipos de actividades presentes en cada una de las tres metodologías presentadas de minería de datos.

Tipo de Etapa	KDD	CRISP-DM	SEMMA
Comprensión del Negocio	NO	SÍ	NO
Selección de Datos	SÍ	SÍ	SÍ
Procesamiento de Datos	SÍ	SÍ	SÍ
Modelado de Patrones	SÍ	SÍ	SÍ
Evaluación de Modelos	SÍ	SÍ	SÍ
Implementación de Mejoras	NO	SÍ	NO

3.1. Descripción de la Metodología

El marco metodológico utilizado en este trabajo es una mezcla entre los métodos PM^2 y SEMMA. El primero compete a minería de procesos, y se ha escogido ya que es capaz de aplicarse en proceso de negocio con diversas características y con cualquier tipo de datos sin problema, además de tener etapas que permiten un entendimiento del negocio hasta la implementación de las posibles mejoras. Por otra parte, se cree que SEMMA es un buen complemento para PM^2 ya que como se mencionó, este método no enfatiza en el negocio ni en las mejoras, sino en la aplicación de la minería de datos, lo que sí hace la metodología de minería de procesos. Entonces pensar en utilizar CRISP-DM sería redundar en las actividades de comprensión del negocio y de los datos, además de implementar mejoras que quizás no aporten nada al proceso, ya que éstas serán deducidas desde un punto de vista de los datos y no de los procedimientos. Así, el método propuesto en este trabajo es la unión de PM^2 con SEMMA.

La figura 12 describe la propuesta de este documento, detallando el flujo entre las actividades que la componen.

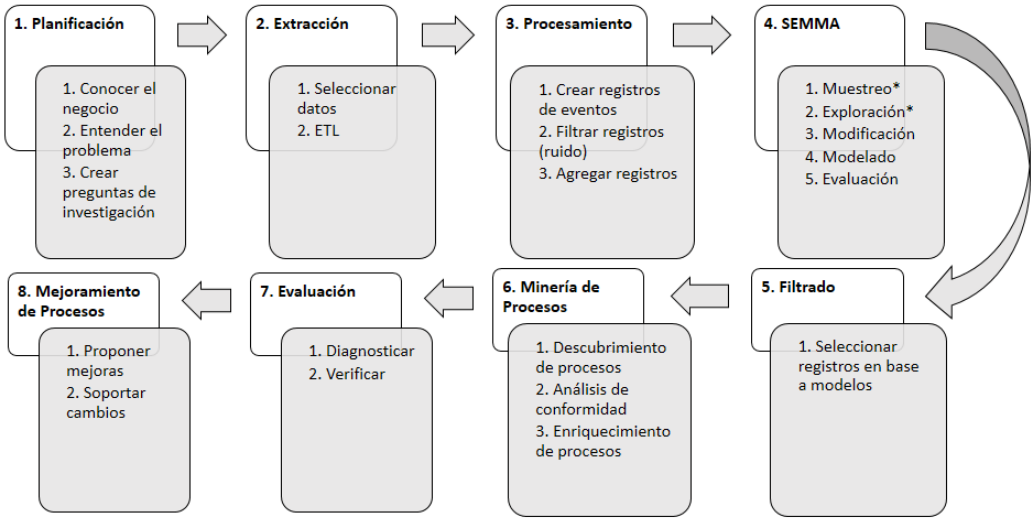

Figura 12: Metodología propuesta para proyectos de minería de procesos apoyado con minería de datos.

Cabe destacar que en la figura 12 la etapa 4 no es idéntica a la metodología SEMMA; en esta ocasión se le ha dado la característica de opcional a la fase de exploración (explore), ya que la comprensión de los datos se hizo previamente y se cree que volver a analizar los datos sólo sería necesario en casos puntuales. Otro aspecto a considerar del diagrama de la figura 12, es que sólo posee muestra el flujo principal. Esto ya que es permitido que desde todas las etapas se pueda volver a alguna anterior, como ocurre con los métodos ya presentados.

3.2. Etapas

A continuación se presentarán con detalle las etapas del método, haciendo énfasis en la fase 4, la que corresponde al trabajo de minería de datos.

3.2.1. Planificación

Esta primera actividad debe obtener el conocimiento necesario para poder trazar todo el trabajo requerido en el proyecto. Lo primero es comprender el negocio y cuál es el proceso que será analizado. En general, serán organizaciones de diversos rubros

las que requerirán de un estudio de minería de procesos, por lo que es de esperar que tengan más de un proceso de negocio relevante; entonces es de suma importancia el entendimiento inicial y la selección del procedimiento adecuado. Luego, se debe conocer el problema que se desea analizar a través de preguntas de investigación, para las cuales se espera encontrar respuesta a partir del proyecto; además es bastante beneficioso, para guiar el desarrollo del trabajo, definir hipótesis posibles relacionadas con el problema. Por último, se deben considerar todos los datos disponibles y validar cuáles serán adecuados para generar los *event logs* necesarios para responder a las preguntas planteadas.

3.2.2. Extracción

Ya con un conocimiento de cuáles son los datos disponibles para el análisis y qué es lo que representan, el primer paso en esta etapa es seleccionar los que serán útiles en el desarrollo del proyecto (los que podrán dar respuesta a las preguntas de investigación). Luego se da inicio a la extracción como tal, en donde se deberán considerar las diferentes fuentes de datos y entender cómo unificar éstas con el fin de obtener registros de eventos que representen el proceso de negocio de forma correcta. De ser necesario, se deberán trabajar los datos extraídos tratando con problemas como valores faltantes, erróneos y no estandarizados, entre otros conflictos. Dependiendo de cómo se procesarán los registros para obtener los *event logs*, se debe tener en cuenta cómo se almacenarán los datos o si se cargarán en alguna nueva plataforma. Es posible pensar en esta etapa como un procedimiento ETL (*Extract, Transform and Load*) para obtener los datos, transformarlos (limpiarlos) y cargarlos donde sea necesario, considerando un previo entendimiento de cuáles serán los requeridos.

3.2.3. Procesamiento

En esta fase se deben establecer los registros de eventos a utilizar en el análisis de minería de procesos. Será requerido cualquier trabajo sobre los datos extraídos que

genere sobre ellos una estructura como la que se presentó en la tabla 1, seleccionando de forma correcta todos los atributos, principalmente las actividades con sus fechas de realización, identificadores de los casos (*case id*) y ejecutores, pudiendo crear diversas vistas de *event logs* para diferentes análisis. Luego, de ser necesario, se deberán filtrar los datos para eliminar casos que no se deseen estudiar, como por ejemplo los que se desarrollaron en determinada fecha o los que están incompletos en el rango de tiempo del cual se extrajeron los datos. También es factible agregar datos para eliminar detalle en ellos o agrupar valores de ciertos atributos para simplificar un poco el posterior análisis.

3.2.4. SEMMA

Esta etapa corresponde a la utilización de minería de datos y todo el marco de trabajo que requiere, que en este caso será la metodología SEMMA. Como se vio en el capítulo anterior, este método está compuesto por 5 fases: muestreo (opcional), exploración, modificación, modelado y evaluación. Todas son de suma importancia en esta nueva propuesta ya que en este tramo del método se deja de lado la perspectiva de proceso y se considera una vista desde los datos, lo que permite el análisis correcto. La figura 13 muestra cómo es el esquema de la etapa SEMMA para el método propuesto en este documento.

El tipo de salida de la fase de exploración es variado, ya que se puede utilizar más de una técnica para el análisis de datos (es lo que se espera al aplicar este método), por lo que cada una de éstas puede requerir una modificación a los registros diferente para la aplicación de sus respectivos algoritmos. Así, todas las etapas que continúan después de SEMMA deberán realizarse para cada una de las técnicas de minería de datos empleadas en esta fase de la metodología.

Además de proponer una metodología, en este trabajo se establecerán cuáles son las

actividades principales a realizar para poder aplicar las diferentes técnicas de minería de datos sobre registros de eventos; específicamente, cómo debe hacerse la etapa de *modificación* con el fin de hacer un procesamiento eficiente sobre los datos y aplicar los algoritmos de forma correcta.

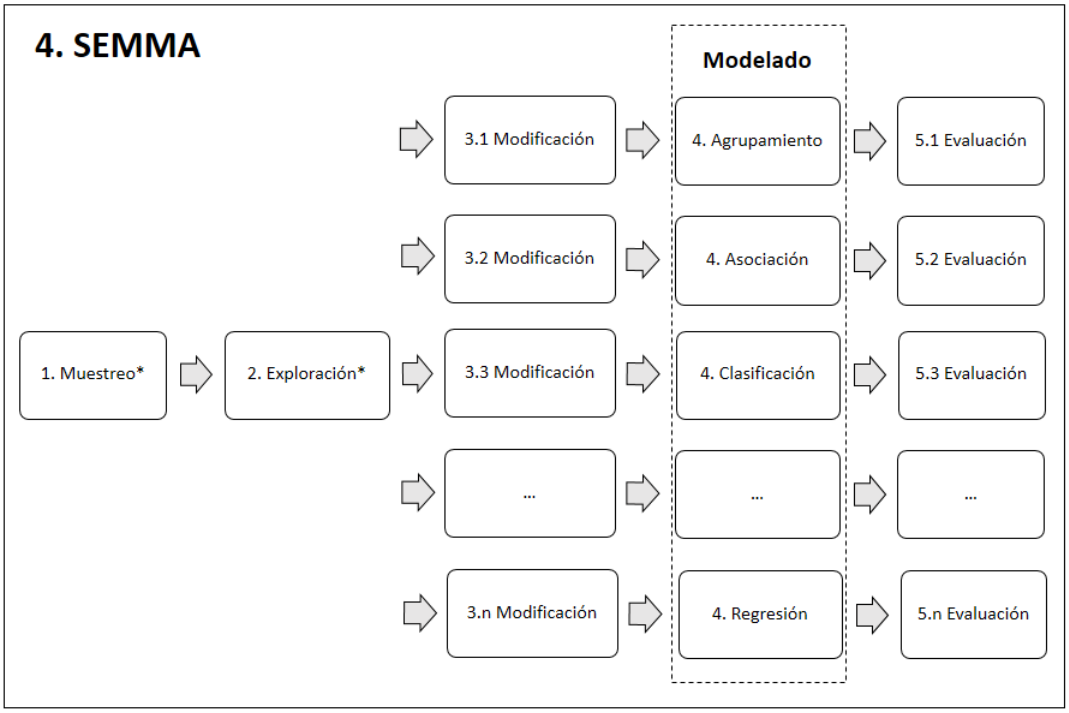

Figura 13: Etapa de análisis de datos de la metodología propuesta, en donde se aplica el marco de trabajo SEMMA.

3.2.5. Filtrado

Luego de los resultados obtenidos en base al análisis de minería de datos, es posible aplicar esa información sobre la perspectiva de procesos de los datos. En otras palabras, el conocimiento adquirido en la etapa anterior es posible reflejarlo en los datos entregados por la etapa de procesamiento, permitiendo continuar el análisis de procedimientos con más información, lo que hace posible aplicar diversos filtros sobre los registros. Así, en la siguiente etapa se podrán realizar análisis específicos del proceso de negocio que no serían posibles sin los patrones detectados por la minería de datos.

3.2.6. Minería de Procesos

El estudio de procedimientos a partir de datos está conformado por tres etapas clave de la minería de procesos: descubrimiento de procesos, análisis de conformidad y enriquecimiento de procesos a partir de atributos no convencionales (los que no son estrictamente requeridos para la minería de procesos). En la fase de descubrimiento se debe encontrar el flujo real del proceso a través de diversos algoritmos especializados, representándolo en modelos, para así poder entender cómo se está ejecutando el procedimiento, qué actividades se suceden, cuáles se ejecutan en paralelo o en un orden excluyente, entre otras características. En el análisis de conformidad se debe contrastar los registros de eventos con algún modelo previamente generado (pudiendo haberse obtenido a partir del descubrimiento o generado de forma manual), esto con el objetivo de estimar si los datos que el proceso genera se ajustan al modelo seleccionado o no. Por último, en la mejora o enriquecimiento, se espera aportar más información sobre el proceso a partir de los atributos de cada registro del *event log*. Por ejemplo, conocer la relación de los ejecutores de las actividades del procedimiento, estimar los tiempos de las actividades con el fin de detectar "cuellos de botella" o conocer probabilidades de enrutamiento del flujo, entre otros aspectos.

3.2.7. Evaluación

Esta etapa está compuesta de dos partes: diagnóstico y verificación. En la primera se espera detectar cuáles son resultados inusuales o interesantes que puedan explicar las preguntas de investigación (las que pueden ser iteradas para hacer el análisis más específico), a partir de una correcta interpretación de los resultados, principalmente a través de los modelos entregados. En la verificación se debe validar lo concluido en el diagnóstico, ya sea con los datos iniciales, datos futuros, actores del proceso o cualquier elemento que esté relacionado con el proceso de forma tal de estimar si la información obtenida es válida (se compara el estado del proceso antes del análisis con el resultado al aplicar el diagnóstico, tarea que deben ejecutar los expertos del negocio junto a los

analistas).

3.2.8. Mejoramiento de Procesos

Al igual que la etapa anterior, esta fase está compuesta por dos actividades: implementación de mejoras y soporte posterior. En base a los resultados de la etapa de evaluación, se deben asentar cambios en el proceso de negocio que impulsen mejoras en él, lo que puede ser realizado a través de técnicas de reingeniería de procesos. La implementación debe concluir con la implantación de estas propuestas. La fase de soporte tiene que ser continua; siempre se debe estar midiendo y evaluando el proceso, con el fin de detectar fallas o comportamientos anómalos y potenciar el buen funcionamiento de éste. En base a esto, son fundamentales sistemas de información de apoyo a los procesos si se desea trabajar con minería de procesos. Cabe destacar que el soporte no sólo tiene que realizarse luego de la implementación de mejoras; debe ser una labor fundamental en el control de procedimientos estructurados siempre.

4. Validación del Método Propuesto

En el presente capítulo se aplicará la metodología antes explicada sobre un caso de estudio real. Se expondrá cómo se llevó adelante cada una de las etapas, haciendo hincapié en la minería de datos y de procesos. Desde la fase 4 (SEMMA) el capítulo se dividirá en 3 secciones diferentes, una para cada una de las técnicas de minería de datos empleadas.

4.1. Planificación

En esta primera etapa lo primero que se debe hacer es comprender el negocio que se estudiará, por lo que se requiere obtener información de éste, de la organización involucrada y del problema que la aqueja.

Conocer el negocio

La entidad seleccionada para realizar el estudio corresponde a una empresa dedicada a la publicidad digital con presencia internacional. Apoya a pequeñas y medianas empresas a encontrar consumidores locales a través de diferentes tecnologías web y medios físicos, por lo que posee diversos productos y servicios. Éstos pueden dividirse en dos categorías: digitales e impresos.

- Servicios Digitales: relacionados a la web, como creación de sitios o mejora en sus contenidos, generación de prospectos, publicidad de navegadores, anuncios en páginas propias, entre otros.
- Servicios Impresos: principalmente directorios de páginas blancas y amarillas, además de marketing directo y revistas propias para los clientes.

Entender el problema

El motivo por el cual la empresa requiere hacer un estudio de sus procesos es la alta tasa de anulación de contratos por parte de sus clientes, en otras palabras, cuando un cliente está negociando un servicio o ya se ha cerrado la venta de éste, por diversas razones el cliente decide anular su compra. Se estima que un 20 % de las órdenes de servicio terminan siendo anuladas, hecho del cual se desconoce su causa.

Crear preguntas de investigación

En base al problema descrito, la principal pregunta de investigación a responder es, **¿por qué los clientes anulan sus órdenes?** Para poder dar una respuesta a ésta, se ha seleccionado el proceso de la empresa "ventas y anulaciones de contratos", el que comienza cuando el cliente ha decidido contratar el servicio, y finaliza cuando la venta es exitosa o el cliente anula la contratación. Con el fin de apoyar el desarrollo del trabajo, se han formulado dos hipótesis que podrían dar respuesta a la pregunta de investigación:

- Tiempos elevados en la confección del contrato provocan anulaciones: el hecho de que el tiempo desde que el cliente acepta el producto hasta que la venta se concreta sea desmedido, puede ocasionar que ellos desistan de la adquisición y terminen anulando el contrato.

- Clientes nuevos en diversos escenarios del proceso tienden a anular el contrato: considerando las diferentes cualidades o atributos del proceso (como por ejemplo modo de compra, ubicación geográfica, monto, entre otras), puede ser que algunas combinaciones de los valores generen, en mayor o menor medida, que los contratos sean anulados por clientes nuevos.

Además es relevante destacar que el proceso "ventas y anulaciones de contratos" está conformado por las siguientes 11 actividades:

- Negociar venta no presencial: acuerdo de venta de un contrato vía telefónica con cliente.
- Confirmar venta no presencial: confirmación por parte del cliente de la contratación vía telefónica.
- Negociar venta: confirmación por parte del cliente de la contratación presencialmente.
- Aumentar contrato: incremento en los productos y valor de contrato.
- Disminuir contrato: reducción en los productos y valor de contrato.
- Modificar contrato: petición de modificación de contrato.
- Validar contrato: revisión del contrato para continuar el proceso.
- Rechazar contrato: rechazo de las condiciones del contrato, requiriendo una posterior validación.
- Vender contrato: concreción de venta de un contrato por ambas partes.
- Sustituir contrato: cambio de un contrato.
- Anular contrato: cancelación del contrato.

La figura 14 muestra la red de Petri que conforma este proceso, identificando las relaciones entre las actividades que lo componen.

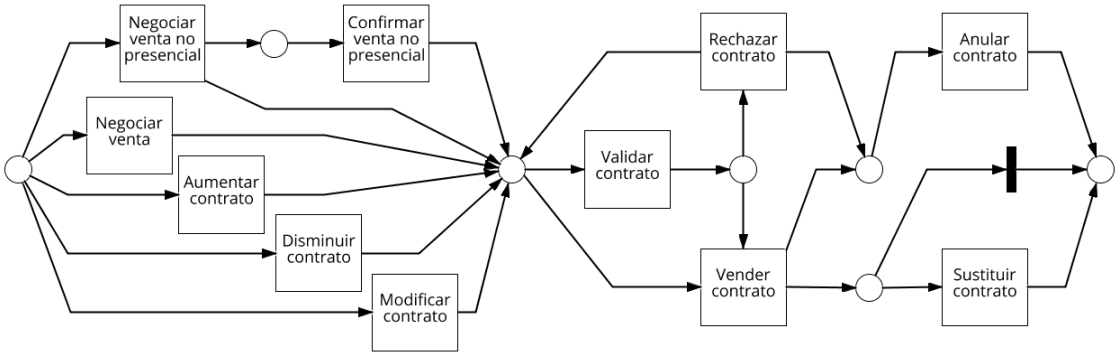

Figura 14: Red de Petri que representa el proceso en estudio.

Su funcionamiento es el siguiente: un contrato puede ser vendido de forma presencial o vía telefónica; en la primera opción se ejecuta la actividad *negociar venta*, y en la segunda las actividades *negociar venta no presencial* (venta telefónica, aún sin confirmar) y *confirmar venta no presencial* (cliente confirma la compra del producto). Luego, se concreta la venta del contrato a través de la firma en *vender contrato* o éste queda pendiente para revisión, en donde se realiza la actividad *validar contrato*. En este último caso se debe validar el contrato, y podrá concretarse la venta (*vender contrato*) o rechazarlo a través de *rechazar contrato*, en donde deberá volver a ejecutarse *validar contrato* (generando un ciclo dentro del proceso o cancelar el servicio con *anular contrato*. Después que el contrato está vendido (o cuando ha sido rechazado), es permitido anular éste con *anular contrato*. Como finalización del proceso el contrato también podrá ser reemplazado por uno anterior realizando la actividad *sustituir contrato*; como inicio del proceso es permitido comenzar con las actividades *aumentar contrato*, *disminuir contrato* o *modificar contrato*, las que aumentan o disminuyen el valor de un contrato, o modifican las características de éste, respectivamente.

4.2. Extracción

Los registros que el proceso genera están almacenados en una base de datos, la cual está compuesta por un gran número de entidades; considerando tablas, planillas, y archivos planos son más de 1.740 elementos. De todos éstos, se han seleccionado 107 tablas para realizar el proyecto luego de comprender el negocio de la organización y el problema a analizar.

Seleccionar datos

A partir del modelo anterior se han escogido 17 atributos de todo el conjunto de datos, los que se listan a continuación, indicando qué tipo de dato es y una breve descripción:

- Número producto (*string*): identificador del contrato y producto (utilizado como *case id*).
- Posición actividad (entero): indica la posición en que se ejecutó la actividad dentro de su instancia respectiva.
- Actividad (categórico): nombre de la actividad.
- Fecha inicio (fecha): fecha en que comenzó a ejecutarse la actividad.
- Fecha fin (fecha): fecha en terminó de ejecutarse la actividad.
- Ejecutor (categórico): responsable de la actividad.
- Monto total (entero): valor total de todos los productos de un contrato.
- Monto anulado (entero): valor que se ha anulado del contrato.
- Forma de pago (categórico): método en que cancela los productos el cliente.
- Tipo cliente (booleano): valor 0 o 1, identificando si el cliente es nuevo o antiguo, respectivamente.
- Producto (categórico): nombre del producto.
- División (categórico): nombre de la división de ventas.
- Identificador vendedor (categórico): identificador de la persona que vende el contrato.
- Tipo cuenta (categórico): tipo de cuenta a la que corresponde el cliente.
- Número caso (entero): número que representa el caso.
- Tipo caso (categórico): tipo de caso de anulación.
- Motivo caso (categórico): motivo por el que anula.

ETL

La extracción como tal de los datos concluyó en 75 mil registros aproximadamente (cada uno con los 17 atributos antes listados), guardados en un archivo CSV (*comma separated values*), los que corresponden a los generados en el primer semestre del año 2015. Sobre la transformación, no se consideró ningún tipo de cambio en los registros más allá de alcances de nombres que podrían tener, pero que no modificaron en nada el posterior estudio. Por último cabe destacar que los datos no fueron cargados en ninguna nueva plataforma de análisis.

4.3. Procesamiento

En esta etapa se busca obtener los registros finales para el posterior análisis, destacando el generar un formato de *event log* sobre éstos.

Crear registros de eventos

Al extraer los datos desde las diversas fuentes, se hizo de forma tal de que los registros de eventos quedaran en su formato estándar (*case id*, actividades, recursos y *timestamps*), por lo que no fue necesario hacer grandes cambios sobre el conjunto de datos. Sí se hizo otro tipo de trabajo, específicamente la creación de tres nuevas columnas:

- Columna *anula*: en base a *monto anulado* se generó un nuevo atributo *anula*, el que señala si el contrato fue anulado o no; esto hizo más sencillo el posterior análisis desde la perspectiva de minería de datos. Se generó este nuevo atributo con la herramienta Excel.

- Columna *horas actividad*: se ha agregado un nuevo atributo *horas actividad*, el que indica cuánto tiempo ha tomado en completarse cada actividad del proceso. Se implementó realizando la diferencia de los atributos *fecha fin* y *fecha inicio*, a través de la herramienta Excel.

- Columna *horas totales*: se ha agregado un nuevo atributo *horas total*, el que indica cuánto tiempo ha tomado en completarse cada instancia del proceso. Fue posible generarlo en base a la anterior columna creada (*horas actividad*), a través de un script desarrollado en el lenguaje de programación Python (ver código en Anexo A).

Filtrar registros (ruido)

Con el *event log* listo, se eliminaron de éste las principales inconsistencias que pudiese haber tenido, las que fueron:

- Instancias incompletas: fueron removidos todos los casos que, tomando en cuenta el intervalo de tiempo en estudio, no tenían todas las actividades con las cuales estaban conformados. Para ello se consideró el atributo *posición actividad*, por lo que todas las instancias que no tenían en alguna de sus actividades el valor 1 fueron eliminadas. Además, sólo se consideraron como actividades terminales *vender contrato*, *anular contrato* y *Sustituir contrato*, por lo que los casos que terminaban con cualquier otra tarea fueron eliminados. Estas operaciones fueron realizadas a través de la herramienta Disco.

- Registros con reclamos pero sin anulaciones: se descartaron todas las instancias en que se registraron reclamos por parte de los clientes, pero que no tenían un monto de anulación. Se entendió esto como una inconsistencia con el objetivo de simplificar el análisis, siendo eliminado sólo el $0,08\,\%$ del total de registros. Esta operación fue realizada a través de la herramienta Excel.

- Valores poco significantes de anulaciones: el atributo *monto anulado* va desde los 0 hasta los 8 millones de pesos aproximadamente, por lo que se consideró que valores ínfimos, en específico 3 y 4 pesos, podrían distorsionar los datos (instancia indica que se anuló el contrato por un monto bajo). Se eliminó aproximadamente al $0,02\,\%$ del total de registros. Esta operación fue realizada a través de la herramienta Excel.

- Registros sin reclamo pero con anulaciones: se ha considerado eliminar todas las instancias que no poseen reclamo, pero sí tienen un *monto anulado* diferente de 0. En esta ocasión el porcentaje de registros anulados sí fue considerable, siendo aproximadamente 6 % del total. Se ha creído pertinente realizar este trabajo para simplificar el análisis, ya que se puede establecer que todas las instancias sin reclamo no tendrán anulaciones. Esta operación fue realizada a través de la herramienta Excel.

Agregar registros

Previo al análisis no se hizo ningún tipo de agregación de datos, ya que se pretende visualizar cuáles podrían ser los pro y contra de trabajar con atributos que puedan tomar muchos o pocos valores (datos categóricos).

4.4. SEMMA

Siguiendo con la aplicación de la metodología, a continuación cada una de las tres secciones estará destinada a las técnicas de minería de datos utilizadas en este desarrollo (agrupamiento, asociación y clasificación), y para cada una se detallarán las fases de SEMMA. Se ha utilizado la herramienta RapidMiner para desarrollar toda esta etapa.

4.4.1. Agrupamiento

En esta sección se utilizaron dos algoritmos de agrupamiento, *k-means* y *k-medoids*, solamente empleando atributos continuos. Previo a realizar cualquier subetapa de SEMMA, fue requerido crear un script (ver código en Anexo A) que permita enlazar los registros agrupados entregados por el modelo generado por RapidMiner a la estructura de *event log* que poseen los registros (por la forma de los datos cargados a RapidMiner, éste asigna un grupo a cada instancia del proceso, no a cada actividad).

Muestreo

Para la ejecución del algoritmo *k-medoids* sí se realizó la fase de muestreo, ya que el tiempo que tomaba éste para completar era muy elevado. Se consideró seleccionar 1.000 instancias que terminaban en contratos vendidos y 1.000 en contratos anulados.

Exploración

Se utilizaron solamente atributos numéricos (continuos) en esta instancia, los que fueron *monto total*, *horas total*, *tipo cliente* y *anula*. Esto con el fin de diferenciar posibles análisis entre la naturaleza de los datos.

Modificación

Los atributos *tipo cliente* y *anula* son booleanos, por lo que fueron transformados a reales; además fue necesario normalizar los atributos para que todos tuviesen el mismo peso en el modelo, lo que se realizó a través de la puntuación estándar (z-*transformation* en RapidMiner).

Modelado

Se ejecutaron ambos algoritmos para 9 valores diferentes de k, desde 2 hasta 10. No se continuó con valores mayores a 10 ya que después de que k fuera igual a 7 la diferencia entre las distancias promedio entre un valor de k y otro fue constante (ver tabla 3).

Evaluación

Para determinar el número adecuado de grupos, para cada ejecución de los algoritmos se midió la distancia promedio de todos los *clusters* (distancias dentro de cada *cluster*) y el índice de Davies Bouldin (DB). Este último valor es un número mayor que cero, y mientras más pequeño sea habrá una mayor cohesión entre miembros de un

grupo y mayor separación entre *clusters*. La tabla 3 muestra los resultados obtenidos para el algoritmo *k-means*, en la que se aprecia que para la ejecución con $k = 6$ se obtuvo el índice DB más bajo, y donde la variación de distancias entre una instancia y la anterior fue la última considerable (luego todas son menores a 0,08).

Tabla 3: Resultados de distancias al aplicar k-means con diferente valores de k.

k	Distancia promedio	Diferencia con k anterior		Davies Bouldin
2	3,371	-		0,567
3	1,774	1,597		0,735
4	1,333	0,441		0,612
5	0,913	0,420		0,584
6	0,472	0,441		0,543
7	0,393	0,079		0,562
8	0,315	0,078		0,580
9	0,327	-0,012		0,572
10	0,249	0,078		0,584

Los valores de los centroides de los seis grupos encontrados se muestran en la tabla 4, con su respectiva gráfica en la figura 15. Existen valores negativos en las coordenadas por la normalización hecha sobre los datos.

Luego, se empleó el algoritmo *k-medoids* sobre una muestra estratificada aleatoria (2.000 registros en total). Así, de igual forma como se hizo para *k-means*, se ejecutaron diversas instancias con diferentes números de grupos, midiendo distancias promedio a centroides y el índice DB, con los resultados que se ven en la tabla 5.

Tabla 4: Valores de coordenadas de los centroides entregados por el algoritmo k-means con k=6.

Atributo	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Monto Total	11,019	-0,218	2,397	-0,200	-0,205	-0,070
Tipo Cliente	0,490	-1,196	0,805	-0,695	-0,944	0,836
Anula	-0,380	-0,495	-0,460	2,019	2,022	-0,495
Horas Total	-0,298	-0,297	-0,327	3,937	0,448	-0,325

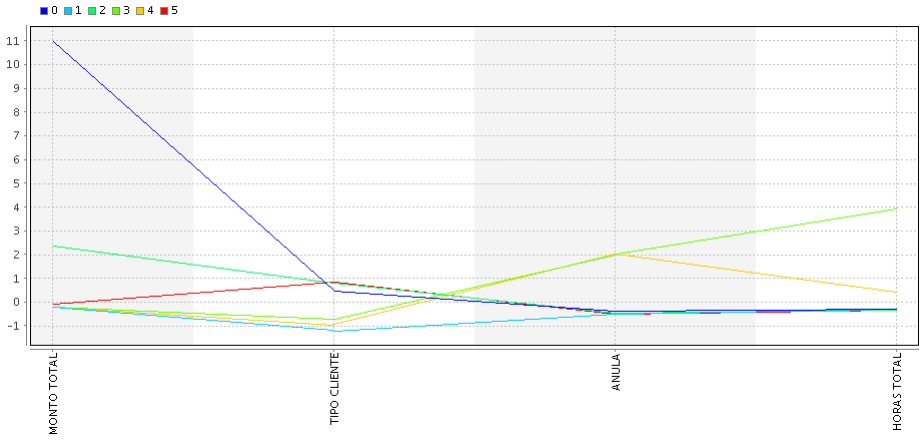

Figura 15: Gráfico de coordenadas de centroides obtenidos con el algoritmo k-means.

Para este caso también se consideró que 6 grupos son el valor con un mejor ajuste, ya que en esta instancia se produce el último decremento significativo en las distancias, además que el valor del índice DB no es de los más elevados de todos. Así, los valores de los centroides de cada atributo se muestran en la tabla 6 con su respectiva gráfica en la figura 16.

Con las figuras 15 y 16 entregados por *k-means* y *k-medoids* (con k = 6 para ambos), desde una perspectiva cualitativa, puede establecerse que el primer algoritmo es capaz de entregar centroides mayormente diferenciados, ya que los valores en el eje vertical

Tabla 5: Resultados de distancias al aplicar k-medoids con diferente valores de k.

k	Distancia promedio	Diferencia con k anterior		Davies Bouldin
2	2,752	-		0,738
3	2,089	0,663		0,928
4	1,840	0,249		0,771
5	1,293	0,547		0,728
6	0,978	0,315		0,858
7	1,114	-0,136		0,785
8	0,884	0,230		1,026
9	0,860	0,024		0,867
10	0,744	0,116		0,952

oscilan con una variación de 3 unidades aproximadamente, mientras que en el segundo la variación es de 2 unidades (recordar que todos los atributos fueron normalizados, por lo que no puede definirse una métrica para estas diferencias), además de que el índice DB es más bajo para *k-means*. Es por esto que el desarrollo continuó con el resultado del primer algoritmo, a través de los seis *clusters* que éste generó. El modelo entregado por esta ejecución es el que se muestra en la tabla 7, además de especificar cuántos contratos fueron vendidos exitosamente y cuántos anulados.

De la tabla 7 es posible inferir que los grupos 0, 1, 2 y 5 corresponden a contratos vendidos, y los grupos 3 y 4 a anulados.

Tabla 6: Valores de coordenadas de los centroides entregados por el algoritmo k-medoids con k=6.

Atributo	Grupo 0	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Monto Total	-0,184	-0,126	-0,330	-0,109	5,674	0,615
Tipo Cliente	-0,858	1,165	1,165	-0,858	1,165	-0,858
Anula	-0,266	-0,082	-0,588	0,865	-0,589	-0,595
Horas Total	1,000	1,000	-1,000	1,000	-1,000	-1,000

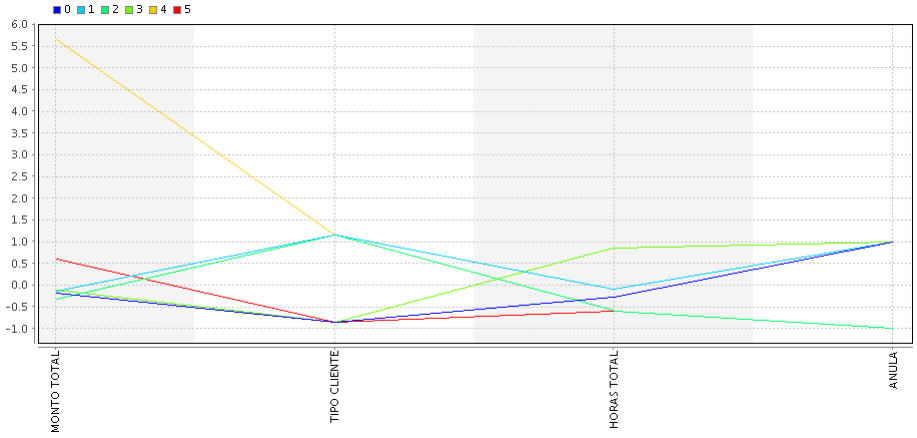

Figura 16: Gráfico de coordenadas de centroides obtenidos con el algoritmo k-medoids sobre atributos continuos.

4.4.2. Asociación

Con esta técnica se busca determinar qué actividades están relacionadas con el hecho de que un contrato pueda terminar siendo anulado.

En este caso no se hizo ningún tipo de **muestreo**; se consideró que todos los datos, sin importar las diversas frecuencias de los diferentes valores que pueden tomar los atributos (como por ejemplo que existan muchas más instancias que terminan de forma exitosa sobre las que no), describen el real comportamiento del proceso por lo que un

Tabla 7: Modelo entregado por el algoritmo k-means, indicando la cantidad de elementos de cada grupo y cuántos contratos fueron vendidos y anulados para cada uno.

Grupo	Cantidad de elementos	Contratos vendidos	Contratos anulados
0	88	84	4
1	4462	4462	0
2	583	575	8
3	872	1	871
4	2707	0	2707
5	9554	9554	0

muestreo de cualquier tipo podría modificar su naturaleza. Tampoco fue requerida la fase de **exploración** ya que no hubo modificación alguna respecto a las etapas anteriores.

Modificación

Tomando en cuenta la estructura de los registros de eventos (ver tabla 1), es factible pensar en generar una matriz de transacciones en base a las actividades del proceso que entregue información de relaciones entre ellas, sobre la cual poder aplicar algoritmos de reglas de asociación y obtener patrones de implicancia entre las tareas. Los que se han utilizado en este trabajo son los algoritmos *apriori* y *FP-growth*.

Lo primero fue crear un script (ver código en Anexo A) que transformara el archivo original que contiene los datos en uno con una estructura de transacciones, como el de la tabla 8, en donde cada columna representa una instancia del proceso y cada fila es una actividad, y la intersección entre una fila y una columna tendrá el valor 1 si dicha actividad se ejecutó en esa instancia, de lo contrario tendrá el valor 0.

Tabla 8: Matriz transpuesta de algunas transacciones (casos) del proceso en estudio, en donde cada columna indica qué actividades se realizaron (1) y cuales no (0) en cada instancia.

Aumentar contrato	0	0	0	0	0
Anular contrato	0	0	1	1	1
Negociar venta no presencial	0	0	0	0	1
Negociar venta	1	1	1	1	0
Rechazar contrato	0	1	1	1	1
Vender contrato	1	1	0	0	0
Disminuir contrato	0	0	0	0	0
Modificar contrato	0	0	0	0	0
Validar contrato	0	1	1	1	1
Sustituir contrato	0	0	0	0	0
Confirmar venta no presencial	0	0	0	0	1

Modelado

El primer algoritmo empleado para obtener reglas de asociación fue *apriori*. Para saber qué valores asignar al soporte y a la confianza, se asignaron diferentes cantidades intentado generar una cantidad significante de reglas relevantes. Se iniciaron ambas métricas en 0,9, y sólo cuando el soporte fue igual a 0,4 y confianza de 0,1 se encontraron 12 reglas, pero todas éstas eran obvias dentro del proceso; entre ellas estaban por ejemplo: si no está *negociar venta no presencial* no estará *confirmar venta no presencial*, o si no está *negociar venta no presencial* sí estará *negociar venta*; además de que en ninguna estaba presenta la actividad *anular contrato*. Esto determinó que todas las reglas que pudieran ser generadas en este conjunto de transacciones, tendrían un soporte relativamente bajo (menor al 40 %), por lo que se disminuyeron ambas métricas a 0,05 para detectar la mayor cantidad de reglas para luego analizarlas. Se obtuvieron 188 reglas (ver resultado en Anexo B), de las cuales sólo en 4 estaba solamente la acti-

vidad *anular contrato* como conclusión (se buscó determinar qué actividades producen la anulación). La tabla 9 muestra los precedentes dichas actividades con sus valores de soporte y confianza respectivos.

De las 4 reglas anteriores, son interesantes de analizar solamente *rechazar contrato* → *anular contrato* y *validar contrato* → *anular contrato*, ya que las actividades que componen su precedente pueden no ser ejecutadas en alguna instancia del proceso (contrario a *vender contrato* y *confirmar venta no presencial*, en donde sí son parte fundamental del flujo del proceso). Así, para comprender qué actividad puede tener alguna relación con la anulación de contratos, se estimó que *rechazar contrato* puede ser la indicada; ya que, por el entendimiento del proceso, podría creerse de antemano que el rechazar un contrato pueda producir la anulación de éste, suposición errada ya que esta actividad se ejecuta para poder adecuar las características del contrato para el cliente o la organización.

Tabla 9: Precedentes de las reglas de asociación detectadas por el algoritmo apriori que tienen como conclusión la actividad anular contrato, señalando sus valores de soporte y confianza.

Actividad precedente	Soporte	Confianza
Vender contrato	0,12	0,44
Rechazar contrato	0,05	0,40
Validar contrato	0,05	0,28
Confirmar venta no presencial	0,05	0,13

Complementando el resultado anterior, se ejecutó el algoritmo *FP-growth* sobre el mismo conjunto de datos; de igual forma se asignó un valor (igual a 0,05) bajo para el soporte y para la confianza, con el fin de obtener una gran cantidad de reglas y seleccionar las que se relacionan con la actividad *anular contrato*. El algoritmo arrojó 564 reglas (ver resultado en Anexo B), de las cuales 31 tienen como conclusión únicamente

dicha actividad. En la mayoría de éstas su precedente está compuesto por un conjunto de actividades (no una única tarea); esto dificulta el análisis, ya que siempre está presente alguna que es indispensable en el flujo del proceso, como lo son las actividades de inicio, a excepción de 3 reglas en donde sus precedentes están compuestos por *validar contrato*, *rechazar contrato* y *vender contrato*. La tabla 10 muestra el soporte y la confianza para éstas reglas.

Tabla 10: Precedentes de las reglas de asociación detectadas por el algoritmo FP-growth que tienen como conclusión la actividad anular contrato, señalando sus valores de soporte y confianza.

Actividad precedente	Soporte	Confianza
Validar contrato	0,079	0,278
Rechazar contrato	0,058	0,398
Vender contrato	0,055	0,992

De igual forma que en los resultados obtenidos a partir de *apriori*, sólo interesará estudiar la relación de las actividades *validar contrato* y *rechazar contrato* con la anulación. Los valores de soporte y confianza para la relación *rechazar contrato* → *anular contrato* prácticamente no variaron entre los resultados de los dos algoritmos, de lo que se podría desprender que esta relación tiende a ser más consistente; además de lo ya mencionado, el que se produzca un rechazo dentro del proceso podría entenderse como una mayor probabilidad de anulación del contrato. Por estas dos razones es que se decidió continuar el análisis a partir de esta regla de asociación.

Evaluación

La regla antes encontrada (*rechazar contrato* → *anular contrato*) podría aportar información valiosa al entendimiento del problema, por lo que se continuó el análisis a través de ella, previa validación de este resultado.

El soporte de dicha regla es de 0,05; este valor puede entenderse simplemente como el porcentaje de veces que la combinación de actividades *rechazar contrato* y *anular contrato* está presente en el conjunto completo de registros, por lo que se podría quitar importancia al hecho de que este valor es bajo. La confianza es de un 0,4, uno de los valores más bajos de todas las reglas; prácticamente todas éstas son relaciones obvias dentro del proceso, es de esperar entonces que tengan una confianza alta (sobre 0,8) en su gran mayoría; la regla analizada no es evidente dentro del proceso, por lo que sí puede permitirse su baja confianza. Otra métrica para validar reglas de asociación es la elevación (*lift*), la cual señala la relación entre el antecedente y el consecuente: mientras más cercano a 1 sea este índice, mayor independencia tendrán los dos elementos de la regla; de forma contraria, mientras más grande sea este valor, mayor es la probabilidad de que exista una relación externa entre antecedente y consecuente. En este caso el *lift* fue de 2,403, valor que se estima bajo considerando que hay reglas con elevación de más de 17 puntos.

En base al argumento anterior, se concluyó que la regla seleccionada sí es válida para continuar el desarrollo del proyecto.

4.4.3. Clasificación Mediante Árboles de Decisión

En esta sección se utilizó la técnica de árboles de decisión para apoyar la etapa SEMMA de la metodología propuesta. Se intentó detectar ciertos valores para los diversos atributos del proceso que determinen cuándo los casos concluyan siendo anulados, en otras palabras, poder clasificar los casos anulados en base a sus características. En esta instancia se consideró no realizar la etapa de **exploración** por los mismos motivos descritos en la técnica anterior.

Muestro

Para no realizar un análisis sesgado de los datos se hizo un muestreo estratificado

considerando la misma cantidad de registros que terminan con el contrato vendido y anulado; en este caso fueron 3.500 casos para los dos posibles términos del proceso.

Modificación

La única modificación sobre el *event log* fue filtrar una fila por cada instancia del proceso, así tener solamente un registro por caso (el hecho de tener varias actividades por traza distorsionaría los resultados). Para ello se seleccionaron todos los registros en donde el atributo *posición actividad* fuese igual a 1, lo que filtra solamente las actividades iniciales de todos los casos.

Modelado

Se ejecutó el algoritmo *random forest* con un número de árboles igual a 20, de los cuales sólo tres presentaron un tamaño adecuado para continuar con la etapa de minería de procesos (los demás poseían únicamente 2 niveles). Considerando dichos árboles, se decidió seleccionar el de la figura 17 por la información que éste entrega en base a la clasificación de anulación (1 como valor final).

Se detectó que cuando el atributo *horas total* se encuentra entre los valores 271,850 y 726,550, además *tipo de cliente* toma el valor 0, el contrato es anulado 663 veces y vendido solamente en 83 instancias. Esto corresponde a un patrón de anulación dentro del proceso que puede ser utilizado para continuar el análisis.

Evaluación

El resultado de *random forest* fue validado con la técnica de validación cruzada, utilizando 10 particiones sobre los registros. Así, los resultados de la matriz de confusión y curva ROC, respectivamente, se muestran en la tabla 11 y la figura 18.

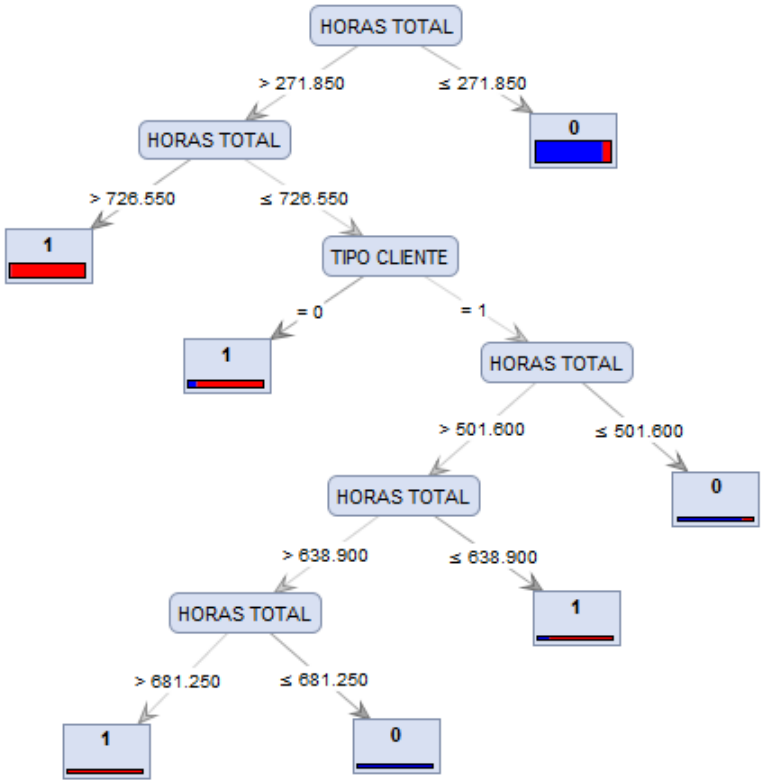

Figura 17: Árboles de decisión obtenidos al aplicar al algoritmo random forest.

Tabla 11: Matriz de confusión obtenida al aplicar validación cruzada sobre el resultado del algoritmo random forest.

Reales / Predecidos	Contratos vendidos	Contratos anulados	Precisión
Contratos vendidos	3.425	61	98,25 %
Contratos anulados	75	3.439	97,87 %
Precisión	97,86 %	98,26 %	

Se aprecia que los resultados son buenos, la clasificación de los datos hecha por el modelo bordea el 98 % de aciertos.

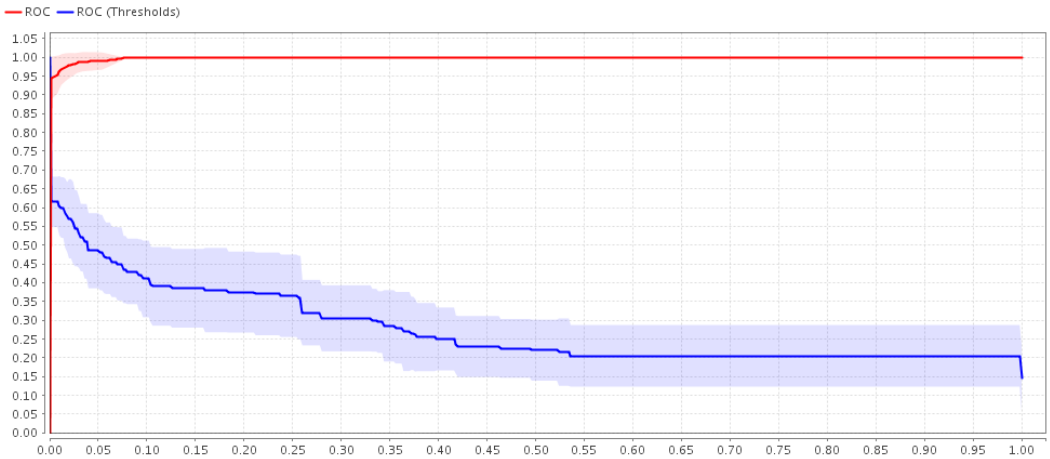

Figura 18: Curva ROC obtenida al aplicar validación cruzada sobre el resultado del algoritmo random forest con un área bajo la curva de 0,998.

El valor del área bajo la curva de la figura 18 fue de 0,998, el que es muy positivo (mientras más cercano a 1, el modelo clasificará de mejor forma los contratos anulados y vendidos). La curva indica que el solapamiento de las dos clases analizadas es bajo, por lo que el modelo no tiende a errar en la clasificación.

Luego de verificar los dos métodos de validación anteriores, se establece que los resultados son apropiados para continuar el análisis con la regla detectada del árbol seleccionado, considerando que la cantidad de registros mal clasificados no es tan elevada.

4.4.4. Clasificación Mediante Reglas de Inducción

Para aplicar esta técnica se ha utilizado el mismo conjunto de datos que en la anterior, pero ahora sobre el módulo *rule induction*, el que entrega como resultado diversas reglas de inducción.

Muestreo

Se realizó el mismo muestreo hecho para el algoritmo anterior, seleccionar la mis-

ma cantidad de registros (3.500) para casos que terminan con el contrato vendido y anulado.

Exploración

Para intentar conseguir reglas que generen información valiosa al caso de estudio, se seleccionó una cantidad reducida de atributos, en donde todos son del tipo nominal. Éstos fueron:*division*, *tipo cliente*, *tipo cuenta* y *anula*, éste último definido como etiqueta del modelo.

Modificación

Se modificaron los datos de igual forma que para aplicar el algoritmo *random forest*; se filtró sólo la actividad inicial para cada instancia del proceso.

Modelado

El modelo generó 73 reglas (ver resultado en Anexo B), de las cuales sólo 1 consideraba más de 1.000 registros anulados. Ésta fue *si tipo cliente = 0 y tipo cuenta = venta directa potencial luego anula = 1*. Ésta posee 333 registros que terminan siendo vendidos y 1.966 que son anulados.

Evaluación

Nuevamente se evaluó el modelo entregado con validación cruzada con 10 particiones. La tabla 12 y la figura 19 muestran los resultados para la matriz de confusión y la curva ROC, respectivamente.

Ambos resultados presentan una bondad inferior a la entregada por el algoritmo *random forest*; aún así son buenos valores considerando que las precisiones de las predicciones del modelo nunca fueron inferiores a 80 %, además de que la curva ROC

presenta un área bajo la curva de 0,877 (valor cercano a 1), indicando que el solapamiento de las clases es leve.

Tabla 12: Matriz de confusión obtenida al aplicar validación cruzada sobre el resultado del algoritmo de reglas de inducción.

Reales / Predecidos	Contratos vendidos	Contratos anulados	Precisión
Contratos vendidos	2.984	627	82,64 %
Contratos anulados	516	2.873	84,77 %
Precisión	85,26 %	82,09 %	

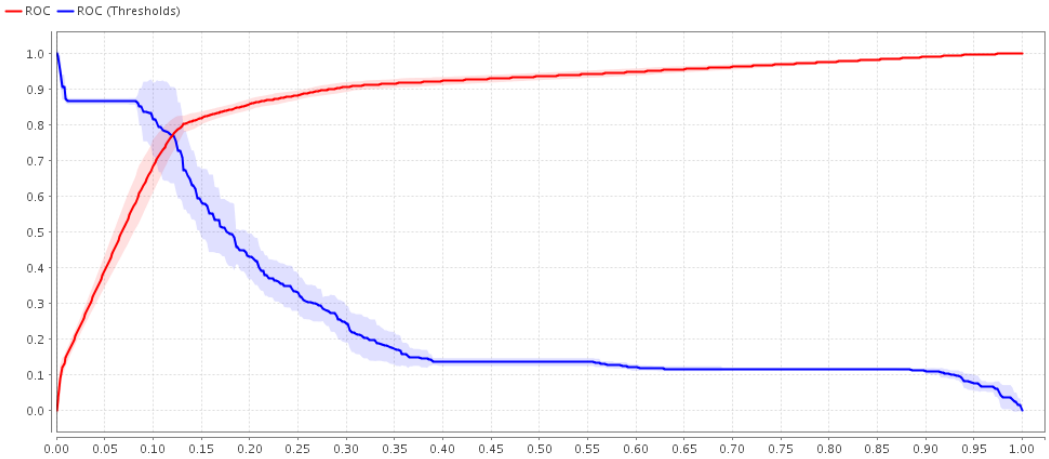

Figura 19: Curva ROC obtenida al aplicar validación cruzada sobre el resultado del algoritmo de reglas de inducción con un área bajo la curva de 0,877.

4.5. Filtrado

En esta fase es necesario seleccionar los registros que sean de interés analizar en base a los resultados obtenidos en la etapa SEMMA. Considerando el desarrollo hecho con minería de datos, a continuación se presenta un resumen del análisis obtenido:

- Agrupamiento: entrega un modelo que divide a los registros en grupos permitiendo ser analizados en forma separada, identificando para cada uno cualidades

diferentes dentro del proceso. En el caso de estudio, examinar el comportamiento de los grupos que terminan de forma exitosa la venta del contrato contra los que concluyen anulando éste. Como propuesta se plantea aplicar minería de procesos a grupos que venden sin anular (*cluster* 5) y al que más anula (*cluster* 4).

- Asociación: entrega reglas que asocian las actividades del proceso, permitiendo saber si la ejecución de una tarea o un conjunto de ellas implicará la realización de otras actividades. En el caso de estudio, determinar qué actividades al realizarse provocan que el contrato sea anulado. Como propuesta se plantea aplicar minería de procesos a las instancias que ejecutan la actividad *rechazar contrato* y a las que no de forma separada.

- Clasificación mediante árboles de decisión: entrega árboles como modelo, los que permiten detectar patrones que sean interesantes incluir en el análisis de minería de procesos a través de la aplicación de filtros. En el caso de estudio, detectar qué combinaciones de valores deben tomar los atributos para que el proceso termine siendo anulado. Como propuesta se plantea aplicar minería de procesos a los registros que fueron filtrados en base a los patrones detectados.

- Clasificación mediante reglas de inducción: entrega reglas de la forma *si X entonces Y* que permiten identificar patrones que sean interesantes incluir en el análisis de minería de procesos a través de la aplicación de filtros. En el caso de estudio, detectar qué combinación de valores deben tomar los atributos para que se produzca la anulación del contrato. Como propuesta se plantea: aplicar minería de procesos a los registros que fueron filtrados en base a las reglas detectadas.

Entendiendo que los dos algoritmos de clasificación utilizados entregaron una pauta similar (patrones dentro del proceso), sólo se continuará el estudio con el resultado de árboles de decisión, además de las otras dos técnicas descriptivas. Para filtrar el *event log* se utilizó la herramienta Disco, la que posee una gran interfaz para realizar este tipo de operaciones. Así, se hicieron 3 diferentes selecciones de datos para cada una

de las técnicas, las que fueron correctamente transformadas en archivos con formato MXML y XES para su uso en ProM versión 5.2 y 6.5, respectivamente. A continuación se presentan los modelos entregados por Disco para la segmentación de datos de cada técnica de minería de datos.

Agrupamiento

Se filtró en base a dos *clusters*, el que más contratos anulados tienen y el que menos, los grupos 4 y 5 (éste tiene 0 contratos anulados igual que el grupo 1, pero posee 5.092 registros más), respectivamente. Las figuras 20 y 21 muestran los modelos generados por Disco para cada caso.

Asociación

El registro de eventos se separó en dos, una parte con todas las trazas que tienen la actividad *rechazar contrato* y otra que no presenta dicha actividad. Los modelos pueden verse reflejados en las figuras 22 y 23. Los porcentajes de anulación de contratos que éstas demuestran respectivamente son de 12,6 % y 36,3 % aproximadamente, por lo que la tasa de anulaciones cuando está presenta la actividad *rechazar contrato* es 3 veces mayor a cuando no se ejecuta dicha tarea.

Clasificación

Se filtró el registro de eventos solamente considerando los casos en que el atributo *horas total* tomase valores entre 271 y 726, además que *tipo cliente* fuese igual a 0. La figura 24 representa el resultado al aplicar este filtro. Es posible notar que aproximadamente el 68 % de los contratos terminan anulados en base al patrón señalado.

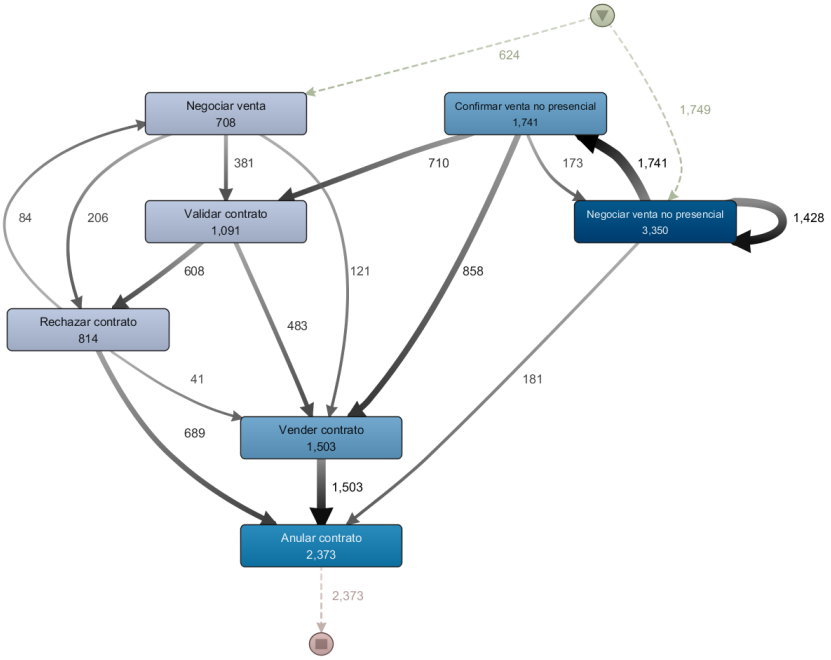

Figura 20: Modelo de proceso para el grupo 4, el cual posee la mayoría de las anulaciones del registro de eventos.

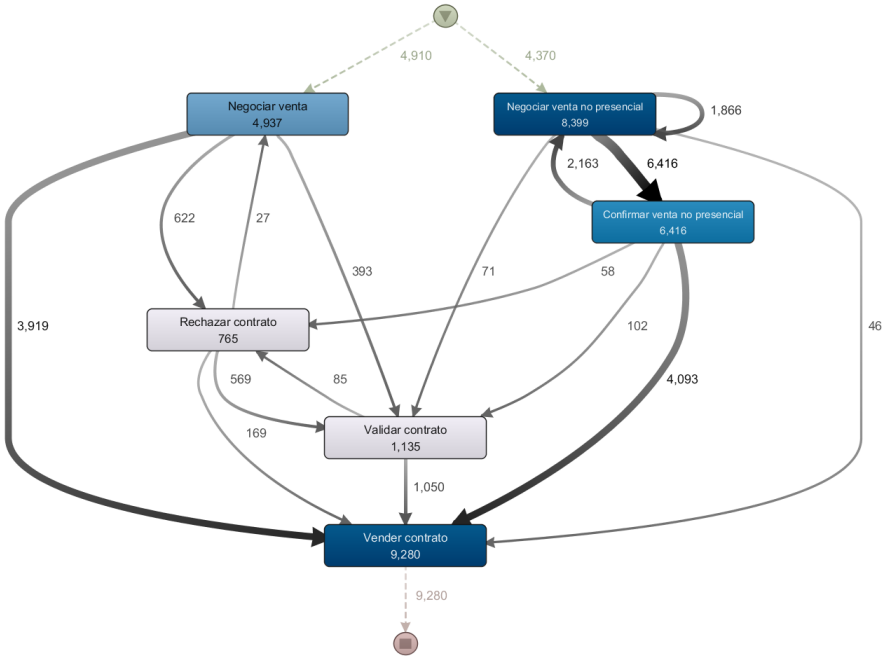

Figura 21: Modelo de proceso para el grupo 5, el cual no posee anulaciones.

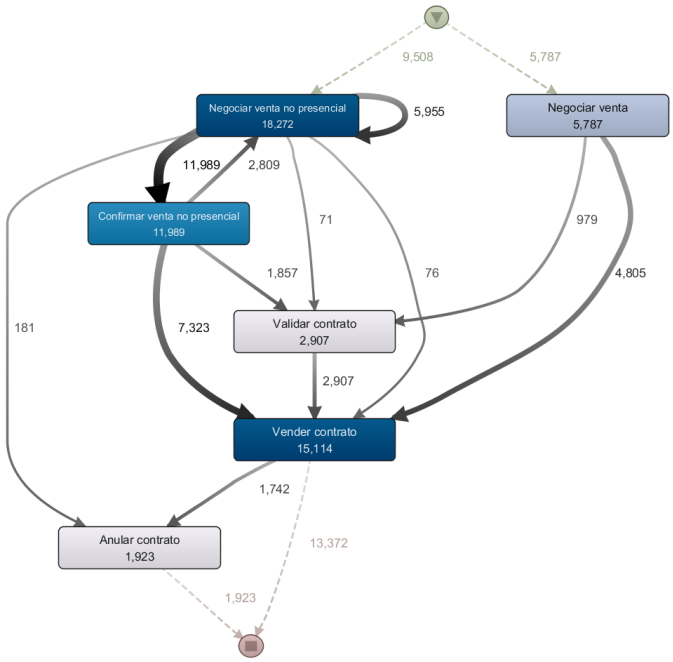

Figura 22: Modelo del proceso entregado por Disco con instancias que no tienen la actividad rechazar contrato.

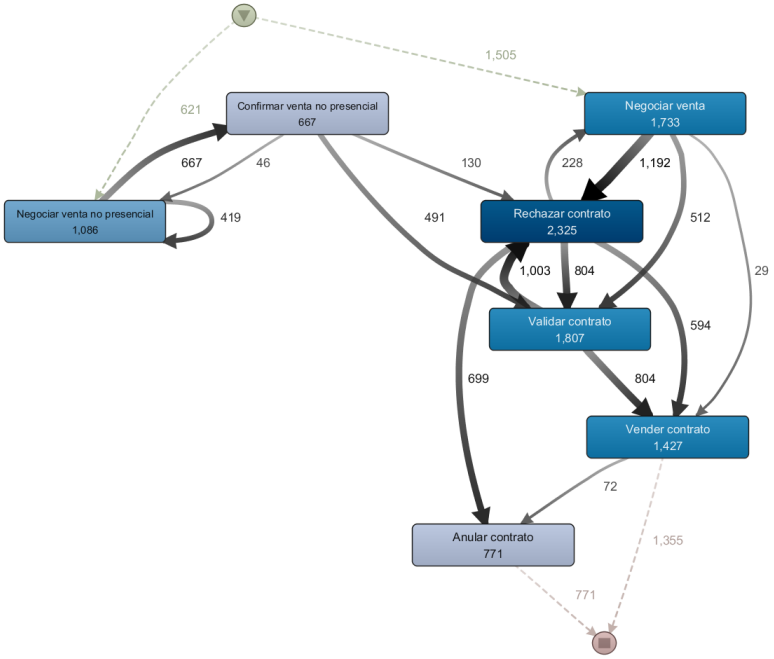

Figura 23: Modelo del proceso entregado por Disco considerando solamente las instancias que tienen en alguna parte de su flujo la actividad rechazar contrato.

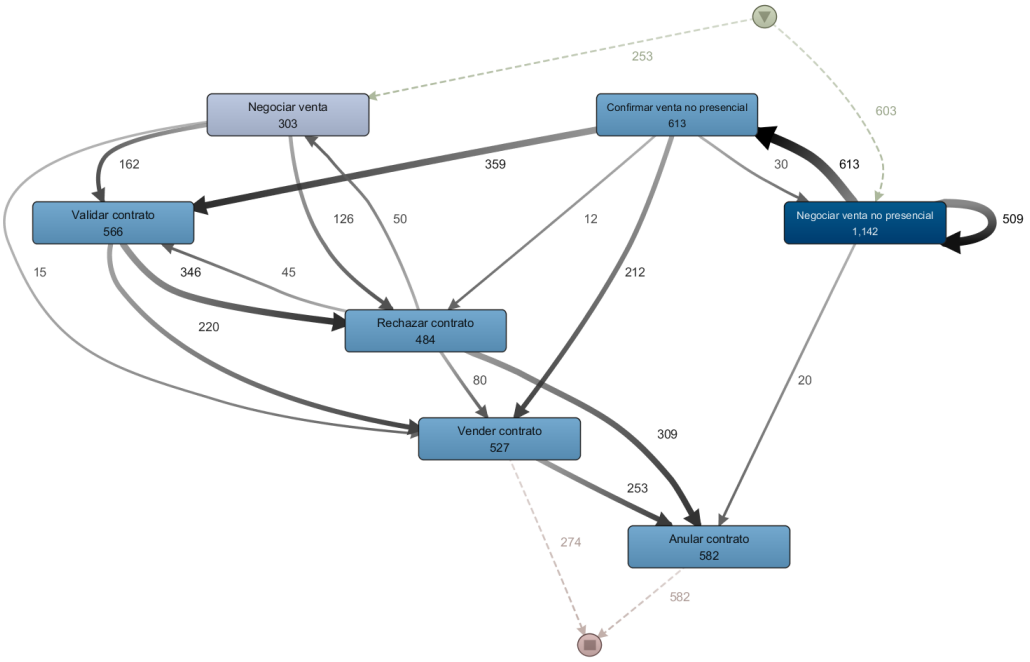

Figura 24: Modelo del proceso filtrado en base al patrón seleccionado del árbol de decisión generado a través del algoritmo random forest.

4.6. Minería de Procesos

Con los registros de eventos de interés filtrados, y generados sus respectivos archivos para su uso en ProM, es posible comenzar el análisis de minería de procesos. Se realizaron en los resultados de las tres técnicas antes descritas las fases de descubrimiento de procesos, análisis de conformidad (a través del modelo esperado del proceso) y el enriquecimiento de procesos, aplicando los mismos criterios.

Descubrimiento de Procesos

Esta tarea se hizo a través de ProM 6.5 utilizando el algoritmo *inductive miner*, el que entregó modelos relativamente sencillos comparados con otras ejecuciones obtenidas como con *genetic* o *heuristic miner*. Las figuras 25 y 26 muestran los modelos para los grupos 4 y 5 obtenidos con el algoritmo *k-means*, respectivamente.

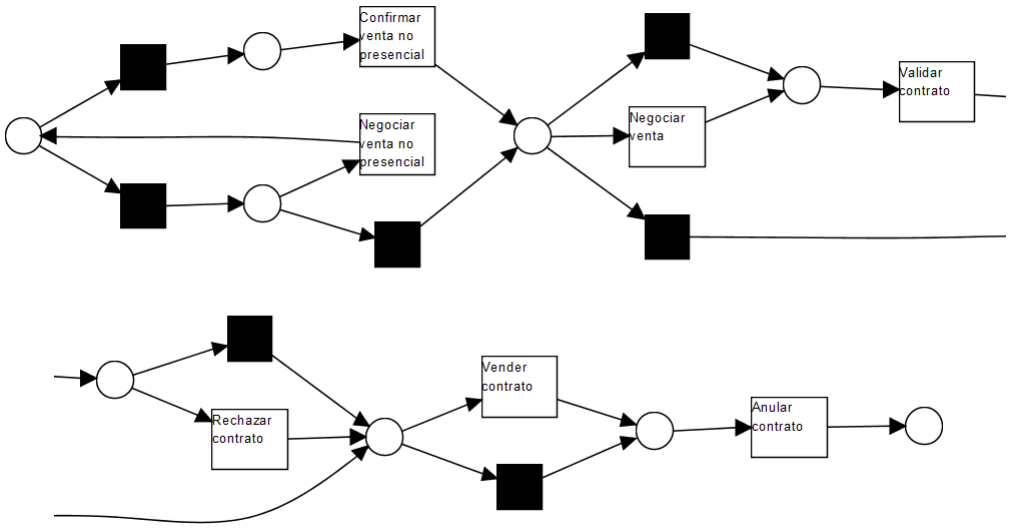

Figura 25: Modelo del proceso encontrado por el algoritmo inductive miner para el event log del cluster 4 entregado por k-means.

Las figuras 27 y 28 muestran los modelos para la segmentación de casos entre los que tienen la actividad *rechazar contrato* y los que no, respectivamente. La figura 29 muestra el modelo del filtrado propuesto por el algoritmo *random forest*, del cual se consideró una ruta entre la raíz del árbol seleccionado y uno de los nodos hoja en donde la probabilidad de anulación del contrato fuese alta.

Análisis de Conformidad

Para realizar este análisis, se utilizó el modelo generado a partir del conocimiento previo del negocio y del proceso (figura 14), en donde se contrastó éste con cada uno de los 5 *event log* que se filtraron en la etapa anterior, obteniendo los indicadores ajuste y precisión para cada uno. La tabla 13 detalla estos valores.

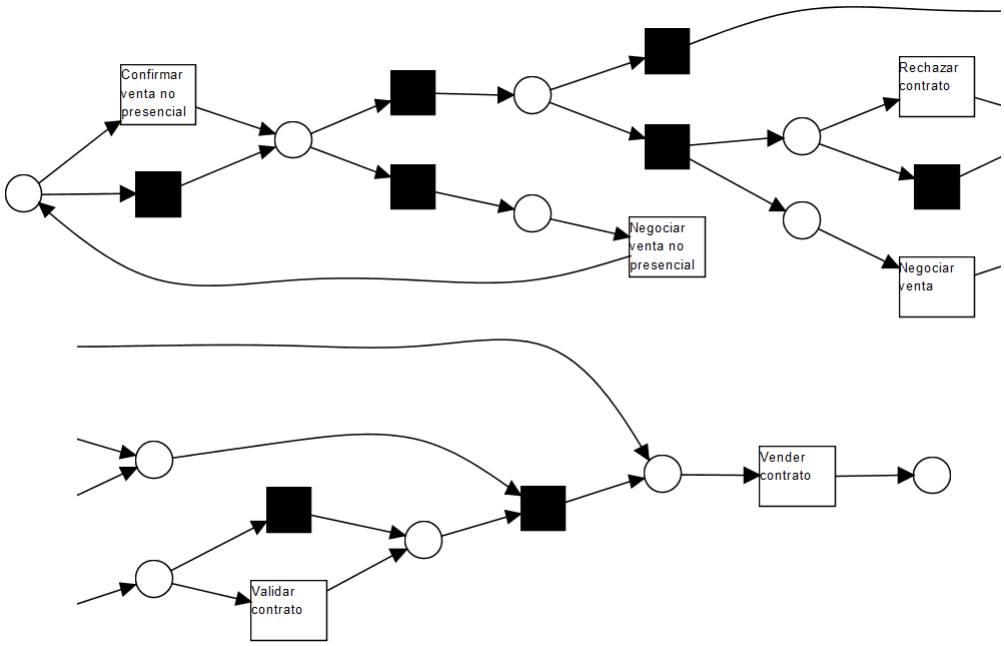

Figura 26: Modelo del proceso encontrado por el algoritmo inductive miner para el event log del cluster 5 entregado por k-means.

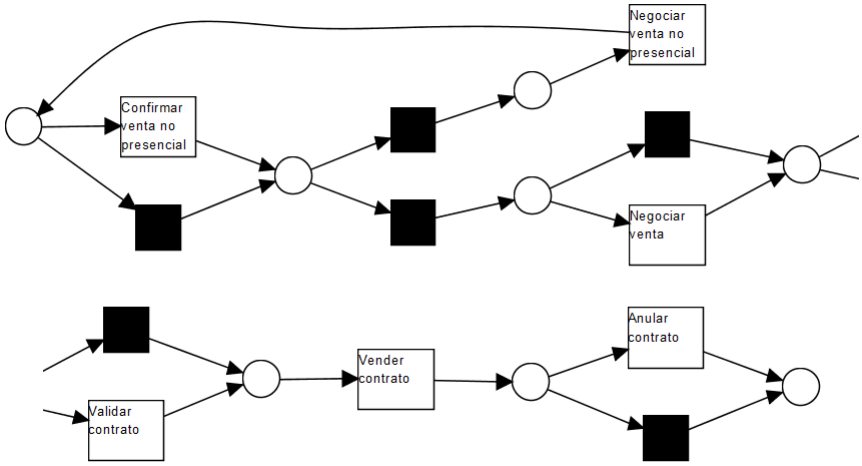

Figura 27: Modelo del proceso encontrado por el algoritmo inductive miner para el event log de casos que no tienen la actividad rechazar contrato.

Tabla 13: Indicadores obtenidos a partir del análisis de conformidad para los 3 resultados de minería de procesos.

Modelo de proceso	Ajuste	Precisión
Agrupamiento, *cluster* 4	0,71	0,79
Agrupamiento, *cluster* 5	0,67	0,48
Asociación, con *rechazar contrato*	0,72	0,81
Asociación, sin *rechazar contrato*	0,68	0,44
Clasificación mediante árboles de decisión	0,71	0,94

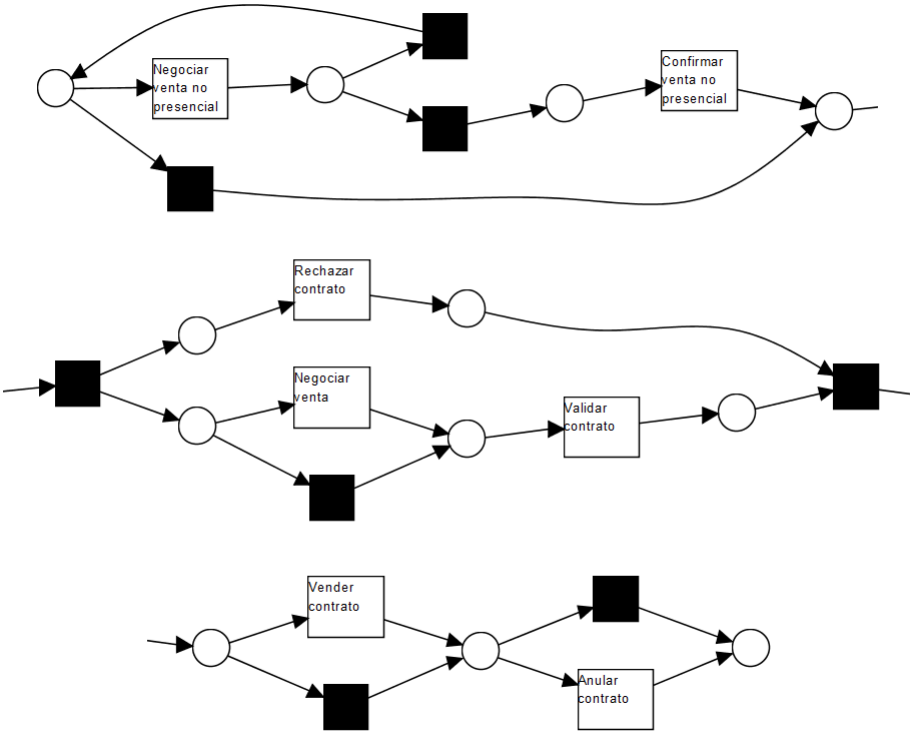

Figura 28: Modelo del proceso encontrado por el algoritmo inductive miner para el event log de casos que tienen en su flujo la actividad rechazar contrato.

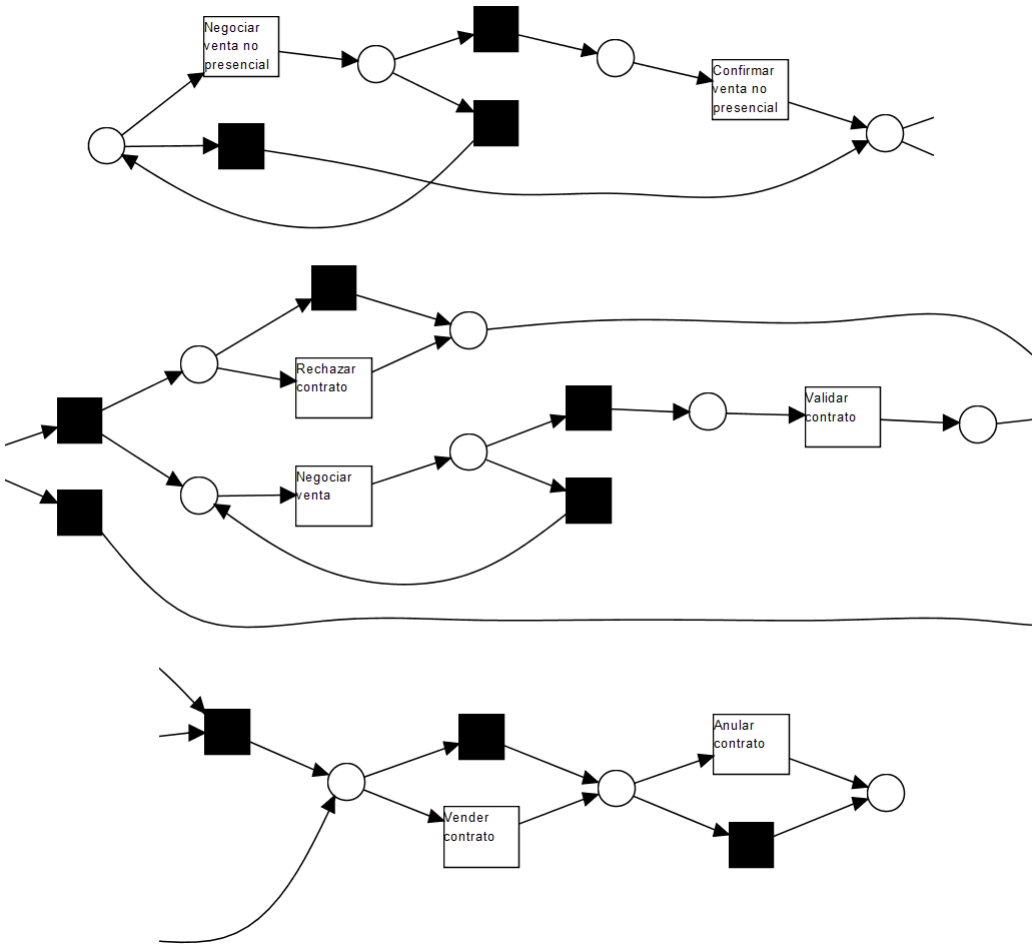

Figura 29: Modelo del proceso encontrado por el algoritmo inductive miner para el event log de casos en donde los atributos toman valores que aumentan la probabilidad de anular el contrato, los que se detectaron con el algoritmo random forest.

Enriquecimiento de Procesos

Por último, se analizó el desempeño temporal de todas las actividades involucradas de cada modelo. Para ello se utilizó la herramienta Disco a través de la vista *performance*, obteniendo el tiempo promedio de todas las tareas de los 5 diferentes modelos. Cabe destacar que el tiempo máximo que debería demorar una actividad en concluir es de 24 horas (pero 12 horas es el prudente). La tabla 14 detalla el número de tareas que tardan más de 1 día en ejecutarse para cada modelo.

Cabe destacar que la actividad *anular contrato* es instantánea (demora 0 segundos), por lo que de las 7 principales actividades (presentes en todos los modelos) sólo para 6 se puede hacer un análisis de desempeño.

Tabla 14: Cantidad de actividades por modelo que demoran más de 24 horas en realizarse.

Modelo	Actividades que demoran más de 24 horas
Agrupamiento, *cluster* 4	*Negociar venta*, *negociar venta no presencial*, *validar contrato*, *rechazar contrato* y *vender contrato*. (5)
Agrupamiento, *cluster* 5	*Negociar venta* y *rechazar contrato*. (2)
Asociación, con *rechazar contrato*	*Negociar venta*, *negociar venta no presencial*, *validar contrato*, *rechazar contrato* y *vender contrato*. (5)
Asociación, sin *rechazar contrato*	*Negociar venta*, *validar contrato* y *vender contrato*. (3)
Clasificación mediante árboles de clasificación	*Negociar venta*, *negociar venta no presencial*, *validar contrato*, *rechazar contrato* y *vender contrato*. (5)

4.7. Evaluación

En esta etapa se obtuvieron conclusiones respecto a todo el análisis previamente hecho. Primero fue necesario diagnosticar posibles respuestas a las preguntas de investigación, para luego verificarlas a partir de datos futuros.

Diagnóstico

Retomando la pregunta de investigación hecha en la primera etapa de la metodología, ¿por qué los clientes anulan sus órdenes?, se validaron las dos hipótesis formuladas de la siguiente forma:

1. Tiempos elevados en la confección del contrato provocan anulaciones: a partir de la tabla 14 queda claro que esta premisa es verdadera. De los resultados del modelo entregado por *k-means*, al comparar los grupos 4 (anula) y 5 (no anula), existe una diferencia de 3 actividades que demoran más de 24 horas, lo que claramente es relevante a la hora de ejecutar el proceso. Lo mismo se valida con el resultado de aplicar reglas de asociación, cuando dentro del flujo del proceso está presente la actividad *rechazar contrato*, 5 tareas se ejecuta en tiempos excesivos, no siendo así en los casos en donde esta actividad no se realiza (sólo 3 actividades tardan). En el trabajo con árboles de decisión también ayudó a validar la hipótesis: cuando los contratos son anulados con valores específicos de los atributos del proceso (en este caso se comprobó a través de la unión de las condiciones *271 < horas total < 726* y *tipo cliente = 0*), todas las actividades demoran más de lo permitido en ejecutarse.

2. Clientes nuevos en diversos escenarios del proceso tienen a anular el contrato: la mejor forma de validar esta premisa es observar el modelo de la figura 29, el cual se obtuvo al filtrar el *event log* en base al resultado de la técnica de clasificación (con *tipo cliente = 0*). Éste tiene solamente una actividad final, que es *anular contrato*, lo que indica que el algoritmo de descubrimiento no fue capaz de detectar otra tarea terminal como *vender contrato*. Esto podría verse reflejado en el *fitness* obtenido para este modelo, el cual no fue un valor alto; puede asumirse que existe un conjunto de trazas no menor que no se ajusta al modelo esperado. Lo anterior se compensa con la precisión conseguida, valor que sí fue alto y refleja que no cualquier traza puede replicarse en el modelo.

Con todo el desarrollo realizado hasta este punto, es posible inferir elementos del funcionamiento del proceso más allá de las hipótesis planteadas. Observando la tabla 13, se aprecia el hecho de que para los modelos con mayores cantidades de contratos anulados (grupo 4 para agrupamiento y casos con contratos rechazados para asociación) el *fitness* y la precisión fueron más altos que para modelos con mayores cantidades de

contratos exitosos. Esto simplemente refleja que el flujo esperado del proceso fomenta la anulación de los contratos. Sobre los dos modelos relacionados a la regla de asociación *rechazar contrato* → *anular contrato*, cuando está presente la actividad *rechazar contrato* (figura 28) el flujo es más largo que cuando no se realiza ésta (figura 27). Considerando el modelo generado para el resultado de la técnica de clasificación $271 < horas\ total < 726$ y $tipo\ cliente = 0$, éste está compuesto por 2 ciclos, de lo que podría inferirse que a mayor cantidad de bucles en la ejecución del proceso, más probabilidades hay de que el contrato termine siendo anulado (este modelo fomenta la anulación). Sobre esto además todos los modelos tienen un ciclo inicial que involucra a la venta no presencial, el cual no debería existir.

En base a lo concluido en el párrafo anterior, a continuación se listan diferentes acciones que deberían generar una menor tasa de contratos anulados:

1. Reducir el tiempo de ejecución del proceso en general. La actividad que se repite en todos los modelos de proceso encontrados es *negociar venta*, por lo que podría deducirse que se anulan más contratos cuando la venta es presencial. Así, una opción para reducir las anulaciones es disminuir la cantidad de ventas presenciales. Además, se destaca que las actividades que en algunas instancias no se realizan (como *validar contrato* y *rechazar contrato*) se repiten constantemente, por lo que reducirlas también podría favorecer la tasa de contratación.

2. Realizar la menor cantidad de veces posible la actividad *rechazar contrato*. Para lograrlo será necesario generar mejores contratos tanto para los clientes como para la empresa desde el comienzo, así evitar futuras revisiones o ajustes de éste.

3. Reducir la cantidad de actividades que se realizan en las instancias del proceso. Propuesta relacionada directamente con las dos anteriores, se esperaría no realizar las actividades que no son fundamentales del proceso, específicamente *validar contrato* y *rechazar contrato*, para lo que será necesario no iterar la

confección del contrato, estando éste listo desde la venta inicial.

4. Evitar ciclos dentro del proceso. El único ciclo permitido dentro del proceso es entre las actividades *validar contrato* y *rechazar contrato*, en donde se busca mejorar el contrato. Ahora bien, considerando el flujo real del proceso se tienen otros ciclos, específicamente: *negociar venta no presencial* (ciclo de largo 1), entre *venta no presencial* y *confirmar venta no presencial*, y por último entre *negociar venta* y *rechazar contrato*. Todos éstos son producto de malas instanciaciones del proceso por parte de los ejecutores de las actividades, por lo que se deberá validar que se siga el flujo esperado por parte de ellos.

Verificación

Todo el análisis anterior se hizo con datos recolectados del primer semestre el año 2015 que el proceso generó, y para poder verificar si las afirmaciones antes presentadas podrían disminuir la cantidad de contratos anulados, se recopilaron los registros del segundo semestre del mismo año. Cabe destacar que este nuevo conjunto de datos posee las mismas características que el anterior, el proceso no sufrió ningún cambio entre los dos periodos del año 2015. La verificación se hizo a través de la herramienta Disco, en donde se consideraron los 4 puntos a mejorar del proceso para que éste no tenga tantos contratos anulados. A continuación se describe lo hecho para cada uno de ellos:

1. Reducir el tiempo de ejecución del proceso: se estimó que 7 actividades para realizar una instancia es un buen número, ya que permite realizar todas las actividades e incluso poder nuevamente revisar el contrato (ciclo entre *pendiente en validación* y *rechazar contrato*). Así, considerando que el tiempo prudente por actividad es de 12 horas (24 horas como máximo), el total de las instancias permitido es de 84 horas. Luego se seleccionaron sólo las instancias del proceso que demoraron esta cantidad de tiempo o menos, obteniendo un porcentaje de 10,1 % de contratos anulados.

2. Realizar la menor cantidad de veces posible la actividad *rechazar contrato*: se descartaron todas las instancias que poseen esta tarea, obteniendo que el 10,5 % de los contratos son anulados.

3. Reducir la cantidad de actividades que se realizan en las instancias del proceso: retomando la cantidad de 7 actividades por caso explicada en el punto 1, se seleccionaron solamente las instancias que tienen este número o menos tareas en su flujo. Se obtuvo un 10,8 % de contratos anulados.

4. Evitar ciclos dentro del proceso: se eliminaron todos los ciclos en el flujo, los que se producen a través de ventas no presenciales (entre *negociar venta no presencial* y *confirmar venta no presencial*) y en mejorar el contrato (entre *validar contrato* y *rechazar contrato*). Se estableció que el 10,2 % de contratos terminan siendo anulados.

Al aplicar estos 4 puntos a mejorar de forma simultánea, se obtuvo un 9,1 % de contratos anulados (diferencia de un 1,3 % respecto al promedio de las 4 por separado). El porcentaje de anulación del conjunto de datos utilizado para el análisis es de 16,7 %, por lo que se establece que la disminución de anulaciones al aplicar las medidas de mejora propuestas es de un 7,6 %. Si bien este número puede parecer pequeño considerando que representa la disminución total de anulaciones, si se contrasta con el conjunto de datos utilizado para el análisis (el que tiene 18.123 casos) se hubiesen anulado 1.379,8 casos menos, lo que corresponde a 7,6 contratos diarios. Llevando el número anterior a pesos chilenos (CLP), el promedio de anulación de un contrato es de CLP 309.124, entonces el aumento de la ganancia diaria sería de CLP 2.349.342 (esto sin considerar los costos que conlleva la venta de un contrato, los que fueron provechosos cuando un contrato se anuló), cifra que sí es considerable. Con esto se verifica que las propuestas concluidas a partir del análisis sí son favorables para la mejora del proceso, en este caso disminuir la cantidad de contratos anulados lo que repercute directamente en las ganancias de la organización.

4.8. Mejoramiento de Procesos

Como se mencionó la sección 4 del capítulo 1, esta etapa de la metodología no se pudo ejecutar ya que no es posible acceder al funcionamiento del proceso. Sí es posible sugerir el desarrollo de las 2 subfases que componen esta etapa.

Implementación de Mejoras

Se propone implementar los 4 puntos de mejora planteados en la etapa anterior. Como ya se comprobó en ésta, al aplicar estos cambios sí hay una disminución significativa en la cantidad de contratos anulados. Además, y en base a lo observado al aplicar las técnicas de clasificación, se recomienda identificar qué otras cualidades del proceso (valores que puedan tomar sus diferentes atributos) tienden a generar anulaciones. En este estudio solamente se analizaron dos de ellas, pudiendo entenderse por la cantidad de atributos que componen el proceso que habrán muchas más por detectar. Identificando estas combinaciones de valores sería posible proponer mejoras específicas para cada una de ellas.

Soporte Posterior

Como soporte continuo se propone, considerando que los datos ya se almacenan de buena forma y posibilitan una fácil extracción del *event log*, cargar los datos en algún BPMS (*Business Process Management System*) que permita realizar evaluaciones de forma periódica (o en tiempo real) para medir el desempeño del proceso, identificando la tasa de anulaciones y ver si ésta disminuye en base a los cambios previamente realizados.

Conclusiones

La metodología propuesta integra análisis de minería de procesos y de datos para lograr un entendimiento acabado de los procesos de negocio que ejecutan las organizaciones. Se complementaron marcos de trabajo de cada una, específicamente el método SEMMA de minería de datos y PM^2 de minería de procesos. Se logró obtener conocimiento del negocio previo al análisis de proceso gracias al estudio de datos, lo que permitió deducir nuevas características del proceso.

Fueron tres las técnicas empleadas en este estudio de lo que respecta a minería de datos: agrupamiento, asociación y clasificación. Sobre agrupamiento, se utilizaron dos de los algoritmos más comunes para el estudio: *k-means* y *k-medoids*; en asociación se ejecutaron los algoritmos *apriori* y *FP-growth*; usando la técnica de clasificación, se emplearon los algoritmos *random forest* y *reglas de inducción*.

Lo que respecta al análisis de minería de procesos, se abarcaron las tres ramas de estudio de esta disciplina: descubrimiento de procesos, análisis de conformidad y enriquecimiento de procesos. En el descubrimiento se aplicó el algoritmo *inductive miner* a los modelos obtenidos a partir de los resultados de minería de procesos; el análisis de conformidad permitió encontrar el ajuste y la precisión de estos modelos con el fin de validarlos; el enriquecimiento de procesos se enfocó en la perspectiva temporal, en donde se detectaron qué actividades eran las que tomaban más del tiempo establecido dentro del proceso.

La propuesta fue aplicada en un caso de estudio, el que consistía en analizar el principal proceso de negocio de una empresa, “ventas y anulaciones de contratos”. Los resultados obtenidos permitieron responder a la pregunta de investigación ¿por qué los clientes anulan sus órdenes?, a través de dos hipótesis planteadas en un comienzo:

“tiempos elevados en la confección del contrato” y “clientes nuevos en diversos escenarios del proceso tienden a anular el contrato”. Esto se concretó a través de cuatro propuestas de mejora al proceso, las que redujeron la tasa de anulación de 16,7 % a 9,1 %, disminución que se consideró muy positiva estimando la cantidad de contratos que pudieron haber sido ventas exitosas, además de la ganancia que éstos hubiesen producido a la organización.

Se lograron cumplir todos los objetivos planteados al comienzo de este trabajo. Sobre el objetivo principal, fue posible establecer una metodología para el análisis de procesos de negocio que mezclara la minería de procesos y de datos, con la que se espera poder entender problemáticas del proceso y poder definir mejoras a éste. Para ello fue necesario cumplir los tres siguientes puntos, los que se propusieron al comienzo del trabajo como objetivos secundarios:

- Que este marco tuviese etapas bien definidas detallando en qué consiste cada una y las subfases que las componen. Esto permitió tener una buena pauta de cómo debe ser implementado un proyecto de análisis de procesos. Se utilizaron metodologías ya existentes de minería de procesos y de datos para crear este método, rescatando las bondades de cada una.

- Detectar técnicas de minería de datos que pudieran ser compatibles con la metodología propuesta. Se utilizaron reglas de asociación y algoritmos de agrupamiento con el fin de describir características del proceso que permitieran continuar el análisis de minería de procesos de buena forma. Además, se usó la técnica de clasificación para detectar patrones que fomentan la anulación; se cree que este tipo de algoritmos fue utilizado de forma descriptiva más que predictiva considerando la naturaleza del análisis: entender el actual funcionamiento de un proceso. De esto se puede plantear que faltó la mirada predictiva del análisis, lo que requeriría un nuevo enfoque del problema, quizás planteando una pregunta de investigación como ¿cuándo un cliente anulará un contrato?

- Aplicar la metodología en otro(s) caso(s) real(es) y poder validarla, además de buscar mejoras a la situación actual del negocio de una organización. Se obtuvieron los datos de una empresa que generaba al ejecutar su principal proceso, con los cuales se realizó todo el análisis. No fue posible obtener datos posteriores a la implementación de la mejora, con lo que se hubiera validado de mejor forma el método. En compensación, se utilizó un conjunto de datos también antiguo pero diferente al de análisis para la validación, al cual se le aplicaron las propuestas de mejoras. Esto permitió aprobar la propuesta del trabajo.

Como trabajo futuro del presente desarrollo, se identificó:

- Realizar la última etapa del método propuesto: en esta oportunidad no fue posible implementar la metodología completamente en la organización en estudio, por lo que queda pendiente poder concluir ésta en algún otro caso de estudio.

- Validar la propuesta en otro tipo de procesos: el proceso de negocio de este caso de estudio es bastante particular, ya que existen dos estados principales de finalización (exitoso y fracasado). Es necesario entender cómo se comporta la metodología propuesta en procesos con otras cualidades, como por ejemplo: que esté compuesto de muchas actividades sin un flujo particular o que tenga un modelo de forma “espagueti” (de difícil lectura a simple vista), entre otros.

- Ejecutar otro tipo de técnicas de minería de datos: solamente se trabajaron tres técnicas de un conjunto amplio de algoritmos. Se espera que en una continuación a este desarrollo pueda detectarse cómo otras técnicas (como la regresión, la que genera modelos matemáticos para predecir valores de nuevos registros) ayudan en la comprensión de un proceso de negocio, además de aportar un análisis predictivo que complemente al de clasificación, yendo más allá del análisis descriptivo en que se enfocó este trabajo.

La ingeniería en informática posee un gran campo de estudio en todo lo que compete al análisis de datos. En este trabajo fueron utilizadas, de forma complementaria, dos

disciplinas que se dedican al entendimiento de grandes conjuntos de datos, la minería de procesos y de datos. Los conocimientos requeridos para lograr lo anterior fueron adquiridos gracias a los estudios previos de esta profesión y la capacidad analítica que ésta entrega.

Referencias Bibliográficas

[1] Arias, M., & Rojas, E. (2014, September). *Deciphering event logs in SharePoint Server: A methodology based on process mining. In Computing Conference (CLEI), 2014 XL Latin American (pp. 1-12), IEEE.*

[2] Barabási, A. L., & Bonabeau, E. (2003). *Redes sin escala-El conocimiento de las leyes fundamentales que rigen la organización de las redes complejas es fundamental, entre otras muchas cosa. Investigación y Ciencia: Edición Española de Scientific American, (322), 58-67.*

[3] Buijs, J. C., van Dongen, B. F., & van der Aalst, W. M. (2012, June). *A genetic algorithm for discovering process trees. In 2012 IEEE Congress on Evolutionary Computation (pp. 1-8). IEEE.*

[4] De Medeiros A. K. A. (2006). *Genetic Process Mining (PhD Thesis). Eindhoven University of Technology, Eindhoven.*

[5] De Weerdt, J., Schupp, A., Vanderloock, A., & Baesens, B. (2013). *Process Mining for the multi-faceted analysis of business processes—A case study in a financial services organization. Computers in Industry, 64(1), 57-67.*

[6] Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners. John Wiley & Sons.*

[7] Devore, J. L. (2008). *Probabilidad y estadística para ingenierías y ciencias. Cengage Learning Editores.*

[8] Gallardo Arancibia, J. A. (2009). *Metodología para la definición de requisitos en proyectos de data mining (Doctoral dissertation, Informatica).*

[9] Gehrke, N., & Werner, M. *Process Mining, 2013. p. 10.*

[10] Gilbert, K., Sánchez, R. R., & Santos, J. C. R. (2006). *Minería de datos: Conceptos y tendencias. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, 10(29), 11-18.*

[11] Günther, C. W., & Van der Aalst, W. M. (2007). *Fuzzy Mining: Adaptive Process Simplification Based on Multi-Perspective Metrics. In G. Alonso, P. Dadam, and M. Rosemann, editors, International Conference on Business Process Management (BPM 2007), volume 4714 of Lecture Notes in Computer Science. Springer, Berlin, 2007, pp 328–343.*

[12] Kodratoff, Y. (2014). *Introduction to machine learning. Morgan Kaufmann.*

[13] Koster, A. M., & Munoz, X. (2009). *Graphs and algorithms in communication networks on seven league boots. In Graphs and Algorithms in Communication Networks (pp. 1-59). Springer Berlin Heidelberg.*

[14] Leemans, S. J., Fahland, D., & van der Aalst, W. M. (2013, June). *Discovering block-structured process models from event logs-a constructive approach. In International Conference on Applications and Theory of Petri Nets and Concurrency (pp. 311-329). Springer Berlin Heidelberg.*

[15] Moine, J. M., Gordillo, S. E., & Haedo, A. S. (2011). *Análisis comparativo de metodologías para la gestión de proyectos de Minería de Datos. In XVII Congreso Argentino de Ciencias de la Computación (CACIC 2011).*

[16] Murata, T. (1989). *Petri nets: Properties, analysis and applications. Proceedings of the IEEE 77.4, pp 541-580.*

[17] Orallo, H., RAMIREZ, J., QUINTANA, C. R., Orallo, M. J. H., Quintana, M. J. R., Ramírez, C. F., ... & Eppen, G. D. (2004). *Introducción a la Minería de Datos. Pearson Prentice Hall,.*

[18] Perez-Castillo, R., Weber, B., Pinggera, J., Zugal, S., de Guzman, I. G. R., & Piattini, M. (2011). *Generating event logs from non-process-aware systems*

enabling business process mining. Enterprise Information Systems, 5(3), 301-335.

[19] Rebuge, Á., & Ferreira, D. R. (2012). *Business process analysis in healthcare environments: A methodology based on process mining. Information Systems, 37(2), 99-116.*

[20] Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). *Process mining in healthcare: A literature review. Journal of biomedical informatics, 61, 224-236.*

[21] Rozinat, A., & van der Aalst, W. M. (2008). *Conformance checking of processes based on monitoring real behavior. Information Systems, 33(1), 64-95.*

[22] Tan, P., Steinbach, M., & Kumar, V. *Introduction to Data Mining. Pearson Education, 2006.*

[23] Trcka, N., Pechenizkiy, M., & Van der Aalst, W. M. P. (2011). *Process mining from educational data (Chapter 9).*

[24] Van Der Aalst, W., Van Dongen, B. F., Herbst, J., Maruster, L., Schimn, G., & Wejters, A. J. M. M. *Workflow mining: A survey of issues and approaches. Data and Knowledge Engineering. 2003.*

[25] Van Der Aalst, W. M., Reijers, H. A., & Song, M. (2005). *Discovering social networks from event logs. Computer Supported Cooperative Work (CSCW), 14(6), 549-593.*

[26] Van Der Aalst, W. M. (2009). *Process-aware information systems: Lessons to be learned from process mining. In Transactions on Petri Nets and Other Models of Concurrency II. Springer Berlin Heidelberg, pp 1-26.*

[27] Van Der Aalst, W. M. (2009). *ProM: The Process Mining Toolkit.*

[28] Van Der Aalst, W. M. *Process mining: discovery, conformance and enhancement of business processes. Springer, 2011.*

[29] Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... , & Burattin, A. (2011, August). *Process Mining Manifesto, In International Conference on Business Process Management (pp. 169-194). Springer Berlin Heidelberg.*

[30] Van Der Heijden, T. H. C. (2012). *Process mining project methodology: Developing a general approach to apply process mining in practice. Master of Science in Operations Management and Logistics, Netherlands TUE, School of Industrial Engineering.*

[31] Van Eck, M. L., Lu, X., Leemans, S. J., & Van Der Aalst, W. M. (2015, June). *PM 2: A Process Mining Project Methodology, In International Conference on Advanced Information Systems Engineering (pp. 297-313). Springer International Publishing.*

[32] Weijters, A. J. M. M., & Van Der Aalst, W. (2003). *Rediscovering workflow models from event-based data using little thumb.*

[33] Weske, M. (2012). *Business process management: concepts, languages, architectures. Springer.*

[34] White, S. A. (2004). *Introduction to BPMN. IBM Cooperation, pp 2008-029.*

[35] Williams, G. J., & Huang, Z. (1996, October). *A case study in knowledge acquisition for insurance risk assessment using a KDD methodology. In Proceedings of the Pacific Rim Knowledge Acquisition Workshop, Dept. of AI, Univ. of NSW, Sydney, Australia (pp. 117-129).*

[36] KNIME *(versión 3.2.1) [software]. (2016). www.knime.org*

[37] Orange *(versión 3.3) [software]. (2016). www.orange.biolab.si*

[38] PYTHON *(versión 2.7) [software]. (2015). www.python.org*

[39] R *(versión 3.3) [software]. (2016). www.r-project.org*

[40] RAPIDMINER STUDIO *(versión 5.3) [software]. (2016). www.rapidminer.com/products/studio*

[41] SAS *(versión 9.4) [software]. (2016). www.sas.com*

[42] SPSS MODELER *(versión 18.0) [software]. (2016). www-03.ibm.com/software/products/es/ibm-spss-modeler*

[43] WEKA *(versión 3.8) [software]. (2016). www.cs.waikato.ac.nz/ml/weka/index.html*

Anexos

Anexo A

Creación Columna *Horas Total*

```
archivo = open("eventlog.csv")
temporal = open("temporal.csv", "w")
primera = True
actual = ""
total = 0.0
pendientes = []
for fila in archivo:
   if primera == True:
      primera = False
      temporal.write(fila[:-1] + ",TIEMPO TOTAL\n")
   else:
      if actual == fila.split(",")[1]:
         total += float(fila.split(",")[6])
         pendientes += [fila]
      else:
         for i in pendientes:
            temporal.write(i[:-1] + "," + str(total) + "\n")
         pendientes = [fila]
         actual = fila.split(",")[1]
         total = float(fila.split(",")[6])
archivo.close()
temporal.close()
```

Creación Matriz de Transacciones con Actividades

```
archivo = open("entrenamiento.csv")
transacciones = open("transacciones_actividades.csv", "w")
actual = ""
primera = [True, True]
hechas = [0,0,0,0,0,0,0,0,0,0,0]
mapeo = {
   "AUMENTAR CONTRATO": 0,
   "ANULAR CONTRATO": 1,
   "NEGOCIAR VENTA NO PRESENCIAL": 2,
   "NEGOCIAR VENTA": 3,
   "RECHAZAR CONTRATO": 4,
   "VENDER CONTRATO": 5,
   "DISMINUIR CONTRATO": 6,
   "MODIFICAR CONTRATO": 7,
   "VALIDAR CONTRATO": 8,
   "SUSTITUIR CONTRATO": 9,
   "CONFIRMAR VENTA NO PRESENCIAL": 10 }
for fila in archivo:
   if primera[0] == True:
      primera[0] = False
      transacciones.write("CASE_ID, \
                           AUMENTAR CONTRATO, \
                           ANULAR CONTRATO, \
                           NEGOCIAR VENTA NO PRESENCIAL, \
                           NEGOCIAR VENTA PRESENCIAL,\
                           RECHAZAR CONTRATO, \
                           VENDER CONTRATO, \
```

```
                                DISMINUIR CONTRATO, \
                                MODIFICAR CONTRATO, \
                                VALIDAR CONTRATO, \
                                SUSTITUIR CONTRATO, \
                                CONFIRMAR VENTA NO PRESENCIAL\n")
   else:
      if actual == fila.split(",")[0]:
         hechas[mapeo[fila.split(",")[1]]] = 1
      else:
         if primera[1] == True:
            primera[1] = False
         else:
            transacciones.write(actual + ", \
            " + str(hechas)[1:-1] + "\n")
            hechas = [0,0,0,0,0,0,0,0,0,0,0]
         actual = fila.split(",")[0]
         hechas[mapeo[fila.split(",")[1]]] = 1
archivo.close()
transacciones.close()
```

Enlace de Grupos y Actividades

```
# ----- CARGA DE DATOS -----
archivo = open("agrupado.csv")
clusters = []
primera = True
for fila in archivo:
   if primera == True:
```

```
      primera = False
   else:
      case = fila.split(";")[4]
      cluster = fila.split(";")[5].replace("\n","")
      clusters += [[cluster, case]]
archivo.close()
grupos = {}
for registro in clusters:
   if not(registro[0] in grupos):
   grupos[registro[0]] = []
for cluster, case in clusters:
   grupos[cluster] += [case.replace("\"","")]

# ----- ENLACE DE INSTANCIAS -----
entrenamiento = open("entrenamiento.csv")
temporal = open("clusterizado.csv", "w")
primera = True
grupo = ""
for fila in entrenamiento:
   if primera == True:
      primera = False
      temporal.write(fila[:-1] + ",CLUSTER\n")
   else:
      caseid = fila.split(",")[0]
      for cluster, cases in grupos.items():
         if caseid in cases:
            grupo = cluster
            break
      temporal.write(fila[:-1] + "," + grupo + "\n")
```

```
      grupo = ""
entrenamiento.close()
temporal.close()
```

Anexo B

Resultado Algoritmo *Apriori*

```
1. Negociar venta no presencial=0 7571 ==> Confirmar venta no
presencial=0 7571    conf:(1)
2. Negociar venta=1 7524 ==> Negociar venta no presencial=0
7524    conf:(1)
3. Negociar venta=1 7524 ==> Confirmar venta no presencial=0
7524    conf:(1)
4. Negociar venta=1 Confirmar venta no presencial=0 7524 ==>
Negociar venta no presencial=0 7524    conf:(1)
5. Negociar venta no presencial=0 Negociar venta=1 7524 ==>
Confirmar venta no presencial=0 7524    conf:(1)
6. Negociar venta=1 7524 ==> Negociar venta no presencial=0
Confirmar venta no presencial=0 7524    conf:(1)
7. Negociar venta no presencial=0 Validar contrato=1 2323 ==>
Confirmar venta no presencial=0 2323    conf:(1)
8. Negociar venta=1 Validar contrato=1 2302 ==> Negociar venta
no presencial=0 2302    conf:(1)
9. Negociar venta=1 Validar contrato=1 2302 ==> Confirmar venta
no presencial=0 2302    conf:(1)
10. Negociar venta=1 Validar contrato=1 Confirmar venta no
presencial=0 2302 ==> Negociar venta no presencial=0 2302
```

conf:(1)

11. Negociar venta no presencial=0 Negociar venta=1 Validar contrato=1 2302 ==> Confirmar venta no presencial=0 2302 conf:(1)

12. Negociar venta=1 Validar contrato=1 2302 ==> Negociar venta no presencial=0 Confirmar venta no presencial=0 2302 conf:(1)

13. Negociar venta no presencial=0 Rechazar contrato=1 1696 ==> Confirmar venta no presencial=0 1696 conf:(1)

14. Negociar venta=1 Rechazar contrato=1 1682 ==> Negociar venta no presencial=0 1682 conf:(1)

15. Negociar venta=1 Rechazar contrato=1 1682 ==> Confirmar venta no presencial=0 1682 conf:(1)

16. Negociar venta=1 Rechazar contrato=1 Confirmar venta no presencial=0 1682 ==> Negociar venta no presencial=0 1682 conf:(1)

17. Negociar venta no presencial=0 Negociar venta=1 Rechazar contrato=1 1682 ==> Confirmar venta no presencial=0 1682 conf:(1)

18. Negociar venta=1 Rechazar contrato=1 1682 ==> Negociar venta no presencial=0 Confirmar venta no presencial=0 1682 conf:(1)

19. Negociar venta no presencial=0 Rechazar contrato=1 Validar contrato=1 1326 ==> Confirmar venta no presencial=0 1326 conf:(1)

20. Negociar venta=1 Rechazar contrato=1 Validar contrato=1 1312 ==> Negociar venta no presencial=0 1312 conf:(1)

21. Negociar venta=1 Rechazar contrato=1 Validar contrato=1 1312 ==> Confirmar venta no presencial=0 1312 conf:(1)

22. Negociar venta=1 Rechazar contrato=1 Validar contrato=1

```
Confirmar  venta no presencial=0 1312 ==> Negociar venta no
presencial=0 1312    conf:(1)
23. Negociar venta no presencial=0 Negociar venta=1 Rechazar
contrato=1
Validar contrato=1 1312 ==> Confirmar venta no presencial=0
1312    conf:(1)
24. Negociar venta=1 Rechazar contrato=1 Validar contrato=1
1312 ==> Negociar venta no presencial=0 Confirmar venta no
presencial=0 1312    conf:(1)
25. Rechazar contrato=1 Confirmar venta no presencial=0 1705
==> Negociar venta no presencial=0 1696    conf:(0.99)
26. Negociar venta no presencial=0 7571 ==> Negociar venta=1
7524    conf:(0.99)
27. Negociar venta no presencial=0 Confirmar venta no
presencial=0 7571 ==> Negociar venta=1 7524    conf:(0.99)
28. Negociar venta no presencial=0 7571 ==> Negociar venta=1
Confirmar  venta no presencial=0 7524    conf:(0.99)
29. Rechazar contrato=1 Validar contrato=1 Confirmar venta no
presencial=0 1335 ==> Negociar venta no presencial=0 1326
conf:(0.99)
30. Vender contrato=0 1014 ==> Anular contrato=1 1006
conf:(0.99)
```

Más 158 reglas.

Resultado Algoritmo *FP-Growth*

```
[Validar contrato] --> [Vender contrato] (confidence: 0.123)
[Validar contrato] --> [Anular contrato, Vender contrato]
(confidence: 0.123)
```

[Validar contrato] --> [Rechazar contrato, Vender contrato] (confidence: 0.123)
[Validar contrato] --> [Anular contrato, Rechazar contrato, Vender contrato] (confidence: 0.123)
[Anular contrato] --> [Confirmar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Anular contrato] --> [Negociar venta, Confirmar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Anular contrato] --> [Negociar venta no presencial, Confirmar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Anular contrato] --> [Negociar venta, Negociar venta no presencial, Confirmar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Anular contrato] --> [Negociar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Anular contrato] --> [Negociar venta, Negociar venta no presencial, Validar contrato, Rechazar contrato] (confidence: 0.162)
[Validar contrato] --> [Anular contrato, Rechazar contrato] (confidence: 0.163)
[Anular contrato] --> [Negociar venta, Validar contrato, Rechazar contrato] (confidence: 0.163)
[Negociar venta, Validar contrato] --> [Confirmar venta no presencial, Anular contrato, Rechazar contrato] (confidence: 0.168)
[Negociar venta, Validar contrato] --> [Negociar venta no presencial, Confirmar venta no presencial, Anular contrato,

Rechazar contrato] (confidence: 0.168)
[Validar contrato] --> [Confirmar venta no presencial, Rechazar contrato] (confidence: 0.169)
[Validar contrato] --> [Negociar venta, Confirmar venta no presencial, Rechazar contrato] (confidence: 0.169)
[Validar contrato] --> [Negociar venta no presencial, Confirmar venta no presencial, Rechazar contrato] (confidence: 0.169)
[Validar contrato] --> [Negociar venta, Negociar venta no presencial, Confirmar venta no presencial, Rechazar contrato] (confidence: 0.169)
[Negociar venta, Validar contrato] --> [Negociar venta no presencial, Anular contrato, Rechazar contrato] (confidence: 0.169)
[Negociar venta no presencial, Validar contrato] --> [Confirmar venta no presencial, Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta no presencial, Validar contrato] --> [Negociar venta, Confirmar venta no presencial, Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta, Negociar venta no presencial, Validar contrato] --> [Confirmar venta no presencial, Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta, Validar contrato] --> [Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta no presencial, Validar contrato] --> [Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta no presencial, Validar contrato] --> [Negociar venta, Anular contrato, Rechazar contrato] (confidence: 0.170)
[Negociar venta, Negociar venta no presencial, Validar contrato]

```
--> [Anular contrato, Rechazar contrato] (confidence: 0.170)
[Validar contrato] --> [Negociar venta no presencial, Rechazar
contrato] (confidence: 0.171)
[Validar contrato] --> [Negociar venta, Negociar venta no
presencial, Rechazar contrato] (confidence: 0.171)
[Anular contrato] --> [Confirmar venta no presencial, Rechazar
contrato] (confidence: 0.173)
[Anular contrato] --> [Negociar venta, Confirmar venta no
presencial, Rechazar contrato] (confidence: 0.173)
```

Más 535 otras reglas.

Resultado Algoritmo *Reglas de Inducción*

```
if TIPO CLIENTE = 0 and TIPO CUENTA = VENTA DIRECTA POTENCIAL
then 1  (333 / 1966)
if TIPO CLIENTE = 1 and TIPO CUENTA = GRANDES CUENTAS POTENCIAL
then 0  (1100 / 139)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and TIPO CLIENTE = 1
then 0  (787 / 107)
if TIPO CUENTA = TELEVENTA POTENCIAL and DIVISION = Santiago then
1  (0 / 131)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and DIVISION = X-XI
Region then 0  (33 / 0)
if TIPO CUENTA = VENTA DIRECTA POTENCIAL and DIVISION = VIII
Region Q4 then 0  (108 / 1)
if TIPO CUENTA = TELEVENTA POTENCIAL and TIPO CLIENTE = 0 then 1
(42 / 223)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and DIVISION =
Santiago then 0  (164 / 61)
```

```
if DIVISION = VIII Region TLV Q4 and TIPO CUENTA = VENTA DIRECTA
POTENCIAL then 0  (101 / 9)
if TIPO CUENTA = DIRECCIoN COMERCIAL POTENCIAL then 1  (8 / 127)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and DIVISION = V
Region then 0  (33 / 6)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and DIVISION = III-IV
Region then 0  (35 / 7)
if TIPO CUENTA = GRANDES CUENTAS POTENCIAL and DIVISION = X-XI
Region then 0  (15 / 4)
if DIVISION = V Region Q4 and TIPO CLIENTE = 1 then 0  (45 / 6)
if DIVISION = Santiago Q4 and TIPO CLIENTE = 0 then 1  (5 / 26)
if TIPO CUENTA = GRANDES CUENTAS POTENCIAL and DIVISION = V
Region then 0  (13 / 2)
if DIVISION = Santiago Externos and TIPO CUENTA = DIRECCIoN
COMERCIAL CARTERA then 1  (5 / 31)
if TIPO CUENTA = DIRECCIoN COMERCIAL CARTERA and DIVISION = II
Region then 0  (24 / 7)
if TIPO CUENTA = GRANDES CUENTAS POTENCIAL and DIVISION = IX
Region then 0  (8 / 0)
if TIPO CUENTA = VENTA DIRECTA CARTERA and DIVISION = Santiago
then 1  (2 / 26)
if TIPO CUENTA = GRANDES CUENTAS POTENCIAL and DIVISION = Santiago
Externos then 0  (43 / 27)
if DIVISION = II Region Q4 and TIPO CUENTA = VENTA DIRECTA CARTERA
then 0  (47 / 7)
if TIPO CUENTA = VENTA DIRECTA CARTERA and TIPO CLIENTE = 0 then
1 (31 / 70)
if TIPO CLIENTE = 0 and DIVISION = VIII Region then 0  (55 / 33)
if TIPO CLIENTE = 1 and DIVISION = VI-VII Region then 1  (0 / 22)
```

```
if DIVISION = I Region TLV Q4 and TIPO CUENTA = VENTA DIRECTA
POTENCIAL then 0  (15 / 1)
if DIVISION = II Region and TIPO CLIENTE = 1 then 1  (0 / 16)
if TIPO CUENTA = VENTA DIRECTA CARTERA and DIVISION = II Region
TLV Q4 then 0  (17 / 2)
if DIVISION = V Region then 1  (1 / 22)
if DIVISION = V Region TLV Q4 and TIPO CUENTA = VENTA DIRECTA
CARTERA then 0  (7 / 0)
if TIPO CLIENTE = 0 and DIVISION = IX Region then 0  (11 / 2)
```

Más 42 otras reglas.