PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

DESARROLLO Y ANÁLISIS DE LA UTILIZACIÓN DE ALGORITMOS DE MINERÍA DE DATOS PARA LA BÚSQUEDA DE ANOMALÍAS Y PATRONES SECUENCIALES EN MINERÍA DE PROCESOS

TOMÁS ENRIQUE DEL CAMPO MONSALVE

Tesis para optar al grado de
Magister en Ciencias de la Ingeniería

Profesor Supervisor:
MARCOS SEPÚLVEDA FERNÁNDEZ

Santiago de Chile, Agosto, 2011

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

DESARROLLO Y ANÁLISIS DE LA UTILIZACIÓN DE ALGORITMOS DE MINERÍA DE DATOS PARA LA BÚSQUEDA DE ANOMALÍAS Y PATRONES SECUENCIALES EN MINERÍA DE PROCESOS

TOMÁS ENRIQUE DEL CAMPO MONSALVE

Tesis presentada a la Comisión integrada por los profesores:

MARCOS SEPÚLVEDA FERNÁNDEZ

KARIM PICHARA BAKSAI(PROFESOR INTEGRANTE)

MARCOS SINGER GONZÁLEZ

RICARDO PAREDES MOLINA

Para completar las exigencias del grado de
Magister en Ciencias de la Ingeniería

Santiago de Chile, Agosto, 2011

A Dios, mi familia y mis amigos,
que me apoyaron y aconsejaron en este camino.

"Los amigos son como la sangre, cuando se está herido acuden sin que se los llame."
Anónimo

AGRADECIMIENTOS

Durante el desarrollo de esta investigación hubo muchas personas que estuvieron a mi lado, sin las cuales hubiera sido muy difícil entregar el trabajo que presento hoy. Estas personas me dieron su cariño, conocimiento técnico, consejos y, por sobre todo, su apoyo:

- Prof. Dr. Marcos Ernesto Sepúlveda Fernández de la Pontificia Universidad Católica de Chile.

- Mi familia, especialmente mi madre Pamela Monsalve Bórquez, Ana Erika Bórquez Osorio, María Elena Parga Gana y Raúl Monsalve Bórquez.

- Ing. Alejandro Fuentes de la Hoz e Ing. Guillermo Calderón Ruiz, investigadores del programa de Doctorado de la Pontificia Universidad Católica de Chile.

- David Young, Joaquín Pérez Alarcón, Fabricio De Abranches Rigotto y Lilian Ramírez del equipo de Finanzas de Pampers para Latinoamérica.

- Tomás Jara Kara, Felipe Marambio Abarca y Nicolás Moreno Köhler, del equipo de Sistemas Comerciales e Investigación de Operaciones de Lan Airlines.

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE ILUSTRACIONES

RESUMEN

Durante los últimos años, la Minería de Procesos ha avanzado y crecido dentro de las áreas de estudio de las Tecnologías de Información. Estos avances han permitido a diversas empresas e instituciones realizar análisis objetivos de los procesos que llevan a cabo, a partir de los registros digitales que tengan de éstos. Sin embargo, dentro de estos avances es escaso el trabajo que se ha realizado enfocado en poder encontrar anomalías y patrones secuenciales en un proceso. Es por esto que se han aplicado dos algoritmos utilizados en el área de Minería de Datos para desarrollar dos metodologías: una para poder encontrar casos anómalos y una segunda para permitir la extracción de patrones secuenciales de un proceso. Ambas metodologías se basan en la observación de un gran número de ejecuciones del proceso analizado. Este trabajo contiene la descripción de dichas metodologías, detallando cómo deben ser procesados los datos antes del análisis, el funcionamiento de los algoritmos utilizados, los resultados obtenidos con ellos y como poder realizar una mejor interpretación de éstos. Con este trabajo se ha conseguido abarcar áreas de Minería de Procesos poco exploradas hasta hoy, entregando procedimientos y herramientas que permiten ir desde el Log de Eventos de un proceso hasta un informe final que incluya las anomalías y/o patrones secuenciales presentes en el caso, con la posibilidad de realizar análisis post-aplicación de los algoritmos. Se espera que las metodologías compartidas aquí, sean aplicadas para el análisis de diversos procesos, como también contribuyan al crecimiento de la Minería de Procesos en la búsqueda de anomalías y patrones, sirviendo para el desarrollo de nuevas herramientas y metodologías en esta área.

Palabras Claves: Minería de Procesos, Minería de Datos, Logs de Eventos, Anomalías, Patrones Secuenciales.

ABSTRACT

During the last years, Process Mining has advanced and grown among the Information Technology research areas. These advances have made possible, to several companies and institutions, the analysis of processes, using objective procedures, based on digital registries of these processes. However, among these advances, there are few investigations that have been done with a focus on finding anomalies or sequential patterns in a process. In this work, we have applied two Data Mining algorithms, to develop two methodologies: one to make possible the search of anomalous cases, and another to detect sequential patterns on a process, both based on the observation of several executions of the process under review. This work contains the description of these methodologies, including how data should be processed before the analysis, how the algorithms work, the results obtained using them and recommendations for a better interpretation of these. With this research, it has been possible to cover Process Mining areas hardly explored until now, handing procedures and tools that make possible to go from a Process Event Log, to a final report that includes the anomalies or the sequential patterns found in the analyzed case, including the possibility to make analysis over the results given by the algorithms. It is expected that the methodologies introduced here, will be applied to analyze several processes. Also, this work aims to contribute in the growth of Process Mining, using the research on anomalies and patterns to develop new tools and methodologies in this area.

Key Words: Process Mining, Data Mining, Event Logs, Anomalies, Sequential Patterns

1. Introducción

1.1 Motivación

El campo de la Minería de Procesos es un área de las Tecnologías de Información que ha tomado gran relevancia para diversas industrias e instituciones académicas, dado que las metodologías y herramientas implementadas permiten un análisis objetivo de los procesos, basado en sus ejecuciones actuales (de Medeiros, 2008). El interés por esta área ha llevado al desarrollo de diversos estudios en el tema, sin embargo la mayoría de éstos se han enfocado en la modelación de comportamientos normales de un proceso (Bezerra *et al*., 2009), dejando un amplio campo de estudio en la detección de anomalías y búsqueda de patrones en registros de procesos que presentan resultados no esperados, negativos o particulares. En tanto, la Minería de Datos se caracteriza por el uso de herramientas y algoritmos para analizar grandes cantidades de datos, con el objetivo de encontrar relaciones y patrones previamente desconocidos entre estos datos (Seifert, 2004; Zamarrón *et al*., 2006). Particularmente, dentro de las posibles relaciones que se pueden encontrar, existe la posibilidad de buscar aquellos eventos que son significativamente diferentes al resto de los datos, los cuales son identificados como anómalos (Lazarevic *et al*., 2008). Estas características muestran el potencial que tienen los algoritmos de Minería de Datos para este estudio.

1.2 Hipótesis

La hipótesis de este trabajo estipula que, con un debido pre-proceso de los registros electrónicos que describen un determinado proceso y, mediante el uso de algoritmos utilizados para la Minería de Datos, se puede encontrar anomalías y patrones secuenciales presentes en el proceso.

1.3 Propuesta de solución y objetivos

De acuerdo a la hipótesis planteada, esta investigación buscará contribuir al campo de la Minería de Procesos, a través del desarrollo y evaluación de dos metodologías: una para buscar anomalías y otra, para encontrar patrones secuenciales, utilizando en ambos casos algoritmos construidos para la Minería de Datos. Esta propuesta medirá su éxito en base a dos objetivos principales. El primero es poder concretar una metodología que permita procesar los datos que representan un proceso, para que puedan ser analizados por algoritmos diseñados para la Minería de Datos. Mientras que el segundo objetivo es poder entregar un análisis detallado del funcionamiento de estos algoritmos, cómo deben ser aplicados y cuáles son los resultados obtenidos al utilizarlos para examinar procesos. Con este análisis se buscará comprobar la efectividad de estas herramientas, comparando los resultados con lo que realmente está ocurriendo en el proceso, y así poder dimensionar el valor agregado que pueden entregar estas técnicas al área de Minería de Procesos.

2. Contexto y estado del arte

En esta investigación se trabajó en cuatro áreas de la Ciencia de Computación, todas fuertemente relacionadas entre sí. Estas cuatro áreas son: Minería de Datos, Minería de Procesos, detección de anomalías y búsqueda de patrones secuenciales. En este capítulo se revisarán investigaciones y trabajos, que permiten poner en contexto de cómo cada una de estas áreas de investigación se relacionan con los objetivos perseguidos en este trabajo, como también se revisarán los principales algoritmos, técnicas y metodologías usados en cada una.

2.1 Minería de Datos

Los avances en tecnología en las últimas décadas, han permitido al hombre almacenar cada vez más cantidad de información. Esto se ve favorecido por el hecho de que año a año se requiere menos espacio, y es menos costoso guardar información en formato digital (Moreira, 2006). Uno de los desafíos más importantes, que ha traído este crecimiento en la cantidad de información disponible, es cómo poder obtener nuevo conocimiento y la información más útil de estas grandes sumas de datos (Sumathi *et al*., 2006). La Minería de Datos es un área de las Tecnologías de Información que se encuentra en continuo desarrollo, y busca contribuir a cumplir con este desafío de obtener información valiosa a partir de grandes cantidades de datos, de manera de poder generar acciones y estrategias a partir de éstos.

El uso de Minería de Datos se puede observar en incontables rubros, y con diversos fines. Es tal el caso de bancos, compañías de seguros, empresas de *retail*, universidades, laboratorios, y muchos más, donde la Minería de Datos se utiliza para reducir costos, conocer mejor a los clientes, definir estrategias de mercado, evitar fraudes, entre otros objetivos (Seifert, 2004).

En el trabajo de Seifert se pueden revisar algunas de las técnicas utilizadas en Minería de Datos. Dentro de las más destacadas, se encuentran análisis de regresión, *clustering*, métodos estadísticos en multivariables, reglas de asociación, entre otros. Estas técnicas han permitido y continúan ayudando a solucionar problemas prácticos en diversos ámbitos.

Si bien los algoritmos de Minería de Datos presentan importantes beneficios, al permitir identificar patrones y relaciones en grandes volúmenes de datos, toda la investigación no puede basarse exclusivamente en la aplicación de estos algoritmos. Es fundamental el análisis que realicen los investigadores sobre todo el proceso de generación de conocimiento con Minería de Datos, tanto antes como después de la

aplicación de un algoritmo. Es aquí donde aparecen con fuerza metodologías como KDD, diseñada para dar un significado más acabado a los datos y los resultados de los algoritmos utilizados (Fayyad *et al*., 1996). Esta metodología, en particular, indica la relevancia que tienen el pre-procesamiento de los datos y el análisis post-aplicación del algoritmo de Minería de Datos. En estas dos etapas es fundamental la interacción hombre-máquina, para poder generar conocimiento valioso a partir de los datos disponibles.

2.2 Minería de Procesos

En la sección anterior se comentó sobre los avances de las Tecnologías de Información, en favor del almacenamiento y análisis de datos. Una de las principales áreas de investigación que se ha visto beneficiada por estos avances, es la que estudia ejecuciones de procesos. Día a día, se hace más relevante para distintas compañías y organizaciones estudiar el comportamiento de sus procesos, de manera que puedan generar acciones y estrategias a partir del nuevo conocimiento sobre éstos (Tiwari *et al*., 2008). Es aquí donde la Minería de Procesos encuentra sus más importantes áreas de aplicación. Esta área de las Tecnologías de Información permite la extracción de información a partir de ejecuciones de procesos.

Antes de poder usar cualquier herramienta o aplicar algún algoritmo de Minería de Procesos, es necesario contar con la información adecuada de las ejecuciones de un proceso. Estos datos se almacenan en archivos conocidos como Logs de Eventos y pueden ser extraídos desde sistemas ERP[1] como SAP, CRM[2] como Microsoft Dynamics CRM, entre otros. Incluso, se pueden encontrar Logs en la memoria de equipos

[1] ERP es el acrónimo para el término en inglés "Enterprise Resource Planning", y se usa para describir sistemas computacionales que permiten el manejo de información a través de toda una compañía, cubriendo tanto datos internos como externos de ésta.
[2] CRM se usa como acrónimo para el término en inglés "Customer Relationship Management", el cual describe sistemas computacionales diseñados principalmente para manejar la interacción de una compañía con sus clientes.

tecnológicos como teléfonos celulares, copiadoras y equipos de rayos X, por mencionar algunos (van der Aalst, 2009).

Como se señaló anteriormente, las herramientas y metodologías de Minería de Procesos buscan permitir extraer información a partir de Logs de Eventos, y con esto facilitar la generación de nuevo conocimiento del proceso estudiado. Esta extracción de información, parte de la base que todo Log entregará en cada registro la ejecución de una tarea, y este registro contará con al menos cinco campos (van der Aalst, 2005):

- Un identificador de la tarea registrada;
- Un identificador que permita saber a qué ejecución del proceso corresponde la tarea;
- Información de quién fue el ejecutor de cada tarea;
- Un campo que informe si el registro corresponde al comienzo o fin de la tarea; y
- Fecha y hora en que se ejecutó la tarea.

Además de estos campos básicos, los Logs de Eventos pueden contar con atributos particulares del proceso que se esté analizando. Por ejemplo, si se trata de un proceso de compras por Internet, se puede encontrar campos que informen la cantidad de artículos solicitados y el valor de éstos. Para más ejemplos como éste, y más detalles e información relacionada a Logs de Eventos, se recomienda revisar el capítulo 4 de este trabajo.

Los resultados de los algoritmos y metodologías que entrega la Minería de Procesos, permiten conocer cómo se están ejecutando en la vida real estos procesos. Además, a través de algoritmos, especialmente diseñados o replicados de otras áreas de la Ciencia de la Computación, se logra obtener nuevo conocimiento de las ejecuciones

disponibles en los Logs. Algunos de los algoritmos que se han aplicado para generar nuevo conocimiento son redes neuronales, *clustering*, cadenas de markov, entre otros. El nuevo conocimiento resultante se puede traducir, por ejemplo, en descubrir dónde se encuentran los cuellos de botella o qué dependencias hay dentro de un proceso (Tiwari *et al*., 2008).

Los resultados que se pueden obtener con herramientas y algoritmos de Minería de Procesos, serán importantes en la medida que estos generen decisiones o acciones por parte de los equipos a cargo del proceso. En la práctica, se ha comprobado que los resultados pueden ser usados para (de Medeiros, 2005):

- Alinear de mejor manera los objetivos de la empresa u organización con la ejecución del proceso;
- Mejorar la eficiencia, efectividad y calidad del proceso; y
- Rediseñar el proceso basándose en sólidos argumentos.

Dentro del área de Minería de Procesos, el equipo liderado por Wil M.P. van de Aalst destaca por sus contribuciones a esta área de las Tecnologías de Información. Este grupo, perteneciente a la Universidad de Tecnología de Eindhoven, ha aportado a la Minería de Procesos con diversas investigaciones y publicaciones, llegando incluso a desarrollar uno de los *software* más reconocidos para evaluar procesos. Esta herramienta computacional se llama ProM y permite extraer el modelo de un proceso, a partir de su Log de Eventos. Con este programa también se pueden descubrir discrepancias entre el modelo esperado del proceso, y aquel que resultó de la observación de los registros de las ejecuciones de éste. Por último, una de las características más importantes que presenta este *software*, es que periódicamente se van agregando herramientas (conocidas como *plug-ins*) para enriquecer el conocimiento que se tiene del proceso. Estas herramientas y sus beneficios han sido probados en diversas organizaciones, como municipalidades, bancos, hospitales, agencias de gobierno, entre otros (van der Aalst,

2009). Ha sido tal la expansión de este software, que el desarrollo de más herramientas para incluir en ProM no sólo se realiza en la Universidad de Eindhoven, contando con colaboraciones de investigadores de Australia, China, Alemania, España, entre otros, llegando a contar con más de 250 *plug-ins* para el análisis de procesos.

2.3 Detección de anomalías

La detección de anomalías, tratada en este trabajo, busca descubrir ejecuciones de un proceso que difieran del comportamiento normal observado. La definición de qué es normal y qué no, resulta del análisis de un importante número de ejecuciones del mismo proceso. Esta es una visión particular e innovadora de la detección de anomalías, ya que apunta al trabajo con procesos, lo cual se encuentra escasamente tratado en investigaciones y herramientas de Minería de Procesos (Bezerra *et al.*, 2009).

Si bien no hay una gran cantidad de material sobre la detección de anomalías en procesos, este es un objetivo que han perseguido numerosos trabajos en Minería de Datos y otras áreas de la Ciencia de Computación. Estos trabajos han abarcado diversas áreas de investigación y distintos dominios de aplicación. Se han probado implementaciones desarrolladas específicamente para un problema, como también a través de soluciones más genéricas (Chandola, 2009).

En el trabajo de Chandola se destaca como la detección de anomalías ha sido útil para detectar intentos de fraude con tarjeta de crédito, evaluaciones de compañías de seguro, detección de intrusos en sistemas virtuales, e incluso aplicaciones militares. Todos estos ejemplos de aplicación de detección de anomalías, han sido exitosos porque los resultados encontrados han permitido generar acciones sobre el caso estudiado. Por ejemplo, si se detecta un cambio brusco en el tráfico de visitas de un sitio Web, respecto a lo visto en las últimas semanas, esto puede deberse a un intento de sobrecarga del sistema por parte de intrusos que quieren acceder a información sensible. Otro caso más tradicional es el análisis de las transacciones realizadas con una tarjeta de crédito. Si

repentinamente se ve un aumento inusual en el número de transacciones o en el monto de éstas, probablemente se trate de un robo de la tarjeta. Con esta información se puede, por ejemplo, tomar medidas para frenar un ataque sobre el sistema de la compañía o se puede contactar al dueño de la tarjeta que presenta transacciones anómalas y así confirmar si se debería bloquear su uso.

La detección de anomalías en Minería de Datos presenta una característica muy importante, ésta es que puede trabajar sin necesidad de tener reglas de negocio o conocimiento previo del caso que se está analizando. Otros sistemas de detección necesitan conocer, de antemano, los patrones o características del caso analizado, para poder encontrar situaciones no deseadas. Un ejemplo de estos sistemas son los *firewalls* y antivirus, los cuales bloquean el acceso a programas, usuarios o sistemas que presenten características específicas en su información de conexión o ejecución. Estos sistemas tienen la ventaja de que pueden capturar cualquier tipo de ataque conocido con un alto nivel de confianza, pero esta ventaja se debilita de manera significativa por el hecho de que implementaciones como éstas, no permiten encontrar nuevos tipo de ataque u anomalías. Por esto la importancia de los algoritmos para detectar anomalías que se han desarrollado en Minería de Datos y otras áreas de investigación de la Ciencia de Computación, como Aprendizaje de Máquina. Estos algoritmos permiten encontrar situaciones anómalas a partir de la observación de los datos disponibles, sin necesidad de tener un conocimiento previo del caso analizado. Para esto, los algoritmos definen características de las situaciones normales, y aquellas que difieran significativamente son consideradas como anómalas. De todas maneras, estas metodologías no están exentas de inconvenientes, entre los que destacan falsas alarmas y la dificultad para determinar qué llevó a la definición de qué es normal y qué no (Patcha *et al*., 2007).

En el trabajo de Patcha se revisan algunas de las técnicas más relevantes que se han implementado para la detección de anomalías en distintas áreas de la Ciencia de la Computación. Estas metodologías incluyen algoritmos estadísticos, Aprendizaje de Máquina, y técnicas reconocidas de Minería de Datos como *clustering*, clasificación y

reglas de asociación. Todas estas metodologías presentan la ventaja de permitir descubrir anomalías, sin necesidad de contar con conocimiento previo del caso analizado. También hay una revisión valiosa de las limitaciones y problemas de las distintas alternativas. En el caso de los algoritmos estadísticos se destacó lo sensibles que son cuando se ha asumido normalidad en la distribución de los datos, además del problema fundamental de definir qué es normal y qué no, donde en diversas oportunidades se ignoran factores como la estacionalidad de los datos. En tanto, al analizar algunos sistemas basados en el aprendizaje de máquina, se destaca el hecho que estos cambian su proceso de ejecución al ir recibiendo nueva información. Uno de los principales problemas que se ve en estas técnicas, es la gran demanda de recursos que requiere el sistema para procesar cada nueva información recibida. Finalmente, al analizar algunos algoritmos de Minería de Datos como redes neuronales, generación inductiva de reglas, entre otros, se destaca como principales problema el tiempo que podría demandar entrenar al algoritmo y la alta tasa de falsas alarmas que se pueden presentar.

2.4 Búsqueda de patrones secuenciales

Al igual que en la detección de anomalías, como se señaló en la sección 2.3, la búsqueda de patrones secuenciales no ha sido mayormente abarcada en el área de Minería de Procesos, pero sí es posible encontrar numerosas investigaciones del tema en Minería de Datos. Estas investigaciones parten de la premisa que hay un gran número de posibles patrones secuenciales dentro de una base de datos (Saleeb, 2008).

Un patrón describe un grupo de elementos (eventos, personas, tiempos, etc.) que se presentan recurrentemente en un gran conjunto de datos. En tanto que, cuando se tiene un patrón compuesto por una secuencia de elementos, se tiene un patrón secuencial.

Como se ha comentado en este trabajo, los avances de la tecnología han permitido a empresas y organizaciones almacenar cada vez más cantidad de datos. Estos datos no tienen mayor valor hasta que se extrae información relevante de ellos. Una de las técnicas usadas para entregar ese valor es la búsqueda de patrones secuenciales (Agrawal *et al*., 1995).

En el trabajo de Agrawal, uno de los primeros en tratar la búsqueda de patrones secuenciales, se plantea el problema de una manera simple y precisa, utilizando como ejemplo una empresa que arrienda diariamente miles de películas. Observando los datos de sus clientes con herramientas de búsqueda de patrones secuenciales, los ejecutivos de la compañía pueden descubrir que si alguien arrienda "La Guerra de las Galaxias", hay una alta probabilidad de que luego (no necesariamente de manera consecutiva) arriende "El Imperio Contraataca" y "El Regreso del Jedi". Estas tres películas pertenecen a una famosa trilogía de George Lucas, y en este ejemplo sirven para describir cómo se pueden extraer patrones secuenciales de una gran cantidad de datos. Como se señaló, estos patrones pueden ser formados por elementos o eventos que no sean consecutivos. Por ejemplo, un cliente puede arrendar cada mes una de las películas de la trilogía, y además podría ir más veces entre cada arriendo y buscar otras películas de su interés. Este cliente no arrendó consecutivamente la trilogía, pero es parte del patrón descrito en un principio.

Para encontrar patrones secuenciales, se plantea que los algoritmos deberían presentar ciertas características (Saleeb, 2008):

- El algoritmo debe ser capaz de encontrar todos los patrones que superen el margen de soporte establecido;

- Debe ser altamente eficiente, escalable y debe involucrar un número bajo de revisiones de la base de datos; y

- Debe ser capaz de incorporar varios tipos de restricciones definidas por el usuario. Además del soporte requerido, se podría definir, por ejemplo, número de elementos del patrón, la confianza de los patrones, reglas de negocio, entre otros.

Algunas de las técnicas usadas para la búsqueda de patrones secuenciales incluyen: aplicación de algoritmo Apriori, o variaciones de éste, para encontrar *itemsets* frecuentes, técnicas de Aumento-de-Patrones (del inglés Pattern-Growth) para búsqueda eficiente de patrones basándose en el análisis fragmentado de datos frecuentes (Han *et al*., 2005), análisis de correlación para la búsqueda de reglas de asociación robustas, y búsqueda basada en restricciones, que considera restricciones entregadas por el usuario para reducir los datos a analizar (Batal, 2007), entre otros.

La búsqueda de patrones secuenciales se enfrenta a características de los datos que pueden perjudicar los resultados de una investigación. Algunas de estas características son la presencia de ruido en los datos y la existencia masiva de patrones que no son de interés para el caso de estudio. Estas situaciones deben ser consideradas, especialmente en las fases de pre-procesamiento de los datos y luego de realizada la búsqueda de patrones.

3. Aporte al área de Minería de Procesos

Esta investigación propone dos metodologías: una para descubrir patrones secuenciales usando el algoritmo Apriori, y otra para encontrar casos anómalos en las ejecuciones de un proceso con el algoritmo Interquartile Range. Estas dos metodologías, como se verá en los capítulos a continuación, aplican la metodología de KDD para la obtención de los resultados requeridos. Al usar KDD para diseñar las soluciones propuestas en este trabajo, se busca contribuir a la ciencia de Minería de Procesos con metodologías que abarcan, además del análisis, la preparación adecuada de los datos y

una interpretación de los resultados, con análisis sobre los resultados entregados por los algoritmos Apriori e Interquartile Range.

KDD, del inglés "Knowledge Discovery in Databases", significa "Descubrimiento de Conocimiento en Bases de Datos". Este concepto propone una metodología donde el descubrimiento de conocimiento útil se obtiene fundamentalmente a partir de cinco pasos. En esta metodología, la aplicación de algoritmos de Minería de Datos es un paso importante, pero que debe ser complementado con otras tareas, como el pre-procesamiento de los datos y la apropiada interpretación de los datos (Fayyad *et al*., 1996). En la Ilustración 3.1 se puede observar los pasos que componen la búsqueda de nuevo conocimiento aplicando KDD. Estos pasos pueden ser retomados iterativamente, en la medida que el caso lo requiera. A continuación, se revisará cómo se aplicó este modelo iterativo, a través de las distintas etapas desarrolladas para buscar anomalías y patrones secuenciales en procesos.

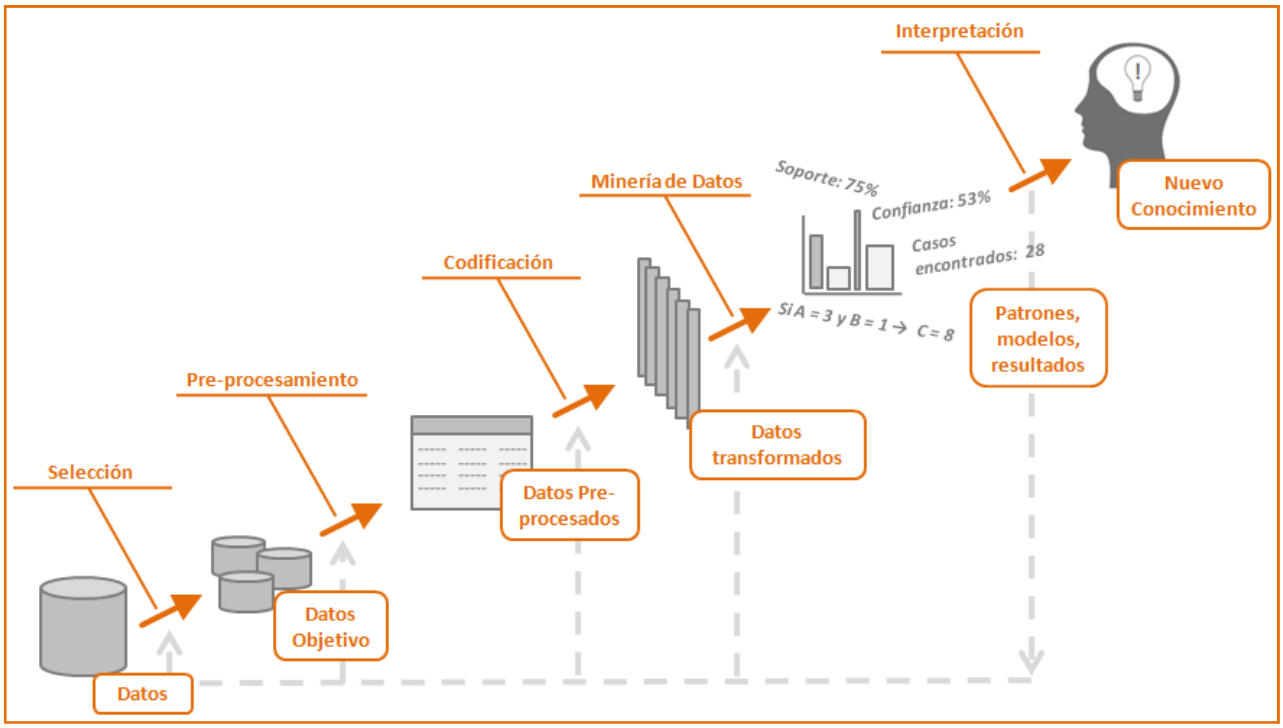


Ilustración 3.1 - Resumen de los pasos que componen la metodología KDD (Fayyad *et al*., 1996)

3.1 Selección

La etapa de selección comprende distintas actividades, entre las cuales se encuentra estudiar los casos a ser analizados. El objetivo es entender lo que están informando los datos que van a ser revisados, comprender el tipo de datos que contienen y el dominio en que se mueven. Este estudio también debe dejar claro cuál es la meta que se persigue al analizar estos datos. Por ejemplo: ¿se está buscando clasificar los datos según un criterio particular?, ¿se desea poder determinar el comportamiento del proceso según los valores de ciertos atributos?, etc. Esta investigación cubre estos puntos con un análisis de cada atributo encontrado en los dos procesos que servirán de casos de estudio.

Además de realizar un análisis de los atributos, la etapa de selección contempla, como bien indica su nombre, una selección de aquellos grupos de datos o atributos que sean más valiosos o afines a los objetivos buscados. Esta selección se realiza como un proceso de limpieza de los datos, donde se puede: eliminar parte de los casos con que se cuenta, eliminar determinados valores de un atributo, o incluso descartar completamente un atributo por no aportar a la investigación.

Como el número de casos a analizar no es excesivo y los atributos se presentan, en general, como relevantes para este estudio, el proceso de limpieza no será necesario al comienzo de esta investigación, pero si será aplicado luego del pre-procesamiento de los datos. Por ejemplo, se tiene un atributo, llamado “caseID”, el cual dada su total variabilidad, no aportará al análisis. Pero, como se verá más adelante, “caseID” se utilizará para poder agrupar la información de cada ejecución, la cual viene originalmente separada por tareas, no como un proceso completo. Debido a esto, se debe mantener “caseID” para el pre-proceso, para luego volver al paso de selección y eliminarlo. Para más detalles de la aplicación de selección en esta investigación, se recomienda revisar los capítulos 4 y 5 de este trabajo.

3.2 Pre-procesamiento y codificación

Las etapas de pre-procesamiento y codificación son abarcadas simultáneamente. Esto se logra mediante el uso de una de las herramientas implementadas en esta investigación, llamada "Pre-procesador Logs", junto con las opciones de pre-procesamiento que entrega Weka[3], y en menor medida en modificaciones manuales que son necesarias sobre determinados atributos.

El principal objetivo que se persigue en las etapas de pre-procesamiento y codificación, es lograr modificar los datos para un óptimo trabajo con los algoritmos de Minería de Datos. Esto se debe realizar siempre cuidando no perder ni alterar el significado y las propiedades de los datos analizados. Buscando este objetivo, se realizan una serie de modificaciones a los datos. La primera modificación aplicada es la estandarización de los ejecutores y tareas que se presentan en los procesos. Dada la variabilidad, y en algunos casos la extensa denominación que presentan estos atributos, se implementó una serie de instrucciones para poder dar un nombre estándar a cada tarea y cada ejecutor. De esta manera, tareas con nombres como "Evaluación Propiedad y Cliente" o "Verificar Disponibilidad", pasan identificarse con un simple T01, T02, etc. Lo mismo se aplica a los ejecutores, los cuales son identificados como E01, E02, etc. Estas modificaciones se realizan pensando en una de las propuestas más relevantes de esta investigación, la cual responde a la necesidad de agrupar los datos, de cada ejecución de un proceso, en un solo registro. Estas modificaciones aparecen como una necesidad fundamental para que los algoritmos a ser utilizados analicen la información como ejecuciones completas de un proceso, y no sólo como tareas individuales. Este proceso, de reunir todos las tareas de una ejecución en un solo registro, viene luego de realizada la estandarización de tareas y ejecutores, momento en que se crean nuevos atributos para agrupar toda la información de cada ejecución del proceso. La

[3] Weka es un *software* de Minería de Datos desarrollado en la Universidad de Waikato de Nueva Zelanda. Está desarrollado en Java y cuenta con herramientas para ejecutar y visualizar algoritmos para el análisis y la generación de modelos predictivos.

implementación de este proceso permitirá eliminar la variabilidad original de los registros correspondientes a cada ejecución, llevando las tareas y ejecutores que se encontraban separados, a un atributo de tareas realizadas y a otro de ejecutores participantes. Cada uno de estos atributos contendrá la secuencia exacta en que se realizaron las tareas y la secuencia en que los participantes las realizaron. De esta manera, se logrará presentar cada ejecución del proceso en un solo registro, a diferencia del número variable en que vienen las ejecuciones originalmente.

Además de proponer la consolidación en un solo registro de cada ejecución, durante el pre-procesamiento y codificación de los datos, se entrega la creación de nuevos atributos a partir de la información que se maneja de los procesos. Uno de estos nuevos atributos entregará el tiempo en que se ejecutó el proceso. Esto se conseguirá a partir de la fecha y hora en que se ejecutó cada tarea del proceso. Sin la creación de este atributo, esta información se perdería y no podría ser parte de los factores a analizar. Otros dos nuevos atributos entregarán una visión simplificada de los ejecutores y las tareas implicadas en cada ejecución del proceso. El orden y la cantidad de repeticiones con que se presentan cada ejecutor y cada tarea, pueden ser diferentes dentro de la ejecución de un proceso. De manera de entregar una alternativa simplificada para analizar estos dos atributos, se crearon atributos donde se consolidarán las tareas y ejecutores, pero donde no se considerará ni el orden ni la cantidad de repeticiones, permitiendo enfocar el análisis en quiénes participaron y qué actividades fueron realizadas en el proceso.

En esta sección se han descrito algunas de las implementaciones más relevantes que abarcan las etapas de pre-procesamiento y codificación. Se recomienda revisar el capítulo 5, para conocer en detalle todas las implementaciones realizadas y el proceso diseñado para lograr las transformaciones mencionadas.

3.3 Minería de Datos

Llegando a la etapa donde ya se cuenta con los datos transformados, como se aprecia en la Ilustración 3.1, el siguiente paso es la aplicación de los algoritmos seleccionados para la investigación. Aquí se propone la utilización del algoritmo Apriori para la búsqueda de patrones secuenciales, e Interquartile Range para extraer los casos anómalos que se puedan presentar en un proceso. La selección de estos algoritmos se basa principalmente en los objetivos que se busca obtener con cada uno de ellos. Ambos algoritmos se basan en la frecuencia de los valores de cada atributo para entregar la información que se está buscando. Apriori deduce reglas de asociación que pueden ser interpretadas como patrones del proceso. Además, este algoritmo presenta otras características importantes, como la opción de que todas las reglas de asociación se construyan para llegar a los distintos valores de un determinado atributo, como ocurre en esta investigación, donde se desea evaluar los factores que llevan a distintos resultados finales de un proceso de venta. Junto a esto, Apriori también permite establecer valores mínimos de soporte y confianza, para así favorecer la búsqueda de reglas más relevantes para la investigación. En cuanto al algoritmo Interquartile Range, éste permite establecer rangos de normalidad, basándose en la posición y frecuencia de cada valor del atributo que se seleccione para buscar los casos anómalos. Este diseño del algoritmo permite que la búsqueda de anomalías no se vea sesgada por la presencia de atributos con valores extremadamente alejados de la media que presenta el proceso. Estos valores se presentarán como anomalías si su posición y frecuencia lo determinan, como también lo harán otros valores que no escapen a la media, pero que su frecuencia sea significativamente distinta al resto de los valores.

Además de seleccionar los algoritmos, en esta investigación se realiza una explicación detallada de cómo estos funcionan, y cuáles son los parámetros más relevantes a la hora de realizar el análisis. Estas explicaciones responden a las indicaciones de esta etapa en la metodología KDD, donde se debe buscar la manera de

seleccionar los modelos y parámetros más apropiados para la búsqueda que se está realizando. Para conocer en detalle cómo funcionan los dos algoritmos propuestos, y los parámetros más importantes de cada uno, se recomienda revisar los capítulos 6 y 7 de este trabajo.

3.4 Interpretación

La metodología KDD indica, como se aprecia en la Ilustración 3.1, que para obtener el óptimo nivel conocimiento de los casos analizados, no basta con aplicar el algoritmo de Minería de Datos y quedarse con esos primeros resultados, sino que es necesario realizar un análisis sobre los modelos, patrones y resultados que haya entregado el algoritmo. Este análisis es posible con las herramientas implementadas para esta investigación. Tanto para la búsqueda de patrones secuenciales, como para la detección de anomalías, se cuenta con herramientas que permiten hacer un último análisis sobre los resultados obtenidos, permitiendo llegar a una interpretación más acabada de estos resultados.

En esta investigación, una vez finalizada la aplicación del algoritmo sobre los datos, se entrega la posibilidad de analizar los resultados obtenidos, observando cómo se distribuyen los valores de los distintos atributos. Estas distribuciones se exponen con estadísticas, tomando como base el total de ejecuciones, junto a otras con base en aquellos casos que indiquen los resultados del algoritmo. Con estos indicadores, se podrá descubrir aquellos valores de los atributos que tengan mayor o menor relevancia en los casos anómalos, o en aquellos definidos por los patrones encontrados. Con esta información se podrá llegar a conclusiones más determinantes y con mayores fundamentos, lo que ayudará a accionar de manera más precisa sobre los resultados obtenidos. El accionar sobre los resultados, es precisamente otro punto importante en la etapa de Interpretación de la metodología KDD, y en esta investigación se apoya esta tarea con el análisis recién descrito y con conclusiones sobre las distintas búsquedas

realizadas. Para ver aplicaciones del análisis mencionado, y las conclusiones obtenidas, se recomienda revisar el capítulo 8 de este trabajo.

4. Fuentes de datos para la investigación

El objetivo principal de esta investigación es analizar el desempeño al aplicar algoritmos de Minería de Datos a conjuntos de datos que describen procesos, para así extraer nueva información a partir de ellos. Estos conjuntos de datos pertenecen a una categoría particular dentro de las Tecnologías de Información y son conocidos como Logs de Eventos.

En este capítulo se revisará las principales características de los Logs y se describirán los conjuntos de datos que serán utilizados para la investigación.

4.1 Log de Eventos: definición y características

Los Logs de Eventos son conjuntos de datos, que pueden ser generados a partir de ejecuciones reales de procesos. Esta generación puede nacer de diversos sistemas de información, por ejemplo sistemas ERP como SAP y sistemas CRM como Microsoft Dynamics CRM. Cada Log presenta diversas ejecuciones de un mismo proceso. Y cada ejecución (de ahora en adelante "caso" en este trabajo) contiene, en su "presentación ideal", todas las etapas (de ahora en adelante "tareas" en este trabajo) que se ejecutaron en ese caso del proceso. Se menciona la "presentación ideal", ya que pueden existir Logs donde sólo algunos casos contengan todas sus tareas, siendo posible encontrar casos incompletos que pueden entorpecer o dificultar cualquier investigación.

Los Logs de Eventos utilizados en este trabajo están almacenados en archivos de extensión .mxml. Estos archivos permiten almacenar de manera estandarizada todos los

casos registrados sobre un proceso, con todos los atributos disponibles para cada tarea desarrollada.

La estructura de un archivo .mxml se basa en la forma utilizada de almacenar datos en archivos .xml (extensión comúnmente utilizada para almacenar datos estructurados). Siguiendo este formato, el atributo se almacena entre llaves "< >". Un ejemplo de esto se aprecia en el registro de la hora en que se realizó la actividad "Registrar Pedido", en el primer caso almacenado en el Log del proceso "Venta de artículos a través de página Web" (extracto de este Log en su formato original disponible en Anexo 1). El registro de esta hora en el Log se muestra así: "<Timestamp>2009-12-31T21:01:00.000+01:00</Timestamp>". Se puede apreciar que la descripción del atributo se encuentra al comienzo y al final del registro, y el valor del atributo se ingresa entre el límite establecido por las llaves que se encuentran al principio y al final de la declaración.

Independiente del caso de negocio analizado, se puede asumir que los Logs presentan características en común al registrar cada tarea, como se resume en la Tabla 4.1. Como se puede apreciar en esta tabla, las características en común están asociadas a ciertos atributos que se asume se encuentran en todos los Logs de Eventos (van der Aalst, 2005). Además de estas características comunes, en algunos casos los Logs también pueden contar con atributos adicionales, no comunes para todos los procesos, que pueden ser considerados como "particulares" o "exclusivos" de cada caso de negocio. Por el momento no se profundizará en este tipo de atributos, ya que no es obligatoria su inclusión en los Logs. Estas características particulares suelen ser ambiguas en su presentación, dado que sus nombres y distintos valores no explicitan el significado de estos campos. Sin embargo, al describir y analizar los Logs utilizados para este trabajo se verán necesariamente algunos ejemplos de estos atributos.

Para conocer con mayor detalle cómo se organizan los datos en un Log y ver más ejemplos de las características descritas en la Tabla 4.1, en el Anexo 1 se pueden

encontrar extractos de los Logs de Eventos utilizados en este trabajo en su formato original (es decir, desde archivos .mxml).

Característica	Descripción
1	Cada tarea registrada está **relacionada a un caso particular del proceso**. Todas las tareas relacionadas a un mismo caso tendrán el mismo atributo numérico, denominado ***caseID*** (p.ej. 1, 15, 23, etc.)
2	Cada tarea es registrada **en el orden que efectivamente se ejecutó**. Esto se puede verificar gracias al atributo ***timestamp,*** él cual informa el momento en que se registró la tarea. El formato usado tradicionalmente en el *timestamp* es: día-mes-año y hora-minutos-segundos
3	Toda tarea registrada en el *log,* describe que tarea se está ejecutando. Este hecho queda registrado en el atributo ***taskID*** (p.ej. Ingresar Pedido, Enviar Factura, Enviar Mail, etc.)
4	Las tareas registradas pueden aparecer más de una vez en un mismo caso, esto puede deberse a que la tarea se repite o a que se detalle el estado de la actividad. Este estado se indica mediante el atributo ***eventtype***, y normalmente se distinguen dos estados: comenzado o completado
5	El ejecutor de cada actividad se registra en el atributo ***originator***, normalmente es una persona o un sistema determinado (p.ej. Caja 1, Vendedor 3, Almacén, etc.)

Tabla 4.1 - Características comunes de las tareas registradas en un Log

4.2 Logs de ejemplo seleccionados

Para la selección y confección de los conjuntos de datos que serían utilizados en esta investigación, se buscó contar con un caso de negocio acorde a cada uno de los dos objetivos buscados: el primero es encontrar casos anómalos entre los distintos casos registrados de un proceso, mientras que el segundo es distinguir patrones secuenciales en éstos. Para cada uno de estos objetivos se contó con la ayuda de un investigador[4], para

[4] Guillermo Calderón y Alejandro Fuentes de la Hoz, investigadores del programa de Doctorado en Ciencia de la Ingeniería de la Pontificia Universidad Católica de Chile, confeccionaron los conjuntos de

idear y construir los datos a ser evaluados. Estos datos debían corresponder a negocios que tuvieran dentro de sus objetivos principales la información ofrecida en este trabajo.

A continuación se revisarán las características de los dos casos de estudio seleccionados, con énfasis en mostrar el porqué fueron escogidos para esta investigación, junto con la descripción de la información encontrada en los Logs (p.ej. qué atributos presenta cada tarea, qué tipos de datos hay, cuántos registros hay disponibles, etc.).

4.2.1 Venta de artículos a través de página Web

El primer conjunto de datos confeccionado para esta investigación, consiste en un Log que contiene información de transacciones realizadas a través de una página Web, la cual vende artículos de entretenimiento general (p.ej. libros, *dvds*, *cds*, etc.). La característica principal de este Log es que cuenta con un número de transacciones que terminaron en una venta exitosa, mientras el resto de los casos concluyeron sin concretar la venta. Esta característica hace muy interesante y atractiva la búsqueda de patrones secuenciales, para así descubrir qué factores fueron determinantes o recurrentes para concluir en una venta exitosa o no.

El Log confeccionado para este caso, consta de un total de mil casos, de los cuales 780 describen ventas exitosas y 220 son registros de ventas que no se concretaron.

Cada una de las tareas, registradas a través de los mil casos, presenta los cinco atributos básicos, o comunes, almacenados en un Log. Además de estos campos básicos, el Log cuenta con tres atributos

datos utilizados para la evaluación de las metodologías de búsqueda de patrones secuenciales y detección de anomalías, respectivamente.

característicos, específicos para este caso de negocio. Estos ocho atributos, detallados en la Tabla 4.2, conforman la información disponible a analizar de cada tarea. En esta tabla se ha consolidado la información fundamental de cada atributo, incluyendo el tipo de dato, junto a una descripción y ejemplos para contextualizar la información que entregan.

Atributo	Tipo de dato	Descripción	Ejemplos
caseID	numérico	Atributo básico. Señala a qué caso corresponde la tarea registrada	1, 35, 578
taskID	texto	Atributo básico. Aquí se entrega la descripción de la tarea ejecutada	Registrar Pedido, Armar Pedido, Procesar Pago
originator	texto	Atributo básico. Informa quién fue el ejecutor de la tarea registrada	system, Almacen1, Caja2
eventtype	texto	Atributo básico. Registra el estado de la tarea	complete
timestamp	fecha	Atributo básico. Indica la fecha y hora en que se registró la ejecución de la tarea	31-12-2009 17:01:00, 08-01-2010 8:21:00
cantidad	nominal	Atributo característico. En este campo del Log se guarda la cantidad de productos solicitado por el cliente. Es nominal dado que tiene un valor que no corresponde a un número: ">3"	"1", "2", "3" o ">3" (mayor a tres)
monto	nominal	Atributo característico. Indica el valor monetario del producto solicitado por el cliente. Es nominal, dado que presenta un valor que no corresponde a un número: ">3000"	"1000", "2000", "3000" o ">3000" (mayor a tres mil)
OK	texto	Atributo caracterítico. Aquí se indica si la venta se concretó con éxito, o si fue cancelada. Este atributo es clave en este conjunto de datos.	OK ó NOK

Tabla 4.2 - Descripción de los ocho atributos que describen las tareas del proceso "Venta de artículos a través de página Web"

4.2.2 Evaluación solicitud de crédito hipotecario

El segundo caso elaborado para este estudio muestra cómo se ejecuta el proceso de evaluación de un crédito hipotecario, para su final aprobación o rechazo. Este proceso es registrado en el Log desde que el cliente hace la solicitud formal al banco, hasta que esta institución entrega la resolución final. Entre el comienzo y el final de este proceso, hay diversas tareas ejecutadas y numerosos ejecutores involucrados. Las tareas y participantes varían constantemente, y es esta variabilidad la que hace altamente interesante este caso para la búsqueda de anomalías. En este análisis de anomalías se buscará responder preguntas como: ¿qué equipos de ejecutores fueron los que menos participaron a través de los casos registrados en el Log? o ¿qué equipos se demoraron más en entregar la resolución final al cliente?, entre otras. Estas y otras preguntas son las que podrían llegar a ser resueltas, mediante la búsqueda de casos anómalos dentro de este caso de negocio.

El conjunto de datos, elaborado para el análisis del proceso de solicitud de un crédito hipotecario, cuenta con un total de 2.485 casos. Estos casos fueron divididos en cuatro grupos, con el objetivo de encontrar distintos tipos de anomalía en cada conjunto. Cada uno de estos conjuntos presenta un número distinto de casos, el primero registra 590 solicitudes para obtener un crédito hipotecario, el segundo 580, el tercero 740, mientras que el cuarto y último cuenta con 575 casos para ser analizados.

Las tareas que componen cada uno de los más de dos mil casos disponibles son descritas, como se detalla en la Tabla 4.3, solamente utilizando los cinco atributos básicos contenidos en un Log. La ausencia de atributos característicos en este caso de negocio se justifica en el hecho

de querer priorizar la búsqueda de anomalías de tipo organizacional, lo que implica enfocarse en quiénes realizaron el trabajo, qué tareas realizaron y el tiempo que les tomó realizar éstas.

Atributo	Tipo de dato	Descripción	Ejemplos
caseID	numérico	Atributo básico. Señala a qué caso corresponde la tarea	15, 28, 157, 268
taskID	texto	Atributo básico. Aquí se registra en qué consistía la tarea a realizar	Solicitud Crédito, Estudio de Títulos, Revisión Final
originator	texto	Atributo básico. Informa quién fue el ejecutor de la tarea	Cliente 1, Ejecutivo 3, Analista de Riesgo 2, Abogado 2
eventtype	texto	Atributo básico. Registra el estado de la tarea	complete
timestamp	fecha	Atributo básico. Indica la fecha y hora en que se registró la ejecución de la tarea	10-02-2010 13:35:43, 10-02-2010 14:59:24

Tabla 4.3 - Descripción de los cinco atributos que describen las tareas del proceso de "Evaluación de solicitud de crédito hipotecario"

5. Pre-procesamiento de los datos

Como se señaló en el capítulo 3, el pre-procesamiento de los datos es una etapa fundamental y crítica para los resultados finales que se obtienen al trabajar con algoritmos de Minería de Datos. Esta es la instancia en que el investigador tiene la posibilidad de filtrar casos y modificar la presentación de los datos, para así favorecer la obtención de mejores resultados con el análisis a ser realizado.

Las modificaciones que se ejecuten en el pre-procesamiento, deben realizarse con cuidado para no perder información valiosa de los datos, enfocándose en el uso de

técnicas de estandarización, filtrado y agrupación de los datos. Estas técnicas, como se apreciará en este capítulo, son aplicadas en distintas etapas del pre-procesamiento diseñado para este trabajo.

5.1 Resumen del procedimiento

Las transformaciones diseñadas para el pre-procesamiento de datos fueron ideadas, e implementadas, buscando obtener como resultado un conjunto de datos que representara mejor y de manera más consolidada el proceso que describe. Este objetivo fue establecido luego de analizar las características de los Logs con que se contaba para analizar los procesos de este trabajo. El formato original de estos procesos, en sus respectivos Logs (ver Anexo 1), muestra sólo una tarea por cada registro capturado. Con esta información se podría buscar anomalías y patrones a nivel de cada tarea, pero no se lograría un análisis a nivel de casos completos de un proceso. Para conseguir este tipo de análisis, el pre-procesamiento cuenta con ocho pasos. Como se describe en la Tabla 5.1, se comienza con el Log original y se termina con una planilla de datos, donde cada fila de datos contiene toda la información disponible de un caso completo del proceso. Lograr reunir los datos en este formato permite, además de contar con cada caso consolidado en una línea, que sean correctamente leídos por herramientas de Minería de Datos.

5.2 Herramientas seleccionadas

A menos que se indique lo contrario, todo el trabajo descrito en este capítulo se desarrolló en Excel, programa computacional para el manejo de datos en planillas. La versión de este programa, usada para este trabajo, es la 2007.

Se escogió trabajar con Excel, para la mayor parte del trabajo de pre-procesamiento, basándose principalmente en cuatro puntos favorables de este programa:

Etapa	Descripción del procesamiento realizado
1	Se exportan datos desde el Log de eventos (de extensión .mxml) a un archivo separado por comas (de extensión .csv)
2	Se ingresan los datos almacenados en el archivo .csv a una planilla Excel para filtrar y procesar los datos
3	Se identifican las tareas y ejecutores presentes a través de todos los casos del proceso
4	Se extraen las secuencias de tareas y ejecutores como efectivamente ocurrieron en cada caso del proceso
5	Se identifican los equipos de trabajo y grupos de tareas consolidadas a través del proceso. A diferencia del paso anterior, aquí no se considera el orden de ejecución de las tareas, ni tampoco las repeticiones de éstos
6	Se obtienen los atributos de tiempo a partir del campo "timestamp"
7	Se procesan, si existen, los atributos "característicos" del proceso de negocio
8	Finalmente, se construye un archivo de de extensión .csv, el cual consolida todos los datos procesados y puede ser analizado con herramientas de Minería de Datos

Tabla 5.1 - Descripción de las etapas que componen el pre-procesamiento de los datos de cada proceso

1. Es una herramienta donde el alumno tesista presenta amplia experiencia;
2. Aún siendo una herramienta computacional pagada, es de acceso y manejo popular, lo cual favorecerá la comunicación y expansión del trabajo aquí expuesto;
3. Permite trabajar con datos de manera ordenada y clara, contando siempre con el apoyo de la interfaz gráfica, para verificar cada una de las etapas de pre-procesamiento realizadas; y
4. Permite trabajar con el lenguaje de programación Visual Basic, facilitando la automatización de tareas, y así poder desarrollar una

herramienta para el análisis de miles de datos de manera rápida, estandarizada e independiente del proceso que se esté analizando.

5.3 Desarrollo del pre-procesamiento

5.3.1 Exportación de datos desde .mxml a .csv[5]

Como se señaló en la sección 4.1, los Logs de Eventos utilizados en este trabajo están almacenados en archivos de extensión .mxml. Se recomienda revisar la sección mencionada para más detalles de las características de estos archivos.

Para que se pudiera trabajar en Excel con los datos almacenados en un archivo .mxml, se debía exportar los datos a un formato compatible con las planillas de cálculo. Esta exportación fue posible gracias a la utilización de una herramienta implementada en ProM (más información de este *software* en la sección 2.2 de este trabajo), identificada como "CSV for Log Exporter" (procedimiento para exportar un archivo .mxml a .csv descrito en Anexo 2). Con esta herramienta, se consiguió almacenar toda la información de los casos de un proceso en un archivo separado por comas (de extensión .csv), que puede ser interpretado por Excel.

5.3.2 Ingresar datos de archivos .csv a planilla para pre-procesamiento de datos

Parte del trabajo de esta investigación consistió en el desarrollo de una herramienta en Excel, donde a través de la programación de Macros en Visual Basic se automatizaron gran parte de las tareas descritas en este

[5] Los archivos .csv (del inglés Comma-Separated Values) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal: España, Francia, Italia, etc.) y las filas por saltos de línea.

capítulo. Esta planilla fue denominada como “Pre-procesador Logs” (instrucciones de descarga archivo .xlsm en Anexo 4).

El primer paso necesario para utilizar esta herramienta consiste en ingresar los datos, contenidos en el archivo .csv, dentro de esta planilla “Pre-procesador Logs” (descripción de cómo ingresar datos en la planilla en Anexo 4). Una vez que la planilla ya cuenta con los datos del proceso, se procede a filtrar y depurar estos datos dentro de la herramienta. El resultado final de este proceso es un archivo que consolida toda la información del proceso, en un formato acorde al utilizado para el análisis con herramientas de Minería de Datos.

5.3.3 Identificar tareas y ejecutores

El primer proceso implementado en la herramienta desarrollada en Excel (de ahora en adelante identificada como “Pre-procesador Logs” en este trabajo) permite identificar, de manera automática, las tareas y ejecutores presentes a través de todo el proceso.

La labor del “Pre-procesador Logs” en esta etapa es revisar todas las tareas y ejecutores que participaron en cada caso registrado del proceso. De todos los datos revisados, la herramienta extrae una única instancia de cada tarea y cada ejecutor. Con esta información se elabora una lista consolidada de ejecutores y una de tareas. El proceso para construir esta última lista se puede apreciar en el ejemplo incluido en la Ilustración 5.1.

En las listas consolidadas se asigna a cada tarea, y cada ejecutor, un identificador único estándar, compuesto de una letra y un número. Estos identificadores se utilizarán para referenciar a las tareas y ejecutores

durante todo el pre-procesamiento y las posteriores etapas de análisis de los datos. El objetivo de los identificadores es estandarizar y simplificar la información contenida originalmente en el Log de Eventos. Estos identificadores serán dos nuevos atributos: para describir las tareas el atributo será nombrado como "ID_Tarea" (por identificador de tarea) y para los ejecutores será "ID_Ejecutor" (por identificador de ejecutor).

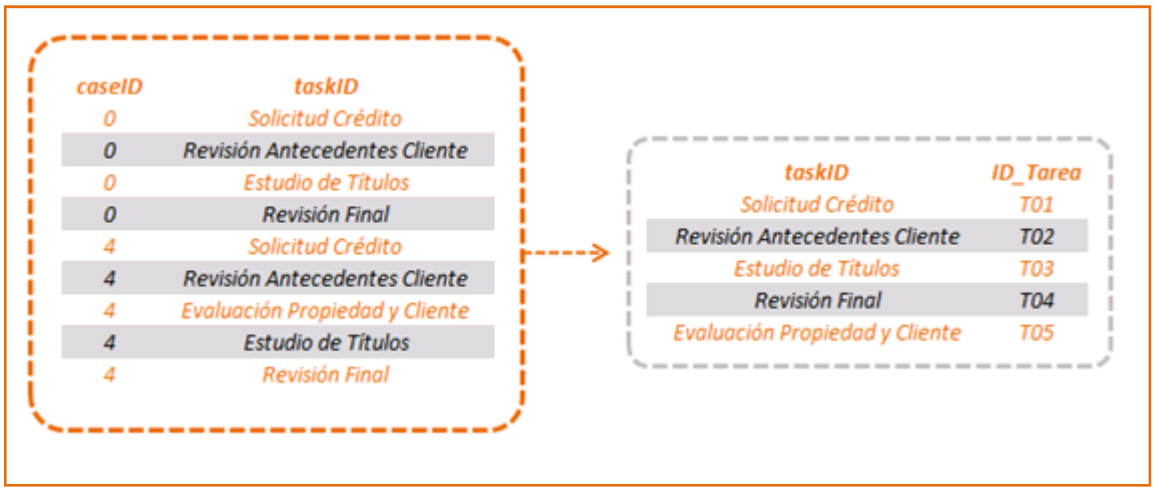


Ilustración 5.1 - Ejemplo del proceso de creación de la tabla consolidada de tareas, donde se aprecia como la herramienta sólo toma una instancia de las tareas ejecutadas en los casos 0 y 4

5.3.4 Identificar secuencias de tareas y de ejecutores

Una de las principales dificultades, que presentan los datos almacenados en el archivo .csv, es que cada caso de un proceso está repartido en varias filas de la planilla, donde cada fila representa una tarea del proceso. Esta representación de los datos es poco eficiente para el análisis de procesos con herramientas de Minería de Datos, ya que los algoritmos toman cada fila como un caso para analizar. Al hacer esto, considerarían cada tarea como un caso independiente, no como parte de un proceso. Para solucionar este inconveniente se implementaron

instrucciones en el “Pre-procesador Logs” para consolidar toda la información de cada caso, en una sola fila. La primera de estas instrucciones consiste en crear un único atributo que describe la secuencia de tareas, y otro único atributo con la secuencia de ejecutores para cada caso registrado del proceso. Para que la identificación de estas secuencias sea exitosa, es necesario que las actividades estén registradas en el Log de Eventos en el orden que efectivamente fueron ejecutadas.

El primer paso para la búsqueda de secuencias de tareas, ejemplificada en la Ilustración 5.2, consiste en ir revisando paso a paso cada ejecución del proceso. Durante la revisión se va anotando, junto a cada tarea, todas las tareas realizadas hasta esa instancia del proceso, identificando cada una con su respectivo “ID_Tarea”, atributo creado en la etapa anterior del pre-procesamiento de datos. Esta información queda registrada en el campo denominado como “Construcción Camino”.

El segundo paso, y final de esta etapa, también ejemplificado en la Ilustración 5.2, se encarga de asignar un identificador a la secuencia final de cada caso. La secuencia final corresponde al último registro encontrado, para cada caso en el campo “Construcción Camino”. El “Pre-procesador Logs” toma todas las secuencias finales y les asigna un identificador denominado “ID_Camino” (por identificador de camino). Hay un identificador único para cada secuencia distinta que se encuentre (es decir, si dos casos del proceso tuvieron la misma secuencia, su “ID_Camino” será el mismo). El “ID_Camino” se compone de la letra “C” (por camino) y un número único para cada secuencia diferente encontrada.

El proceso para construir las secuencias de ejecutores es completamente análogo al recién descrito para las secuencias de tareas.

Las diferencias radican en que, en vez de tomar los identificadores encontrados en "ID_Tarea", se toman los datos bajo el atributo "ID_Ejecutor", mientras que el registro de todos los participantes hasta cada instancia del proceso se va registrando en el campo "Construcción Equipo", para finalmente identificar cada secuencia completa de ejecutores bajo el atributo denominado como "ID_Equipo" (por identificador de equipo).

Ilustración 5.2 - Ejemplo de los pasos realizados para construir el atributo "ID_Camino" de cuatro casos del proceso "Evaluación solicitud de crédito hipotecario"

Es importante destacar que tanto las secuencias de tareas, como las de ejecutores, consideran el orden exacto con que se ejecutó cada caso. Por esto, también se registrarán las repeticiones de tareas y ejecutores que se puedan presentar. Esto se realiza de esta manera, para no perder información del caso y conservar las propiedades que

originalmente presentan los datos de cada proceso. Luego, también durante el pre-procesamiento, se armarán secuencias más simples, que no consideran repeticiones ni orden.

5.3.5 Identificar conjuntos consolidados de tareas y ejecutores

Parte de los objetivos del pre-procesamiento de los datos es poder generar, a partir de los datos disponibles, atributos que favorezcan la obtención de información relevante en el posterior proceso de análisis de los datos. Como se mencionó en el capítulo 3, siguiendo la metodología KDD, la generación de estos atributos puede nacer en las etapas de Pre-procesamiento y Codificación. Para esto, se puede utilizar, por ejemplo: filtrado, normalización o consolidación de datos. Este último método fue aplicado en el paso anterior y también se utiliza en la etapa a ser descrita en esta sección, donde se mostrará el proceso para obtener grupos consolidados de las tareas y ejecutores de cada proceso. Estos grupos consolidados mostrarán las tareas y ejecutores, suprimiendo dos características que se consideraban en los grupos creados en el paso anterior. Estas características suprimidas son las repeticiones y el orden en que aparecen las tareas o ejecutores.

La consolidación de tareas y ejecutores permitirá asociar casos donde se hayan ejecutado las mismas actividades, o donde hayan participado el mismo grupo de ejecutores, sin que necesariamente el orden o la cantidad de repeticiones dentro de cada caso haya sido la misma. Se espera que esta simplificación en el criterio de creación de grupos, permita disminuir la cantidad de grupos existentes y así asociar más casos a un mismo equipo de ejecutores o a una misma serie de tareas.

Los grupos consolidados de tareas son registrados bajo el atributo denominado “ID_TareasConsolidadas” (por identificador de tareas consolidadas), mientras que el atributo para los grupos consolidados de ejecutores se denomina “ID_EquipoConsolidado” (o identificador de equipo consolidado). Ambos atributos diferenciarán los distintos grupos, haciendo uso de un identificador formado por un número antecedido por las siglas “TC” (por tareas consolidadas) en el caso de las tareas, y por “EC” (por equipo consolidado) en el caso de los ejecutores.

Para construir los grupos consolidados se implementó en el “Pre-procesador Logs” una rutina automática que, tomando todas las tareas y ejecutores de un caso, elimina las repeticiones en ambos grupos de datos, y ordena en orden creciente los elementos, de acuerdo al número de su identificador: “ID_Tarea” en el caso de las tareas y “ID_Ejecutor” para ordenar los ejecutores. Esta rutina es la que permite lograr asociaciones que se habrían perdido teniendo sólo los atributos “ID_Camino” y “ID_Equipo”. Por ejemplo, si se observan dos ejecuciones del caso “Evaluación solicitud de crédito hipotecario”, como se ejemplifica en la Ilustración 5.3, aunque el número de tareas no es igual en ambos casos, los participantes son los mismos. Pero, ya que en uno de los casos (caseID = 6) un miembro del equipo actuó más de una vez (ID_Ejecutor = Ejecutivo 1), no tienen el mismo “ID_Equipo”. Esta diferencia hará que, al analizar los datos, estos casos sean considerados como totalmente distintos, cuando la realidad es que existe una relación entre ellos, y el atributo “ID_EquipoConsolidado” se encargará de establecer esa conexión dando el mismo identificador a ambos casos. Este mismo procedimiento de consolidación se realiza, de la misma manera, analizando las tareas para crear el atributo “ID_TareasConsolidadas”.

Ilustración 5.3 - Ejemplo de clasificación de un equipo con los atributos "ID_Equipo" y "ID_EquipoConsolidado"

5.3.6 Identificar atributos de tiempo

Los Logs de Eventos cuentan con un atributo de tiempo denominado "timestamp". Este atributo registra el año, mes, día, hora y minutos en que ocurrió el evento. Estos datos podrían ser de gran utilidad para la etapa de análisis, pero también necesitan de un apropiado pre-procesamiento.

Aprovechando la existencia del atributo "timestamp", se programó un procedimiento en el "Pre-procesador Logs" para obtener automáticamente el tiempo transcurrido entre dos tareas, y a partir de ese cálculo, el tiempo acumulado en cada tarea desde el comienzo del proceso. Con este último atributo también se obtiene el tiempo total del proceso. Los resultados de estos cálculos se registran en dos nuevos atributos de la información registrada de cada tarea: "Tiempo_Desde_Ultima_Tarea" y "Tiempo_Desde_Comienzo".

Para calcular el tiempo entre dos tareas se toma el "timestamp" de una tarea ("n" en la descripción genérica de la Ilustración 5.4) y se le resta el "timestamp" de la tarea inmediatamente anterior ("n-1" en la descripción genérica) o, en otras palabras, la última tarea ejecutada previamente. El resultado de la resta queda expresado en días. Para llevar este resultado a minutos, unidad de tiempo seleccionada para este trabajo, es necesario multiplicar el resultado obtenido por 1440 (1 día = 1440 minutos).

Para la primera tarea del proceso (n = 1)

$$Tiempo\_Desde\_Ultima\_Tarea(t_1) = 0$$

Para la segunda y demás tareas (n > 1)

$$Tiempo\_Desde\_Ultima\_Tarea(t_n) = (timestamp(t_n) - timestamp(t_{n-1})) * 1.440$$

Ilustración 5.4 – Descripción genérica de las fórmulas utilizadas para calcular el tiempo transcurrido desde la última tarea

Para calcular el tiempo que ha pasado, desde el comienzo del proceso hasta cualquiera de las tareas, primero se busca el tiempo transcurrido desde la tarea anterior. A esta cifra se debe sumar el tiempo transcurrido, desde el comienzo del proceso hasta la tarea anterior. Este proceso iterativo, el cual se detalla en la descripción genérica en la Ilustración 5.5, entregará el tiempo transcurrido desde que comenzó el proceso hasta cada una de las tareas que se ejecutan, con lo cual también se obtendrá el tiempo total de cada caso en el proceso analizado.

Para la primera tarea del proceso (n = 1)

$$Tiempo\_Desde\_Comienzo(t_1) = 0$$

Para la segunda y demás tareas (n > 1)

$$Tiempo\_Desde\_Comienzo(t_n) = Tiempo\_Desde\_Ultima\_Tarea(t_n) + Tiempo\_Desde\_Comienzo(t_{n-1})$$

Ilustración 5.5 - Descripción genérica de las fórmulas utilizadas para calcular el tiempo transcurrido desde el comienzo del proceso hasta la actividad o tarea "n"

5.3.7 Procesamiento de atributos característicos

Al trabajar con Logs de Eventos, se asume que se encontrarán los cinco atributos básicos como descriptores de cada tarea. Además de estos atributos, no es requisito ni se puede asumir la existencia de otros campos con información "característica" o "exclusiva" del proceso analizado. A pesar de esta situación, es importante contar con un procedimiento que permita integrar estos atributos "característicos" a la información del proceso, ya que cuando estos existen su análisis podría entregar importantes resultados para el estudio.

Los valores de los atributos "característicos" de un proceso, como se puede apreciar en la Ilustración 5.6, pueden no aparecer en todas las filas que describen cada caso, ya que podrían estar asociados exclusivamente a una tarea particular del proceso. Dada esta situación, se programó el "Pre-procesador Logs" para que revisara una a una las filas

de cada proceso extrayendo los valores “característicos” que aparecieran en cualquiera de las tareas ejecutadas, para finalmente reunirlos todos en una única fila que contiene todos los atributos de cada ejecución del proceso.

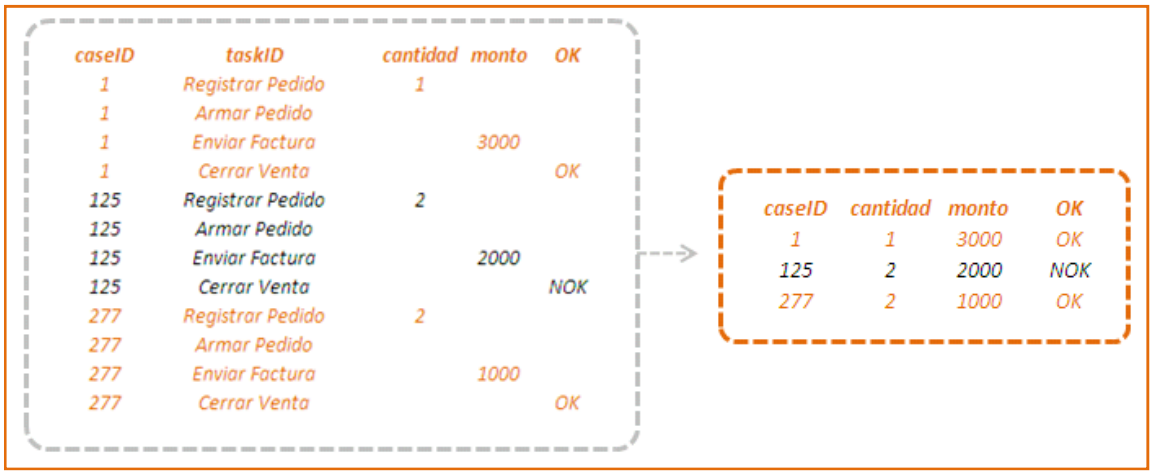


Ilustración 5.6 - Ejemplos de cómo el "Pre-procesador Logs" consolida los atributos “cantidad”, “monto” y “OK” en tres extractos de casos del proceso “Venta de artículos a través de página Web”

5.3.8 Consolidación pre-procesamiento

Para concluir la etapa de pre-procesamiento y continuar con el análisis de los datos, es necesario construir una planilla donde se consoliden los datos obtenidos a través de las distintas etapas descritas en este capítulo. Esta consolidación debe registrarse en un formato adecuado para que los datos puedan ser interpretados por algoritmos de Minería de Datos.

La consolidación de los datos se puede realizar de manera automática gracias a una rutina programada en el “Pre-procesador Logs”. Este procedimiento recorre las distintas hojas contenidas en el archivo

Excel, extrayendo y consolidando la información de todos los casos en una única planilla de datos, que muestra un caso en cada fila. En esta planilla se almacenan, como se puede apreciar en la Tabla 5.2, los campos "caseID", "ID_Camino", "ID_Equipo", "Tiempo_Desde_Comienzo", "ID_TareasConsolidadas" y "ID_EquipoConsolidado", más los atributos "característicos" que presente el caso de negocio analizado. Al unir todos estos atributos se consigue describir, en una sola línea, la información disponible para cada ejecución del proceso.

caseID	ID_Camino	ID_Equipo	Tiempo_Desde_Comienzo	ID_Tareas Consolidadas	ID_Equipo Consolidado	cantidad	monto	patrón_prom	OK
1	C01	EQ01	8800	TC01	EC01	1	3000	0	OK
10	C02	EQ02	11800	TC01	EC02	1	3000	1	OK
100	C03	EQ03	11800	TC01	EC11	1	3000	1	OK
1000	C04	EQ04	5800	TC02	EC03	1	1000	2	OK
101	C05	EQ05	23800	TC01	EC06	1	3000	3	OK
102	C02	EQ06	11800	TC01	EC06	1	3000	1	OK
103	C04	EQ07	5800	TC02	EC03	1	3000	2	OK
104	C06	EQ08	14800	TC01	EC01	1	3000	4	OK
105	C04	EQ09	5800	TC02	EC17	1	3000	2	OK
106	C07	EQ10	32800	TC01	EC25	1	3000	5	OK
107	C01	EQ11	8800	TC01	EC19	1	3000	0	OK
108	C04	EQ12	5800	TC02	EC10	1	3000	2	OK
109	C04	EQ13	5800	TC02	EC03	1	3000	2	OK
11	C01	EQ14	8800	TC01	EC08	1	3000	0	OK
110	C01	EQ15	8800	TC01	EC02	1	3000	0	OK

Tabla 5.2 - Extracto de la planilla final obtenida luego del pre-procesamiento, donde se muestra la información de quince casos del proceso "Venta de artículos a través de página Web"

6. Búsqueda de patrones secuenciales

6.1 Resumen del procedimiento

El proceso de búsqueda de patrones secuenciales se puede resumir en dos etapas. En la primera, se buscan los patrones mediante la búsqueda de asociaciones frecuentes, observando los valores de cada atributo del caso mediante el uso del algoritmo Apriori. Luego, en la segunda parte, se utilizan los patrones encontrados para revisar el comportamiento de los distintos atributos presentes. El objetivo, en esta segunda etapa, es revisar qué valores toman estos datos en el marco de una situación considerada como recurrente dentro de las ejecuciones del proceso.

6.2 Herramientas seleccionadas

En este trabajo se seleccionó el *software* Weka para la aplicación de algoritmos de Minería de Datos sobre los datos del proceso analizado, en particular Apriori. Weka, además de ya contar con el algoritmo implementado, permite cambiar algunos de los parámetros más importantes de este algoritmo y también presenta, dentro de los resultados del análisis, los cálculos más relevantes realizados durante la búsqueda de patrones.

Para la segunda parte de la metodología propuesta se seleccionó el *software* Excel. El objetivo era implementar una herramienta que permitiera observar la distribución de frecuencias que muestran los valores de cada atributo, principalmente en los casos asociados a los patrones encontrados. La herramienta se desarrolló en Excel, dado que por medio del uso de Macros se puede automatizar las ejecuciones necesarias para llegar al resultado esperado. Junto a esto, el uso de una planilla de Excel permite dar una interfaz cómoda y simple para cualquier investigación.

6.3 Desarrollo del análisis

Como se señaló previamente, el proceso de búsqueda de patrones secuenciales se compone de dos etapas. A continuación, se revisarán los pasos que hay que seguir para

completar exitosamente cada etapa (tres pasos en la primera, uno en la segunda), y así obtener los mejores resultados en la investigación. También se entregarán detalles del diseño de cada etapa, y cómo contribuye cada paso al resultado final.

Esta sección se divide en los cuatro pasos mencionados previamente. Los tres primeros, apuntan a aplicar el algoritmo de búsqueda de patrones sobre los datos del proceso, mientras que el cuarto paso, describe cómo se puede analizar el comportamiento de los distintos atributos del proceso, observando específicamente los conjuntos de datos que coinciden con los patrones encontrados.

6.3.1 Exportar datos desde planilla "Pre-procesador Logs" a archivo .csv

Una vez concluido el pre-procesamiento de los datos, es necesario exportar la hoja de datos ya consolidada, a un formato compatible con el programa Weka. El formato escogido fue csv, dado que Excel permite guardar en esta extensión, sin pérdida de información o propiedades de los datos, mientras que Weka puede leer datos en este formato sin problemas.

Se debe tener cuidado con el formato en que Excel exporta el archivo csv. Dependiendo de la versión que esté utilizando (por ej. versión en inglés o versión en español), el formato puede variar. Para que la posterior importación del archivo csv en Weka sea exitosa, es importante que el separador de valores sea el signo "," (coma) y no ";" (punto y coma). Si los datos se presentan separados con este último signo, Weka no reconocerá los distintos atributos y tomará cada línea completa como un solo dato. Si este es el caso, con cualquier procesador de texto, se puede reemplazar rápidamente todos los signos, para lograr una lectura correcta de los datos.

6.3.2 Importar datos a Weka y uso de “Preprocess”

Para poder importar un archivo de extensión .csv al programa Weka es necesario encontrarse en la ventana principal del explorador de este *software* (consejos de navegación en Weka en Anexo 3). Encontrándose aquí, bajo la pestaña “Preprocess”, se encuentra el botón “Open file…”, el cual permite abrir diversos tipos de archivos soportados por Weka, entre ellos el que se usará para la búsqueda de patrones en procesos: csv.

Una vez abierto el archivo csv en la pestaña “Preprocess”, el programa mostrará los distintos atributos del proceso analizado, y los distintos valores que presenta cada uno de estos campos. Esta primera visualización de los datos se ejemplifica en la Ilustración 6.1, donde se aprecia el listado de atributos. Mientras que la Ilustración 6.2, muestra el detalle de los valores, de un determinado atributo. Estas ilustraciones muestran la interfaz gráfica, que permite realizar un segundo pre-procesamiento de los datos, mediante el análisis de frecuencia, variabilidad y tipo de dato que muestra cada atributo.

El análisis de frecuencia y variabilidad de los datos, en la pestaña “Preprocess”, puede ayudar a descartar inmediatamente ciertos atributos. O por el contrario, puede destacar atributos o valores de éstos, que tengan una alta incidencia en el resultado final. Un ejemplo de estos análisis se puede revisar en la Ilustración 6.2, dónde se ha escogido el atributo “OK”, el cual es el más importante del caso analizado, ya que indica si la venta fue exitosa o no. Este atributo muestra sólo dos posibles valores, a través de las mil instancias que presenta el caso de ventas por Internet. La baja variabilidad de este atributo, indica la alta probabilidad que este atributo pueda ser utilizado para establecer un patrón de comportamiento

del proceso. Análisis de este tipo, y otros más, serán revisados en la sección 8.1 de este trabajo, donde se discutirán los resultados de la búsqueda de patrones secuenciales.

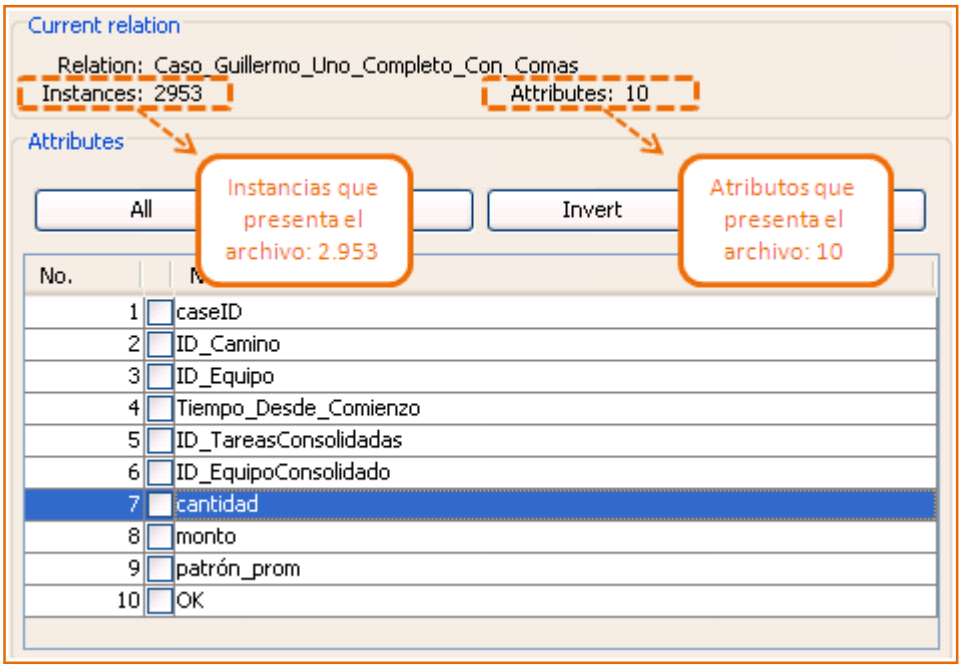


Ilustración 6.1 - Listado de atributos del proceso "Venta de artículos a través de página Web" mostrados en la pestaña "Preprocess" de Weka

6.3.3 Analizar los datos con algoritmo Apriori

Una vez importados los datos a Weka, y habiendo realizado posibles modificaciones en la sección "Preprocess", el siguiente paso es exponer los datos a análisis mediante el uso del algoritmo Apriori. Este algoritmo se puede seleccionar, y aplicar, en la sección "Associate" de Weka.

La pregunta clave a responder en esta sección es: ¿cómo funciona el algoritmo Apriori?

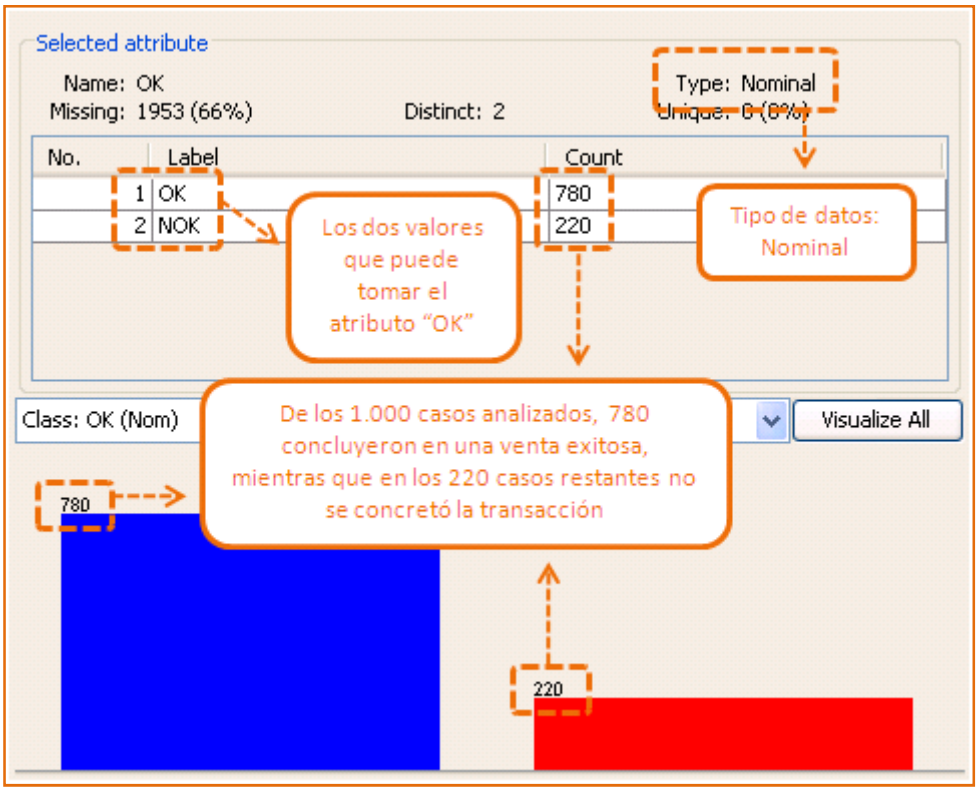


Ilustración 6.2 - Sección de la pestaña "Preprocess" que permite visualizar de manera numérica y gráfica los valores que puede tomar cada atributo del caso analizado (en el ejemplo se ha seleccionado el atributo "OK")

Apriori es un algoritmo de asociación, el cual se divide, principalmente, en dos etapas. En la primera etapa se buscan grupos de datos, los cuales deben cumplir con que su valor se repita con una frecuencia igual, o superior, a la mínima necesaria para cumplir con el nivel de soporte requerido. Estos grupos de datos son conocidos como *itemsets* frecuentes[6]. Los *itemsets* pueden tener desde uno hasta "n" atributos, siendo "n" el número de atributos que tiene cada proceso analizado. La segunda etapa consiste en evaluar si cada uno de estos

[6] Un *itemset* reúne desde uno a "n" atributos de un caso, donde cada uno de estos atributos está asociado a uno de sus posibles valores. Un *itemset* frecuente es aquel que muestra un nivel de soporte mayor al definido como parámetro del algoritmo Apriori (Hamilton, 2009).

itemsets cumple con el nivel de confianza requerido por el investigador. Lo que se busca en este punto es verificar qué *itemsets* tienen un porcentaje importante de apariciones asociado a un determinado resultado del proceso. Con estos dos pasos, se van armando iterativamente los candidatos a ser reglas de asociación, las que definirán finalmente los patrones del proceso.

El algoritmo Apriori, comienza armando *itemsets* con un sólo atributo, luego con dos atributos, continuando iterativamente hasta abarcar todas las combinaciones posibles. Finalmente, el algoritmo toma todos los *itemsets* resultantes del proceso y los ordena de mayor a menor, de acuerdo a la confianza que entrega cada *itemset*.

A continuación, se revisará un ejemplo para ilustrar como se construyen los *itemsets* de valores frecuentes, y como se van eliminando al no cumplir con el soporte o la confianza necesarias. Para este ejemplo, se tomará un extracto simplificado de diez ejecuciones del caso "Venta de artículos a través de página Web". Este extracto simplificado, como se aprecia en la Tabla 6.1, considera sólo tres atributos: "ID_EquipoConsolidado", "cantidad" y "OK".

Luego de ingresar los diez casos en Weka, se debe seleccionar la pestaña "Associate", donde se debe escoger Apriori como el asociador (o *associator*). Luego de hacer esta selección, se deberá ingresar los parámetros que regirán la ejecución del algoritmo. De los parámetros disponibles, los más importantes para la investigación son tres, los cuales se han destacado en la Ilustración 6.3, donde además se muestra los valores por defecto de los distintos campos que se pueden modificar.

ID_EquipoConsolidado	cantidad	OK
EC01	C1.0	OK
EC02	C2.0	OK
EC01	C3.0	OK
EC03	C1.0	NOK
EC01	C1.0	OK
EC03	C1.0	NOK
EC04	C2.0	OK
EC01	C1.0	NOK
EC02	C2.0	NOK
EC03	C1.0	OK

Tabla 6.1 - Tabla con los diez registros que componen el ejemplo simplificado del proceso "Venta de artículos a través de página Web"

Como se señaló previamente, de todas las opciones que ofrece el algoritmo Apriori, los parámetros más relevantes identificados en esta investigación fueron tres:

- **CAR:** acrónimo de "Class Association Rules". Este parámetro permite restringir el formato de las reglas de asociación. Su valor puede ser Verdadero o Falso, siendo la primera opción con la cual se activa. Al activar esta opción, se indica a Weka que busque reglas que sólo muestren conclusiones referidas a un atributo, identificado como clasificador. Esto es justamente lo que se necesita, tanto para el ejemplo de esta sección como para la búsqueda completa de patrones del caso "Venta de artículos a través de página Web". En esta investigación se optó por utilizar esta funcionalidad, usando el atributo "OK" como clasificador, ya que es el más relevante al indicar si la venta terminó de manera exitosa o no.

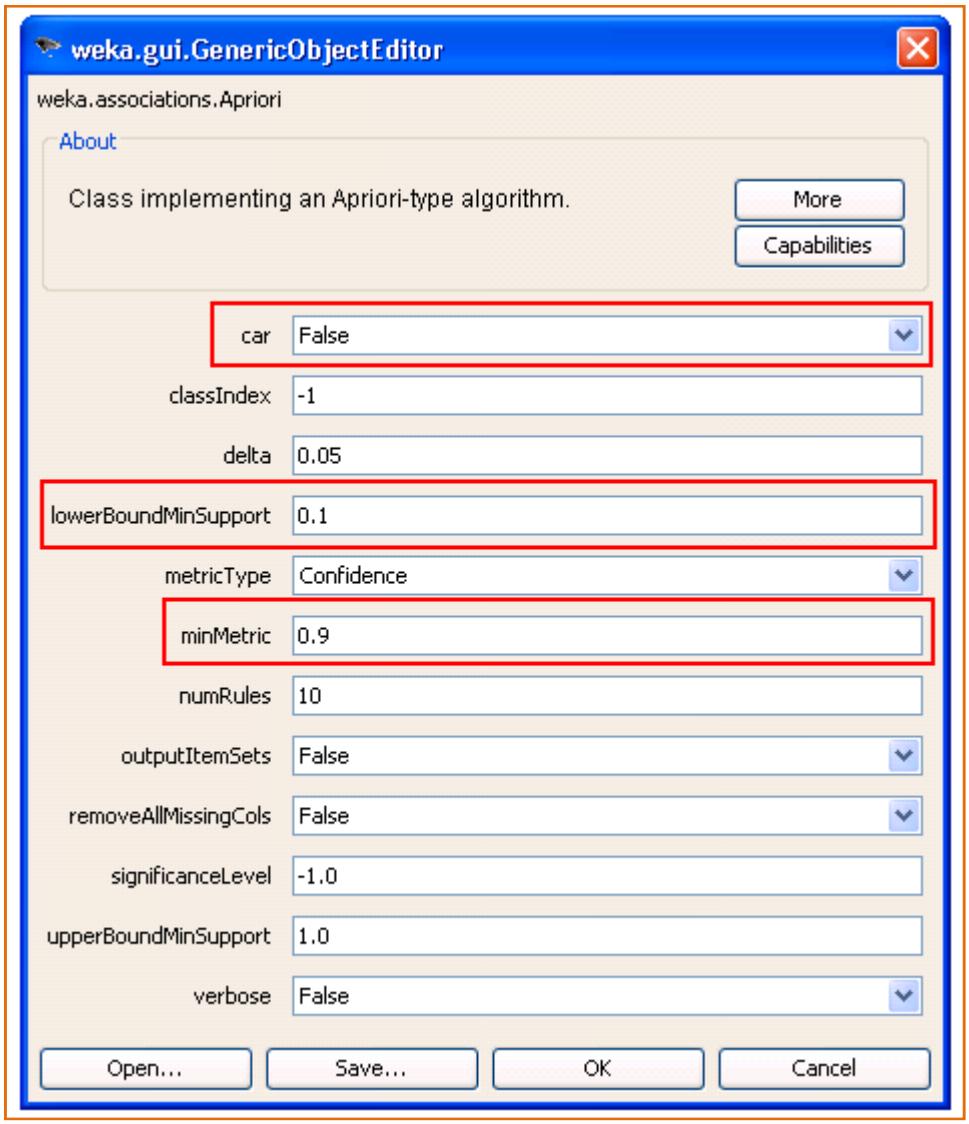


Ilustración 6.3 - Cuadro de Weka que permite modificar los parámetros a ser utilizados en la ejecución del algoritmo Apriori

El usar la opción "car" permite centrar el análisis sobre el atributo que se considere más importante en el caso. Por ejemplo, como ocurre en este caso, un atributo que describe el resultado final del proceso. Para definir el atributo que será usado como clasificador, se debe modificar el parámetro

"classIndex" que se aprecia en la Ilustración 6.3. Este parámetro por defecto tiene el valor "-1", el cual corresponde al último atributo de cada grupo de datos. En esta investigación, al igual que en este ejemplo, es justamente el último atributo el que se desea utilizar como clasificador, por lo que se dejará el "-1" como valor de este parámetro.

- **LowerBoundMinSupport:** parámetro clave para influenciar sobre la cantidad de reglas resultantes. Este parámetro permite definir el nivel de soporte requerido para que un conjunto de datos califique como un *itemset* frecuente. En otras palabras, al cambiar el valor de este parámetro, se está estableciendo la cantidad mínima de valores iguales que se debe encontrar, para que un grupo de casos sea considerado como un *itemset* importante o de alta frecuencia. Estos *itemsets* frecuentes son los que finalmente serán candidatos a ser parte de un patrón. La cantidad de candidatos disminuirá en la medida que se aumente el valor del parámetro "LowerBoundMinSupport". Por ejemplo, si se está trabajando con un grupo de datos que incluye mil ejecuciones de un proceso, al ingresar 0,05 como valor de este parámetro, se le está indicando al programa que se necesita, al menos, 50 instancias con el mismo valor, para que se considere como un grupo de ítems importante. En otras palabras, se le está indicando a Weka que sólo se quede con aquellos grupos de datos que se repitan una cantidad de veces igual, o mayor, al 5% del total de casos analizados. En este caso, 5% sobre mil casos analizados. Es importante destacar que la aplicación de este parámetro, en el algoritmo Apriori, varía al activar la opción "car", como se hará en esta investigación. Como se le indica a Weka que sólo entregue

reglas de asociación que lleven a valores del atributo indicado como clasificador, la evaluación de soporte de cada *itemset* se realiza observando la cantidad de veces que se repite el *itemset* para cada valor del atributo clasificador. Continuando con el ejemplo, donde los *itemsets* pasarían la evaluación de soporte solamente si tenían 50 o más instancias. Ahora, al agregar la activación de la opción "car", se requiere que esas 50 o más instancias estén asociadas a 50 o más instancias de un determinado valor del atributo elegido como clasificador. Esta característica del algoritmo Apriori se verá con mayor claridad al revisar el ejemplo de diez casos diseñado para esta sección.

- **MinMetric:** este parámetro, al igual que "LowerBoundMinSupport", puede impactar fuertemente en el número de patrones encontrados. Esto dado que, modificando el valor de "minMetric", se logra que se considere un grupo particular de las reglas de asociación encontradas. Cada una de las reglas de este grupo particular debe entregar, al menos, el valor indicado por "minMetric" en la métrica utilizada para descartar parte de las asociaciones encontradas. Como en esta investigación se ha optado por trabajar con la opción "car", una consecuencia inmediata es que la métrica para descartar una regla es estrictamente la confianza que muestre ésta. Por lo tanto, el valor que se ingrese en "minMetric" corresponderá al mínimo nivel de confianza que debe tener una regla, para ser parte del grupo de reglas final. Por ejemplo, si en un proceso de mil ejecuciones, se encuentra una regla de asociación con un *itemset* que se repite 200 veces y el valor de "minMetric" es 0,6. Entonces, para que esta regla de asociación quede entre

los resultados finales, al menos el 60% de las 200 instancias deben llevar a un mismo resultado.

Ya conocidos los parámetros más importantes que se deben utilizar para la búsqueda de patrones, es más sencillo continuar con el ejemplo de diez casos que se utilizará en este capítulo para ilustrar el funcionamiento del algoritmo Apriori.

Los tres parámetros revisados recientemente fueron ajustados con los siguientes valores: la opción "car" se activó seleccionando la opción "True", mientras que "lowerBoundMinSupport" quedó en 0,2 y "minMetric" en 0,6. Con estos parámetros, se encontraron cinco reglas de asociación para los diez casos preparados para este ejemplo. El proceso que llevó a estas reglas finales puede ser descrito en tres etapas.

La primera etapa consistió en la búsqueda de aquellos *itemsets* frecuentes formados con un sólo atributo. Estos *itemsets* debían cumplir con el nivel de soporte requerido. Finalizada esta búsqueda, se encontraron cinco *itemsets* frecuentes.

Del atributo "ID_EquipoConsolidado" dos valores no fueron clasificados como *itemsets* frecuentes. Estos corresponden a tres casos del ejemplo, los cuales han sido destacados en la Tabla 6.2. El equipo EC04 no fue considerado por no presentar el soporte necesario, ya que, en las 10 ejecuciones del proceso, sólo se presentó una vez. En tanto, el equipo EC02 también fue descartado de los *itemsets* de un sólo atributo. Aunque este equipo presentó el soporte requerido, con dos apariciones, estas no correspondían al mismo valor del atributo "OK". Como se puede apreciar en la Tabla 6.2, las dos ejecuciones del equipo EC02 llevaron a dos resultados diferentes. Por esto, el equipo no entregó el soporte requerido

en este ejemplo, ya que sólo presentó un soporte de 0,1 para cada valor de "OK".

ID_EquipoConsolidado	cantidad	OK
EC01	C1.0	OK
EC02	C2.0	OK
EC01	C3.0	OK
EC03	C1.0	NOK
EC01	C1.0	OK
EC03	C1.0	NOK
EC04	C2.0	OK
EC01	C1.0	NOK
EC02	C2.0	NOK
EC03	C1.0	OK

Tabla 6.2 - Tabla con los diez casos diseñados para el ejemplo, donde se destacan los tres valores del atributo "ID_EquipoConsolidado" que no clasificaron como *itemset* frecuente

Además de los tres valores descartados observando el atributo "ID_EquipoConsolidado", también se descartó un valor al observar el atributo "cantidad", el cual está destacado en la Tabla 6.3. Este descarte se aplicó sobre el único caso que presenta una cantidad igual a tres artículos. Fue justamente la unicidad de este valor lo que provocó la no consideración del valor, por no presentar el soporte que se indicó en "lowerBoundMinSupport".

Descartados los cuatro valores mencionados, quedaron cinco *itemsets* frecuentes, donde se encuentran los valores EC01 y EC03 del atributo "ID_EquipoConsolidado", y las cantidades iguales a uno o dos artículos (C1.0 y C2.0). Ocurrió algo particular con uno de estos valores, el cual presentó dos *itemsets*, a diferencia de los demás que sólo

entregaron un grupo de datos cada uno. Esto ocurrió porque las instancias, en que se vendió un sólo producto (C1.0), alcanzaron para dar soporte a los dos valores que puede presentar el atributo "OK". Estos dos *itemsets*, y los otros tres, se pueden observar en la Ilustración 6.4, donde se aprecia un extracto de la salida que muestra Weka al ejecutar el algoritmo Apriori.

ID_EquipoConsolidado	cantidad	OK
EC01	C1.0	OK
EC02	C2.0	OK
EC01	C3.0	OK
EC03	C1.0	NOK
EC01	C1.0	OK
EC03	C1.0	NOK
EC04	C2.0	OK
EC01	C1.0	NOK
EC02	C2.0	NOK
EC03	C1.0	OK

Tabla 6.3 - Tabla con los diez casos diseñados para el ejemplo, donde se destacan el único valor del atributo "cantidad" que no clasificó como *itemset* frecuente

En la Ilustración 6.4 se pueden observar los cinco *itemsets* que pasaron la evaluación de soporte aplicada por el algoritmo Apriori. En esta imagen se han destacado los dos *itemsets* que fueron encontrados a partir del mismo valor del atributo "cantidad". Esto ocurrió con el valor C1.0, que corresponde a un artículo. Se destacaron estos dos casos por dos razones: primero, explicar la nomenclatura que entrega Weka para describir los *itemsets* y segundo, explicar en detalle porque un mismo valor entregó dos *itemsets*.

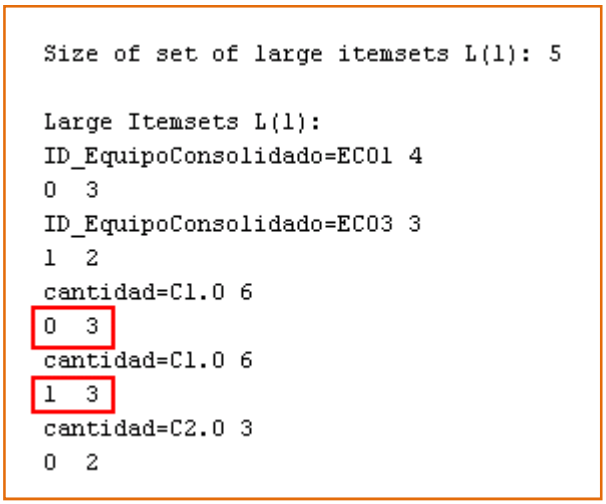


Ilustración 6.4 - Extracto de la salida de Weka al ejecutar Apriori, donde se detallan los *itemsets* de nivel 1, y en particular el valor C1.0 que presentó dos *itemsets*

La nomenclatura que usa Weka para describir los *itemsets* frecuentes no es trivial. Como se puede ver en la Ilustración 6.4, cada *itemset* está acompañado de tres números, uno a la derecha y dos debajo del *itemset*. El número a la derecha indica el total de apariciones del grupo de registros, en tanto que los indicadores debajo detallan la cantidad de apariciones que mostró el *itemset* para cada valor del atributo clasificador. El primer número debajo del *itemset* corresponde a un identificador estándar del valor del atributo clasificador. Aquí se asigna un valor de 0 a "n" a cada valor del atributo clasificador, siendo "n" la cantidad de valores que tiene este atributo. En este caso 0 corresponde a OK (cuando la venta termino exitosamente) y 1 es NOK (cuando la venta no se concretó). Si el atributo "OK" hubiese tenido más valores, se podrían haber encontrado más números, además del 0 y 1. Conocido el valor del atributo clasificador a que se está haciendo referencia en cada *itemset*, sólo falta saber la cantidad de instancias en que el *itemset*

presentó ese valor. Esta información la entrega el segundo número que se encuentra debajo de todo *itemset* que despliega Weka.

Como se había mencionado, había dos razones para destacar los dos *itemsets* en la Ilustración 6.4. La segunda era explicar por qué un mismo valor del atributo entregó dos *itemsets*. Ya explicada la nomenclatura, esto será más simple de explicar. Como se mencionó al detallar los parámetros utilizados en este ejemplo, se utilizó la opción "car" con el atributo "OK" como clasificador. Por esta razón, el algoritmo Apriori realizó la búsqueda de *itemsets* frecuentes inmediatamente revisando que se cumpliera con el nivel de soporte requerido, y no tan sólo a nivel total, sino que a nivel de cada valor del atributo "OK". El único *itemset* que pudo dar el soporte solicitado a los dos valores del atributo "OK", fue aquel que indicaba que la cantidad de artículos involucrados en la venta era igual a uno. Los seis casos que presentan esta cantidad de artículos están repartidos de igual manera en los dos valores que puede mostrar el resultado final. De esta manera, como cada resultado tiene tres casos asociados, se cumple con el soporte requerido y se crean los dos *itemsets*.

La segunda etapa ejecutada por el algoritmo Apriori es análoga a la primera etapa ya descrita. La diferencia radica en que ahora se buscaron los *itemsets* formados por dos atributos, y sólo con los valores que quedaron luego de la primera búsqueda. Como se vio anteriormente, la búsqueda sobre *itemsets* con un solo atributo dejó solamente a los equipos EC01 y EC03, y las cantidades C1.0 y C2.0. Con estos equipos y cantidades, se presentaron seis registros como candidatos a ser clasificados como *itemsets* frecuentes, los cuales se aprecian en la Tabla 6.4. En esta tabla hay dos registros destacados, los cuales fueron los únicos que no pasaron la evaluación de soporte. Estos registros fueron

eliminados porque presentaron sólo una aparición, lo cual no es suficiente para el soporte que exige, al menos, dos repeticiones.

ID_EquipoConsolidado / cantidad	OK
EC01 / C1.0	OK
EC03 / C1.0	NOK
EC01 / C1.0	OK
EC03 / C1.0	NOK
EC01 / C1.0	NOK
EC03 / C1.0	OK

Tabla 6.4 - Tabla con los seis *itemsets* de dos atributos, candidatos a ser identificados como *itemsets* frecuentes

El resultado en Weka de la búsqueda de *itemsets* de dos atributos se puede observar en la Ilustración 6.5, donde se detallan los dos *itemsets* que resultaron seleccionados. Además, se incluye la información de cuantas veces se repitió cada *itemset*, tanto en el total de ejecuciones, como sólo en aquellos casos correspondientes a un valor determinado del atributo "OK". Con esta información, se puede apreciar que los *itemset* EC01/C1.0 y EC03/C1.0 se repitieron tres veces cada uno, pero pasaron la evaluación del soporte con lo mínimo solicitado, ya que solamente en dos ocasiones llegaron a un mismo resultado. Los dos casos que no se observan de cada *itemset*, son justamente los casos que fueron destacados como eliminados en la Tabla 6.4. Estos registros eliminados presentaron los mismos *itemsets* que quedaron seleccionados, pero llegaron al otro resultado del atributo "OK".

Teniendo los siete *itemsets* frecuentes seleccionados, la tercera y última etapa que ejecuta el algoritmo Apriori, es evaluar cuáles de estos *itemsets* cumplen con el nivel de confianza requerido para que se

consideren como una regla de asociación, y de esta manera un patrón del proceso estudiado. Los resultados de esta evaluación se pueden apreciar en la Ilustración 6.6. En esta ilustración, la confianza se aprecia junto a cada regla de asociación, con el denominador "conf". Esta métrica se calcula mediante una división. Primero, se toma la cantidad de instancias que el *itemset* llegó al valor de "OK" indicado en la regla, y luego se divide por la cantidad de repeticiones total que tuvo el *itemset* en las diez ejecuciones que componen este ejemplo.

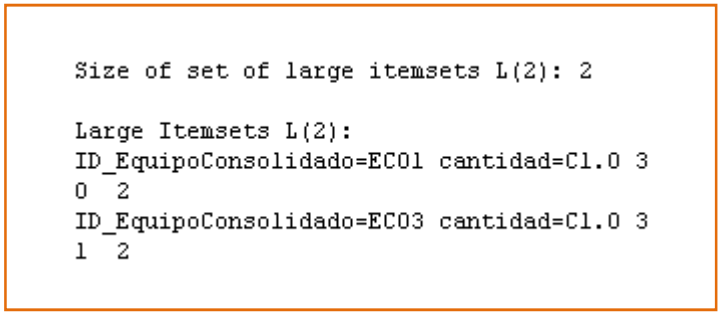


Ilustración 6.5 - Extracto de la salida de Weka al ejecutar Apriori, donde se detallan los *itemsets* de nivel 2

La primera regla de asociación de la Ilustración 6.6, indica que se puede concluir, con un 75% de confianza, que si el equipo ejecutor del proceso es EC01, el resultado de la venta será exitoso. Este es el resultado que se obtiene al dividir las tres oportunidades en que el equipo EC01 mostró un resultado positivo, sobre el total de cuatro ejecuciones que mostró el *itemset* a través de las diez ejecuciones del ejemplo.

Los dos *itemset* que no calificaron como reglas de asociación se aprecian en la Ilustración 6.7, y corresponden a los dos *itemset* de un sólo atributo que tenían el valor de cantidad igual a C1.0. Estos dos *itemsets* entregaron la cantidad de soporte requerido (tres repeticiones), pero como sólo la mitad de sus apariciones (tres de seis) se asignan a un determinado

valor de “OK”, su confianza sólo alcanza un 50%, que es insuficiente para el 60% solicitado en los parámetros del algoritmo.

```
Best rules found:

 1. ID_EquipoConsolidado=EC01 4 ==> OK=OK 3    conf:(0.75)
 2. ID_EquipoConsolidado=EC03 3 ==> OK=NOK 2    conf:(0.67)
 3. cantidad=C2.0 3 ==> OK=OK 2    conf:(0.67)
 4. ID_EquipoConsolidado=EC01 cantidad=C1.0 3 ==> OK=OK 2    conf:(0.67)
 5. ID_EquipoConsolidado=EC03 cantidad=C1.0 3 ==> OK=NOK 2    conf:(0.67)
```

Ilustración 6.6 - Extracto de la salida de Weka al ejecutar Apriori, donde se detallan las reglas de asociación encontradas por el algoritmo en el grupo de diez casos que componen el ejemplo

```
cantidad=C1.0 6
0  3
cantidad=C1.0 6
1  3
```

Ilustración 6.7 - Extracto de la salida de Weka al ejecutar Apriori, donde se detallan los dos *itemset* que no formaron parte de las reglas de asociación

6.3.4 Analizar relación entre patrones encontrados y los atributos del caso

Los resultados del algoritmo Apriori se pueden complementar con un segundo análisis, entregando la posibilidad de obtener información más completa y valiosa de los datos disponibles. Para realizar esto, se deben llevar los datos del archivo csv utilizado en Weka a la planilla “Analiza Patrones” implementada para este trabajo (consejos de uso sobre “Analiza Patrones” en Anexo 5). En la hoja “01. Datos” de esta planilla se deben filtrar los datos, de acuerdo a las condiciones determinadas por

las reglas (o patrones) encontradas en Weka. Este filtrado se realiza dado que el objetivo no es sólo analizar cómo se relacionan los atributos a través de todos los casos disponibles, sino que hay más interés en aquellos casos que tengan una frecuencia importante, y que por ello tienen más probabilidades de influir en el resultado final del proceso. Luego, antes de proceder a ver los resultados, se debe seleccionar en la hoja "02. Metricas" cuál es el atributo que se quiere analizar. Realizada esta selección, la planilla muestra una lista con todos los valores que presenta el atributo escogido a través de los casos analizados. Para cada valor se muestran cuatro indicadores, los cuales fueron escogidos para mostrar la relevancia que tiene cada valor a través del proceso estudiado. Los cuatro indicadores son:

1. **Frecuencia del valor en los casos filtrados:** como indica su nombre, entrega el número de instancias que se repite el valor en los casos filtrados.
2. **Peso de este valor en los casos filtrados:** este indicador indica el porcentaje de oportunidades en que se repitió cada valor, tomando como base sólo los casos filtrados. Un bajo o alto porcentaje de este indicador, entrega inmediatamente información clara de qué valores son importantes, o cuáles se presentan con menor frecuencia en los casos descritos por las reglas encontradas.
3. **Frecuencia del valor en todos los casos (sin filtrar):** como indica su nombre, entrega la cantidad de veces que se repite el valor a través de todos los casos del proceso analizado (sin filtrar según reglas).
4. **Peso de este valor en todos los casos (sin filtrar):** este indicador presenta aquellos valores del atributo que tienen mayor o menor relevancia a través de todos los casos, y por lo

tanto son importantes de analizar en mayor profundidad a la hora de definir patrones.

Además de analizar individualmente cada uno de los cuatro indicadores recién descritos, es recomendable también revisar cómo varían los porcentajes de cada valor al cambiar la base de casos que se está analizando. En otras palabras, lo que se aconseja es comparar el peso que tiene un valor en los casos filtrados contra el total de instancias disponibles. Con este análisis, el investigador podrá verificar si la frecuencia de un valor se comporta de manera similar bajo ambas perspectivas, o si por el contrario muestra un comportamiento diferente, dependiendo del conjunto de datos que se utilice como base. Si se observa una distribución de frecuencia significativamente distinta al comparar ambas bases de observación, se pueden establecer tendencias de determinados valores de un atributo a un resultado particular del proceso.

7. Búsqueda de anomalías

7.1 Resumen del procedimiento

La búsqueda de anomalías se puede resumir en dos etapas. En la primera, se toman todos los casos que se pre-procesaron a partir del archivo csv del proceso, y se aplica el algoritmo Interquartile Range sobre éstos. El resultado de esta ejecución entregará los casos anómalos del proceso. Éstos son aquellos que presentaron un comportamiento significativamente distinto al mostrado por el total de casos analizados. Luego de identificar estas anomalías, se procede a una segunda etapa de análisis donde se revisará la frecuencia, de los distintos valores de cada atributo, a través de los casos que fueron considerados como anómalos. El objetivo de esta segunda etapa es poder

complementar la información que se tiene de los casos anómalos, entendiendo que valores se presentan con mayor o menor frecuencia en situaciones de anomalía.

7.2 Herramientas seleccionadas

A diferencia de lo expuesto para la búsqueda de patrones, donde se utilizaban Excel y Weka, en el caso de las anomalías sólo se trabajará con un archivo, y en un solo programa: Excel. Con esta herramienta se podrá encontrar aquellos casos que presenten un comportamiento significativamente distinto a las demás ejecuciones del proceso. Se seleccionó este programa, ya que permite automatizar los pasos del algoritmo Interquartile Range mediante el uso de Macros. Además, usando una planilla de Excel se entrega una interfaz cómoda y simple para trabajar, donde se pueden modificar los parámetros del algoritmo, y también se pueden observar los principales cálculos de éste.

7.3 Desarrollo del análisis

Aunque el proceso implementado para la búsqueda de anomalías trabaja sólo con Excel, los pasos a seguir en el desarrollo conservan una estructura muy similar a lo presentado para la búsqueda de patrones secuenciales. Como se podrá ver a continuación, y como se comentó en el resumen al comienzo de este capítulo, la búsqueda de anomalías se compone de dos etapas. La primera etapa se desarrolla a través de cuatro pasos, donde se toman los datos entregados por la planilla "Pre-procesador Logs", y se buscan los casos anómalos de acuerdo a criterios y parámetros establecidos antes de realizar el análisis. Con estos resultados se puede continuar con la segunda etapa del análisis, donde se puede observar qué valores toman los distintos atributos en los casos considerados como anómalos.

7.3.1 Exportar datos desde planilla "Pre-procesador Logs" a planilla "Busca Anomalías"

La primera etapa de este análisis comienza luego que haya finalizado el pre-procesamiento de los datos. Aquí se debe importar los datos consolidados a la planilla “Busca Anomalías” (consejos de uso de la planilla “Busca Anomalías” en Anexo 6). Estos datos se pueden encontrar en la hoja llamada “CSV” de la herramienta “Pre-procesador Logs”, la cual contiene en cada fila de datos, toda la información disponible de cada caso del proceso analizado.

Como ambas planillas (“Pre-procesador Logs” y “Busca Anomalías”) contienen datos almacenados en Excel, basta con copiar los datos presentes en el “Pre-procesador Logs” y pegarlos en la hoja inicio de la planilla “Busca Anomalías”, la cual tiene el nombre “01.Datos”.

7.3.2 Definir tipo de datos de cada atributo

La búsqueda de anomalías se realiza aplicando el algoritmo Interquartile Range sobre un atributo específico del caso, el cual es seleccionado al momento de realizar el análisis. Antes de seleccionar este atributo, sobre el cual se buscarán los casos anómalos, se debe indicar qué tipo de datos están contenidos en cada campo usado para describir los casos analizados. Existen dos posibles clasificaciones de tipo de dato para cada atributo:

1. De tipo **NUMÉRICO**, como por ejemplo el caso del atributo “Tiempo_Desde_Comienzo” que muestra valores como 0,083333342, 0,116666673, etc.
2. De tipo **NOMINAL**, donde se puede mencionar algunos casos como el atributo “ID_Equipo”, con valores EQ01, EQ05, etc, o “ID_Camino” con valores como C03, C08, etc.

La razón de porque es necesario indicar si un atributo es numérico o nominal radica en el hecho que dependiendo del tipo de dato, el proceso de búsqueda de anomalías funciona de manera levemente diferente.

El algoritmo Interquartile Range, como se verá más en detalle en este capítulo, ocupa la mediana y otros datos estadísticos para establecer el rango de normalidad de los datos. Por esta razón el algoritmo necesita recibir datos numéricos, de manera que pueda, por ejemplo, ordenar los datos en orden ascendente y así buscar la mediana, que es el valor que se encuentra justo en la mitad de todos los casos. Dado este requerimiento del algoritmo, cuando se presentan datos nominales no se puede aplicar el algoritmo Interquartile Range directamente sobre estos valores. Para solucionar esto, en este trabajo se optó por calcular la frecuencia de los distintos valores nominales que toma el atributo que se desee utilizar como referencia para buscar casos anómalos. De esta manera, siempre se puede construir un conjunto de datos numéricos que puede ser analizado con el algoritmo Interquartile Range.

La indicación de qué tipo de dato describe cada atributo se realiza en la hoja “02. Atributos” de la planilla “Busca Anomalías”. Una vez completada la información necesaria, se debe escoger qué tipo de atributo se quiere utilizar para buscar las anomalías. Esto por el mismo motivo mencionado anteriormente: dependiendo del tipo de dato, el algoritmo trabajará de manera distinta.

7.3.3 Selección de parámetros

Encontrándose en el tercer paso de la primera etapa del análisis, ya se habrá definido sobre qué tipo de atributo se desea buscar las anomalías, ya sea de tipo nominal o numérico. Dependiendo lo que se haya escogido,

el análisis continuará en la hoja “03. Parametros”, o en “03. Parametros B”. Si se escogió buscar casos anómalos a través de un atributo numérico, se llegará a la hoja “03. Parametros”. En cambio, si se optó por buscar los casos anómalos a partir de la observación de un atributo nominal, el análisis continuará en la hoja “03. Parametros B”.

Una vez en la hoja diseñada para ingresar los parámetros necesarios para la búsqueda de anomalías (cualquiera de las dos previamente mencionadas), lo primero que se deberá hacer es escoger el atributo que servirá de referencia para buscar los casos anómalos. Al escoger un atributo como referencia, se está estableciendo que se buscarán aquellos casos que muestren, en ese atributo, una frecuencia de aparición significativamente distinta al resto de los casos. Para encontrar estos casos, significativamente distintos, se utilizará el algoritmo Interquartile Range. A continuación se explicará cómo funciona este algoritmo y cuáles son los elementos más importantes que hay que conocer para trabajar correctamente con esta metodología.

El algoritmo Interquartile Range trabaja con cuartiles, es decir, toma todos los valores que muestra el atributo seleccionado y los divide en cuatro grupos. La medida más conocida para definir uno de los límites, necesarios para definir estos cuartiles, es la mediana. Este límite indica el valor que separa en dos a todos los valores encontrados. Los casos con un valor menor a la mediana representarán el 50% del total de casos analizados, mientras que la otra mitad de instancias serán valores con un indicador mayor a la mediana encontrada. Además de la mediana, existen dos límites importantes para el trabajo del algoritmo:

1. **Primer cuartil:** los casos analizados con un valor menor al indicado por el límite superior del primer cuartil representan

cerca del 25% del total de casos. Como se puede deducir de esta definición, este 25% de los datos corresponde a los de menor valor dentro de todas las instancias capturadas del proceso estudiado. Para calcular este límite, Excel ordena de menor a mayor todos los datos y busca el valor que se encuentre en la posición dada por la fórmula detallada en la Ilustración 7.1. Si el resultado de esta fórmula no es entero, se realiza una interpolación lineal para obtener la posición correspondiente al punto donde termina el primer cuartil.

$$Posición\ del\ primer\ cuartil = \frac{1}{4} * (n + 3)$$

donde n es el número de casos analizados

Ilustración 7.1 - Fórmula utilizada por Excel para encontrar la posición donde se encuentra el límite superior del primer cuartil

2. **Tercer cuartil:** los casos con valor mayor al límite superior del tercer cuartil representan cerca de un 25% del total de casos analizados del proceso. De esta manera, los valores que se encuentren en el rango definido por este cuartil, corresponden a los casos que presentan los valores más altos del proceso. Para calcular este límite, Excel ordena de menor a mayor todos los datos y busca el valor que se encuentre en la posición dada por la fórmula indicada en la Ilustración 7.2. Como en el cálculo del primer cuartil, si el resultado de esta fórmula no es entero, se realiza una interpolación lineal para obtener la posición correspondiente al punto donde termina el tercer cuartil.

$$Posición\ del\ tercer\ cuartil = \frac{1}{4} * (3n + 1)$$

donde n es el número de casos analizados

Ilustración 7.2 - Fórmula utilizada por Excel para encontrar la posición donde se encuentra el límite superior del tercer cuartil

Los rangos definidos por el primer y tercer cuartiles no siempre contendrán exactamente el 25% de los casos, respectivamente. Las fórmulas indicadas previamente buscan efectivamente la posición para estar justo en el primer y tercer cuartos de la lista, respectivamente. Pero el límite no se define con la posición, sino con el valor que se muestre en esa posición. Si el valor encontrado, en las posiciones que entregan las fórmulas, se repite a través de la lista, la posición del límite se moverá a la primera instancia en que aparece dicho valor. Esta es la razón porqué puede variar el porcentaje de casos que abarque cada cuartil.

Por defecto, el algoritmo Interquartile Range considera a los valores contenidos entre el primer y tercer cuartil como el grupo de casos normales. La magnitud de este rango (resultado de restar el valor del primer cuartil al tercer cuartil) se conoce como "rango entre cuartiles".

Si se desea disminuir o agrandar el rango de normalidad definido por defecto, se pueden calcular nuevos límites usando el "rango entre cuartiles" y un "multiplicador de rango entre cuartiles" definido por el investigador. En la Ilustración 7.3 se pueden apreciar las fórmulas para calcular los nuevos límites de normalidad.

$$Limite\ inferior\ de\ normalidad = (valor\ primer\ cuartil) - (multiplicador\ de\ rango\ entre\ cuartiles) \times (rango\ entre\ cuartiles)$$

$$Limite\ superior\ de\ normalidad = (valor\ tercer\ cuartil) + (multiplicador\ de\ rango\ entre\ cuartiles) \times (rango\ entre\ cuartiles)$$

Ilustración 7.3 - Fórmulas implementadas para encontrar los límites de normalidad, de acuerdo a los parámetros entregados

Dependiendo del "multiplicador de rango entre cuartiles" que se defina, el rango de normalidad abarcará más o menos valores. Si este multiplicador fuera cero, los nuevos límites de normalidad se mantendrían en los valores determinados por el primer y tercer cuartil. Mientras que si se utiliza un valor mayor a cero, el rango de normalidad aumentará y como consecuencia se encontrarán menos casos anómalos. Por el contrario, si se escoge un multiplicador menor a cero, se obtendrá un rango de normalidad más pequeño y más casos serán considerados como anómalos. En el capítulo de resultados se revisará la sensibilidad de la cantidad de casos anómalos encontrados respecto al valor de este multiplicador.

Ya conocida la manera en que trabaja el algoritmo Interquartile Range, se hace más simple comprender los parámetros que se deben escoger antes de buscar las anomalías.

Estando en la hoja "03. Parametros" o "03. Parametros B", y luego de escoger el atributo de referencia, se deberá definir el "multiplicador de rango entre cuartiles" y el tipo de anomalías que se quiere observar en el

resultado. El objetivo del multiplicador fue explicado previamente, mientras que los tipos de anomalías deseados se definen según la ubicación de los casos anómalos, respecto a los límites de normalidad. Esta opción permite escoger si se desea buscar todos los casos anómalos o si se prefiere limitar la búsqueda a sólo un tipo de anomalía, ya sea datos que estén bajo el límite inferior de normalidad (sólo umbral inferior) o sobre el límite superior de normalidad (sólo umbral superior).

Luego de definir los parámetros requeridos, la herramienta mostrará automáticamente los indicadores que regirán la búsqueda de anomalías.

Cuando se esté trabajando en el ingreso de parámetros, tanto con un atributo numérico o nominal de referencia para buscar las anomalías, se podrán observar nueve indicadores. Estos indicadores son prácticamente iguales para ambos casos, y sólo se diferencian en un punto. Como se puede observar en la Tabla 7.1, la diferencia de trabajar con un atributo numérico o nominal, determinará si se observarán los valores del atributo seleccionado, o la frecuencia de éstos. La frecuencia se utilizará solamente cuando el análisis esté basado en un atributo nominal.

7.3.4 Resultado búsqueda de anomalías

En el cuarto y último paso de la primera etapa del análisis, la planilla “Busca Anomalías” procesará los parámetros ingresados, y de acuerdo a esta información buscará qué casos no coinciden con el criterio de normalidad establecido. Los resultados de esta búsqueda se entregan en la hoja “04. Anomalías” si se busca anomalías sobre un atributo

numérico, o en la hoja “04. Anomalías B” si es que se optó por buscar los casos anómalos observando los datos de un atributo nominal.

Indicador	Descripción
Total de casos a analizar	Si el investigador opta por no restringir el conjunto de datos para el análisis, el total de casos será igual a la cantidad de casos que fueron ingresados en la hoja "01. Datos"
Valor (o frecuencia) mínimo encontrado en los casos	La herramienta buscará, a través de todos los casos seleccionados para el análisis, cuál es el menor valor (o frecuencia) que presenta el atributo usado como referencia para la búsqueda de anomalías
Valor (o frecuencia) máximo encontrado en los casos	Igual que el caso anterior, pero ahora se busca el caso con el valor (o frecuencia) más alto
Valor (o frecuencia) de la mediana calculada sobre los datos	Corresponde al valor (o frecuencia) que se encuentra justo en la mitad de todos los casos. Para encontrar este valor se observan todos los casos, ordenándolos de menor a mayor, según el valor (o frecuencia) que muestren. Una vez ordenados, se busca el valor que se encuentre en la mitad de este listado
Valor (o frecuencia) del primer cuartil	Corresponde al valor (o frecuencia) encontrado en la posición que calcula Excel como límite superior del primer cuartil. Los casos con un valor (o frecuencia) menor al mostrado por este indicador, corresponden a cerca del 25% de los casos.
Valor (o frecuencia) del tercer cuartil	Corresponde al valor (o frecuencia) encontrado en la posición que calcula Excel como límite superior del tercer cuartil. Los casos con un valor (o frecuencia) mayor al mostrado por este indicador, también corresponden a cerca del 25% de los casos
Rango entre cuartiles	Indica el resultado de restar el valor (o frecuencia) del primer cuartil, al valor (o frecuencia) del tercer cuartil. Este cálculo corresponde a la magnitud del rango de normalidad por defecto.
Umbral inferior	Aquí se indica el valor (o frecuencia) calculado como límite inferior de normalidad. Los casos que muestren un valor (o frecuencia) menor al indicado por el umbral inferior, serán considerados como anómalos
Umbral superior	El último indicador entrega el valor (o frecuencia) calculado como límite superior de normalidad. Los casos que muestran un valor (o frecuencia) mayor al indicado por el umbral superior, serán considerados como anómalos

Tabla 7.1 - Indicadores entregados por la planilla “Busca Anomalías” luego que el investigador ingresa los parámetros de búsqueda

En cualquiera de las hojas donde se detallan los casos anómalos (ya sea “04. Anomalias” o “04. Anomalias B”), se podrá revisar todos aquellos casos que hayan sido clasificados como anómalos. Además del detalle de cada caso, en estas hojas se indica el número total de anomalías, como también el atributo que sirvió de referencia, y los umbrales de normalidad establecidos previamente.

7.3.5 Analizar relación entre anomalías encontradas y los atributos del caso

La segunda etapa del análisis consiste en revisar cómo se distribuyen los distintos valores de cada atributo del caso. Principalmente, a través de los casos considerados como anómalos, pero también a través de todas las ejecuciones del proceso. Este análisis apunta a obtener información más completa y valiosa de los resultados recién obtenidos, observando la frecuencia que toman los distintos valores de cada atributo. Este análisis se puede realizar en las hojas “05. Metricas” o “05. Metricas B”, dependiendo del tipo de atributo que haya sido escogido para encontrar las anomalías.

La segunda funcionalidad de análisis implementada en la planilla “Busca Anomalías”, permite analizar la importancia de cada valor, de cada atributo del proceso. Para realizar este análisis, lo único que se debe seleccionar es el atributo que se quiere observar, el cual debe ser de tipo nominal. Luego de seleccionado este atributo, la planilla entregará una lista con todos los valores que presenta el atributo escogido, a través de los casos analizados. En esta misma lista también se incluyen cuatro indicadores, implementados para favorecer el análisis de relevancia que tiene cada valor del atributo seleccionado. Estos cuatro indicadores son:

1. **Frecuencia del valor:** entrega la cantidad de instancias que se repite el valor del atributo, a través de todos los casos disponibles para el análisis (considerando casos anómalos y normales).
2. **Anomalías encontradas con este valor:** entrega la cantidad de veces que se repite el valor, observando solamente los casos anómalos.
3. **Peso anomalías v/s frecuencia del valor:** presenta qué porcentaje, del total de apariciones de un valor determinado, fue parte de un caso anómalo. Si el valor de este porcentaje es muy alto, indica al investigador que existe una alta relación entre ese determinado valor y la ocurrencia de una anomalía.
4. **Peso anomalías v/s total anomalías encontradas:** aquí se detalla el porcentaje de apariciones que tiene cada valor del atributo analizado, usando como base solamente aquellos casos considerados como anómalos. Observando este indicador, se puede comprender cómo es la distribución de los valores de cada atributo a través de los casos anómalos encontrados, centrándose seguramente en aquellos casos con más altos porcentajes.

8. Resultados

8.1 Resultados búsqueda de patrones secuenciales

La búsqueda de patrones secuenciales se realizó observando mil casos, que describen procesos de compra a través de una página Web. El proceso se registra desde el momento que el cliente hace el pedido, hasta la finalización de la interacción entre el cliente y el sistema, que puede terminar en una venta exitosa o en la cancelación de ésta.

Los mil casos estudiados son descritos a través de 9.077 filas de datos. Cada fila se compone de ocho columnas, con los cinco atributos básicos obtenidos de un proceso, más tres atributos característicos que son registrados por necesidad del caso de negocio. Para conocer más detalles sobre el caso analizado, se recomienda revisar la sección 4.2.1 de este trabajo.

8.1.1 Resultados pre-procesamiento

El pre-procesamiento de los datos, en la búsqueda de patrones, tiene como objetivo principal exportar y consolidar datos. Esto se traduce en llevar los datos del proceso, entregados por ProM, a un formato que pueda ser leído por el programa computacional Weka, donde cada registro contenga toda la información de una ejecución del proceso. Además de cumplir estos objetivos, cada etapa del pre-procesamiento entrega la posibilidad de comprender de manera sintetizada algunas características relevantes de los mil casos analizados.

Se recomienda, antes de continuar la lectura de estos resultados, revisar el capítulo 5 de este documento, para así conocer en profundidad el paso a paso que se realiza en la planilla “Pre-procesador Logs”, desarrollada para este trabajo.

La primera información que entrega la planilla “Pre-procesador Logs” es un listado consolidado de las distintas tareas y ejecutores que participan en el proceso. Estas listas se pueden apreciar en la Ilustración 8.1, donde se muestran nueve tareas y siete ejecutores, cada uno con un identificador único y estándar, para facilitar el trabajo con estos datos. Estas listas muestran como el “Pre-procesador Logs” permite resumir los datos contenidos a través de las 9.077 filas de información, entregando

una imagen clara de cuántas tareas y ejecutores participan a través de los mil casos.

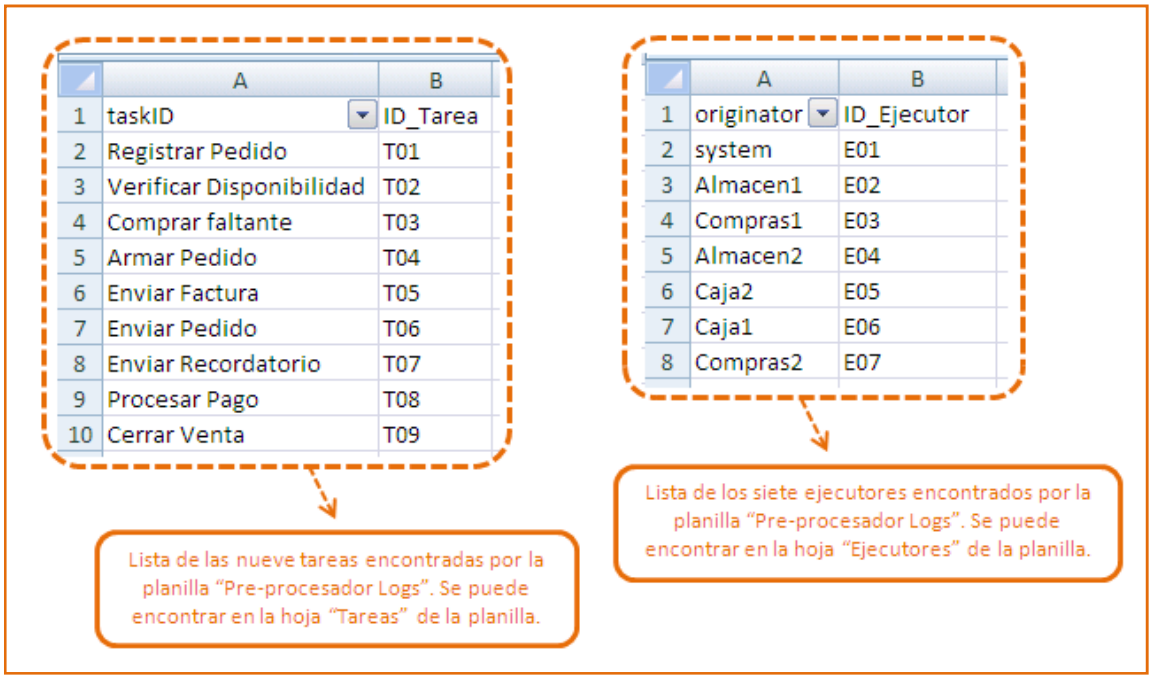


Ilustración 8.1 - Listas de tareas y ejecutores, encontrados con el "Pre-procesador Logs", del proceso "Venta de artículos a través de página Web"

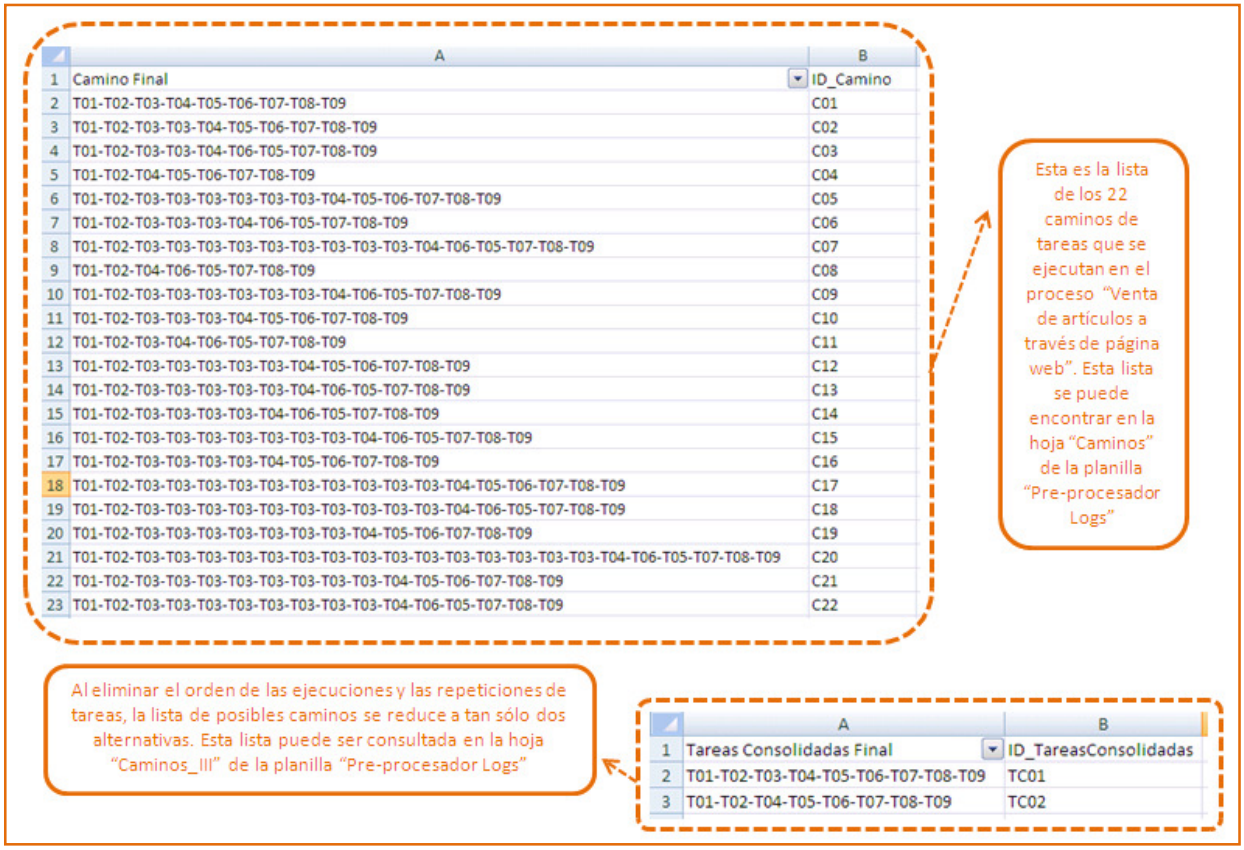
En los dos siguientes pasos que realiza el "Pre-procesador Logs", se observan las fuertes diferencias que se pueden obtener al consolidar las tareas desde dos perspectivas diferentes. Primero, se reúne las tareas de cada caso según el orden en que se ejecutan, mientras que en el siguiente paso se realiza una consolidación más permisiva, donde no se considera el orden de ejecución de las tareas, ni la repetición de éstas. La implementación de esta segunda perspectiva se realizó buscando armar grupos de tareas con mayor frecuencia, sacrificando la información de orden y repetición a favor de obtener reglas con mayor soporte.

Luego de consolidar las tareas por medio de las dos perspectivas descritas previamente, se obtienen dos listas con las distintas agrupaciones de tareas construidas por la planilla “Pre-procesador Logs”. Los resultados de ambas perspectivas, y la importante diferencia entre éstas, se pueden observar en la Ilustración 8.2. De estos resultados, destaca que a pesar de que el proceso muestra 22 caminos distintos de ejecución, una mirada más amplia o permisiva, indica que estos mismos caminos pueden ser resumidos en sólo dos grupos. Estos 22 caminos, con excepción de dos casos, presentan las mismas actividades. La existencia de los demás 20 casos encontrados con el primer método, se justifica por el hecho de existir casos con un orden distinto en la ejecución de las actividades o que repiten más veces una determinada tarea.

De manera similar a los resultados recién descritos, la identificación de equipos de trabajo se realiza primero de manera directa, y luego eliminando el orden y las repeticiones de los distintos ejecutores. Los resultados obtenidos con la agrupación directa no se pueden incluir en este documento, ya que en los mil casos analizados se encontraron 524 grupos distintos (estos resultados se pueden revisar en el Anexo 7). En cambio, los resultados de la agrupación que no considera orden ni repeticiones se pueden apreciar en la Tabla 8.1.

Las 35 posibles agrupaciones de ejecutores, encontradas con el segundo método y mostradas en la Tabla 8.1, difieren considerablemente de los 524 grupos distintos encontrados considerando orden y repeticiones de participación. Esta gran diferencia refleja la importancia de incluir el método consolidado para armar equipos. Si no se realizara la consolidación según esta segunda perspectiva, los equipos de ejecutores serían probablemente descartados como parte del análisis en Weka. Este descarte ocurriría casi con total seguridad, dado que el algoritmo Apriori

trabaja buscando grupos de alta frecuencia de aparición. Y contando con más de quinientos equipos distintos en sólo mil casos, sería muy difícil encontrar equipos con una frecuencia significativamente alta. Al consolidar eliminando las repeticiones y orden como factores de agrupación, se puede disminuir fuertemente el número de equipos, como ocurrió en este caso de estudio. Esta perspectiva distinta abre la posibilidad de poder analizar los equipos como parte de los patrones.



	A	B
1	Camino Final	ID_Camino
2	T01-T02-T03-T04-T05-T06-T07-T08-T09	C01
3	T01-T02-T03-T03-T04-T05-T06-T07-T08-T09	C02
4	T01-T02-T03-T03-T04-T06-T05-T07-T08-T09	C03
5	T01-T02-T04-T05-T06-T07-T08-T09	C04
6	T01-T02-T03-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C05
7	T01-T02-T03-T03-T03-T04-T06-T05-T07-T08-T09	C06
8	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C07
9	T01-T02-T04-T06-T05-T07-T08-T09	C08
10	T01-T02-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C09
11	T01-T02-T03-T03-T03-T04-T05-T06-T07-T08-T09	C10
12	T01-T02-T03-T04-T06-T05-T07-T08-T09	C11
13	T01-T02-T03-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C12
14	T01-T02-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C13
15	T01-T02-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C14
16	T01-T02-T03-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C15
17	T01-T02-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C16
18	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C17
19	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C18
20	T01-T02-T03-T03-T03-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C19
21	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C20
22	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T04-T05-T06-T07-T08-T09	C21
23	T01-T02-T03-T03-T03-T03-T03-T03-T03-T03-T04-T06-T05-T07-T08-T09	C22

	A	B
1	Tareas Consolidadas Final	ID_TareasConsolidadas
2	T01-T02-T03-T04-T05-T06-T07-T08-T09	TC01
3	T01-T02-T04-T05-T06-T07-T08-T09	TC02

Ilustración 8.2 - Resultado de las dos perspectivas usadas para la consolidación de caminos de tareas sobre el proceso "Venta de artículos a través de página Web"

Equipos Consolidados Final	ID_EquipoConsolidado
E01-E02-E03-E04-E05-E06	EC01
E01-E02-E04-E05-E06-E07	EC02
E01-E02-E04-E05-E06	EC03
E01-E02-E03-E05-E06-E07	EC04
E01-E02-E04-E05	EC05
E01-E02-E03-E04-E05-E06-E07	EC06
E01-E02-E04-E06	EC07
E01-E02-E03-E05-E06	EC08
E01-E04-E05-E06-E07	EC09
E01-E02-E05-E06	EC10
E01-E02-E03-E04-E05	EC11
E01-E02-E03-E04-E06	EC12
E01-E04-E05	EC13
E01-E02-E05	EC14
E01-E02-E04-E06-E07	EC15
E01-E03-E04-E06-E07	EC16
E01-E04-E05-E06	EC17
E01-E03-E04-E05-E06-E07	EC18
E01-E02-E05-E06-E07	EC19
E01-E04-E05-E07	EC20
E01-E02-E03-E04-E06-E07	EC21
E01-E02-E03-E05-E07	EC22
E01-E02-E06	EC23
E01-E02-E04-E05-E07	EC24
E01-E02-E03-E04-E05-E07	EC25
E01-E02-E03-E06-E07	EC26
E01-E02-E06-E07	EC27
E01-E03-E04-E05-E07	EC28
E01-E03-E04-E06	EC29
E01-E03-E04-E05-E06	EC30
E01-E04-E06-E07	EC31
E01-E04-E06	EC32
E01-E02-E03-E05	EC33
E01-E02-E03-E06	EC34
E01-E03-E04-E05	EC35

Tabla 8.1 - Lista de equipos consolidados del proceso "Venta de artículos a través de página Web" (no se considera orden ni repeticiones de participación)

Luego de completar la consolidación de tareas y ejecutores, se procede a calcular el tiempo que transcurrió, desde el comienzo hasta el término de cada una de las mil ejecuciones contenidas en este caso de

estudio. Estos cálculos se almacenan en la hoja “Tiempo” de la planilla “Pre-procesador Logs”. Los dos nuevos campos de información que se encuentran en esta hoja son “Tiempo_Desde_Ultima_Tarea” y “Tiempo_Desde_Comienzo”.

Con los pasos descritos hasta este punto, ya se habrá completado la revisión de los atributos fundamentales, o básicos, del caso “Ventas a través de página Web”. Con esta información ya se puede generar la base de los datos que serán exportados a Weka. La característica principal del formato de estos datos es el hecho que cada ejecución sólo ocupará una fila en el registro. Con esta consolidación de los datos, se logra un cambio de una base de 9.077 filas a tan sólo mil. Aún más importante, es que se entrega a Weka la información en el formato necesario para que analice cada caso como un proceso completo, evitando que analice cada tarea como una actividad independiente del proceso del cual forma parte.

Finalmente, luego de armar la base de los datos que serán llevados a Weka, se procede a incluir los tres atributos característicos. Estos atributos se usan para complementar la descripción de las ventas a través de la página Web, y se pueden incluir en la hoja de datos que se exportará a Weka usando el procedimiento semi-automático implementado en la hoja “Otros Atributos”. También se incluyó un cuarto atributo característico de manera manual. Este atributo lleva el nombre de “patrón_prom”, y no se pudo incluir a través de la hoja “Otros Atributos” ya que no se obtiene en el archivo de extensión .mxml que entrega ProM.

8.1.2 Resultados pre-procesamiento en Weka

Al ingresar los datos a Weka, el investigador tiene la posibilidad de complementar el pre-procesamiento recién realizado. Las herramientas

que entrega Weka para esto, han sido utilizadas en este trabajo principalmente para descartar ciertos atributos y normalizar los valores de otros. No se utiliza como herramienta exclusiva para normalizar, dado que Weka no entrega clasificadores fáciles de comprender. Para un detalle de las distintas herramientas utilizadas en este trabajo para pre-procesar en Weka, se recomienda revisar la sección 6.3.2 de este documento.

Al revisar los distintos atributos del caso en la pestaña "Preprocess" de Weka, se optó por eliminar los atributos "caseID" y "ID_Equipo".

La situación del atributo "caseID" era conocida antes del análisis en Weka. Sólo se dejó como parte de los datos para entregar un ejemplo extremo, donde los datos del atributo muestran una enorme variabilidad, por lo cual no pueden aportar al estudio del proceso. Mientras más alta la variabilidad de los valores de un atributo, se vuelven más específicas e irregulares las apariciones de cada uno de estos, produciéndose la situación totalmente contraria a la deseada para definir patrones a través de los casos. El atributo "caseID" presenta el extremo que se puede encontrar en variabilidad, al tener una cantidad de valores igual a la cantidad de casos, haciendo imposible agrupar estos datos, ya que no existen datos con el mismo valor.

La eliminación del atributo "ID_Equipo" es consecuente con los resultados obtenidos al buscar los grupos de equipos en la planilla "Pre-procesador Logs". Al encontrar 524 equipos distintos, era muy baja la probabilidad de que este atributo entregara información relevante respecto al caso. Esto se confirma al analizar las frecuencias de cada equipo, donde

ninguno supera las trece repeticiones y sólo tres superan las diez apariciones.

8.1.3 Resultado normalización en Excel

Como el algoritmo Apriori disponible en Weka sólo recibe atributos nominales, se procede a normalizar los dos atributos numéricos que presenta el caso de ventas por Internet. Estos atributos son "Tiempo_Desde_Comienzo" y "patrón_prom". Este último atributo no debería ser considerado como numérico, ya que usa números para identificar cada tipo de patrón distinto encontrado en ProM. Pero, como el atributo sólo fue catalogado con números, sin ninguna letra, Weka lo interpretó como un atributo numérico.

Antes de utilizar Excel, se probó la normalización de datos con Weka sobre el atributo "Tiempo_Desde_Comienzo". El resultado, además de presentarse en un formato poco amigable, entrega grupos que no representan la real distribución de los valores de este atributo. Dado este resultado se optó por realizar la normalización usando fórmulas en Excel. Lo que se realizó en este programa fue simplemente agregar una letra T delante del tiempo de cada caso, quedando doce grupos. El resultado de este procedimiento permite agrupar correctamente los valores del atributo, asignando a cada tiempo la frecuencia que le corresponde. Esta normalización realizada con Excel se compara contra lo obtenido en primera instancia con Weka en la Ilustración 8.3. A través de esta comparación, se observa la simplificación en los denominadores usados para describir cada valor. Primero, se eliminan los rangos que genera Weka, optando por un formato de más fácil lectura, que es el que se genera utilizando Excel. Además de este cambio, también se puede observar que cada normalización muestra una cantidad distinta de valores

que puede tomar el atributo. La normalización entregada por Weka sólo muestra nueve valores, ya que al normalizar junta tiempos de ejecución, que no eran iguales, en un mismo rango. Esto se puede verificar al observar el valor '(-inf-9550]' con una frecuencia de 731, el cual juntó en un sólo rango los casos que demoraron 5.800 o 8.800 minutos, los cuales tienen frecuencias de 491 y 240 respectivamente.

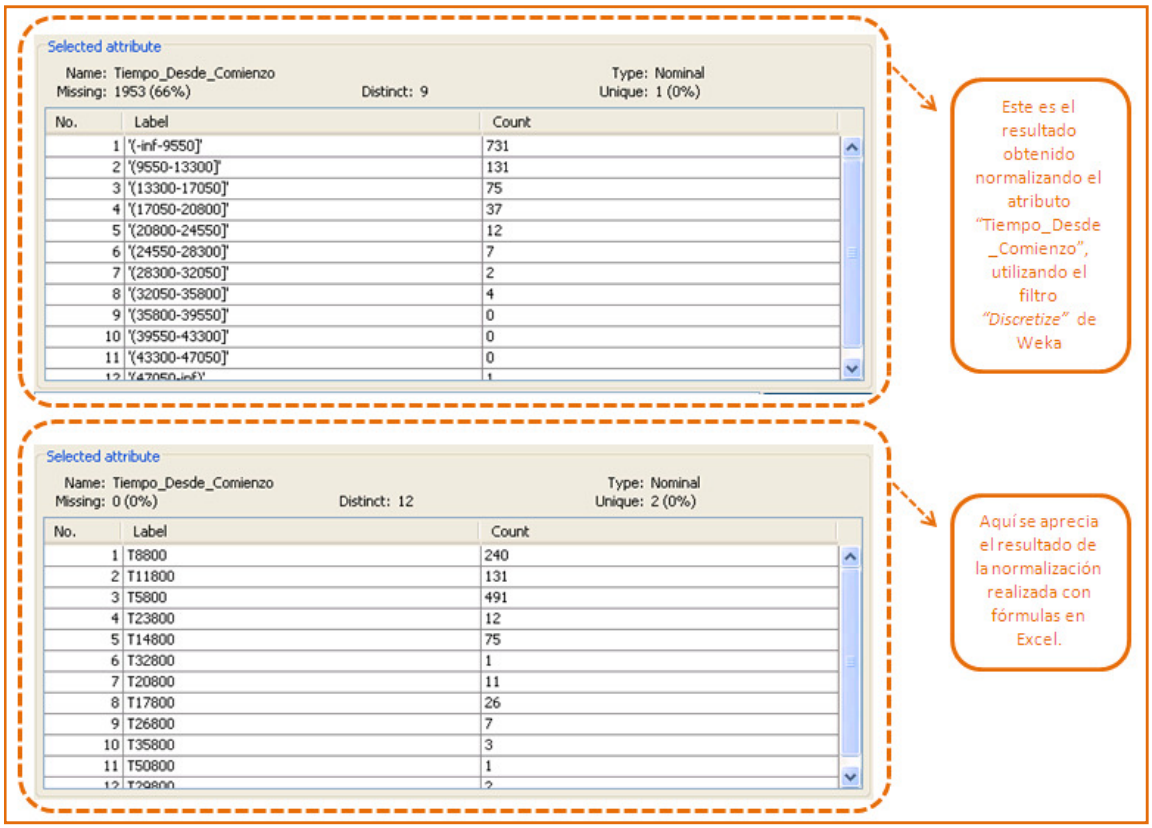


Ilustración 8.3 - Comparación de la normalización del atributo "Tiempo_Desde_Comienzo", primero realizada con Weka y luego con Excel

También se aplicó la herramienta de normalización de Weka sobre los datos del atributo "patrón_prom". El proceso, cuyos resultados se aprecian en la Ilustración 8.4, arrojó correctamente la cantidad de valores

distintos y la frecuencia de cada uno. Sin embargo, el denominador que utiliza Weka para nombrar los distintos grupos no tiene un formato de fácil lectura. Esta situación sumada al hecho que no se encontró un gran número de patrones distintos (sólo doce), llevaron a optar por realizar la normalización a través de fórmulas en Excel. La edición realizada en este programa consistió en agregar al identificador, de cada valor encontrado en el atributo "patrón_prom", una letra P al comienzo.

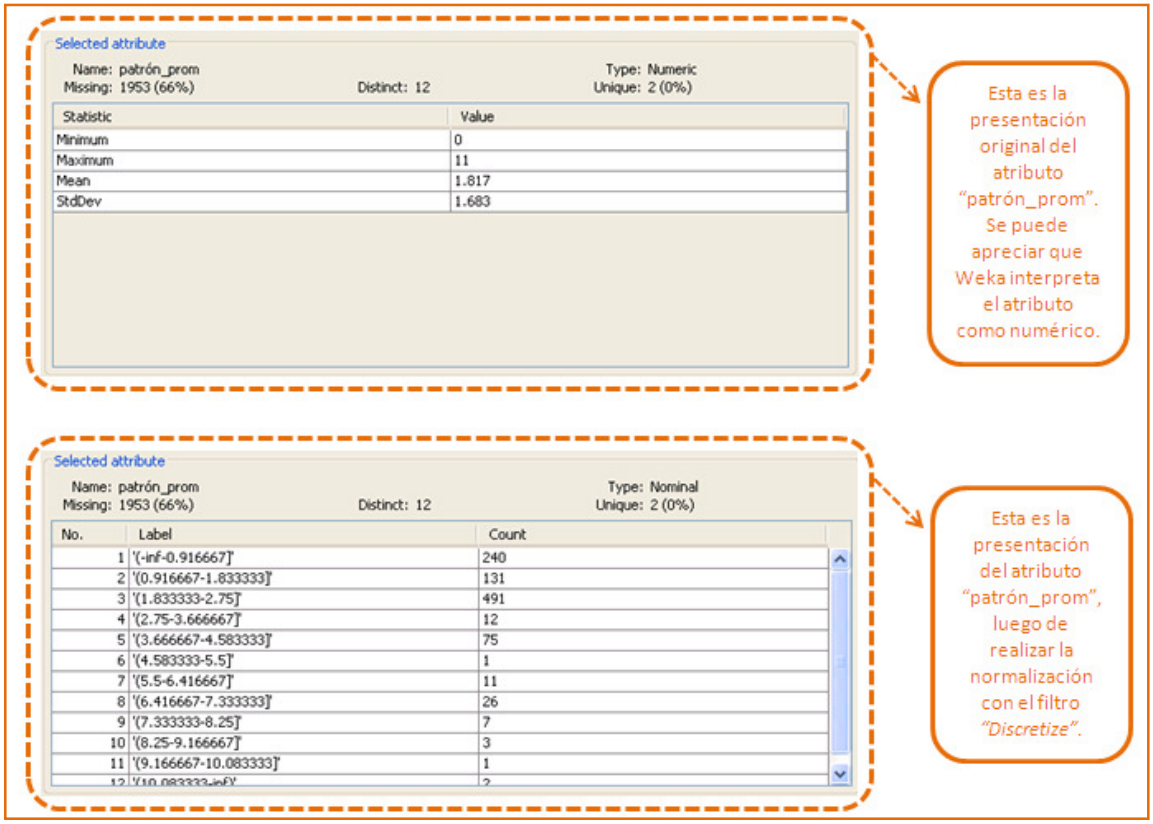


Ilustración 8.4 - Presentación de los datos del atributo "patrón_prom", antes y después de normalizar en Weka

8.1.4 Resultados algoritmo Apriori

Luego de normalizar los atributos numéricos en Excel, se cuenta con la base de datos completamente procesada para continuar con el

análisis de los casos mediante el algoritmo Apriori. Antes de ejecutar el algoritmo, se deben ingresar algunos parámetros que pueden tener gran influencia en los resultados que se obtengan del procedimiento. Para mayor información de los parámetros del algoritmo Apriori, y cómo configurarlos, se recomienda revisar la sección 6.3.3.

En los resultados que se apreciarán a continuación se utilizaron el método "car", junto a un "LowerBoundMinSupport" igual a 0.1, "minMetric" de 0.9 y "numRules" igual a 50 (para limitar el número de reglas que se pueden obtener). Con esta configuración el programa entrega 45 reglas. De este total, hay seis reglas que se destacan por sobre el resto.

Hay cuatro reglas de asociación, entregadas por el algoritmo Apriori, que cubren el 100% de los casos exitosos:

- **Regla 1:** Monto = 2000 [285] –> OK = OK [285]
- **Regla 2:** Monto = 1000 [270] –> OK = OK [270]
- **Regla 3:** Monto = 3000 [125] –> OK = OK [125]
- **Regla 4:** cantidad=2.0 monto=>3000 [100]–> OK=OK [100]

Estas reglas muestran resultados que destacan por la completa confianza que entregan. Con estos resultados, se puede indicar que en todos los casos que la venta alcanzó montos de 1.000, 2.000 o 3.000, el resultado fue una venta exitosa. Con esta conclusión ya se ha cubierto un 87% de los casos exitosos contenidos en el caso. La cuarta regla en tanto, indica que las cien ventas donde el monto fue mayor a 3.000 y la cantidad de artículos igual a dos, también concluyeron en una venta exitosa. Reuniendo estas cuatro reglas se cubren completamente los casos

exitosos, con una confianza del 100%. Estos resultados están entregando distintos patrones sobre el comportamiento del proceso. Estos patrones deben ser contemplados a la hora de tomar decisiones o efectuar acciones respecto al proceso.

A continuación se revisará un ejemplo de cómo los resultados, recién expuestos, deberían impactar sobre las decisiones y acciones que se tomen sobre el proceso:

"El patrón más relevante está indicando que hay altas probabilidades de terminar exitosamente una venta, si el monto no supera los 3.000. Con esta información, las acciones lógicas deberían apuntar a lograr que los consumidores armen su paquete de compra sin superar los 3.000, sabiendo que así se estaría potenciando el poder concretar satisfactoriamente la transacción a través de la página Web. Tomando esta decisión, se establece la base para crear proyectos que apunten a este objetivo: análisis de la cartera de productos, promociones de marketing, descuentos, etc."

Como las reglas analizadas hasta ahora están basadas principalmente en un sólo atributo, se hace llamativo el poder ver como se comportan los demás atributos en los casos descritos por los patrones encontrados. Este análisis se realizará en la siguiente sección de este trabajo, buscando responder interrogantes como: ¿qué frecuencias muestran los valores de cada atributo en los casos exitosos?, ¿qué atributos presentan mayor o menor variabilidad en su valor, si se observan solamente los casos exitosos?, entre otras preguntas.

Hasta ahora se han revisado cuatro reglas de asociación, y se había mencionado que 6 de las 45 reglas encontradas presentaban resultados

importantes para la investigación. Las dos reglas restantes describen el 100% de los casos no exitosos:

- **Regla 5:** Cantidad = 3 y Monto => 3000 [120] –> OK = NOK [120]

- **Regla 6:** Cantidad => 3 y Monto => 3000 [100] –> OK = NOK [100]

Como las cuatro reglas descritas anteriormente, estos resultados sobresalen del resto por la alta confianza que muestran, y por el hecho que los 220 casos no exitosos, de un total de mil, pueden ser agrupados o relacionados en tan sólo dos reglas. Los patrones definidos por estas reglas, indican una alta probabilidad de que si la venta supera los 3.000 en el monto y considera tres o más artículos, el resultado será una venta cancelada (o no exitosa).

8.1.5 Resultado análisis de atributos en grupos determinados de casos

El algoritmo Apriori entregó resultados importantes, que permiten entender con alta seguridad como se comportan los casos de ventas a través de Internet. Sin embargo, el hecho que los resultados mostrados por Weka estuvieran basados en uno o dos atributos, mostró la necesidad de implementar una herramienta que permitiera analizar los valores que toman los demás atributos en los casos descritos por los patrones. Por esto se desarrolló una herramienta en Excel, que permite realizar un análisis de atributos sobre grupos determinados de casos. Esta herramienta es llamada “Analiza Patrones", y con ella se analizaron tanto los patrones que describen los casos exitosos, como aquellos que mostraron casos donde no se concretó la transacción.

Los casos exitosos lograron ser descritos a través de cuatro reglas, basadas principalmente en el atributo “Monto”. Para analizar más acabadamente el comportamiento de los demás atributos en estos casos exitosos, se llevó los datos del caso a la planilla “Analiza Patrones", donde se filtraron todos los casos de manera de dejar solamente aquellos 780 en que efectivamente se concretó la venta.

Analizando el atributo “Tiempo_Desde_Comienzo” en los 780 casos exitosos, se encontró que cuatro valores, de un total de doce, concentraban el 94% de estas instancias exitosas. Los cuatro valores fueron:

- **T5800** con 375 apariciones
- **T8800** con 193 apariciones
- **T11800** con 103 apariciones
- **T14800** con 59 apariciones

Es importante destacar que la aparición de varias ejecuciones del proceso con un mismo tiempo, responde a como fue confeccionado el caso de estudio. De manera de simplificar los atributos de tiempo y poder obtener grupos de mayor frecuencia, se consolidaron los tiempos de ejecución en los doce grupos que se observan a través de los datos. De esta manera el grupo que representa, por ejemplo, T5800 son aquellas ejecuciones que demoraron entre 0 y 5.800 minutos, en tanto que T8800 comprende a aquellas ejecuciones que tomaron más de 5.800 minutos, hasta 8.800 minutos.

Los 730 casos abarcados por los cuatro valores anteriormente listados, coinciden con los tiempos más cortos de ejecución encontrados a

través de todas las instancias exitosas del proceso. No obstante, no se puede establecer una relación entre el tiempo de ejecución y el resultado del proceso. Esto se comprueba al comparar la distribución de estos valores en los casos exitosos, contra los mil casos que se analizaron. Con esta comparación, la cual se detalla en la Tabla 8.2, se observa que las distribuciones son muy similares, impidiendo establecer una conexión entre ciertos valores y el resultado de la transacción.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
T11800	103	13%	131	13%
T14800	59	8%	75	8%
T17800	18	2%	26	3%
T20800	9	1%	11	1%
T23800	10	1%	12	1%
T26800	7	1%	7	1%
T29800	2	0%	2	0%
T32800	1	0%	1	0%
T35800	2	0%	3	0%
T50800	1	0%	1	0%
T5800	375	48%	491	49%
T8800	193	25%	240	24%

Tabla 8.2 - Comparación de la distribución de los valores del atributo "Tiempo_Desde_Comienzo", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos exitosos

A continuación, se analizó el atributo "ID_Camino" en los 780 casos exitosos. El resultado mostró que siete, de los 22 caminos posibles de tareas, son los que presentan mayor frecuencia. Estos siete caminos son:

- **C04** con 189 apariciones
- **C08** con 186 apariciones
- **C11** con 98 apariciones

- **C01** con 95 apariciones
- **C03** con 62 apariciones
- **C02** con 41 apariciones
- **C10** con 38 apariciones

Los caminos de tareas recién listados, en conjunto, cubren 709 de los 780 casos exitosos, lo que representa cerca de un 91% del grupo de transacciones donde se concretó la venta.

Al igual que como se realizó en el caso del atributo "Tiempo_Desde_Comienzo", se comparó la distribución de los valores de "ID_Camino" sobre dos bases: observando los casos exitosos y el total de casos disponibles para el análisis. Observando esta comparación, detallada en la Tabla 8.3, se constata que el peso de los caminos, dentro de los casos exitosos, es muy similar al peso o participación que muestran en el total de casos. Por esta razón, no se puede establecer una relación o patrón, que relacione a las secuencias de tareas ejecutadas con el resultado del proceso. Esta situación también sirve para confirmar porqué el algoritmo Apriori no entregó una regla de asociación que incluyera el atributo "ID_Camino".

Si luego se analiza el atributo "ID_TareasConsolidadas", el cual muestra los caminos de tareas sin considerar el orden de las tareas ni las repeticiones de éstas, se aprecia que los dos únicos valores que toma este atributo tienen distribuciones prácticamente idénticas. Esta situación, como se puede apreciar en la Tabla 8.4, se da tanto en los casos exitosos como en el total de ejecuciones. Esto indica que no hay una relación entre los valores de "ID_TareasConsolidadas" y el resultado del proceso, y que

tampoco hay una tendencia de alguno de sus valores a aparecer con más frecuencia que el otro.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
C01	95	12%	118	12%
C02	41	5%	62	6%
C03	62	8%	69	7%
C04	189	24%	246	25%
C05	7	1%	9	1%
C06	21	3%	24	2%
C07	1	0%	1	0%
C08	186	24%	245	25%
C09	3	0%	3	0%
C10	38	5%	51	5%
C11	98	13%	122	12%
C12	5	1%	6	1%
C13	4	1%	5	1%
C14	5	1%	13	1%
C15	2	0%	2	0%
C16	13	2%	13	1%
C17	1	0%	2	0%
C18	1	0%	1	0%
C19	5	1%	5	1%
C20	1	0%	1	0%
C21	1	0%	1	0%
C22	1	0%	1	0%

Tabla 8.3 - Comparación de la distribución de los valores del atributo "ID_Camino", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos exitosos.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
TC01	405	52%	509	51%
TC02	375	48%	491	49%

Tabla 8.4 - Comparación de la distribución de los valores del atributo "ID_TareasConsolidadas", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos exitosos

Continuando con el análisis, al observar la frecuencia de los distintos valores que toma el atributo "ID_EquipoConsolidado", se aprecia, como se destaca en la Tabla 8.5, que cuatro grupos de ejecutores

se reparten un 58% de los casos exitosos. El 42% restante se divide en 31 grupos de ejecutores, cada uno con frecuencias que no superan los 40 casos. Los cuatro equipos mencionados como más relevantes son:

- **EC03** con 221 apariciones
- **EC06** con 87 apariciones
- **EC01** con 78 apariciones
- **EC02** con 66 apariciones

Luego, al comparar la distribución de los equipos recién listados, colocando como base los casos exitosos contra el total de casos para analizar, se observa que la situación no cambia según qué tipo de base se esté analizando. Como sucedió en los análisis previos de otros atributos, el porcentaje de aparición de cada equipo se mantiene prácticamente idéntico al observar solamente los casos exitosos o el total de mil casos disponible.

La tendencia de mostrar la misma proporción en los casos exitosos y el total de ejecuciones, se quiebra al analizar el atributo “cantidad”. Este atributo particularmente presenta dos valores que, como se muestra en la Tabla 8.6, tienen una mayor tasa de participación cuando hubo resultado exitoso contra lo que muestra su tasa en el total de casos analizados. De hecho, todas las apariciones de estos dos valores, coinciden con una venta exitosa. Estos dos valores son:

- **2.0** con 250 apariciones
- **1.0** con 245 apariciones

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
EC01	78	10%	97	10%
EC02	66	8%	83	8%
EC03	221	28%	294	29%
EC04	12	2%	12	1%
EC05	37	5%	43	4%
EC06	87	11%	110	11%
EC07	29	4%	34	3%
EC08	17	2%	20	2%
EC09	8	1%	12	1%
EC10	37	5%	46	5%
EC11	8	1%	10	1%
EC12	14	2%	15	2%
EC13	3	0%	8	1%
EC14	4	1%	5	1%
EC15	10	1%	15	2%
EC16	3	0%	4	0%
EC17	31	4%	43	4%
EC18	15	2%	19	2%
EC19	15	2%	16	2%
EC20	3	0%	3	0%
EC21	16	2%	21	2%
EC22	1	0%	1	0%
EC23	11	1%	14	1%
EC24	10	1%	14	1%
EC25	10	1%	17	2%
EC26	2	0%	3	0%
EC27	1	0%	3	0%
EC28	6	1%	7	1%
EC29	1	0%	2	0%
EC30	13	2%	15	2%
EC31	3	0%	3	0%
EC32	2	0%	4	0%
EC33	2	0%	2	0%
EC34	2	0%	3	0%
EC35	2	0%	2	0%

Tabla 8.5 - Comparación de la distribución de los valores del atributo "ID_EquipoConsolidado", extraída directamente de la planilla "Analiza Patrones", donde además se destacan los valores que acaparan más de la mitad de los casos. Los casos filtrados corresponden a los casos exitosos

A partir de la tasa de distribución de los valores recién mencionados se puede deducir que habrá más posibilidades de tener éxito en el proceso, cuando la orden involucre uno o dos artículos. En tanto que si el cliente solicita tres o más artículos existirá mayor riesgo de no

concretar la venta. Esto se refleja en la Tabla 8.6 en como la tasa de participación de estas cantidades se ve disminuida cuando se observa solamente los 780 casos exitosos.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
1.0	245	31%	245	25%
2.0	250	32%	250	25%
3.0	135	17%	255	26%
Más de 3	150	19%	250	25%

Tabla 8.6 - Comparación de la distribución de los valores del atributo "cantidad", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos exitosos

Al analizar el atributo "monto", como se puede observar en la Tabla 8.7, el proceso tiende al éxito siempre que el monto de la transacción no supere los 3.000. Esta fue la conclusión obtenida de Weka con el algoritmo Apriori y se confirma a través de este segundo análisis.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
1000.0	270	35%	270	27%
2000.0	285	37%	285	29%
3000.0	125	16%	125	13%
Más de 3000	100	13%	320	32%

Tabla 8.7 - Comparación de la distribución de los valores del atributo "monto", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos exitosos

Finalmente, el último análisis de los casos exitosos lleva a observar cómo se comportan los valores del atributo "patrón_prom". Realizando este análisis, se puede identificar, como se destaca en la Tabla 8.8, que hay tres clasificaciones de patrón entregadas por ProM que resaltan por sobre el resto al acaparar más del 85% de los casos exitosos.

Sin embargo, al observar la misma Tabla 8.8, se puede concluir que los patrones entregados por ProM no muestran una incidencia importante sobre el resultado del proceso, ya que muestran la misma tasa de participación a través de los 780 casos exitosos y en los mil casos disponibles para la investigación.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
P0	193	25%	240	24%
P1	103	13%	131	13%
P10	1	0%	1	0%
P11	2	0%	2	0%
P2	375	48%	491	49%
P3	10	1%	12	1%
P4	59	8%	75	8%
P5	1	0%	1	0%
P6	9	1%	11	1%
P7	18	2%	26	3%
P8	7	1%	7	1%
P9	2	0%	3	0%

Tabla 8.8 - Comparación de la distribución de los valores del atributo "patrón_prom", extraída directamente de la planilla "Analiza Patrones", donde además se destacan los tres valores que acaparan la mayor cantidad de casos. Los casos filtrados corresponden a los casos exitosos

Ya cubierto el análisis de todos los atributos sobre los casos exitosos, se procedió a trabajar solamente en aquellos casos donde no se concretó la venta. Para esto se tomó la misma herramienta usada anteriormente, la planilla "Analiza Patrones". Aquí se filtró los datos disponibles de manera de dejar solamente los 220 casos no exitosos. Los resultados que se encuentran al analizar estos casos confirman el análisis realizado previamente sobre los casos exitosos: a excepción de los atributos "cantidad" y "monto", la distribución de los valores no varía si se cambia la base de observación, ya sea ésta los casos no exitosos o todos los casos con que se cuenta sobre el proceso.

Cuando se trabaja con el atributo "cantidad" en la planilla "Analiza Patrones", los resultados son evidentes. Tal cual se obtuvo en Weka con el algoritmo Apriori y como se detalla en la Tabla 8.9, en los casos no exitosos sólo se puede apreciar presencia de cantidades de artículos iguales o mayores a tres.

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
1.0	0	0%	245	25%
2.0	0	0%	250	25%
3.0	120	55%	255	26%
Más de 3	100	45%	250	25%

Tabla 8.9 - Comparación de la distribución de los valores del atributo "cantidad", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos no exitosos

Finalmente, al trabajar con el atributo "monto" como último análisis destacado de los casos no exitosos, se vuelve a confirmar los resultados entregados por el algoritmo Apriori. Como se puede apreciar en la Tabla 8.10, la distribución de valores del atributo "monto" en los casos no exitosos se inclina de manera total a un solo valor: "Más de 3.000".

Valor	Frecuencia del valor en los casos filtrados	Peso de este valor en los casos filtrados	Frecuencia del valor en todos los casos (sin filtrar)	Peso de este valor en todos los casos (sin filtrar)
1000.0	0	0%	270	27%
2000.0	0	0%	285	29%
3000.0	0	0%	125	13%
Más de 3000	220	100%	320	32%

Tabla 8.10 - Comparación de la distribución de los valores del atributo "monto", extraída directamente de la planilla "Analiza Patrones". Los casos filtrados corresponden a los casos no exitosos

8.1.6 Análisis de sensibilidad algoritmo Apriori

En la sección 8.1.4 de este capítulo se revisaron los resultados del algoritmo Apriori, utilizando una determinada combinación de los parámetros más relevantes del algoritmo. La combinación utilizada fue con la opción "car" activada, junto con un "lowerBoundMinSupport" igual a 0,1, "minMetric" de 0,9 y "numRules" igual a 50. Con esta configuración de parámetros, se obtuvieron 45 reglas. Con seis de éstas, se logró describir patrones relevantes para los casos del proceso, tanto para los exitosos, como aquellos en que no se logró concretar la venta. A continuación, se revisará como habría cambiado la cantidad de reglas si se hubiese utilizado otros valores en los parámetros mencionados. Junto con ver la variabilidad en el número de reglas, se buscó también analizar las características de las nuevas reglas resultantes, para revisar si guardaban relación con los resultados obtenidos al analizar los valores de cada atributo, con respecto a los patrones encontrados en la primera búsqueda de patrones.

Para realizar el análisis de sensibilidad se ejecutó el algoritmo Apriori combinando diversos valores de los parámetros "lowerBoundMinSupport" y "minMetric". Se dejó la opción "car" activada para todas las ejecuciones, ya que uno de los objetivos principales era poder establecer patrones que lleven únicamente a los distintos resultados finales del proceso. En tanto, en "numRules" se ingresó el valor 1.000, para no restringir la cantidad de reglas que se pudieran encontrar. En la Tabla 8.11 se encuentra un resumen de los resultados obtenidos a través de este análisis, apreciándose en la esquina superior derecha de la matriz las 45 reglas resultantes de la aplicación de Apriori detallada en la sección 8.1.4 de este capítulo. A continuación, se comentará porqué este resultado es uno de los cuatro casos destacados dentro de la tabla.

Una de las conclusiones que se pueden obtener de esta matriz, es que la cantidad de reglas se ve mucho más afectada al variar el valor de "lowerBoundMinSupport", que por cambios en "minMetric". Particularmente, llaman la atención dos grandes saltos que se aprecian en la cantidad de reglas. Estas grandes variaciones se producen al cambiar "minMetric" de 0,9 a 0,7, y al cambiar "lowerBoundMinSupport" de 0,1 a 0,3. En estos casos se aprecia que, al cambiar en la misma magnitud los dos parámetros, la cantidad de reglas cambia en mayor magnitud con "lowerBoundMinSupport" que con "minMetric".

Otra de las conclusiones que se puede obtener al revisar la Tabla 8.11 es que ningún valor, de ningún atributo, mostró una frecuencia para dar más de un 30% de soporte a los patrones encontrados. Esta es la razón de porque no se encuentran reglas en las últimas tres filas de la tabla.

lowerBoundMinSupport \ minMetric	0,1	0,3	0,5	0,7	0,9
0,1	127	119	116	112	45
0,3	8	8	8	8	0
0,5	0	0	0	0	0
0,7	0	0	0	0	0
0,9	0	0	0	0	0

Tabla 8.11 - Tabla que indica la cantidad de reglas encontradas en el caso "Venta de artículos a través de página Web", usando distintas combinaciones de los parámetros "lowerBoundMinSupport" (entre 0,1 y 0,9) y "minMetric" (entre 0,1 y 0,9)

Dado el importante cambio en el número de reglas, que se produjo al cambiar "minMetric" de 0,9 a 0,7, y al cambiar "lowerBoundMinSupport" de 0,1 a 0,3., se revisó en profundidad como se mueven la cantidad de reglas dentro de los rangos mencionados. Los resultados de este análisis se pueden apreciar en la Tabla 8.12.

lowerBoundMinSupport \ minMetric	0,7	0,75	0,8	0,85	0,9
0,1	112	97	51	45	45
0,15	44	44	12	6	6
0,2	20	20	4	4	4
0,25	11	11	3	3	3
0,3	8	8	0	0	0

Tabla 8.12 - Tabla que indica la cantidad de reglas encontradas en el caso "Venta de artículos a través de página Web", usando distintas combinaciones de los parámetros "lowerBoundMinSupport" (entre 0,1 y 0,3) y "minMetric" (entre 0,7 y 0,9)

A continuación, se revisarán algunas de las nuevas reglas que fueron surgiendo al ir bajando el nivel de confianza requerido (modificando el parámetro "minMetric"). Para esta revisión, se dejará el nivel de soporte requerido siempre en 0,1, lo cual corresponde a la primera fila de la Tabla 8.12.

Al revisar el caso en que "minMetric" es igual a 0,8, se observa que el número de reglas resultantes aumenta de 45 a 51, respecto a cuándo se utilizó 0,9 como nivel de confianza requerido. Estas seis

nuevas reglas se pueden apreciar en la Ilustración 8.5. Los principales *itemsets* que aparecen en estas nuevas reglas, son aquellos casos que presentaron un tiempo de ejecución igual a T8800, TC01 como grupo de tareas consolidado, o P0 como patrón entregado por ProM. Estos *itemsets*, y las reglas de asociación resultantes al combinarlos, presentan altos niveles de soporte y confianza. De todas maneras, estos *itemsets* y reglas no destacan por sobre las primeras 45 reglas encontradas. Esto debido principalmente a que, las primeras 45 reglas presentaron un 100% de confianza, contra el 80% que entregan estas nuevas reglas. Además, en la sección 8.1.5, se logró destacar que estos valores tenían alta presencia en los casos exitosos, pero no eran determinantes sobre el resultado final. Esto se confirmó al revisar la distribución de cada uno de los atributos en el total de casos, considerando éxitos y fracasos, donde no se observaron cambios importantes respecto a analizar solamente los casos exitosos. En otras palabras, los valores de los atributos encontrados en las seis nuevas reglas no tienen una tendencia marcada a alguno de los resultados del proceso, por lo cual no se pueden considerar como patrones importantes.

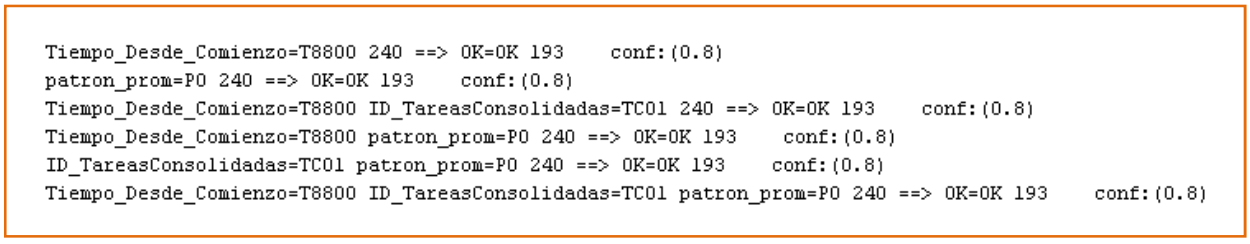


Ilustración 8.5 - Extracto de la salida de Weka al ejecutar Apriori, donde se detallan las nuevas reglas de asociación encontradas por el algoritmo, al bajar el parámetro "minMetric" de 0,9 a 0,8

El segundo grupo de reglas nuevas que fue analizado, es aquel que se produjo al bajar el nivel de confianza requerido de 0,8 a 0,75. Como se aprecia en la primera fila de la Tabla 8.12, con este cambio la cantidad de

reglas aumenta de 51 a 97. Estas 46 nuevas reglas (detalle de las reglas en Anexo 8) presentan importantes niveles de soporte respecto a los casos positivos del proceso, superando normalmente las 150 instancias, llegando incluso sobre las 200 (de un total de 1.000 instancias). Sin embargo, sólo aparecieron al fijar el nivel de confianza aceptado bajo el 80%. Sumado a esta baja en el nivel de confianza, hay otro factor que influye en que estas nuevas reglas no sean de gran relevancia para la investigación. Este factor es que las 46 reglas tienen alta presencia de atributos que fueron evaluados en la sección 8.1.5 de este trabajo, mostrando distribuciones muy similares al comparar los porcentajes de cada valor en los casos exitosos y en todas las ejecuciones del proceso. Algunos ejemplos que destacan son los caminos C08 y C04, el grupo de tareas consolidado TC02 o los casos con un tiempo de ejecución igual a T5800. Todos presentan altos niveles de soporte para los resultados positivos del proceso. Pero, al observar su distribución en el total de ejecuciones, se concluye que no presentan una tendencia marcada a un resultado particular.

Los resultados del análisis de sensibilidad aquí expuesto, demuestran la utilidad de las herramientas usadas para obtener la información expuesta en la sección 8.1.5. Allí, se mostró la distribución de cada valor en todos los atributos del proceso, logrando definir qué atributos presentaban valores con tendencia marcada a algún resultado final del proceso. El tener ese análisis permitió entender de manera más justificada, porqué las nuevas reglas que aparecían en este análisis de sensibilidad no son de mayor relevancia para la investigación, como sí lo son las 45 de la primera búsqueda detallada en la sección 8.1.4, en particular las seis reglas que se describieron en extenso en esa sección.

8.2 Resultados detección de anomalías

Para poner a prueba el método de búsqueda de anomalías diseñado en este trabajo, se analizaron 2.485 casos, distribuidos en cuatro grupos de datos. Todos estos casos corresponden a ejecuciones de un proceso de solicitud de crédito hipotecario. La cantidad de casos no es igual a través de los cuatro grupos, el detalle es el siguiente:

- **Grupo de datos A** cuenta con 590 casos
- **Grupo de datos B** cuenta con 580 casos
- **Grupo de datos C** cuenta con 740 casos
- **Grupo de datos D** cuenta con 575 casos

Se separó los casos en cuatro grupos de datos, ya que interesaba ver como se comportaba el algoritmo implementado ante distintas situaciones de anomalía, conservando los mismos tipos de atributo a través de todos los casos. Estos cuatro grupos de datos describen cada caso utilizando solamente los cinco atributos básicos que se pueden obtener de un proceso: “caseID”, “taskID”, “originator”, “eventType” y “timestamp”. Dada esta situación, la búsqueda de anomalías se centró, principalmente, en encontrar anomalías observando los caminos de tareas realizados y los equipos de trabajo que participaron en cada grupo de datos.

Para más detalles sobre la estructura de los casos de prueba utilizados en este capítulo, se recomienda revisar la sección 4.2.2 de este trabajo.

Para encontrar los casos anómalos de cada conjunto de datos, primero se debe pre-procesar los datos entregados por ProM, para luego ir directamente en búsqueda de las anomalías a través del algoritmo Interquartile Range implementado en la planilla “Busca Anomalías".

8.2.1 Resultados pre-procesamiento

Antes de poder descubrir las anomalías presentes en cada uno de los grupos de datos, el pre-procesamiento permite reunir en una sola fila de datos toda la información disponible para cada caso. Como se ha mencionado en distintos puntos de este trabajo, los datos que entrega ProM describen cada caso del proceso en un número variable de filas, el cual depende de la cantidad de tareas que se ejecuten en cada caso. Si se analizara directamente estos datos, sin realizar el pre-procesamiento adecuado, se estaría analizando cada tarea como una ejecución independiente, ignorando el hecho que forman parte de un conjunto mayor, que es el proceso donde se ejecutan. Para llevar la información de cada caso a una sola fila se usa la planilla denominada "Pre-procesador Logs", la cual identifica y consolida toda la información que entrega ProM.

Los principales resultados de la consolidación de información, realizada sobre cada uno de los cuatro grupos de datos, se detallan en la Tabla 8.13. La primera parte de esta tabla muestra la significativa disminución de filas a ser procesadas en la búsqueda de anomalías. En todos los casos, la cantidad de filas luego del pre-procesamiento fue igual a la cantidad de casos de cada grupo de datos. En tanto, la segunda parte de la tabla entrega un resumen de los atributos encontrados por la planilla "Pre-procesador Logs". Lo más importante de esta segunda sección de resultados es que, a diferencia de lo ocurrido en el pre-procesamiento de los casos de ventas por Internet, en este proceso no se consiguió disminuir el número de caminos de tareas o de equipos con la segunda consolidación. Esto se observa en el caso de las tareas, al revisar la cantidad de valores encontrados para los atributos "ID_Camino" y "ID_TareasConsolidadas". Estos dos atributos muestran las mismas cifras

a través de los cuatro grupos de datos. Lo mismo ocurre en el caso de los equipos al observar los atributos "ID_Equipo" y "ID_EquipoConsolidado". Estas observaciones llevan a la conclusión de que no se presentaron caminos de tareas que presentaran las mismas tareas, pero con un orden distinto o con más o menos repeticiones de algunas de éstas. Análogamente, se llega a la misma conclusión respecto a los equipos de ejecutores. Por lo tanto, cómo no hubo una simplificación de los caminos de tareas o de los equipos de ejecutores, el investigador puede realizar su trabajo solamente observando los atributos "ID_Camino" y "ID_Equipo", descartando "ID_TareasConsolidadas" y "ID_EquipoConsolidado", por ser redundantes a través de los cuatro conjunto de datos.

Número de filas de información	Grupo de datos A	Grupo de datos B	Grupo de datos C	Grupo de datos D
Cantidad de filas provenientes desde ProM	2.861	2.806	3.468	2.638
Cantidad de filas luego del pre-procesamiento	590	580	740	575
Número de valores identificados para cada atributo				
ID_Tarea	5	5	5	5
ID_Ejecutor	15	20	20	20
ID_Camino	3	3	3	3
ID_TareasConsolidadas	3	3	3	3
ID_Equipo	6	6	6	6
ID_EquipoConsolidado	6	6	6	6

Tabla 8.13 - Resultados pre-procesamiento aplicado a los cuatro grupos de datos del proceso "Evaluación solicitud de crédito hipotecario"

8.2.2 Resultados algoritmo Interquartile Range

Luego de consolidar los datos en la planilla “Pre-procesador Logs”, la investigación continuó en la planilla “Busca Anomalías”. En esta planilla se ingresaron, por separado, todos los datos de cada uno de los cuatro casos disponibles para el análisis.

Originalmente, los datos entregados por el pre-procesador incluían seis campos:

- “caseID”
- “ID_Camino”
- “ID_Equipo”
- “Tiempo_Desde_Comienzo”
- “ID_TareasConsolidadas”
- “ID_EquipoConsolidado”

De estos seis campos, inmediatamente se eliminaron tres atributos en los cuatro casos analizados:

- **caseID:** Este atributo no aporta ninguna información para la búsqueda de anomalías, ya que sólo identifica el número del caso que se está analizando. Por esto, este dato no entrega ninguna característica relevante del proceso, por lo cual se eliminó para higienizar la información a procesar.
- **ID_TareasConsolidadas:** Como se comentó en los resultados del pre-procesamiento, al hacer la consolidación de tareas sin considerar repeticiones ni orden, no se consiguió disminuir la cantidad respecto a los caminos de tareas registrados en

"ID_Camino". Por esta razón, conservar este atributo es redundante y no aporta nueva información para el análisis. Dada esta situación, se procedió a eliminar este atributo antes de ejecutar la búsqueda de anomalías.

- **ID_EquipoConsolidado:** Completamente análogo a lo comentado para las tareas consolidadas. Los equipos de trabajo presentaron la misma situación que los caminos de tareas, por lo cual se procedió a eliminar el campo.

Eliminados los atributos recién mencionados, se procedió a ingresar los parámetros y características de los datos, los cuales son necesarios para que el algoritmo Interquartile Range se pueda ejecutar. Para más detalles sobre el funcionamiento de este algoritmo y sus parámetros, se recomienda revisar el capítulo 7.3 de este trabajo.

La primera información requerida es el tipo de datos de cada uno de los atributos disponibles. Esta información debe ser ingresada en la hoja "02. Atributos" de la planilla "Busca Anomalías". Aquí se debe señalar si cada atributo es de tipo numérico o nominal. Como se describió anteriormente en este trabajo, el algoritmo Interquartile Range ocupa la mediana y otros datos estadísticos para establecer el rango de normalidad de los datos. Por esta razón, el algoritmo se modifica levemente si se desea realizar el análisis sobre atributos nominales. De esta manera, con la información entregada, la planilla selecciona el método para buscar las anomalías, acorde al tipo de datos sobre el cual se desea realizar el análisis. En los ejemplos expuestos en este capítulo, como los cuatro casos analizados contaban con los mismos tres atributos, la declaración del tipo de dato, ejemplificada en la Ilustración 8.6, fue igual para todos.

Ilustración 8.6 - Extracto de la hoja "02. Atributos" de la planilla "Busca Anomalías" donde se declaró el tipo de datos de cada uno de los atributos del proceso para obtener un crédito hipotecario

Una vez ingresados el tipo de datos de cada uno de los atributos disponibles para la investigación, cualquiera fuera la selección sobre el tipo de dato a ser utilizado para buscar las anomalías, el siguiente paso en el análisis debía ser ingresar los tres parámetros señalados en la Ilustración 8.7. Para más detalles sobre cada uno de estos parámetros, se recomienda revisar la sección 7.3.3 de este trabajo.

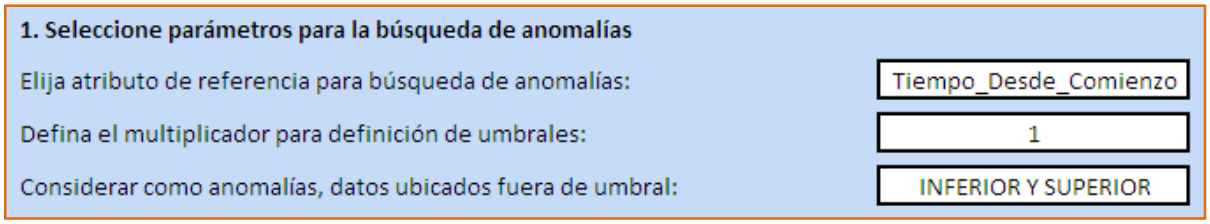


Ilustración 8.7 - Extracto de la planilla "Busca Anomalías" que se puede encontrar en las hojas "03. Parametros" y "03. Parametros B". En esta sección se debe ingresar los parámetros que serán utilizados en la búsqueda de anomalías

Cambiando los tres parámetros detallados en la Ilustración 8.7, se realizaron diversas búsquedas de anomalías sobre cada uno de los cuatro grupos de datos disponibles para el análisis. En la Tabla 8.14 se incluye las búsquedas que dieron los resultados más relevantes en cada grupo de

datos. La relevancia se definió comparando las anomalías encontradas con el algoritmo Interquartile Range contra lo esperado por el diseñador de los casos de estudio.

Parámetros para la búsqueda de anomalías	Grupo de datos A	Grupo de datos B	Grupo de datos C	Grupo de datos D
Atributo de referencia	ID_Equipo	ID_Equipo	ID_Camino	ID_Camino
Multiplicador para definición de umbrales	1,0	1,0	-0,1	-0,1
Umbrales analizados para la búsqueda de anomalías	Inferior	Inferior	Inferior	Inferior
Resultados de la búsqueda de anomalías				
Número total de casos	590	580	740	575
Número esperado de casos anómalos	89	111	260	283
Número encontrado de casos anómalos	89	111	232	237
Porcentaje de casos anómalos encontrados	**100%**	**100%**	**89%**	**84%**

Tabla 8.14 - Resultados de la búsqueda de anomalías más relevante para cada grupo de datos del caso "Evaluación solicitud de crédito hipotecario"

A partir de los resultados mostrados en la Tabla 8.14 y algunos análisis complementarios, se obtuvieron las siguientes conclusiones respecto a cada caso:

- En el **caso A** se encontró el mismo número de anomalías que esperaba el diseñador del caso. Estos 89 casos anómalos también se repitieron al buscar anomalías fuera del umbral superior, por lo cual se concluye que todas las anomalías se encuentran bajo el umbral inferior.

- En el **caso B** también se consiguió encontrar las mismas anomalías esperadas para el caso. Además del análisis sobre el atributo "ID_Equipo" que entregó las 111 anomalías, también

hubo otro análisis importante sobre el atributo “ID_Camino” que entregó 94 anomalías. Se constató que estas 94 anomalías ya eran parte de las 111 anomalías encontradas con el atributo “ID_Equipo”. Las dos búsquedas permitieron tener una mejor idea de las características de los casos anómalos, lo cual se observará en la sección 8.2.3 de este trabajo, cuando se describa el análisis de comportamiento de cada atributo en los casos anómalos.

- Al analizar el **caso C**, utilizando el atributo “ID_Camino”, se logró encontrar gran parte de las anomalías esperadas, faltando sólo 28 de los 260 casos detectados por el diseñador del caso. Estos casos que no fueron detectados observando el atributo “ID_Camino”, se encontraron al analizar el caso usando como referencia para la búsqueda de anomalías el atributo “ID_Equipo”. Las características de éstos y los demás 232 casos anómalos, como también la necesidad de utilizar un multiplicador negativo para el análisis, serán parte de los temas a tratar en la sección 8.2.3 de este trabajo.

- En el **caso D**, al igual como ocurrió en el caso C, no se logró encontrar todas las anomalías con un sólo análisis. Pero sí se llegó a las anomalías esperadas complementando el análisis sobre el atributo “ID_Camino”, que entregó 237 anomalías, con un análisis sobre el atributo “ID_Equipo”. Este segundo análisis entregó 105 anomalías, de las cuales 59 ya habían aparecido usando “ID_Camino” como referencia, en tanto que las 46 anomalías restantes resultaron ser las anomalías que faltaban por identificar. En la sección 8.2.3 se revisarán las características más importantes de estas anomalías, como

también se analizará las razones que llevaron a utilizar un multiplicador negativo y los factores que influyeron en que la cantidad de anomalías fuese tan alta en este conjunto de datos.

8.2.3 Resultado análisis de atributos en grupos determinados de casos

Además de la posibilidad de buscar los casos con un comportamiento significativamente distinto a la mayoría de las ejecuciones de un proceso, en la planilla "Busca Anomalías" se implementó una segunda herramienta para analizar los distintos atributos de cada grupo de datos. Esta segunda herramienta se encuentra implementada en las hojas “05. Metricas” y “05. Metricas B”, y su objetivo es permitir observar la frecuencia que presentó cada valor de cada atributo del proceso, tanto en los casos anómalos, como en el total de ejecuciones registradas. Esta información permitirá establecer, si las hubiera, relaciones entre valores de un atributo y la clasificación de los casos.

En las hojas “05. Metricas” o “05. Metricas B” se encontrará una vista similar al ejemplo que se muestra en la Ilustración 8.8. Luego de revisar la información que se entrega en estas hojas, sólo se tiene que escoger cuál es el atributo del cual se desea obtener el desglose con la frecuencia de sus valores. Para más información sobre las características e indicadores implementados en las hojas “05. Metricas” y “05. Metricas B”, se recomienda revisar la sección 7.3.5 de este trabajo.

Probando distintos atributos, se realizó el análisis de frecuencias en cada uno de los cuatro conjunto de de datos con que se trabajó. Estos análisis permitieron definir de manera más precisa las características de los casos anómalos, como también se logró obtener más información para

justificar el uso de ciertos parámetros y los resultados de la búsqueda de anomalías.

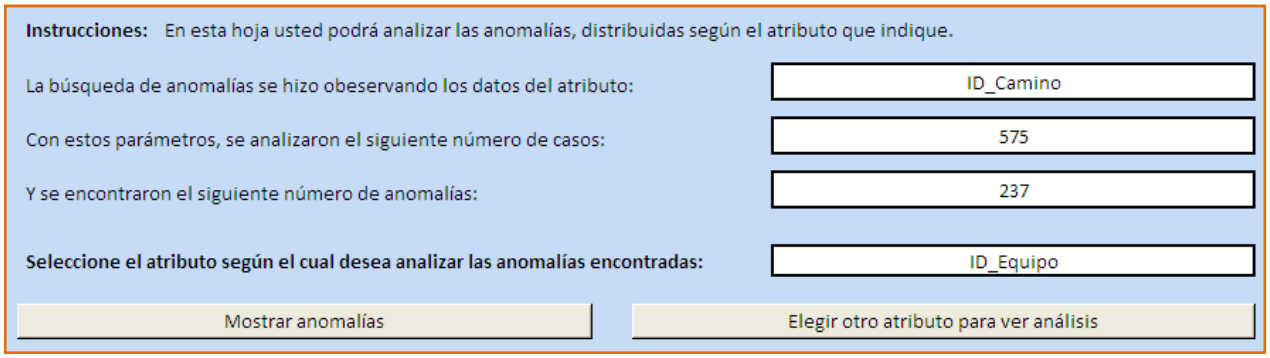


Ilustración 8.8 - Extracto de la hoja "05. Metricas B" de la planilla "Busca Anomalías", donde se aprecia la información que entrega sobre la búsqueda de anomalías realizada y las opciones para escoger un atributo para ser analizado

En el **caso A**, donde se encontraron 89 anomalías, se aplicó el análisis de frecuencias sobre los atributos "ID_Equipo" y "ID_Camino". La información obtenida permitió entender qué características comunes presentaban los casos anómalos encontrados. Primero, se analizó el atributo "ID_Equipo". Como se puede apreciar en la Tabla 8.15 sólo tres de los seis equipos participaron en los casos anómalos, y más importante aún, es que en el 100% de los casos que aparecieron estos equipos, el caso fue catalogado como anómalo. Estos tres equipos (EQ01, EQ04 y EQ05) tienen en común ser los equipos que presentaron las frecuencias más bajas de participación a través de las 590 ejecuciones analizadas.

Al analizar el atributo "ID_Camino" para el **caso A**, se logró descubrir más características de los casos anómalos. Al observar la Tabla 8.16 se puede apreciar que los casos anómalos se concentran en los caminos C01 y C02, mientras que el camino C03 con una frecuencia

mucho mayor no presenta casos anómalos. Este resultado, junto con el análisis realizado sobre el atributo "ID_Equipo", lleva a una conclusión final. Ésta es que las anomalías del **caso A** corresponden a equipos que presentaron menor frecuencia de participación, y que además siguieron los caminos menos recurrentes para realizar el proceso.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	18	18	100%	20%
EQ02	162	0	0%	0%
EQ03	179	0	0%	0%
EQ04	31	31	100%	35%
EQ05	40	40	100%	45%
EQ06	160	0	0%	0%

Tabla 8.15 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos A, del caso "Evaluación solicitud de crédito hipotecario"

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	60	60	100%	67%
C02	29	29	100%	33%
C03	501	0	0%	0%

Tabla 8.16 - Resultados del análisis de frecuencias del atributo "ID_Camino", para el conjunto de datos A, del caso "Evaluación solicitud de crédito hipotecario"

En el **caso B** se obtuvieron dos búsquedas de anomalías con resultados importantes. Al buscar anomalías observando el atributo "ID_Camino" se encontraron 94 anomalías, mientras que al usar "ID_Equipo" se encontraron las mismas anomalías más 17 nuevos casos anómalos, totalizando 111 casos.

Al realizar el análisis de frecuencias del atributo “ID_Camino” en las 94 anomalías encontradas con este mismo atributo, se encuentran los resultados mostrados en la Tabla 8.17. En esta tabla se aprecia que los casos anómalos corresponden a aquellos caminos que presentaron la menor frecuencia de aparición, los cuales fueron C02 y C03. Al sumar las frecuencias de estos dos caminos el resultado es que apenas cubren un 19% de los 580 casos analizados. Este resultado se complementó realizando el mismo análisis de frecuencias sobre el atributo “ID_Equipo”. Este análisis, cuyos resultados se pueden apreciar en la Tabla 8.18, mostró como característica común la baja frecuencia de los equipos que cayeron en la categoría de anómalos. Esta característica se repite en los equipos EQ02, EQ04, EQ06 y EQ08, los cuales juntos no alcanzan a abarcar más del 16% de los 580 casos del **conjunto de datos B**. También hay que destacar que el equipo EQ03, aún presentando una baja tasa de participación (con 17 apariciones como se destaca en la Tabla 8.18), no tuvo casos anómalos asociados. Esto ocurrió debido a que este equipo ejecutó el proceso siguiendo el orden de tareas descrito por el camino C01, el cual fue el más utilizado con 486 casos. Como la búsqueda de anomalías se basó en el atributo “ID_Camino”, el equipo EQ03 se mantuvo fuera de la categoría de anomalía por seguir el camino considerado como normal.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	486	0	0%	0%
C02	53	53	100%	56%
C03	41	41	100%	44%

Tabla 8.17 - Resultados del análisis de frecuencias del atributo “ID_Camino”, para el conjunto de datos B, del caso “Evaluación solicitud de crédito hipotecario”. Este análisis se realizó en base a las 94 anomalías encontradas utilizando el atributo “ID_Camino” como referencia

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	153	0	0%	0%
EQ02	17	17	100%	18%
EQ03	17	0	0%	0%
EQ04	3	3	100%	3%
EQ05	160	0	0%	0%
EQ06	40	40	100%	43%
EQ07	156	0	0%	0%
EQ08	34	34	100%	36%

Tabla 8.18 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos B, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 94 anomalías encontradas utilizando el atributo "ID_Camino" como referencia

Con los dos análisis recién descritos, se logró explicar la existencia de las 94 anomalías resultantes de la primera búsqueda sobre el **caso B**. En esta primera búsqueda faltaron 17 casos anómalos que aparecieron al buscar las anomalías usando el atributo "ID_Equipo" como referencia. Por esta razón, también se realizó el análisis de frecuencias sobre los resultados mostrados por la búsqueda de anomalías usando el atributo "ID_Equipo".

Al buscar la distribución de frecuencias del atributo "ID_Camino", en las 111 anomalías encontradas, destaca la aparición del camino C01, con 17 casos anómalos como se puede apreciar en la Tabla 8.19. Estos 17 casos habían quedado afuera en la primera búsqueda de anomalías, pero ahora aparecen al estar tomando los equipos como referencia para buscar las anomalías. De esta manera se entiende, y como se confirmará a continuación, que las anomalías del camino C01 no se deben a la frecuencia del camino mismo, sino a la frecuencia del equipo que siguió esta secuencia de actividades sólo en 17 oportunidades. Esto se confirma al observar los resultados de analizar las frecuencias de los

valores que toma el atributo “ID_Equipo”. Estos resultados se muestran en la Tabla 8.20, donde destaca el EQ03 con 17 casos anómalos que no habían aparecido en la anterior búsqueda de anomalías. Este equipo es el causante de que el camino de actividades más usado (C01) también presente casos anómalos, dado que es un equipo con una muy baja tasa de participación respecto a las 580 ejecuciones del proceso.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	486	17	3%	15%
C02	53	53	100%	48%
C03	41	41	100%	37%

Tabla 8.19 - Resultados del análisis de frecuencias del atributo “ID_Camino”, para el conjunto de datos B, del caso “Evaluación solicitud de crédito hipotecario”. Este análisis se realizó en base a las 111 anomalías encontradas utilizando el atributo “ID_Equipo” como referencia

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	153	0	0%	0%
EQ02	17	17	100%	15%
EQ03	17	17	100%	15%
EQ04	3	3	100%	3%
EQ05	160	0	0%	0%
EQ06	40	40	100%	36%
EQ07	156	0	0%	0%
EQ08	34	34	100%	31%

Tabla 8.20 - Resultados del análisis de frecuencias del atributo “ID_Equipo”, para el conjunto de datos B, del caso “Evaluación solicitud de crédito hipotecario”. Este análisis se realizó en base a las 111 anomalías encontradas utilizando el atributo “ID_Equipo” como referencia

Luego de los cuatro análisis de frecuencia realizados sobre el **caso B**, se puede concluir que las 111 anomalías encontradas corresponden a los equipos con más baja tasa de participación. Éstos son: EQ02, EQ03,

EQ04, EQ06 y EQ08. Estos 111 casos anómalos no son necesariamente asociados a los caminos de actividades menos usados, ya que hay un equipo (EQ03) que usó el camino de mayor frecuencia del proceso (C01). La mayoría de los casos anómalos corresponden a caminos con baja frecuencia (94 de los 111), pero EQ03 con sus 17 apariciones se convirtió en la excepción a esa tendencia.

El **caso C** necesitó de dos análisis para poder encontrar todas las anomalías definidas por el diseñador del caso. Usando el atributo "ID_Camino" como referencia para la búsqueda de anomalías, se encontraron 232 de las 260 anomalías esperadas. Las 28 anomalías faltantes se obtuvieron al realizar la búsqueda de anomalías tomando como referencia los equipos de ejecutores que participaron en el proceso. Este segundo análisis arrojó 108 anomalías, de las cuales 80 también aparecieron en la primera búsqueda, quedando precisamente los 28 casos que faltaba encontrar. Para complementar estos resultados se procedió a realizar el análisis de frecuencias sobre ambas búsquedas.

La Tabla 8.21 muestra el resultado de buscar las frecuencias de cada uno de los caminos de actividades, que se usaron en los 740 casos del proceso, mostrando también esta información, particularmente, para los 232 casos anómalos. Estos casos anómalos fueron encontrados observando el atributo "ID_Camino" como referencia. En la tabla mencionada se puede apreciar que todos los casos en que se siguió el camino C02 o C03 fueron clasificados como anómalos debido a su baja tasa de aparición en comparación con el camino C01. Además de estos resultados, la tabla también entrega información que justifica el haber utilizado un multiplicador negativo para la búsqueda de anomalías. Esto se debió a que, como se aprecia en la Tabla 8.21, el camino C01 tiene concentrados 508 de los 740 casos que componen el **conjunto de datos**

C. Pero, aún con esta significativa presencia, no alcanza a dejar fuera del segundo cuartil a los caminos C02 y C03. Esto se debe a que los 508 casos que recorrieron el camino C01, representan cerca del 69% del total de ejecuciones, y para haber dejado fuera del límite del primer umbral a los demás caminos, debía tener al menos un 75% de los casos. Es tal la situación, que basta con aumentar muy levemente el multiplicador utilizado, para conseguir un cambio drástico en la cantidad de anomalías encontradas. Por ejemplo, si se escogiera un multiplicador igual a cero, el camino C03 con sus 117 apariciones quedaría dentro del rango de normalidad, quedando sólo 115 anomalías, correspondientes al camino C02. Mientras que, si se tomara un multiplicador igual a 0,1, entraría en el rango de normalidad cualquier camino con una frecuencia igual o superior a 78, con lo cual no habrían aparecido casos anómalos. Finalmente con el multiplicador igual a -0,1 se consigue dejar fuera a los dos caminos que se encontraban en el borde del primer cuartil, pero que dada su gran diferencia respecto a la frecuencia de C01, debían ser considerados como anómalos.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	508	0	0%	0%
C02	115	115	100%	50%
C03	117	117	100%	50%

Tabla 8.21 - Resultados del análisis de frecuencias del atributo "ID_Camino", para el conjunto de datos C, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 232 anomalías encontradas utilizando el atributo "ID_Camino" como referencia

Luego de estudiar en profundidad la distribución de valores del atributo "ID_Camino", se realizó un análisis sobre la frecuencia de cada uno de los valores del atributo "ID_Equipo". La Tabla 8.22 muestra los

resultados mostrados, por la planilla "Busca Anomalías", al solicitar la frecuencia de aparición de cada uno de los equipos que participaron en las ejecuciones del proceso. En esta tabla se han destacado dos de los ocho equipos, por las características particulares que presentan. El primero, el equipo EQ05, es uno de los tres grupos de ejecutores con más frecuencia a través de las 740 ejecuciones. Aún así, todas las instancias en que participó el equipo EQ05 se consideraron como casos anómalos, debido a que este grupo de ejecutores nunca siguió el camino de actividades C01, definido como el normal. En cambio, el segundo equipo destacado: EQ06, resalta por no mostrar casos anómalos, siendo que presenta una de las tres frecuencias más bajas de aparición. Esta situación se explica porque el equipo EQ06 realizó todas sus ejecuciones según las actividades mostradas por el camino C01, considerado como el trazado de actividades normal para el proceso. En definitiva, estos 28 casos de EQ06 serán parte de los casos anómalos, como se esperaba. Esto ocurrirá solamente cuando, cómo se verá a continuación, se haga la búsqueda de anomalías tomando el atributo "ID_Equipo" como referencia.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	180	0	0%	0%
EQ02	20	20	100%	9%
EQ03	119	0	0%	0%
EQ04	21	21	100%	9%
EQ05	152	152	100%	66%
EQ06	28	0	0%	0%
EQ07	39	39	100%	17%
EQ08	181	0	0%	0%

Tabla 8.22 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos C, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 232 anomalías encontradas utilizando el atributo "ID_Camino" como referencia

En definitiva, las 232 anomalías encontradas, tomando el atributo "ID_Camino" como referencia con un multiplicador igual a -0,1, representan todos aquellos casos donde no se siguió el camino C01, el cual presenta una alta diferencia de frecuencia respecto a C02 y C03. Estas anomalías no corresponden necesariamente a los equipos con menor frecuencia de participación, cómo se aprecia en la situación del equipo EQ05 en la Tabla 8.22.

Tal cual como se analizó la frecuencia de valores, sobre los resultados mostrados por la búsqueda de anomalías sobre el **grupo de datos C**, usando como referencia el atributo "ID_Camino", también se realizaron los mismos análisis sobre los 108 casos anómalos, encontrados al hacer la búsqueda observando el atributo "ID_Equipo". El primer análisis, como se aprecia en la Tabla 8.23, se realizó para verificar qué equipos aparecían más dentro de los 108 casos anómalos. De los cuatro equipos que presentan anomalías, se destaca EQ06, con las 28 anomalías que no habían aparecido en la primera búsqueda de anomalías de este caso, y con lo cual se pudo completar las 260 anomalías esperadas.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	180	0	0%	0%
EQ02	20	20	100%	19%
EQ03	119	0	0%	0%
EQ04	21	21	100%	19%
EQ05	152	0	0%	0%
EQ06	28	28	100%	26%
EQ07	39	39	100%	36%
EQ08	181	0	0%	0%

Tabla 8.23 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos C, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 108 anomalías encontradas utilizando el atributo "ID_Equipo" como referencia

El segundo análisis que se realizó en base a las 108 anomalías del **grupo de datos C**, consistió en verificar la cantidad de apariciones de cada camino de actividades. Como se aprecia en la Tabla 8.24, la mayoría de las anomalías (80 de 108) corresponden a los caminos menos utilizados, con excepción del resaltado caso del camino C01. Los 28 casos anómalos que muestra este camino, corresponden al equipo EQ06, que siguió el trazado de actividades más utilizado del proceso, pero dada su baja participación a través de los 740 casos, también fue considerado dentro del grupo de datos anómalos.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	508	28	6%	26%
C02	115	45	39%	42%
C03	117	35	30%	32%

Tabla 8.24 - Resultados del análisis de frecuencias del atributo "ID_Camino," para el conjunto de datos C, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 108 anomalías encontradas utilizando el atributo "ID_Equipo" como referencia

Con los análisis recién planteados, se puede concluir que las 108 anomalías encontradas, corresponden a los equipos que mostraron una menor tasa de participación dentro del proceso. De este total, 80 anomalías coinciden con la primera búsqueda de anomalías, al seguir los caminos con menor frecuencia. En tanto que los 28 casos restantes, presentan a un equipo que siguió el camino más frecuente, pero que tuvo una de las menores tasas de participación del proceso, logrando así completar las 260 anomalías del caso. En definitiva, estas 260 anomalías que presentó el **conjunto de datos C**, corresponden principalmente a casos de equipos que siguieron los caminos de actividades menos utilizados a través de las 740 ejecuciones del proceso, sin que

necesariamente estos equipos hayan presentado una baja frecuencia (ej. EQ05). Estos casos corresponden a las 232 anomalías encontradas usando el atributo "ID_Camino" como referencia. Y como se observó en los últimos párrafos de este capítulo, estas 232 anomalías se complementaron con un solo grupo de casos donde se siguió el camino C01, el cual fue el más utilizado del proceso. Este grupo de casos corresponden a 28 ejecuciones del equipo EQ06, el único de los cuatro equipos, con menor tasa de participación, que siguió el camino C01.

Al igual que lo ocurrido en la búsqueda de anomalías en el caso C, en el **conjunto de datos D** fue necesario realizar dos análisis para encontrar el número de anomalías esperado. En primera instancia, se buscó qué casos presentaban un comportamiento significativamente distinto observando la frecuencia del atributo "ID_Camino". Como se comentó en la sección 8.2.2, esta búsqueda arrojó 237 casos anómalos. Este resultado se complementó con las anomalías encontradas al observar la frecuencia de cada uno de los equipos que participaron en el proceso. Esto se realizó usando el atributo "ID_Equipo" como referencia, con lo cual se encontraron 105 anomalías, de las cuales precisamente sólo 46 eran nuevas, ya que las demás 59 ya habían sido parte del resultado de la primera búsqueda usando "ID_Camino" como referencia. De esta manera, con las nuevas 46 anomalías y las 237 anteriormente encontradas, se obtuvieron las 283 anomalías esperadas por el diseñador del caso.

Para comprender con mayor profundidad las características del **conjunto de datos D** y sus 283 anomalías, se procedió a analizar la distribución de frecuencias que mostraron sus distintos atributos en las dos búsquedas realizadas. En el primer análisis se observó la frecuencia de cada valor de "ID_Camino" e "ID_Equipo", tanto en el total de ejecuciones, como en las 237 ejecuciones anómalas de la primera

búsqueda. Al realizar este análisis sobre el atributo "ID_Camino", cuyos resultados se exponen en la Tabla 8.25, se encontró que las 237 anomalías estaban concentradas en los caminos C02 y C03, debido a la baja tasa de participación que presentaban respecto al camino C01. Para encontrar estas 237 anomalías se debió utilizar un multiplicador igual a -0,1. Esto fue necesario dado que, aunque el camino C01 muestra una tasa de participación mucho más alta que los demás caminos, las 338 veces que se utilizó este trazado no alcanzan a abarcar el 75% de los casos. El camino C01 representa cerca del 59% de los casos, quedando sobre el límite superior del primer cuartil algunos casos del camino C03. Esto dejaba al camino C03 dentro del rango de normalidad, definido por los límites del primer y tercer cuartil. Pero, como la diferencia de frecuencia es muy alta entre C03 y C01, se optó por usar un multiplicador negativo que redujera este rango. De esta manera, las 237 anomalías encontradas describen a aquellos caminos con menor porcentaje de participación en el proceso.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	338	0	0%	0%
C02	108	108	100%	46%
C03	129	129	100%	54%

Tabla 8.25 - Resultados del análisis de frecuencias del atributo "ID_Camino", para el conjunto de datos D, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 237 anomalías encontradas utilizando el atributo "ID_Camino" como referencia

Ya conocidas las frecuencias de los distintos caminos, también interesaba conocer qué cantidad de apariciones tuvo cada uno de los equipos, dentro de todas las ejecuciones y, particularmente, en los 237 casos anómalos encontrados en la primera búsqueda. A través de los

resultados de este análisis, detallados en la Tabla 8.26, destacan tres equipos por la particularidad de sus casos. El primer equipo que resalta es EQ01 que, aún presentando la segunda tasa de participación más baja, no tiene asociación a casos anómalos. Esto es producto de que este equipo ejecutó el proceso siempre siguiendo el camino C01, considerado como normal. Misma situación ocurre con otro equipo que presentó muy pocas apariciones. Este es EQ06 que, con sólo 32 apariciones, no tiene casos anómalos asociados, por haber ejecutado las actividades también según el camino C01. Por último, el tercer equipo que llama la atención es EQ05, el cual presenta 178 apariciones, y todas han sido consideradas como casos anómalos del proceso. Este equipo concentró un 75% de las anomalías encontradas en la primera búsqueda, dado que nunca siguió el camino más utilizado por el resto de los equipos que ejecutaron el proceso. Esta situación provocó que, sin importar la alta participación del equipo, este igual cayera en el grupo anómalo. Con este nuevo conocimiento, se puede concluir que las 237 anomalías, corresponden a los caminos menos utilizados, sin que necesariamente sean también los equipos menos participativos, como es el caso de EQ05.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	14	0	0%	0%
EQ02	1	1	100%	0%
EQ03	136	0	0%	0%
EQ04	24	24	100%	10%
EQ05	178	178	100%	75%
EQ06	32	0	0%	0%
EQ07	156	0	0%	0%
EQ08	34	34	100%	14%

Tabla 8.26 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos D, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 237 anomalías encontradas utilizando el atributo "ID_Camino" como referencia

Definidas las características de las primeras 237 anomalías, el último paso fue analizar los mismos atributos, "ID_Camino" e "ID_Equipo", pero ahora observando la frecuencia de cada valor, en las 105 anomalías encontradas en la segunda búsqueda, donde se usó la frecuencia de los equipos para definir los casos anómalos. El primer análisis de frecuencia que se realizó sobre las 105 anomalías, se ejecutó sobre el atributo "ID_Camino". Los resultados de este análisis, los cuales se pueden apreciar en la Tabla 8.27, muestran las nuevas anomalías encontradas con el segundo análisis (destacadas en amarillo). Las 46 nuevas anomalías corresponden a casos donde se siguió el camino más recurrente (C01) pero que, como se verá a continuación, fueron clasificados como anómalos por corresponder a aquellos equipos que tuvieron una tasa de participación más baja que el resto. Las otras 59 anomalías asociadas a los caminos C02 y C03 no demandan mayor análisis ya que forman parte de los casos anómalos encontrados en la primera búsqueda.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
C01	338	46	14%	44%
C02	108	28	26%	27%
C03	129	31	24%	30%

Tabla 8.27 - Resultados del análisis de frecuencias del atributo "ID_Camino", para el conjunto de datos D, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 105 anomalías encontradas utilizando el atributo "ID_Equipo" como referencia

Finalmente, se observó la distribución de frecuencias del atributo "ID_Equipo", a través de las 105 anomalías. Al observar los resultados que se detallan en la Tabla 8.28, se explica el porqué de las nuevas 46 anomalías. Estos nuevos casos anómalos, destacados en la tabla

previamente mencionada, corresponden a los equipos EQ01 y EQ06, los cuales siguieron el camino C01 (el más recurrente del proceso) pero que, al presentar dos de las tasas más bajas de participación, entran al grupo anómalo.

Valor	Frecuencia del valor	Anomalías encontradas con este valor	Peso anomalías v/s frecuencia del valor	Peso anomalías v/s total anomalías encontradas
EQ01	14	14	100%	13%
EQ02	1	1	100%	1%
EQ03	136	0	0%	0%
EQ04	24	24	100%	23%
EQ05	178	0	0%	0%
EQ06	32	32	100%	30%
EQ07	156	0	0%	0%
EQ08	34	34	100%	32%

Tabla 8.28 - Resultados del análisis de frecuencias del atributo "ID_Equipo", para el conjunto de datos D, del caso "Evaluación solicitud de crédito hipotecario". Este análisis se realizó en base a las 105 anomalías encontradas utilizando el atributo "ID_Equipo" como referencia

Con las observaciones realizadas sobre las últimas dos tablas, el investigador ya puede entender las características comunes, y más importantes, de las 105 anomalías. De esta manera se puede concluir que, al buscar los casos anómalos tomando el atributo "ID_Equipo" como referencia, los casos considerados fuera de lo normal, resultaron ser aquellos en que los equipos presentaron bajas tasas de participación. En particular, los 46 nuevos casos anómalos cumplen con esta característica y además coinciden en haber usado el camino C01 de tareas, lo cual evitó que aparecieran en los casos anómalos de la primera búsqueda.

Con esto concluye el análisis de las anomalías encontradas a través de los cuatro conjuntos de datos desarrollados para probar la efectividad del algoritmo Interquartile Range. A través de lo visto en esta sección se ha podido complementar los resultados obtenidos con el

algoritmo, logrando describir de manera más acaba las características de los casos anómalos, llegando a definiciones que permitirían tomar acciones respecto a ciertos equipos o caminos de tareas que están entregando estadísticas fuera de lo normal.

8.2.4 Análisis de sensibilidad del algoritmo Interquartile Range

Al analizar las búsquedas de anomalías descritas en la sección 8.2.2, se puede concluir que son tres factores los que más influyen sobre el número de casos anómalos encontrados. Estos factores son: los datos que presente el proceso, el atributo que se seleccione para buscar las anomalías, y el multiplicador que se ingrese como parámetro para la ejecución del algoritmo Interquartile Range. Este último factor es el que se analizará en este capítulo para ver cómo varía el número de anomalías encontradas, en la medida que se aumente o disminuya el valor de este parámetro del algoritmo. Para analizar estas variaciones, se revisará el número de anomalías encontradas en los umbrales superior e inferior por separado, para así no mezclar estos casos y ver mejor como se mueve el número de anomalías en cada extremo de los casos analizados.

Para realizar el análisis de sensibilidad sobre el algoritmo Interquartile Range, se tomaron dos de las búsquedas que se mostraron en la sección 8.2.2. Primero, se tomó la búsqueda de anomalías que se realizó sobre el caso A. Los resultados de esta búsqueda, entregaron 89 casos anómalos. Esta cantidad de anomalías fue encontrada usando el atributo “ID_Equipo” como referencia y un multiplicador igual a 1,0. A partir de ese resultado se fue aumentando y disminuyendo el valor del multiplicador, para llegar a los resultados que se aprecian en la Tabla 8.29.

	Umbral en que se buscaron casos anómalos	
Valor multiplicador	Inferior	Superior
-8,5	590	590
...	...	...
-8,0	590	572
...	...	...
-7,5	590	541
-7,4	590	541
...	...	...
-1,1	590	501
-1,0	411	341
-0,9	411	341
...	...	...
-0,2	411	179
-0,1	249	179
...	...	...
1,0	89	0
...	...	...
6,0	89	0
6,4	49	0
...	...	...
6,8	18	0
...	...	...
7,5	0	0

Tabla 8.29 - Tabla que indica la cantidad de anomalías encontradas en el caso A del proceso "Evaluación solicitud crédito hipotecario", usando distintos valores para el multiplicador que modifica el rango de normalidad

El primer punto que se puede destacar de la Tabla 8.29 es que refleja claramente la función y el efecto del multiplicador sobre los resultados del algoritmo. Como se mencionó en diversos puntos del trabajo, el multiplicador permite reducir o ampliar el rango de normalidad

definido por el algoritmo Interquartile Range. Como se aprecia en la Tabla 8.29, a medida que el multiplicador va tomando valores más negativos, el rango de normalidad se va reduciendo, y en consecuencia, menos casos son considerados como normales y el número de anomalías comienza a crecer. Lo opuesto ocurre al ir aumentando el valor del multiplicador. A medida que este valor se aleja del 1,0, utilizado inicialmente, la cantidad de anomalías va disminuyendo. Esto ocurre porque el rango de normalidad original se va expandiendo, hasta llegar a considerar a todos los casos como normales, como ocurre al utilizar un multiplicador igual a 7,5.

La Tabla 8.29 también muestra que cambios pequeños en el multiplicador original no causan cambios sobre la cantidad de anomalías encontradas. El ejemplo más claro es que se necesitó llegar un multiplicador igual a 6,4, para lograr ampliar el rango de normalidad a un punto que permitiera disminuir las 89 anomalías encontradas. Esto confirma lo lejos que se encontraban los casos anómalos del comportamiento normal del proceso, justificando así su clasificación. Lo señalado se puede confirmar al observar los cálculos del algoritmo Interquartile Range. De los 590 casos que componen el grupo de datos disponibles para el caso A, la mediana se ubicó en el equipo con una frecuencia igual a 162. Mientras que el límite del primer cuartil se ubicó en 160 y el del tercero en 179. Con estos resultados, hay que alejarse 120 puntos del umbral inferior, para que el rango de normalidad incluya al siguiente equipo que muestra la frecuencia más alta bajo 160, la cual corresponde apenas a 40.

Continuando el análisis sobre la Tabla 8.29, se aprecia que, al ocupar valores negativos en el multiplicador, el número de anomalías encontradas cambia más rápidamente que al probar números positivos.

Como la mayoría de los casos están concentrados en un rango muy pequeño (sólo 19 puntos de diferencia entre los umbrales inferior y superior), reducciones pequeñas del rango de normalidad, pueden dejar a muchos casos fuera de lo considerado como comportamiento normal. Esto se aprecia, en la Tabla 8.29, al usar un multiplicador igual a -0,1, donde el número de anomalías cambia bruscamente de 89 a 428 (249 bajo el umbral inferior y 179 sobre el umbral superior). Lo que ocurrió es que al reducir el rango de normalidad, se excluyeron de éste los casos que se encontraban en los extremos, quedando como normal únicamente aquel equipo que se presentó en 162 oportunidades, que corresponde justamente a la mediana del proceso. Los valores negativos más grandes que se aprecian en la tabla, entregaron más casos anómalos. Pero estos casos no merecen mayor análisis, ya que al usar valores negativos de mayor o igual magnitud a -0,9, el umbral inferior quedó en 178, lo cual es superior a la mediana, mientras que el umbral inferior quedó en 161, bajo la mediana. Esto significa que el rango de normalidad fue totalmente eliminado, consiguiendo que ningún caso califique como normal.

La segunda búsqueda que se escogió para realizar el análisis de sensibilidad del algoritmo Interquartile Range, fue una de las que se aplicaron sobre el conjunto de datos C. La búsqueda escogida fue aquella que se ejecutó con el atributo “ID_Camino” como referencia, junto a un multiplicador igual a -0,1. Con esta configuración se buscaron las anomalías que se encontraran solamente bajo el umbral inferior, obteniendo 232 de las 260 anomalías del caso. En la Tabla 8.30 se puede observar cómo estas 232 anomalías aumentan o disminuyen bruscamente, de acuerdo al valor que se ingrese como multiplicador.

La Tabla 8.30 es mucho más breve que la analizada anteriormente en esta sección, lo que demuestra que, además del multiplicador, los

propios datos de cada caso van a afectar fuertemente en el número de anomalías encontradas y, en la sensibilidad frente a cambios en el multiplicador. Otro punto a destacar a partir de esta tabla, es que los 232 casos anómalos encontrados se encuentran justo en el límite inferior del rango de normalidad. Esto se demuestra al ver cómo se modifica bruscamente el número de anomalías al cambiar el multiplicador de -0,1 a 0 y luego a 0,1. Con sólo mover 0,2 puntos el multiplicador, las anomalías llegan a cero. Esto indica que basta con ampliar mínimamente el rango de normalidad, para que el algoritmo indique que no hay casos anómalos. Al analizar los cálculos del algoritmo Interquartile Range se puede entender el porqué de estas importantes variaciones, al cambiar de manera mínima el multiplicador. Al revisar las 740 ejecuciones que componen el grupo de datos del caso C, el algoritmo encontró la mediana en el camino que presentaba una frecuencia igual a 508, con el umbral superior en el mismo valor y el inferior en 117. Dejando el rango de normalidad por defecto (multiplicador igual a cero), se encontraron sólo 115 anomalías como se aprecia en la Tabla 8.30, que es justamente el camino que presentaba esta cantidad de apariciones. Pero, con tan sólo reducir mínimamente el rango de normalidad (con un multiplicador igual a -0.1), fue posible excluir el otro camino que se encontraba en el borde del umbral inferior, con 117 apariciones. De esta manera, se llega a los 232 casos que difieren de manera importante respecto al tercer camino que se ejecutó 508 veces.

El análisis de sensibilidad recién expuesto, junto con el realizado sobre la búsqueda de anomalías en el caso A, demuestran la importancia de analizar los cálculos realizados por el algoritmo Interquartile Range, junto a diversos valores para el multiplicador usado como parámetro. Este análisis, como fue descrito aquí, permite comprender por qué ciertos

casos quedan fuera o dentro del rango de normalidad, que tan lejos se encuentran los casos anómalos de este rango, y la amplitud que tiene éste.

Valor multiplicador	Umbral en que se buscaron casos anómalos	
	Inferior	Superior
-1,1	740	740
...	...	...
-0,1	232	508
0	115	0
0,1	0	0

Tabla 8.30 - Tabla que indica la cantidad de anomalías encontradas en el caso C del proceso "Evaluación solicitud crédito hipotecario", usando distintos valores para el multiplicador que modifica el rango de normalidad

9. Conclusiones

La hipótesis planteada en este trabajo estipulaba que en esta investigación, a partir de Logs de Eventos, se podría encontrar patrones secuenciales y anomalías en procesos, utilizando algoritmos de Minería de Datos. Esto se consiguió, llegando incluso más allá de la búsqueda de patrones y la detección de anomalías. El resultado final expuesto en este trabajo, entrega metodologías completas para cumplir con éxito con los objetivos planteados. Junto a esto se desarrollaron un grupo de herramientas que permiten aplicar estas metodologías de manera ágil, sin importar el caso de negocio que se esté analizando.

El aporte más importante que entrega este trabajo al área de Minería de Procesos es el diseño y aplicación de dos metodologías de análisis de procesos: una para buscar

patrones secuenciales y otra para la detección de anomalías. Una de las características más relevantes de las técnicas descritas para enfrentar estas tareas, es que se basan en el modelo KDD. Este fue diseñado para problemas de Minería de Datos, y siguiendo sus pasos se logró abarcar no tan sólo la aplicación de los algoritmos sobre datos de procesos, sino que también se diseñaron pasos para el pre-procesamiento de los datos y el análisis post-búsqueda de anomalías y patrones. Al incluir estos pasos en las metodologías propuestas, se busca enfatizar la importancia de la interacción hombre-máquina que debe haber para obtener los mejores resultados de la investigación. Cada uno de los pasos diseñados siguiendo el modelo KDD mostró su relevancia al ir obteniendo los resultados de la investigación.

La selección y pre-procesamiento de los datos permitieron preparar los datos para su correcta lectura y máximo aprovechamiento, logrando, entre diversos beneficios:

- Consolidar las distintas tareas que componen la ejecución de un proceso, de manera que cada registro a ser analizado contara con toda la información disponible de cada ejecución del proceso. Sin esta consolidación, no hubiese sido posible realizar un análisis a nivel de proceso como un todo, sino que se habría tenido que analizar la información a nivel de tareas individuales.

- Limpiar los datos eliminando aquellos atributos que no serían de utilidad por su alta variabilidad.

- Analizar los atributos del proceso desde distintas perspectivas, para poder contar con la mayor cantidad de información útil para cada análisis. Esto fue evidente en la sección 8.1.1 de este trabajo, donde no hubiera sido posible evaluar el impacto de los ejecutores sobre el proceso, si no se hubiese contado con una visión más global de los equipos participantes en el caso “Venta de artículos a través de página Web”. Al consolidar los ejecutores según el orden exacto en que participaron, se llegaba a un número de equipos distintos superior a la mitad de los casos analizados. Con esto, los ejecutores quedaban

fuera del análisis, pero una segunda consolidación del pre-procesamiento entregó una visión más simple de los equipos, donde se omiten las repeticiones y el orden de los participantes, agrupación con la que sí se pudo tomar a los ejecutores como parte del análisis.

La aplicación de los algoritmos de Minería de Datos entregaron resultados muy positivos en su desempeño:

- En la búsqueda de patrones secuenciales usando el algoritmo Apriori, se logró describir los mil casos analizados con un 100% de confianza. Las reglas que permitieron llegar a este nivel de confianza describieron patrones importantes del proceso, alineados con las expectativas del diseñador del caso "Venta de artículos a través de página Web", y que además entregan características del proceso para poder tomar decisiones y generar acciones, como se planteó en la sección 8.1.4 de este trabajo.

- Al poner en práctica la detección de anomalías, en los cuatro casos diseñados para probar el algoritmo Interquartile Range, se logró en la búsqueda más relevante de cada caso un porcentaje promedio de éxito cercano al 93%. El porcentaje se calculó comparando las anomalías detectadas contra las esperadas por el diseñador del caso "Evaluación solicitud de crédito hipotecario". Estos resultados expuestos en la sección 8.2.2 también mostraron que, en aquellos casos donde no se logró un 100% de efectividad, se logró detectar el total de anomalías al complementar con una segunda búsqueda, usando el algoritmo Interquartile Range y otro atributo como referencia.

El análisis post-búsqueda de patrones secuenciales y post-detección de anomalías, permitió dar mayor valor a los resultados obtenidos y extraer aún más conocimiento de los procesos:

- Terminada la búsqueda de patrones secuenciales, se realizó un análisis de frecuencia de los distintos valores de cada atributo, poniendo especial énfasis en la distribución de valores mostrada sobre aquellos casos que formaron parte de los patrones. Este análisis permitió comprobar, como se detalló en la sección 8.1.5, la información obtenida con el algoritmo Apriori, obteniendo también más detalle de por qué ciertos atributos tenían influencia en el resultado final y otros no.

- La detección de anomalías se realiza observando la frecuencia de un sólo atributo del proceso analizado. Por esto, era fundamental poder analizar qué ocurría con los demás atributos en los conjuntos de datos clasificados como anómalos. Esto se logró con el análisis de frecuencia de los distintos valores de cada atributo, especialmente en los casos anómalos, Estos resultados se detallan en la sección 8.2.3, y permitieron llegar a definir qué características en común presentaban las anomalías en cada uno de los cuatro grupos de datos analizados, lo que es finalmente el objetivo más importante del análisis.

Además de desarrollar las metodologías necesarias para los objetivos planteados en este trabajo, se entregan herramientas para facilitar cada una de las tareas definidas para la búsqueda de patrones y la detección de anomalías. El pre-procesamiento del Log de Eventos, la detección de anomalías aplicando el algoritmo Interquartile Range, y el análisis de patrones secuenciales, están apoyados por herramientas que fueron desarrolladas en esta investigación. Estas herramientas fueron implementadas utilizando Macros y fórmulas en planillas Excel, cuidando que éstas fueran desarrolladas de manera genérica para que, independiente del proceso analizado, se pudieran reutilizar. El desarrollo de estas herramientas fue complementado con el apoyo en dos *software*: ProM para poder exportar la información de procesos almacenados en archivos de extensión .mxml a archivos .csv; y Weka, para aplicar el algoritmo Apriori en búsqueda de reglas de asociación que permitieran establecer patrones secuenciales de un proceso.

Como se señaló anteriormente, los resultados del trabajo fueron satisfactorios. Sin embargo, el desarrollo de esta investigación no estuvo exento de complicaciones en su desarrollo. Dentro de las dificultades enfrentadas durante el trabajo, destaca la gran demanda de trabajo que significó la construcción de la metodología y la herramienta para pre-procesar los datos de un proceso. Esta fue la parte más compleja de desarrollar, dado el volumen de información que se debía manejar y la importante cantidad de pasos que se incluyeron en la planilla "Pre-procesador Logs". El resultado fue positivo, pero al no contar con el *output* de esta etapa, no era posible probar distintas alternativas de algoritmos de Minería de Datos. Se espera que para futuras investigaciones el pre-procesador implementado aquí permita evitar esta complicación, llegando rápidamente a la aplicación de algoritmos. Otro problema que se enfrentó en este trabajo fue la dificultad para obtener casos de ejemplo con los cuales probar las metodologías propuestas. Las políticas de seguridad de las empresas consultadas y lo complejo que es generar conjuntos de datos artificiales son obstáculos a la hora de contar con una buena cantidad de casos para trabajar. Se recomienda, para cualquier investigación de este tipo, tener este punto como una prioridad desde el comienzo del proyecto, para que llegada la fase de pruebas, no sea la falta de casos la razón de no avanzar en la investigación.

Tal como en la hipótesis se planteó el desafío de implementar metodologías para la búsqueda de anomalías y patrones en procesos, este trabajo también fijó otro desafío que es poder contribuir al crecimiento de la Minería de Procesos. Para esto se han entregado técnicas y herramientas que podrían ser usadas en diversas áreas de investigación o rubros de negocio. Se espera que empresas e investigadores utilicen lo presentado en este trabajo. Por esto es importante, para usar las metodologías y herramientas desarrolladas en esta investigación, conocer algunas recomendaciones y limitaciones sobre éstas. Primero, es necesario revisar los datos a ser usados en el análisis. Las planillas no cuentan con procesos de validación de los datos, y se encuentran diseñadas asumiendo que siempre recibirán, al menos, los cinco atributos básicos de un proceso. Si el conjunto de datos a ser analizado presenta casos vacíos o no cuenta con los cinco atributos esperados, los resultados del pre-proceso y posteriores

análisis no serán de confianza, e incluso podrían llevar a error en la ejecución de las herramientas implementadas. También hay que tener precaución con la cantidad de datos que se ingresen en las herramientas desarrolladas en Excel. Si bien con diez mil o veinte mil registros funcionan sin problemas, al aumentar y llegar cerca a cerca de los treinta mil datos, las planillas implementadas presentan problemas de ejecución. Por último, respecto a las metodologías planteadas en este trabajo, hay que tener en cuenta que cualquier análisis que se haga sobre un proceso, será realizado observando cada ejecución de manera completa. Dada esta situación, no se puede esperar resultados que involucren segmentos del proceso (fase o subprocesos) o tareas específicas del proceso.

Como se mencionó previamente, uno de los objetivos principales de este trabajo es generar más interés sobre el área de Minería de Procesos, llevando a más investigadores, empresas y otras organizaciones a trabajar en ella. Si este interés se traduce en trabajos que busquen extender las metodologías y herramientas entregadas aquí, se esperaría que en un principio las mejoras apuntaran a eliminar o disminuir las limitaciones presentes hoy en lo expuesto en este trabajo. Dentro de las posibles mejoras, se ve un enorme potencial para el trabajo con herramientas ETL (del inglés *Extract Transform Load*, que significa Extraer, Transformar, Cargar), especialmente como alternativa a la actual planilla "Pre-procesador Logs". Implementar las instrucciones de esta planilla con *software* de ETL, posiblemente apoyándose en Bases de Datos, aumentaría la capacidad de procesamiento de miles a millones de casos por proceso. Además de la ganancia en capacidad de datos analizados, se ganaría rapidez y un proceso más robusto al no estar sujeto a fórmulas ni posiciones de celdas en Excel. La planilla pre-procesadora tiene las características ideales para ser implementada con un *software* de ETL, ya que para cada proceso que se quiera analizar sólo se debe pre-procesar una vez los datos. Idealmente, este proceso debería consistir en entregar el archivo mxml o csv a un programa, y que éste entregue como salida los datos en el formato necesario para poder ser analizados con algoritmos de Minería de Datos. En cuanto a las planillas "Analiza Patrones" y "Busca Anomalías", la situación es distinta. Estas herramientas apuntan a un trabajo de constante interacción con el investigador.

Esta interacción se debe traducir en decenas o centenas de pruebas sobre un mismo grupo de datos, cambiando los parámetros de los algoritmos para ver como varían los resultados. Dada esta necesidad es que se esperaría que estas planillas cambiaran de plataforma a un ambiente generado en un lenguaje de programación como Java. La idea es obtener una herramienta como las que se pueden usar en Weka, o como los cientos de *plug-ins* disponibles en ProM. Al igual que lo propuesto para la planilla pre-procesadora, un cambio como éste permitiría aumentar a millones la capacidad de datos a ser analizados, y la velocidad de respuesta seguramente también se vería favorecida de manera significativa.

A modo de síntesis, la investigación entregó dos metodologías que permiten realizar búsquedas de patrones secuenciales y detectar anomalías en procesos. Estas metodologías describen un camino completo, desde el pre-procesamiento hasta el análisis luego de aplicar un algoritmo. Y dado que su implementación es genérica, se favorece la capacidad de análisis sobre una gran variedad de tipos de procesos.

Se espera que estas metodologías sean aplicadas en diversos rubros, y que se pueda extender lo entregado aquí con futuras investigaciones realizadas en el área de Minería de Procesos.

BIBLIOGRAFIA

de Medeiros A.K.A. (2008). Minería de Procesos: Más allá de los Modelos de Procesos. En *BPM Chile*. Recuperado de: http://sites.google.com/site/bpmchile/entrevista30-06-08

Bezerra F., Wainer J. y van der Aalst W.M.P. (2009). Anomaly Detection using Process Mining. *Lecture Notes in Business Information Processing*, *29*(1), 149-161. doi: 10.1007/978-3-642-01862-6_13

Seifert J.W. (2004). Data Mining An Overview. Recuperado de: http://www.fas.org/irp/crs/RL31798.pdf

Zamarrón C., García V., Calvo U., Pichel F. y Rodríguez J.R. (2006). *Aplicación de la Minería de Datos al estudio de las alteraciones respiratorias durante el sueño.* Recuperado de: http://www.sogapar.info/index.php?id=85&cid=107&fid=32&task=download&option=com_flexicontent&Itemid=71

Lazarevic, A. (2008). *Data Mining for Anomaly Detection*. Recuperado de: http://cs.kangwon.ac.kr/~ysmoon/courses/2011_1/grad_mining/slides/15.pdf

Sumathi S. y Sivanandam S.N. (2006). Introduction to Data Mining and its Applications. *Studies in Computational Intelligence, 29*. Recuperado de: http://www.csbdu.in/pdf/Introduction%20to%20Data%20Mining%20and%20its%20Applications.pdf

Fayyad U., Piatetsky-Shapiro G. y Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine*, *17*(3), 37-54. doi: 10.1.1.42.1071

Tiwari A. y Turner C.J. (2008). A review of business process mining state of the art and future trends. *Business Process Management Journal*., *14*(1), 5-22. doi: 10.1108/14637150810849373

van der Aalst W.M.P., van Dongen B.F., Günther C., Rozinat A., Verbeek H.M.W. y Weijters A.J.M.M. (2009). *ProM The Process Mining Toolkit*. Recuperado de: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-489/paper3.pdf

van der Aalst W.M.P. (2005). *Process Mining in CSCW Systems*. Recuperado de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.7215&rep=rep1&type=pdf

de Medeiros A.K.A. y Günther W. (2005). *Process Mining Using CPN Tools*. Recuperado de: http://www.daimi.au.dk/CPnets/workshop05/cpn/slides/cpn05_cpnlogs_slides.pdf

Chandola V., Banerjee A., y Kumar V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*, *41*(3). doi: 10.1145/1541880.1541882

Patcha A. y Park J. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448-3470. doi: 10.1016/j.comnet.2007.02.001

Saleeb H. (2008). Mining sequence patterns in transactional databases. Recuperado de: cs.nyu.edu/courses/spring08/G22.3033-003/8timeseries.ppt

Han J., Pei J. y Yan X. (2005). Sequential Pattern Mining by Pattern-Growth Principles and Extensions. *StudFuzz*, *180*, 183-220. Recuperado de: http://www.cs.sfu.ca/~jpei/publications/seqpat-05.pdf

Batal I. (2007). Association Rule Mining. Recuperado de: http://www.cs.pitt.edu/~iyad/AR.pdf

Moreira F. (2006). Ten-year Forecast of Storage Evolution. Recuperado de: http://prestospace.org/project/deliverables/D12-5.pdf

ANEXOS

ANEXO 1: EXTRACTOS DE LOGS DE EVENTOS

A modo de ejemplo para revisar cómo se almacena la información de un proceso, aquí se incluyen algunos extractos de un archivo de extensión .mxml.

El siguiente código describe la ejecución id="1" del proceso "Venta de artículos a través de página Web":

```
<ProcessInstance id="1" description="Simulated process instance">
        <AuditTrailEntry>
        <Data><Attribute name = "cantidad">1 </Attribute>
        </Data><WorkflowModelElement>Registrar Pedido</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2009-12-31T21:01:00.000+01:00</Timestamp>
        <Originator>system</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Verificar Disponibilidad</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-01T13:41:00.000+01:00</Timestamp>
        <Originator>Almacen1</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Comprar faltante</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-02T14:41:00.000+01:00</Timestamp>
        <Originator>Compras1</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Armar Pedido</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-04T16:41:00.000+01:00</Timestamp>
        <Originator>Almacen2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <Data><Attribute name = "monto">3000 </Attribute>
        </Data><WorkflowModelElement>Enviar Factura</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-05T17:41:00.000+01:00</Timestamp>
        <Originator>Caja2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Enviar Pedido</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-05T17:41:00.000+01:00</Timestamp>
        <Originator>Almacen2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Enviar Recordatorio</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-06T10:21:00.000+01:00</Timestamp>
```

```
        <Originator>Caja1</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <Data><Attribute name = "OK">OK </Attribute>
        </Data><WorkflowModelElement>Procesar Pago</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-06T18:41:00.000+01:00</Timestamp>
        <Originator>Caja2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Cerrar Venta</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-06T23:41:00.000+01:00</Timestamp>
        <Originator>system</Originator>
        </AuditTrailEntry>
</ProcessInstance>
```

El siguiente código describe la ejecución id="10" del proceso "Venta de artículos a través de página web":

```
<ProcessInstance id="10" description="Simulated process instance">
        <AuditTrailEntry>
        <Data><Attribute name = "cantidad">1 </Attribute>
        </Data><WorkflowModelElement>Registrar Pedido</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2009-12-31T21:01:00.000+01:00</Timestamp>
        <Originator>system</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Verificar Disponibilidad</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-01T13:41:00.000+01:00</Timestamp>
        <Originator>Almacen1</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Comprar faltante</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-02T14:41:00.000+01:00</Timestamp>
        <Originator>Compras2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Comprar faltante</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-04T16:41:00.000+01:00</Timestamp>
        <Originator>Compras2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <WorkflowModelElement>Armar Pedido</WorkflowModelElement>
        <EventType >complete</EventType>
        <Timestamp>2010-01-06T18:41:00.000+01:00</Timestamp>
        <Originator>Almacen2</Originator>
        </AuditTrailEntry>
        <AuditTrailEntry>
        <Data><Attribute name = "monto">3000 </Attribute>
```

```
</Data><WorkflowModelElement>Enviar Factura</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-07T19:41:00.000+01:00</Timestamp>
<Originator>Caja1</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Enviar Pedido</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-07T19:41:00.000+01:00</Timestamp>
<Originator>Almacen2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Enviar Recordatorio</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-08T12:21:00.000+01:00</Timestamp>
<Originator>Caja2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<Data><Attribute name = "OK">OK </Attribute>
</Data><WorkflowModelElement>Procesar Pago</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-08T20:41:00.000+01:00</Timestamp>
<Originator>Caja2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Cerrar Venta</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-09T01:41:00.000+01:00</Timestamp>
<Originator>system</Originator>
</AuditTrailEntry>
</ProcessInstance>
```

El siguiente código describe la ejecución id="100" del proceso "Venta de artículos a través de página web":

```
<ProcessInstance id="100" description="Simulated process instance">
<AuditTrailEntry>
<Data><Attribute name = "cantidad">1 </Attribute>
</Data><WorkflowModelElement>Registrar Pedido</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2009-12-31T21:01:00.000+01:00</Timestamp>
<Originator>system</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Verificar Disponibilidad</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-01T13:41:00.000+01:00</Timestamp>
<Originator>Almacen1</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Comprar faltante</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-02T14:41:00.000+01:00</Timestamp>
<Originator>Compras1</Originator>
</AuditTrailEntry>
```

<AuditTrailEntry>
<WorkflowModelElement>Comprar faltante</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-04T16:41:00.000+01:00</Timestamp>
<Originator>Compras1</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Armar Pedido</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-06T18:41:00.000+01:00</Timestamp>
<Originator>Almacen2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Enviar Pedido</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-07T19:41:00.000+01:00</Timestamp>
<Originator>Almacen2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<Data><Attribute name = "monto">3000 </Attribute>
</Data><WorkflowModelElement>Enviar Factura</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-07T19:41:00.000+01:00</Timestamp>
<Originator>Caja2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Enviar Recordatorio</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-08T12:21:00.000+01:00</Timestamp>
<Originator>Caja2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<Data><Attribute name = "OK">OK </Attribute>
</Data><WorkflowModelElement>Procesar Pago</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-08T20:41:00.000+01:00</Timestamp>
<Originator>Caja2</Originator>
</AuditTrailEntry>
<AuditTrailEntry>
<WorkflowModelElement>Cerrar Venta</WorkflowModelElement>
<EventType >complete</EventType>
<Timestamp>2010-01-09T01:41:00.000+01:00</Timestamp>
<Originator>system</Originator>
</AuditTrailEntry>
</ProcessInstance>

ANEXO 2: PROCEDIMIENTO PARA EXPORTAR UN ARCHIVO .MXML A .CSV

Aquí se entregarán las instrucciones paso a paso para poder llevar un archivo de extensión .mxml a un archivo de estructura .csv.

Las instrucciones aquí compartidas fueron probadas en la versión 5.2 del *software* ProM.

Instrucciones: ¿Cómo exportar un archivo .mxml a un .csv usando ProM?

1. Abrir ProM 5.2, ir a "File" y abrir el archivo deseado usando la opción "Open MXML Log file".
2. Automáticamente se abrirá la ventana de análisis del proceso seleccionado, pero ésta no será necesaria para la tarea que se desea realizar aquí.
3. Ir a la opción "Exports" en el menú principal de ProM, y seleccionar la opción "Raw + nombre archivo (unfiltered)", y del menú desplegable que se abre seleccionar "CSV for log Exporter".
4. Finalmente, sólo queda elegir la ubicación donde se desea guardar el archivo y la tarea estará lista.

ANEXO 3: CONSEJOS PARA TRABAJAR CON WEKA

En este trabajo se utilizó Weka para la búsqueda de patrones secuenciales, haciendo uso del algoritmo Apriori implementado en este *software*. Además, se aprovecharon algunas de las funciones de pre-procesamiento que entrega Weka. Aquí se explicará cómo llegar a estas herramientas y las características más importantes de cada una.

Las instrucciones aquí compartidas fueron probadas en la versión 3.6.2 del programa.

Instrucciones: ¿Cómo llegar a las herramientas de pre-procesamiento en Weka?

1. Abrir Weka 3.6.2 y escoger "Explorer" en las opciones de aplicaciones posibles.
2. Estando en "Explorer", la pantalla de inicio corresponderá automáticamente a la sección de pre-procesamiento, llamada "Preprocess".
3. Estando en "Preprocess" se debe ingresar un archivo para comenzar a ver la información a través de las distintas visualizaciones que presenta Weka. Esto se realiza mediante la opción "Open file…".
4. Una vez abierto un archivo, en la sección "Current relation" se podrá revisar la información básica del archivo, detallando el nombre, cantidad de instancias y número de atributos.
5. La sección "Attributes" mostrará el listado de atributos que presenta el caso. Mientras que al seleccionar cualquiera de los atributos listados, se podrá apreciar en "Selected attribute" todos los valores que toma el atributo, junto a la frecuencia de cada uno y un gráfico resumen.
6. Por último, en la sección "Filter" se podrán seleccionar herramientas para pre-procesar los datos.

Instrucciones: ¿Cómo se puede aplicar el algoritmo Apriori sobre un set de datos de un proceso?

1. Una vez concluido el pre-procesamiento de los datos en “Preprocess”, el siguiente paso es ir a la pestaña “Associate” del “Explorer”.
2. Haciendo clic en el botón “Choose” de la sección “Associator” se puede elegir el algoritmo Apriori, o el que se piense sea mejor para el caso.
3. Luego de seleccionado el algoritmo, haciendo clic en la descripción de éste, se puede acceder a configurar los parámetros que se utilizarán al ejecutar el análisis.
4. Finalizada la configuración de los parámetros, haciendo clic en “Start” se dará comienzo a la ejecución del algoritmo, cuyos resultados serán mostrados en la sección “Associator output”.
5. Se pueden revisar las distintas ejecuciones realizadas, gracias al registro histórico que se va creando en la sección “Result list”.

ANEXO 4: PRE-PROCESADOR LOGS

La planilla “Pre-procesador Logs” fue implementada como parte del trabajo de la investigación descrita en este documento. Esta herramienta tiene como objetivo consolidar la información de cada ejecución de un proceso en un solo registro. Los Logs de Eventos entregan, en cada registro, la información que tienen de cada tarea realizada. Para este trabajo se necesitaba tener una visión a nivel de proceso y no solamente a nivel de tareas. Aquí se explicará cómo se puede llegar a consolidar la información de un proceso haciendo uso de la planilla “Pre-procesador Logs”.

Instrucciones: ¿Cómo se usa la planilla “Pre-procesador Logs” para consolidar la información de un proceso?

1. Lo primero que se debe realizar es ir a la hoja “Inicio” de la planilla y hacer clic en la opción “¡REINICIAR!”. Con esto se logrará eliminar cualquier dato de otro proceso que haya quedado cargado y que podría entorpecer el pre-procesamiento.
2. Luego, hay que seguir el paso a paso indicado en la hoja “Inicio”.
3. Los datos del proceso deben ser ingresados en la hoja “Log”.
4. Los resultados de la identificación de Tareas y Ejecutores se puede revisar en las hojas que llevan el mismo nombre.
5. La creación de los caminos de actividades se puede revisar en las hojas “At_Tareas”, “Caminos”, “Caminos_II” y “Caminos_III”. Lo mismo se puede realizar con los equipos de ejecutores, a través las hojas “At_Ejecutores”, “Equipos”, “Equipos_II” y “Equipos_III”.
6. Luego, se calculan los atributos de tiempo, los cuales se pueden verificar en la hoja “Tiempo”.
7. Con el procedimiento realizado hasta este punto, ya se cuenta con los datos fundamentales para completar la planilla final que luego será analizada. Esta planilla de datos se puede revisar en la hoja “CSV”.

8. De manera que no se pierda información de un caso, también se ha implementado la posibilidad de agregar atributos característicos (propios de cada caso de negocio) a la planilla de datos final. Para esto, hay que seleccionar en la hoja “Inicio” la opción “AGREGAR ATRIBUTOS ESPECIALES A LA HOJA CSV” y luego seguir las instrucciones indicadas en la hoja “Otros Atributos”.

Instrucciones: ¿Dónde se puede descargar la planilla “Pre-procesador Logs”?

La planilla puede ser descargada desde el siguiente link:

http://www.megaupload.com/?d=DKP1CN9V

ANEXO 5: ANALIZA PATRONES

La planilla "Analiza Patrones" es una herramienta desarrollada en Excel, como parte del trabajo realizado en la búsqueda de patrones secuenciales. Por medio de Macros escritas en Visual Basic se buscó facilitar el análisis de frecuencia de los valores de cada atributo que presente un proceso. El objetivo es poder analizar el comportamiento de cada atributo, especialmente en aquellos casos que coincidan con las definiciones entregadas por los patrones.

Instrucciones: ¿Cómo se puede realizar un análisis de frecuencia, de los atributos de un proceso, haciendo uso de la planilla "Analiza Patrones"?

1. Al abrir la planilla "Analiza Patrones" la hoja de inicio corresponderá a aquella llamada "01. Datos".
2. Estando en la hoja "01. Datos" hay que seguir las instrucciones señaladas en la planilla, partiendo por seleccionar la opción "Reiniciar".
3. Luego, se deben copiar en la planilla los datos que van a ser utilizados para el análisis. Hay que cuidar que los nombres de los atributos queden en la fila 16, y los datos queden ingresados a partir de la fila 17.
4. A continuación, se debe seleccionar la opción "Alistar para filtrar", que coloca las herramientas necesarias para poder filtrar los casos, las cuales permitirán enfocar el trabajo en aquellos casos que coincidan con los patrones entregados por el algoritmo Apriori.
5. Una vez filtrados los casos, se podrá observar en la hoja "01. Datos" la cantidad total de casos, y la cantidad luego de aplicar el filtro.
6. Para proceder al análisis de frecuencia de los atributos hay que seleccionar la opción "Ver más", la cual trasladará al investigador a la hoja "02. Metricas".
7. Estando en la hoja "02. Metricas" se podrá seleccionar cualquier atributo nominal, para luego ver sus distintos valores y la frecuencia que presentaron, tanto en los casos filtrados, como en el total de instancias que aparecieron.

Instrucciones: ¿Dónde se puede descargar la planilla “Analiza Patrones”?

La planilla puede ser descargada desde el siguiente link:

http://www.megaupload.com/?d=G6EYFNJN

ANEXO 6: BUSCA ANOMALÍAS

La planilla "Busca Anomalías" es parte de las herramientas desarrolladas entre los aportes de esta investigación. Esta herramienta implementada en Excel busca cumplir dos objetivos: primero, encontrar los casos anómalos dentro de un proceso y segundo, entregar la posibilidad de analizar el comportamiento de los distintos atributos, especialmente en aquellos casos que hayan sido clasificados como anómalos.

Instrucciones: ¿Cómo se puede realizar la búsqueda de anomalías en un proceso, haciendo uso de la planilla "Busca Anomalías"?

1. Al abrir la planilla "Busca Anomalías" la hoja de inicio corresponderá a aquella llamada "01. Datos".
2. Estando en la hoja "01. Datos" hay que seguir las instrucciones señaladas en la planilla, partiendo por seleccionar la opción "Reiniciar".
3. Luego, se deben copiar en la planilla los datos que van a ser utilizados para el análisis. Hay que cuidar que los nombres de los atributos queden en la fila 15, y los datos queden ingresados a partir de la fila 16.
4. A continuación, se debe seleccionar la opción "Definir atributos", que llevará al investigador a la hoja "02. Atributos", donde se deberá ingresar el tipo de dato de cada uno de los atributos que comprendan el proceso analizado.
5. Finalizado el ingreso de los tipos de datos, se debe seleccionar sobre qué tipo de atributo se desea realizar la búsqueda de anomalías. Dependiendo de esta decisión, el proceso continuará en la hoja "03. Parametros" o "03. Parametros B".
6. Estando en cualquiera de las hojas de parámetros, la tarea es la misma: seleccionar los tres parámetros necesarios para ejecutar el algoritmo Interquartile Range. Estos tres parámetros son: el atributo de referencia para la búsqueda de anomalías, el multiplicador para modificar el rango de normalidad y los umbrales en que se desea buscar las anomalías.

7. Luego de ingresar los parámetros, se puede revisar los distintos casos anómalos haciendo clic en “Mostrar anomalías”. Esto llevará al investigador a la hoja “04. Anomalias” o a “04. Anomalias B”, dependiendo de la selección que haya hecho en el punto 5 de este anexo.

Instrucciones: ¿Cómo se puede realizar el análisis de frecuencias de un proceso, haciendo uso de la planilla “Busca Anomalías”?

1. Encontrándose revisando las anomalías en las hojas “04. Anomalias” o “04. Anomalias B”, es muy simple pasar al análisis de frecuencias.
2. Hay que seleccionar la opción “Para ver más info de las anomalías, haga clic acá”. Esto llevará al investigador a la hoja “05. Metricas” o “05. Metricas B”, dependiendo de la hoja en que se haya encontrado.
3. Estando en cualquiera de las hojas “Metricas” se puede ver un resumen de los resultados obtenidos en la búsqueda de anomalías, como también están los campos para poder seleccionar un atributo y ver su distribución de frecuencias, tanto en los casos anómalos como en el total de ejecuciones.

Instrucciones: ¿Dónde se puede descargar la planilla “Busca Anomalías”?

La planilla puede ser descargada desde el siguiente link:

http://www.megaupload.com/?d=3RV5C55E

ANEXO 7: RESULTADOS IDENTIFICACIÓN EQUIPOS CASO "VENTA DE ARTÍCULOS A TRAVÉS DE PÁGINA WEB"

Aquí se entrega el detalle de los 524 equipos encontrados a través de los mil casos del proceso.

Equipo Final	ID_Equipo
E01-E02-E03-E04-E05-E04-E06-E05-E01	EQ01
E01-E02-E07-E07-E04-E06-E04-E05-E05-E01	EQ02
E01-E02-E03-E03-E04-E04-E05-E05-E05-E01	EQ03
E01-E04-E02-E05-E04-E06-E06-E01	EQ04
E01-E04-E07-E07-E07-E03-E07-E03-E04-E06-E02-E05-E06-E01	EQ05
E01-E04-E07-E03-E02-E06-E04-E05-E06-E01	EQ06
E01-E04-E02-E06-E02-E05-E05-E01	EQ07
E01-E04-E03-E03-E03-E02-E02-E06-E05-E05-E01	EQ08
E01-E04-E04-E05-E04-E05-E06-E01	EQ09
E01-E02-E03-E07-E03-E07-E03-E07-E07-E03-E07-E04-E04-E05-E05-E05-E01	EQ10
E01-E02-E07-E02-E06-E02-E05-E05-E01	EQ11
E01-E02-E02-E05-E02-E06-E05-E01	EQ12
E01-E04-E02-E06-E04-E05-E06-E01	EQ13
E01-E02-E03-E02-E05-E02-E06-E05-E01	EQ14
E01-E04-E07-E02-E06-E04-E05-E05-E01	EQ15
E01-E04-E02-E02-E05-E06-E05-E01	EQ16
E01-E02-E03-E03-E03-E07-E03-E03-E04-E02-E05-E05-E05-E01	EQ17
E01-E04-E04-E06-E02-E05-E06-E01	EQ18
E01-E02-E07-E07-E07-E04-E05-E04-E06-E05-E01	EQ19
E01-E04-E04-E05-E02-E06-E05-E01	EQ20
E01-E04-E04-E06-E02-E06-E06-E01	EQ21
E01-E02-E03-E04-E04-E06-E06-E06-E01	EQ22
E01-E02-E04-E02-E05-E05-E06-E01	EQ23
E01-E04-E07-E04-E05-E04-E06-E05-E01	EQ24
E01-E04-E02-E02-E05-E05-E06-E01	EQ25
E01-E04-E07-E03-E04-E06-E04-E05-E06-E01	EQ26
E01-E02-E04-E06-E02-E05-E06-E01	EQ27
E01-E04-E03-E07-E07-E07-E07-E04-E06-E04-E05-E05-E01	EQ28
E01-E02-E02-E06-E02-E06-E06-E01	EQ29
E01-E04-E07-E03-E02-E04-E05-E06-E05-E01	EQ30
E01-E02-E02-E02-E06-E06-E06-E01	EQ31
E01-E02-E03-E07-E07-E04-E04-E06-E05-E05-E01	EQ32

E01-E02-E03-E04-E05-E02-E05-E05-E01	EQ33
E01-E02-E03-E04-E06-E04-E05-E05-E01	EQ34
E01-E02-E04-E06-E02-E06-E06-E01	EQ35
E01-E02-E07-E03-E03-E02-E06-E02-E06-E06-E01	EQ36
E01-E02-E07-E02-E06-E02-E06-E06-E01	EQ37
E01-E04-E04-E02-E05-E06-E05-E01	EQ38
E01-E02-E07-E04-E02-E06-E06-E06-E01	EQ39
E01-E04-E07-E02-E05-E02-E05-E05-E01	EQ40
E01-E04-E03-E07-E02-E05-E02-E05-E06-E01	EQ41
E01-E02-E04-E05-E02-E05-E06-E01	EQ42
E01-E02-E07-E04-E04-E06-E06-E05-E01	EQ43
E01-E04-E07-E07-E02-E05-E04-E05-E06-E01	EQ44
E01-E04-E02-E06-E02-E06-E06-E01	EQ45
E01-E02-E02-E04-E05-E06-E05-E01	EQ46
E01-E02-E02-E06-E02-E05-E06-E01	EQ47
E01-E02-E04-E02-E06-E06-E06-E01	EQ48
E01-E04-E03-E02-E04-E05-E06-E05-E01	EQ49
E01-E04-E04-E04-E06-E05-E05-E01	EQ50
E01-E02-E07-E04-E04-E06-E05-E05-E01	EQ51
E01-E02-E04-E05-E04-E06-E06-E01	EQ52
E01-E04-E07-E07-E02-E05-E04-E06-E06-E01	EQ53
E01-E04-E07-E03-E07-E07-E03-E04-E02-E06-E05-E06-E01	EQ54
E01-E02-E02-E04-E05-E05-E06-E01	EQ55
E01-E02-E04-E04-E06-E05-E05-E01	EQ56
E01-E04-E03-E02-E06-E04-E05-E05-E01	EQ57
E01-E02-E07-E03-E04-E06-E02-E06-E06-E01	EQ58
E01-E02-E02-E04-E06-E05-E05-E01	EQ59
E01-E02-E07-E04-E04-E05-E05-E05-E01	EQ60
E01-E04-E02-E04-E05-E06-E05-E01	EQ61
E01-E04-E03-E02-E04-E05-E06-E06-E01	EQ62
E01-E02-E04-E06-E04-E06-E06-E01	EQ63
E01-E04-E03-E03-E04-E06-E02-E05-E06-E01	EQ64
E01-E02-E02-E05-E02-E05-E06-E01	EQ65
E01-E02-E03-E03-E07-E04-E05-E02-E06-E05-E01	EQ66
E01-E04-E07-E03-E04-E06-E04-E06-E05-E01	EQ67
E01-E04-E07-E07-E04-E06-E04-E05-E06-E01	EQ68
E01-E04-E04-E06-E02-E06-E05-E01	EQ69
E01-E02-E02-E02-E06-E05-E06-E01	EQ70
E01-E04-E03-E07-E03-E02-E06-E04-E06-E05-E01	EQ71
E01-E02-E07-E07-E03-E07-E03-E07-E04-E06-E02-E05-E06-E01	EQ72

E01-E02-E07-E03-E04-E06-E04-E05-E06-E01	EQ73
E01-E02-E04-E04-E06-E06-E05-E01	EQ74
E01-E04-E03-E02-E05-E02-E05-E06-E01	EQ75
E01-E02-E03-E04-E05-E04-E05-E05-E01	EQ76
E01-E04-E07-E07-E03-E07-E02-E02-E06-E05-E06-E01	EQ77
E01-E02-E07-E07-E07-E03-E03-E04-E02-E05-E06-E05-E01	EQ78
E01-E04-E07-E03-E02-E05-E02-E05-E05-E01	EQ79
E01-E04-E02-E05-E02-E05-E06-E01	EQ80
E01-E02-E04-E06-E02-E06-E05-E01	EQ81
E01-E04-E04-E04-E06-E05-E06-E01	EQ82
E01-E04-E04-E05-E04-E05-E05-E01	EQ83
E01-E04-E07-E02-E06-E04-E06-E06-E01	EQ84
E01-E04-E03-E03-E02-E06-E04-E06-E06-E01	EQ85
E01-E02-E04-E04-E06-E05-E06-E01	EQ86
E01-E04-E04-E06-E04-E05-E05-E01	EQ87
E01-E04-E07-E04-E04-E05-E06-E06-E01	EQ88
E01-E04-E02-E06-E02-E05-E06-E01	EQ89
E01-E04-E04-E04-E05-E05-E05-E01	EQ90
E01-E02-E07-E03-E04-E04-E05-E05-E06-E01	EQ91
E01-E02-E03-E07-E03-E03-E04-E02-E06-E06-E05-E01	EQ92
E01-E04-E04-E06-E04-E05-E06-E01	EQ93
E01-E04-E07-E02-E05-E04-E06-E05-E01	EQ94
E01-E02-E03-E02-E05-E04-E06-E06-E01	EQ95
E01-E04-E03-E03-E02-E02-E06-E05-E05-E01	EQ96
E01-E04-E03-E03-E07-E04-E06-E02-E05-E05-E01	EQ97
E01-E04-E04-E05-E02-E05-E06-E01	EQ98
E01-E02-E04-E02-E06-E05-E06-E01	EQ99
E01-E04-E04-E05-E02-E06-E06-E01	EQ100
E01-E04-E07-E02-E02-E06-E06-E06-E01	EQ101
E01-E04-E02-E04-E06-E05-E05-E01	EQ102
E01-E04-E07-E07-E04-E05-E02-E06-E05-E01	EQ103
E01-E02-E03-E03-E02-E02-E06-E06-E05-E01	EQ104
E01-E04-E07-E02-E06-E04-E06-E05-E01	EQ105
E01-E02-E02-E04-E05-E05-E05-E01	EQ106
E01-E04-E02-E02-E06-E05-E06-E01	EQ107
E01-E04-E07-E04-E06-E02-E05-E05-E01	EQ108
E01-E02-E03-E07-E07-E04-E05-E04-E05-E06-E01	EQ109
E01-E04-E07-E03-E04-E05-E02-E05-E05-E01	EQ110
E01-E02-E07-E02-E05-E04-E05-E05-E01	EQ111
E01-E04-E03-E02-E05-E02-E05-E05-E01	EQ112

E01-E02-E03-E07-E07-E03-E03-E03-E04-E05-E04-E06-E06-E01	EQ113
E01-E04-E02-E06-E04-E06-E05-E01	EQ114
E01-E04-E03-E03-E07-E03-E04-E04-E05-E06-E06-E01	EQ115
E01-E02-E02-E06-E04-E05-E05-E01	EQ116
E01-E02-E04-E06-E04-E06-E05-E01	EQ117
E01-E04-E02-E04-E05-E05-E05-E01	EQ118
E01-E04-E04-E02-E05-E06-E06-E01	EQ119
E01-E02-E04-E04-E05-E05-E05-E01	EQ120
E01-E04-E03-E07-E07-E04-E05-E04-E05-E05-E01	EQ121
E01-E02-E04-E02-E05-E06-E06-E01	EQ122
E01-E04-E07-E04-E04-E06-E06-E05-E01	EQ123
E01-E04-E03-E04-E06-E04-E06-E06-E01	EQ124
E01-E02-E04-E05-E04-E06-E05-E01	EQ125
E01-E04-E03-E07-E07-E02-E05-E04-E05-E05-E01	EQ126
E01-E02-E07-E07-E03-E04-E06-E02-E06-E05-E01	EQ127
E01-E04-E07-E07-E07-E02-E05-E04-E05-E06-E01	EQ128
E01-E02-E02-E02-E05-E06-E06-E01	EQ129
E01-E04-E04-E04-E05-E05-E06-E01	EQ130
E01-E02-E03-E03-E03-E02-E05-E04-E06-E05-E01	EQ131
E01-E02-E03-E07-E03-E03-E03-E04-E02-E06-E05-E06-E01	EQ132
E01-E04-E07-E04-E02-E05-E05-E06-E01	EQ133
E01-E04-E07-E03-E07-E02-E06-E02-E06-E06-E01	EQ134
E01-E04-E07-E02-E04-E06-E06-E06-E01	EQ135
E01-E04-E02-E02-E06-E06-E05-E01	EQ136
E01-E04-E03-E04-E04-E05-E06-E06-E01	EQ137
E01-E02-E02-E05-E02-E05-E05-E01	EQ138
E01-E02-E07-E07-E02-E04-E05-E06-E05-E01	EQ139
E01-E04-E03-E04-E02-E05-E06-E05-E01	EQ140
E01-E02-E04-E06-E04-E05-E05-E01	EQ141
E01-E04-E03-E02-E05-E04-E06-E06-E01	EQ142
E01-E04-E04-E05-E02-E05-E05-E01	EQ143
E01-E02-E07-E07-E02-E02-E06-E05-E06-E01	EQ144
E01-E02-E02-E02-E05-E05-E05-E01	EQ145
E01-E02-E07-E04-E05-E02-E05-E06-E01	EQ146
E01-E02-E03-E04-E05-E02-E05-E06-E01	EQ147
E01-E02-E07-E02-E05-E04-E06-E06-E01	EQ148
E01-E04-E04-E04-E06-E06-E05-E01	EQ149
E01-E04-E02-E04-E06-E06-E05-E01	EQ150
E01-E02-E07-E03-E03-E02-E06-E04-E06-E06-E01	EQ151
E01-E02-E03-E07-E07-E02-E06-E02-E06-E05-E01	EQ152

E01-E02-E03-E04-E04-E05-E06-E05-E01	EQ153
E01-E04-E02-E05-E02-E06-E05-E01	EQ154
E01-E02-E07-E04-E02-E06-E05-E06-E01	EQ155
E01-E02-E07-E07-E02-E02-E06-E06-E05-E01	EQ156
E01-E02-E02-E02-E06-E06-E05-E01	EQ157
E01-E04-E07-E04-E02-E06-E06-E05-E01	EQ158
E01-E02-E02-E05-E04-E06-E06-E01	EQ159
E01-E02-E07-E03-E03-E02-E06-E02-E05-E06-E01	EQ160
E01-E04-E07-E03-E02-E04-E06-E05-E05-E01	EQ161
E01-E04-E07-E03-E02-E04-E05-E05-E06-E01	EQ162
E01-E02-E03-E03-E02-E02-E05-E06-E05-E01	EQ163
E01-E04-E07-E07-E07-E02-E06-E02-E06-E05-E01	EQ164
E01-E04-E07-E02-E04-E06-E05-E05-E01	EQ165
E01-E04-E07-E02-E06-E02-E05-E05-E01	EQ166
E01-E04-E03-E07-E02-E06-E04-E05-E06-E01	EQ167
E01-E04-E02-E05-E04-E05-E05-E01	EQ168
E01-E02-E07-E07-E03-E02-E05-E02-E06-E06-E01	EQ169
E01-E04-E07-E02-E02-E05-E06-E05-E01	EQ170
E01-E04-E07-E02-E04-E05-E06-E06-E01	EQ171
E01-E02-E04-E06-E04-E05-E06-E01	EQ172
E01-E02-E03-E04-E06-E04-E06-E06-E01	EQ173
E01-E02-E04-E04-E05-E06-E05-E01	EQ174
E01-E02-E07-E07-E03-E04-E05-E04-E05-E06-E01	EQ175
E01-E04-E03-E02-E02-E05-E05-E05-E01	EQ176
E01-E04-E07-E04-E02-E06-E05-E06-E01	EQ177
E01-E04-E04-E02-E06-E05-E06-E01	EQ178
E01-E04-E07-E02-E04-E05-E05-E05-E01	EQ179
E01-E04-E07-E02-E05-E02-E06-E05-E01	EQ180
E01-E04-E03-E03-E04-E05-E04-E06-E06-E01	EQ181
E01-E02-E04-E02-E05-E05-E05-E01	EQ182
E01-E04-E02-E06-E04-E05-E05-E01	EQ183
E01-E04-E07-E03-E03-E02-E02-E06-E06-E06-E01	EQ184
E01-E04-E07-E03-E04-E02-E05-E06-E05-E01	EQ185
E01-E02-E02-E06-E04-E06-E05-E01	EQ186
E01-E02-E07-E04-E02-E05-E06-E06-E01	EQ187
E01-E04-E03-E02-E06-E02-E06-E06-E01	EQ188
E01-E02-E03-E03-E04-E02-E06-E06-E05-E01	EQ189
E01-E02-E03-E07-E03-E03-E07-E07-E04-E06-E04-E06-E05-E01	EQ190
E01-E04-E03-E04-E02-E06-E06-E06-E01	EQ191
E01-E02-E04-E02-E05-E06-E05-E01	EQ192

E01-E04-E07-E07-E03-E03-E03-E02-E02-E06-E06-E06-E01	EQ193
E01-E04-E02-E04-E06-E06-E06-E01	EQ194
E01-E02-E02-E05-E02-E06-E06-E01	EQ195
E01-E04-E03-E04-E02-E06-E05-E05-E01	EQ196
E01-E04-E04-E02-E06-E06-E05-E01	EQ197
E01-E04-E03-E07-E07-E04-E04-E06-E05-E06-E01	EQ198
E01-E02-E07-E02-E06-E04-E06-E06-E01	EQ199
E01-E04-E03-E02-E02-E06-E05-E05-E01	EQ200
E01-E02-E03-E02-E02-E06-E05-E06-E01	EQ201
E01-E04-E07-E04-E04-E06-E06-E06-E01	EQ202
E01-E04-E03-E04-E02-E05-E05-E06-E01	EQ203
E01-E04-E02-E02-E05-E06-E06-E01	EQ204
E01-E02-E02-E06-E04-E05-E06-E01	EQ205
E01-E04-E04-E05-E04-E06-E05-E01	EQ206
E01-E04-E04-E04-E05-E06-E05-E01	EQ207
E01-E02-E03-E03-E03-E03-E07-E03-E03-E04-E04-E05-E06-E06-E01	EQ208
E01-E04-E04-E02-E05-E05-E06-E01	EQ209
E01-E02-E03-E04-E02-E06-E06-E06-E01	EQ210
E01-E02-E07-E07-E04-E04-E05-E06-E05-E01	EQ211
E01-E04-E03-E03-E07-E03-E04-E05-E04-E05-E06-E01	EQ212
E01-E04-E02-E04-E05-E06-E06-E01	EQ213
E01-E02-E04-E05-E02-E06-E05-E01	EQ214
E01-E04-E03-E07-E07-E02-E04-E05-E06-E06-E01	EQ215
E01-E04-E07-E03-E07-E04-E06-E04-E06-E06-E01	EQ216
E01-E04-E03-E04-E04-E06-E06-E05-E01	EQ217
E01-E04-E03-E03-E04-E06-E02-E06-E05-E01	EQ218
E01-E02-E07-E03-E07-E04-E02-E05-E06-E05-E01	EQ219
E01-E02-E03-E07-E04-E05-E04-E06-E06-E01	EQ220
E01-E04-E07-E07-E07-E07-E07-E03-E07-E03-E03-E07-E04-E05-E02-E06-E06-E01	EQ221
E01-E02-E03-E03-E03-E03-E04-E06-E04-E06-E06-E01	EQ222
E01-E04-E03-E07-E07-E02-E05-E02-E06-E05-E01	EQ223
E01-E02-E03-E02-E06-E04-E05-E05-E01	EQ224
E01-E02-E07-E02-E05-E04-E06-E05-E01	EQ225
E01-E02-E03-E04-E05-E04-E06-E06-E01	EQ226
E01-E02-E03-E03-E07-E03-E03-E07-E04-E06-E04-E06-E05-E01	EQ227
E01-E02-E03-E04-E05-E02-E06-E06-E01	EQ228
E01-E02-E03-E02-E05-E04-E05-E06-E01	EQ229
E01-E04-E02-E05-E02-E06-E06-E01	EQ230
E01-E04-E02-E06-E02-E06-E05-E01	EQ231
E01-E04-E04-E02-E05-E05-E05-E01	EQ232

E01-E02-E07-E07-E02-E04-E06-E05-E06-E01	EQ233
E01-E04-E02-E06-E04-E06-E06-E01	EQ234
E01-E04-E07-E03-E07-E03-E02-E06-E02-E05-E06-E01	EQ235
E01-E02-E04-E05-E02-E05-E05-E01	EQ236
E01-E02-E02-E05-E04-E05-E05-E01	EQ237
E01-E04-E07-E07-E07-E04-E05-E02-E05-E06-E01	EQ238
E01-E04-E04-E06-E04-E06-E05-E01	EQ239
E01-E02-E02-E02-E06-E05-E05-E01	EQ240
E01-E04-E04-E06-E02-E05-E05-E01	EQ241
E01-E04-E03-E07-E07-E07-E03-E02-E05-E04-E06-E06-E01	EQ242
E01-E02-E07-E02-E02-E06-E06-E06-E01	EQ243
E01-E04-E07-E02-E02-E05-E05-E06-E01	EQ244
E01-E02-E03-E04-E04-E06-E05-E06-E01	EQ245
E01-E04-E03-E07-E07-E04-E06-E04-E06-E06-E01	EQ246
E01-E04-E07-E07-E02-E02-E06-E06-E06-E01	EQ247
E01-E04-E03-E02-E06-E02-E05-E05-E01	EQ248
E01-E02-E03-E02-E05-E02-E06-E06-E01	EQ249
E01-E02-E03-E07-E04-E05-E04-E05-E06-E01	EQ250
E01-E04-E03-E07-E04-E02-E05-E05-E05-E01	EQ251
E01-E02-E02-E02-E05-E05-E06-E01	EQ252
E01-E02-E07-E02-E02-E05-E06-E06-E01	EQ253
E01-E04-E03-E03-E02-E05-E02-E05-E05-E01	EQ254
E01-E02-E02-E06-E04-E06-E06-E01	EQ255
E01-E02-E07-E02-E06-E04-E06-E05-E01	EQ256
E01-E02-E04-E05-E04-E05-E06-E01	EQ257
E01-E04-E07-E02-E05-E04-E05-E05-E01	EQ258
E01-E02-E04-E02-E06-E05-E05-E01	EQ259
E01-E02-E03-E04-E04-E06-E05-E05-E01	EQ260
E01-E04-E07-E02-E05-E04-E05-E06-E01	EQ261
E01-E02-E07-E07-E04-E04-E06-E05-E06-E01	EQ262
E01-E04-E02-E05-E02-E05-E05-E01	EQ263
E01-E02-E07-E03-E02-E06-E04-E05-E06-E01	EQ264
E01-E02-E07-E04-E05-E04-E05-E06-E01	EQ265
E01-E02-E07-E04-E05-E02-E06-E05-E01	EQ266
E01-E04-E07-E03-E02-E02-E06-E06-E05-E01	EQ267
E01-E02-E04-E05-E04-E05-E05-E01	EQ268
E01-E04-E04-E02-E06-E06-E06-E01	EQ269
E01-E04-E03-E07-E02-E04-E05-E06-E05-E01	EQ270
E01-E04-E07-E07-E02-E04-E05-E06-E05-E01	EQ271
E01-E02-E04-E05-E02-E06-E06-E01	EQ272

E01-E02-E03-E03-E07-E07-E02-E04-E05-E06-E05-E01	EQ273
E01-E04-E03-E03-E04-E02-E05-E06-E05-E01	EQ274
E01-E04-E07-E07-E02-E04-E05-E05-E06-E01	EQ275
E01-E04-E07-E07-E07-E02-E04-E05-E06-E05-E01	EQ276
E01-E02-E07-E02-E02-E05-E06-E05-E01	EQ277
E01-E02-E07-E03-E04-E02-E05-E06-E05-E01	EQ278
E01-E04-E03-E07-E04-E06-E02-E05-E06-E01	EQ279
E01-E04-E07-E04-E05-E04-E05-E05-E01	EQ280
E01-E04-E03-E07-E02-E02-E05-E05-E06-E01	EQ281
E01-E04-E07-E07-E04-E04-E05-E05-E05-E01	EQ282
E01-E04-E07-E04-E04-E05-E06-E05-E01	EQ283
E01-E04-E07-E02-E06-E02-E06-E06-E01	EQ284
E01-E04-E03-E07-E03-E03-E02-E06-E02-E06-E05-E01	EQ285
E01-E04-E03-E03-E02-E04-E06-E06-E05-E01	EQ286
E01-E04-E03-E07-E04-E04-E05-E06-E05-E01	EQ287
E01-E04-E07-E03-E03-E07-E03-E07-E04-E05-E04-E05-E06-E01	EQ288
E01-E04-E03-E03-E07-E07-E03-E02-E05-E02-E06-E06-E01	EQ289
E01-E02-E07-E03-E07-E07-E02-E04-E06-E05-E06-E01	EQ290
E01-E02-E03-E02-E02-E05-E06-E06-E01	EQ291
E01-E02-E03-E03-E03-E07-E02-E06-E02-E06-E06-E01	EQ292
E01-E02-E02-E06-E02-E05-E05-E01	EQ293
E01-E02-E07-E02-E04-E05-E05-E05-E01	EQ294
E01-E04-E07-E03-E04-E04-E06-E05-E06-E01	EQ295
E01-E04-E07-E03-E03-E04-E02-E06-E05-E05-E01	EQ296
E01-E04-E03-E02-E06-E02-E06-E05-E01	EQ297
E01-E04-E04-E06-E04-E06-E06-E01	EQ298
E01-E02-E07-E04-E04-E05-E06-E06-E01	EQ299
E01-E04-E07-E03-E03-E07-E04-E05-E02-E06-E06-E01	EQ300
E01-E04-E03-E04-E06-E02-E05-E05-E01	EQ301
E01-E02-E03-E07-E03-E03-E04-E06-E04-E05-E05-E01	EQ302
E01-E04-E03-E07-E07-E02-E04-E06-E05-E05-E01	EQ303
E01-E02-E03-E03-E02-E04-E06-E06-E05-E01	EQ304
E01-E02-E07-E03-E04-E02-E06-E05-E06-E01	EQ305
E01-E02-E07-E07-E02-E05-E02-E05-E06-E01	EQ306
E01-E02-E03-E02-E02-E05-E05-E05-E01	EQ307
E01-E02-E07-E03-E02-E05-E02-E05-E06-E01	EQ308
E01-E04-E04-E04-E06-E06-E06-E01	EQ309
E01-E04-E03-E07-E04-E06-E02-E06-E05-E01	EQ310
E01-E02-E03-E02-E04-E05-E06-E05-E01	EQ311
E01-E02-E07-E02-E05-E02-E06-E06-E01	EQ312

E01-E02-E07-E02-E06-E02-E05-E06-E01	EQ313
E01-E04-E07-E07-E04-E04-E05-E05-E06-E01	EQ314
E01-E02-E03-E07-E07-E04-E04-E05-E06-E06-E01	EQ315
E01-E04-E03-E03-E07-E03-E04-E04-E06-E06-E05-E01	EQ316
E01-E04-E07-E04-E04-E06-E05-E06-E01	EQ317
E01-E02-E03-E02-E06-E04-E06-E06-E01	EQ318
E01-E04-E07-E03-E07-E02-E05-E04-E05-E05-E01	EQ319
E01-E04-E03-E03-E07-E03-E07-E02-E05-E02-E05-E05-E01	EQ320
E01-E04-E07-E02-E02-E05-E06-E06-E01	EQ321
E01-E04-E02-E04-E05-E05-E06-E01	EQ322
E01-E02-E03-E02-E02-E06-E05-E05-E01	EQ323
E01-E04-E07-E02-E05-E02-E05-E06-E01	EQ324
E01-E04-E03-E07-E07-E04-E02-E06-E06-E06-E01	EQ325
E01-E02-E07-E03-E04-E02-E06-E06-E05-E01	EQ326
E01-E04-E03-E03-E02-E02-E05-E06-E06-E01	EQ327
E01-E04-E03-E07-E03-E04-E05-E02-E06-E06-E01	EQ328
E01-E04-E03-E03-E02-E05-E02-E05-E06-E01	EQ329
E01-E02-E07-E03-E03-E02-E05-E04-E06-E06-E01	EQ330
E01-E04-E03-E04-E04-E05-E06-E05-E01	EQ331
E01-E02-E07-E07-E03-E02-E05-E04-E06-E05-E01	EQ332
E01-E04-E07-E04-E02-E05-E06-E06-E01	EQ333
E01-E04-E03-E07-E03-E03-E07-E03-E03-E07-E03-E07-E02-E05-E04-E05-E05-E01	EQ334
E01-E04-E07-E02-E04-E06-E06-E05-E01	EQ335
E01-E02-E07-E04-E06-E02-E06-E06-E01	EQ336
E01-E04-E03-E07-E07-E07-E02-E04-E05-E05-E05-E01	EQ337
E01-E02-E03-E04-E02-E06-E05-E05-E01	EQ338
E01-E02-E03-E07-E04-E04-E05-E06-E06-E01	EQ339
E01-E02-E03-E03-E04-E06-E04-E05-E05-E01	EQ340
E01-E04-E03-E03-E02-E05-E02-E06-E06-E01	EQ341
E01-E04-E07-E07-E04-E05-E02-E06-E06-E01	EQ342
E01-E02-E03-E02-E06-E02-E06-E06-E01	EQ343
E01-E02-E03-E03-E03-E02-E06-E02-E05-E05-E01	EQ344
E01-E04-E03-E02-E04-E06-E06-E05-E01	EQ345
E01-E02-E04-E02-E06-E06-E05-E01	EQ346
E01-E04-E07-E03-E04-E02-E06-E06-E06-E01	EQ347
E01-E02-E02-E06-E02-E06-E05-E01	EQ348
E01-E02-E03-E07-E04-E06-E04-E06-E05-E01	EQ349
E01-E02-E03-E03-E07-E04-E06-E02-E06-E05-E01	EQ350
E01-E04-E07-E04-E02-E05-E05-E05-E01	EQ351
E01-E02-E03-E02-E06-E04-E05-E06-E01	EQ352

E01-E04-E07-E03-E03-E07-E04-E04-E06-E06-E06-E01	EQ353
E01-E02-E04-E06-E02-E05-E05-E01	EQ354
E01-E04-E07-E07-E03-E04-E06-E02-E06-E06-E01	EQ355
E01-E02-E03-E03-E03-E04-E04-E05-E05-E06-E01	EQ356
E01-E02-E03-E07-E02-E05-E04-E06-E05-E01	EQ357
E01-E04-E03-E02-E02-E05-E06-E06-E01	EQ358
E01-E04-E03-E03-E07-E02-E05-E02-E05-E06-E01	EQ359
E01-E04-E02-E05-E04-E05-E06-E01	EQ360
E01-E04-E03-E07-E03-E07-E02-E02-E06-E05-E06-E01	EQ361
E01-E04-E03-E04-E04-E06-E05-E05-E01	EQ362
E01-E04-E04-E05-E04-E06-E06-E01	EQ363
E01-E04-E07-E03-E04-E05-E04-E05-E06-E01	EQ364
E01-E02-E07-E04-E06-E04-E05-E06-E01	EQ365
E01-E02-E07-E02-E02-E05-E05-E06-E01	EQ366
E01-E02-E07-E07-E02-E06-E04-E06-E06-E01	EQ367
E01-E04-E03-E07-E02-E05-E02-E05-E05-E01	EQ368
E01-E04-E02-E02-E06-E05-E05-E01	EQ369
E01-E02-E03-E07-E02-E06-E04-E06-E06-E01	EQ370
E01-E04-E07-E07-E02-E02-E05-E05-E06-E01	EQ371
E01-E04-E07-E04-E04-E05-E05-E05-E01	EQ372
E01-E02-E03-E03-E02-E06-E02-E05-E05-E01	EQ373
E01-E04-E03-E02-E05-E04-E06-E05-E01	EQ374
E01-E02-E07-E02-E05-E02-E06-E05-E01	EQ375
E01-E04-E04-E02-E06-E05-E05-E01	EQ376
E01-E02-E02-E04-E06-E05-E06-E01	EQ377
E01-E04-E07-E07-E02-E02-E06-E06-E05-E01	EQ378
E01-E04-E03-E02-E05-E02-E06-E05-E01	EQ379
E01-E02-E03-E04-E05-E02-E06-E05-E01	EQ380
E01-E04-E02-E05-E04-E06-E05-E01	EQ381
E01-E04-E03-E07-E04-E02-E06-E05-E05-E01	EQ382
E01-E04-E07-E07-E03-E04-E05-E04-E05-E05-E01	EQ383
E01-E02-E07-E07-E07-E03-E03-E07-E03-E07-E03-E07-E04-E04-E05-E05-E06-E01	EQ384
E01-E04-E07-E04-E06-E02-E06-E06-E01	EQ385
E01-E04-E03-E07-E02-E04-E06-E06-E06-E01	EQ386
E01-E02-E07-E04-E02-E05-E05-E06-E01	EQ387
E01-E02-E07-E03-E02-E02-E05-E05-E06-E01	EQ388
E01-E02-E07-E04-E05-E02-E06-E06-E01	EQ389
E01-E04-E07-E02-E02-E06-E06-E05-E01	EQ390
E01-E02-E03-E07-E02-E04-E06-E05-E06-E01	EQ391
E01-E02-E03-E07-E02-E06-E02-E06-E06-E01	EQ392

E01-E04-E07-E07-E03-E02-E02-E06-E06-E06-E01	EQ393
E01-E04-E03-E07-E04-E04-E06-E05-E05-E01	EQ394
E01-E04-E07-E07-E02-E04-E05-E05-E05-E01	EQ395
E01-E02-E03-E07-E02-E04-E06-E05-E05-E01	EQ396
E01-E02-E03-E03-E03-E03-E02-E05-E04-E06-E05-E01	EQ397
E01-E04-E03-E07-E02-E04-E05-E06-E06-E01	EQ398
E01-E04-E07-E02-E04-E05-E05-E06-E01	EQ399
E01-E04-E03-E03-E03-E02-E05-E04-E06-E06-E01	EQ400
E01-E04-E03-E03-E04-E06-E04-E05-E05-E01	EQ401
E01-E04-E07-E04-E06-E04-E06-E06-E01	EQ402
E01-E04-E02-E02-E05-E05-E05-E01	EQ403
E01-E02-E02-E04-E06-E06-E06-E01	EQ404
E01-E04-E02-E04-E06-E05-E06-E01	EQ405
E01-E02-E03-E02-E05-E04-E06-E05-E01	EQ406
E01-E02-E02-E05-E04-E06-E05-E01	EQ407
E01-E02-E03-E02-E04-E06-E05-E05-E01	EQ408
E01-E02-E07-E03-E04-E04-E06-E06-E06-E01	EQ409
E01-E04-E07-E07-E03-E07-E04-E05-E04-E05-E05-E01	EQ410
E01-E02-E03-E07-E02-E04-E06-E06-E06-E01	EQ411
E01-E04-E03-E07-E02-E05-E04-E05-E05-E01	EQ412
E01-E02-E07-E04-E05-E04-E06-E05-E01	EQ413
E01-E02-E03-E07-E03-E07-E03-E02-E02-E05-E05-E06-E01	EQ414
E01-E02-E07-E07-E03-E04-E06-E04-E06-E05-E01	EQ415
E01-E04-E07-E07-E03-E07-E03-E03-E03-E02-E05-E02-E05-E05-E01	EQ416
E01-E04-E07-E04-E04-E06-E05-E05-E01	EQ417
E01-E02-E07-E04-E04-E05-E06-E05-E01	EQ418
E01-E02-E07-E07-E02-E02-E05-E06-E05-E01	EQ419
E01-E04-E07-E03-E04-E04-E05-E05-E06-E01	EQ420
E01-E04-E03-E07-E03-E07-E07-E07-E07-E03-E03-E03-E07-E07-E03-E03-E07-E02-E02-E06-E05-E05-E01	EQ421
E01-E02-E03-E04-E06-E04-E06-E05-E01	EQ422
E01-E02-E07-E04-E02-E06-E05-E05-E01	EQ423
E01-E02-E07-E03-E02-E05-E02-E06-E06-E01	EQ424
E01-E04-E03-E07-E07-E07-E03-E07-E07-E02-E06-E04-E05-E05-E01	EQ425
E01-E02-E03-E07-E03-E02-E06-E02-E05-E05-E01	EQ426
E01-E02-E02-E04-E06-E06-E05-E01	EQ427
E01-E02-E07-E07-E07-E04-E02-E05-E05-E05-E01	EQ428
E01-E04-E03-E07-E03-E02-E04-E05-E05-E06-E01	EQ429
E01-E04-E03-E04-E05-E04-E05-E06-E01	EQ430
E01-E02-E07-E04-E02-E05-E06-E05-E01	EQ431

E01-E02-E03-E07-E02-E04-E05-E06-E05-E01	EQ432
E01-E04-E07-E02-E05-E04-E06-E06-E01	EQ433
E01-E02-E07-E03-E02-E06-E02-E06-E05-E01	EQ434
E01-E02-E07-E04-E06-E02-E05-E05-E01	EQ435
E01-E04-E03-E07-E04-E02-E06-E05-E06-E01	EQ436
E01-E02-E03-E02-E05-E02-E05-E06-E01	EQ437
E01-E02-E07-E03-E04-E04-E06-E05-E06-E01	EQ438
E01-E02-E03-E07-E02-E05-E02-E06-E05-E01	EQ439
E01-E04-E03-E02-E02-E06-E06-E06-E01	EQ440
E01-E04-E03-E02-E05-E02-E06-E06-E01	EQ441
E01-E04-E03-E03-E07-E07-E07-E03-E02-E02-E06-E05-E05-E01	EQ442
E01-E02-E03-E02-E04-E06-E06-E06-E01	EQ443
E01-E04-E03-E07-E03-E04-E06-E02-E06-E05-E01	EQ444
E01-E04-E07-E07-E03-E04-E06-E04-E05-E06-E01	EQ445
E01-E02-E03-E04-E02-E05-E06-E05-E01	EQ446
E01-E02-E03-E02-E04-E05-E05-E06-E01	EQ447
E01-E02-E03-E07-E03-E02-E05-E04-E06-E06-E01	EQ448
E01-E04-E03-E03-E03-E07-E03-E04-E06-E04-E05-E05-E01	EQ449
E01-E02-E07-E02-E04-E06-E06-E05-E01	EQ450
E01-E04-E03-E03-E02-E02-E05-E06-E05-E01	EQ451
E01-E04-E07-E07-E07-E03-E03-E07-E07-E03-E04-E06-E02-E05-E05-E01	EQ452
E01-E04-E03-E02-E04-E06-E05-E06-E01	EQ453
E01-E02-E07-E03-E03-E02-E05-E02-E05-E06-E01	EQ454
E01-E04-E07-E03-E03-E04-E06-E04-E05-E05-E01	EQ455
E01-E02-E03-E07-E07-E03-E04-E06-E02-E06-E06-E01	EQ456
E01-E04-E03-E07-E02-E06-E02-E05-E05-E01	EQ457
E01-E04-E03-E03-E04-E04-E06-E05-E05-E01	EQ458
E01-E02-E03-E07-E07-E02-E04-E06-E06-E05-E01	EQ459
E01-E04-E07-E02-E06-E04-E05-E06-E01	EQ460
E01-E02-E03-E04-E06-E02-E05-E06-E01	EQ461
E01-E04-E03-E07-E02-E06-E04-E06-E06-E01	EQ462
E01-E04-E03-E02-E04-E05-E05-E05-E01	EQ463
E01-E02-E03-E07-E07-E07-E07-E07-E07-E02-E05-E04-E06-E06-E01	EQ464
E01-E02-E07-E03-E03-E04-E06-E02-E06-E05-E01	EQ465
E01-E02-E02-E04-E05-E06-E06-E01	EQ466
E01-E04-E07-E03-E02-E05-E04-E05-E06-E01	EQ467
E01-E02-E07-E03-E03-E02-E05-E02-E05-E05-E01	EQ468
E01-E02-E03-E03-E03-E04-E05-E04-E05-E06-E01	EQ469
E01-E02-E04-E04-E05-E05-E06-E01	EQ470
E01-E02-E07-E04-E02-E06-E06-E05-E01	EQ471

E01-E02-E03-E07-E04-E02-E06-E06-E05-E01	EQ472
E01-E04-E03-E04-E02-E06-E06-E05-E01	EQ473
E01-E02-E07-E07-E03-E02-E05-E04-E05-E06-E01	EQ474
E01-E04-E07-E07-E03-E04-E02-E06-E06-E06-E01	EQ475
E01-E02-E03-E04-E02-E06-E06-E05-E01	EQ476
E01-E02-E07-E03-E03-E04-E02-E05-E05-E06-E01	EQ477
E01-E04-E03-E07-E07-E04-E05-E02-E06-E06-E01	EQ478
E01-E02-E03-E03-E04-E06-E04-E06-E06-E01	EQ479
E01-E04-E02-E02-E06-E06-E06-E01	EQ480
E01-E04-E07-E04-E05-E02-E06-E06-E01	EQ481
E01-E02-E03-E03-E03-E04-E02-E06-E06-E06-E01	EQ482
E01-E04-E07-E03-E07-E07-E04-E04-E06-E06-E06-E01	EQ483
E01-E04-E07-E03-E02-E06-E04-E06-E06-E01	EQ484
E01-E04-E07-E02-E06-E02-E06-E05-E01	EQ485
E01-E04-E07-E02-E05-E02-E06-E06-E01	EQ486
E01-E02-E03-E03-E04-E05-E02-E06-E05-E01	EQ487
E01-E04-E07-E03-E07-E04-E06-E02-E06-E06-E01	EQ488
E01-E04-E07-E03-E03-E07-E07-E07-E04-E05-E04-E05-E05-E01	EQ489
E01-E02-E07-E03-E07-E07-E02-E06-E04-E06-E05-E01	EQ490
E01-E02-E07-E03-E02-E06-E04-E05-E05-E01	EQ491
E01-E04-E07-E07-E07-E03-E07-E04-E06-E02-E05-E05-E01	EQ492
E01-E02-E03-E03-E02-E02-E06-E06-E06-E01	EQ493
E01-E04-E03-E04-E04-E05-E05-E05-E01	EQ494
E01-E04-E07-E03-E07-E07-E07-E03-E02-E06-E02-E06-E06-E01	EQ495
E01-E04-E07-E03-E03-E02-E06-E04-E05-E05-E01	EQ496
E01-E02-E03-E04-E02-E05-E06-E06-E01	EQ497
E01-E02-E04-E04-E05-E06-E06-E01	EQ498
E01-E04-E07-E03-E04-E05-E04-E05-E05-E01	EQ499
E01-E04-E03-E07-E07-E04-E04-E05-E05-E06-E01	EQ500
E01-E04-E07-E07-E03-E07-E02-E06-E04-E06-E06-E01	EQ501
E01-E02-E03-E03-E07-E04-E04-E06-E06-E06-E01	EQ502
E01-E02-E07-E03-E04-E05-E04-E05-E05-E01	EQ503
E01-E04-E07-E04-E06-E02-E06-E05-E01	EQ504
E01-E02-E07-E03-E03-E04-E05-E04-E05-E06-E01	EQ505
E01-E04-E03-E03-E03-E07-E03-E07-E03-E07-E02-E04-E06-E06-E05-E01	EQ506
E01-E02-E03-E04-E06-E02-E06-E05-E01	EQ507
E01-E02-E03-E07-E07-E07-E07-E03-E02-E06-E02-E06-E05-E01	EQ508
E01-E02-E03-E07-E03-E03-E07-E07-E07-E02-E04-E05-E06-E06-E01	EQ509
E01-E04-E03-E04-E04-E05-E05-E06-E01	EQ510
E01-E04-E07-E03-E02-E05-E04-E06-E05-E01	EQ511

E01-E02-E03-E03-E03-E07-E02-E06-E04-E05-E05-E01	EQ512
E01-E04-E03-E07-E03-E07-E03-E07-E02-E04-E05-E06-E05-E01	EQ513
E01-E04-E03-E03-E07-E03-E03-E03-E03-E04-E05-E04-E06-E06-E01	EQ514
E01-E02-E03-E03-E07-E03-E04-E04-E05-E05-E06-E01	EQ515
E01-E02-E07-E07-E07-E07-E03-E07-E07-E02-E05-E04-E05-E05-E01	EQ516
E01-E04-E03-E03-E02-E06-E02-E06-E05-E01	EQ517
E01-E04-E07-E03-E04-E06-E02-E06-E05-E01	EQ518
E01-E04-E03-E04-E05-E04-E06-E05-E01	EQ519
E01-E02-E03-E07-E03-E07-E04-E02-E05-E05-E05-E01	EQ520
E01-E02-E04-E04-E06-E06-E06-E01	EQ521
E01-E04-E03-E03-E03-E02-E02-E05-E06-E05-E01	EQ522
E01-E04-E07-E03-E07-E04-E06-E02-E06-E05-E01	EQ523
E01-E04-E03-E04-E06-E02-E05-E06-E01	EQ524

ANEXO 8: REGLAS DE ASOCIACIÓN CASO "VENTA DE ARTÍCULOS A TRAVÉS DE PÁGINA WEB" (ANÁLISIS DE SENSIBILIDAD)

Dentro de los análisis de sensibilidad del algoritmo Apriori, se destacó un punto en que las reglas aumentaban de 51 a 97, producto de la disminución del nivel de confianza requerido, precisamente cuando se bajó de de 0,8 a 0,75.

Las reglas encontradas con esta configuración fueron:

1. monto=2000.0 285 ==> OK=OK 285 conf:(1)

2. monto=1000.0 270 ==> OK=OK 270 conf:(1)

3. cantidad=2.0 250 ==> OK=OK 250 conf:(1)

4. cantidad=1.0 245 ==> OK=OK 245 conf:(1)

5. cantidad=>3 monto=2000.0 150 ==> OK=OK 150 conf:(1)

6. cantidad=2.0 monto=1000.0 150 ==> OK=OK 150 conf:(1)

7. ID_TareasConsolidadas=TC01 monto=2000.0 144 ==> OK=OK 144 conf:(1)

8. Tiempo_Desde_Comienzo=T5800 monto=2000.0 141 ==> OK=OK 141 conf:(1)

9. ID_TareasConsolidadas=TC02 monto=2000.0 141 ==> OK=OK 141 conf:(1)

10. monto=2000.0 patron_prom=P2 141 ==> OK=OK 141 conf:(1)

11. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 monto=2000.0 141 ==> OK=OK 141 conf:(1)

12. Tiempo_Desde_Comienzo=T5800 monto=2000.0 patron_prom=P2 141 ==> OK=OK 141 conf:(1)

13. ID_TareasConsolidadas=TC02 monto=2000.0 patron_prom=P2 141 ==> OK=OK 141 conf:(1)

14. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 monto=2000.0 patron_prom=P2 141 ==> OK=OK 141 conf:(1)

15. ID_TareasConsolidadas=TC01 monto=1000.0 140 ==> OK=OK 140 conf:(1)

16. cantidad=3.0 monto=2000.0 135 ==> OK=OK 135 conf:(1)

17. ID_TareasConsolidadas=TC01 cantidad=2.0 131 ==> OK=OK 131 conf:(1)

18. Tiempo_Desde_Comienzo=T5800 monto=1000.0 130 ==> OK=OK 130 conf:(1)

19. ID_TareasConsolidadas=TC01 cantidad=1.0 130 ==> OK=OK 130 conf:(1)

20. ID_TareasConsolidadas=TC02 monto=1000.0 130 ==> OK=OK 130 conf:(1)

21. monto=1000.0 patron_prom=P2 130 ==> OK=OK 130 conf:(1)

22. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 monto=1000.0 130 ==> OK=OK 130 conf:(1)

23. Tiempo_Desde_Comienzo=T5800 monto=1000.0 patron_prom=P2 130 ==> OK=OK 130 conf:(1)

24. ID_TareasConsolidadas=TC02 monto=1000.0 patron_prom=P2 130 ==> OK=OK 130 conf:(1)

25. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 monto=1000.0 patron_prom=P2 130 ==> OK=OK 130 conf:(1)

26. monto=3000.0 125 ==> OK=OK 125 conf:(1)

27. cantidad=1.0 monto=3000.0 125 ==> OK=OK 125 conf:(1)

28. cantidad=1.0 monto=1000.0 120 ==> OK=OK 120 conf:(1)

29. cantidad=3.0 monto=>3000 120 ==> OK=NOK 120 conf:(1)

30. Tiempo_Desde_Comienzo=T5800 cantidad=2.0 119 ==> OK=OK 119 conf:(1)

31. ID_TareasConsolidadas=TC02 cantidad=2.0 119 ==> OK=OK 119 conf:(1)

32. cantidad=2.0 patron_prom=P2 119 ==> OK=OK 119 conf:(1)

33. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 cantidad=2.0 119 ==> OK=OK 119 conf:(1)

34. Tiempo_Desde_Comienzo=T5800 cantidad=2.0 patron_prom=P2 119 ==> OK=OK 119 conf:(1)

35. ID_TareasConsolidadas=TC02 cantidad=2.0 patron_prom=P2 119 ==> OK=OK 119 conf:(1)

36. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 cantidad=2.0 patron_prom=P2 119 ==> OK=OK 119 conf:(1)

37. Tiempo_Desde_Comienzo=T5800 cantidad=1.0 115 ==> OK=OK 115 conf:(1)

38. ID_TareasConsolidadas=TC02 cantidad=1.0 115 ==> OK=OK 115 conf:(1)

39. cantidad=1.0 patron_prom=P2 115 ==> OK=OK 115 conf:(1)

40. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 cantidad=1.0 115 ==> OK=OK 115 conf:(1)

41. Tiempo_Desde_Comienzo=T5800 cantidad=1.0 patron_prom=P2 115 ==> OK=OK 115 conf:(1)

42. ID_TareasConsolidadas=TC02 cantidad=1.0 patron_prom=P2 115 ==> OK=OK 115 conf:(1)

43. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 cantidad=1.0 patron_prom=P2 115 ==> OK=OK 115 conf:(1)

44. cantidad=>3 monto=>3000 100 ==> OK=NOK 100 conf:(1)

45. cantidad=2.0 monto=>3000 100 ==> OK=OK 100 conf:(1)

46. Tiempo_Desde_Comienzo=T8800 240 ==> OK=OK 193 conf:(0.8)

47. patron_prom=P0 240 ==> OK=OK 193 conf:(0.8)

48. Tiempo_Desde_Comienzo=T8800 ID_TareasConsolidadas=TC01 240 ==> OK=OK 193 conf:(0.8)

49. Tiempo_Desde_Comienzo=T8800 patron_prom=P0 240 ==> OK=OK 193 conf:(0.8)

50. ID_TareasConsolidadas=TC01 patron_prom=P0 240 ==> OK=OK 193 conf:(0.8)

51. Tiempo_Desde_Comienzo=T8800 ID_TareasConsolidadas=TC01 patron_prom=P0 240 ==> OK=OK 193 conf:(0.8)

52. ID_TareasConsolidadas=TC01 509 ==> OK=OK 405 conf:(0.8)

53. Tiempo_Desde_Comienzo=T11800 131 ==> OK=OK 103 conf:(0.79)

54. patron_prom=P1 131 ==> OK=OK 103 conf:(0.79)

55. Tiempo_Desde_Comienzo=T11800 ID_TareasConsolidadas=TC01 131 ==> OK=OK 103 conf:(0.79)

56. Tiempo_Desde_Comienzo=T11800 patron_prom=P1 131 ==> OK=OK 103 conf:(0.79)

57. ID_TareasConsolidadas=TC01 patron_prom=P1 131 ==> OK=OK 103 conf:(0.79)

58. Tiempo_Desde_Comienzo=T11800 ID_TareasConsolidadas=TC01 patron_prom=P1 131 ==> OK=OK 103 conf:(0.79)

59. ID_Camino=C04 246 ==> OK=OK 189 conf:(0.77)

60. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 246 ==> OK=OK 189 conf:(0.77)

61. ID_Camino=C04 ID_TareasConsolidadas=TC02 246 ==> OK=OK 189 conf:(0.77)

62. ID_Camino=C04 patron_prom=P2 246 ==> OK=OK 189 conf:(0.77)

63. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 246 ==> OK=OK 189 conf:(0.77)

64. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 patron_prom=P2 246 ==> OK=OK 189 conf:(0.77)

65. ID_Camino=C04 ID_TareasConsolidadas=TC02 patron_prom=P2 246 ==> OK=OK 189 conf:(0.77)

66. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 patron_prom=P2 246 ==> OK=OK 189 conf:(0.77)

67. Tiempo_Desde_Comienzo=T5800 491 ==> OK=OK 375 conf:(0.76)

68. ID_TareasConsolidadas=TC02 491 ==> OK=OK 375 conf:(0.76)

69. patron_prom=P2 491 ==> OK=OK 375 conf:(0.76)

70. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 491 ==> OK=OK 375 conf:(0.76)

71. Tiempo_Desde_Comienzo=T5800 patron_prom=P2 491 ==> OK=OK 375 conf:(0.76)

72. ID_TareasConsolidadas=TC02 patron_prom=P2 491 ==> OK=OK 375 conf:(0.76)

73. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 patron_prom=P2 491 ==> OK=OK 375 conf:(0.76)

74. ID_Camino=C08 245 ==> OK=OK 186 conf:(0.76)

75. ID_Camino=C08 Tiempo_Desde_Comienzo=T5800 245 ==> OK=OK 186 conf:(0.76)

76. ID_Camino=C08 ID_TareasConsolidadas=TC02 245 ==> OK=OK 186 conf:(0.76)

77. ID_Camino=C08 patron_prom=P2 245 ==> OK=OK 186 conf:(0.76)

78. ID_Camino=C08 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 245 ==> OK=OK 186 conf:(0.76)

79. ID_Camino=C08 Tiempo_Desde_Comienzo=T5800 patron_prom=P2 245 ==> OK=OK 186 conf:(0.76)

80. ID_Camino=C08 ID_TareasConsolidadas=TC02 patron_prom=P2 245 ==> OK=OK 186 conf:(0.76)

81. ID_Camino=C08 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 patron_prom=P2 245 ==> OK=OK 186 conf:(0.76)

82. ID_Camino=C04 ID_EquipoConsolidado=EC03 151 ==> OK=OK 114 conf:(0.75)

83. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_EquipoConsolidado=EC03 151 ==> OK=OK 114 conf:(0.75)

84. ID_Camino=C04 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 151 ==> OK=OK 114 conf:(0.75)

85. ID_Camino=C04 ID_EquipoConsolidado=EC03 patron_prom=P2 151 ==> OK=OK 114 conf:(0.75)

86. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 151 ==> OK=OK 114 conf:(0.75)

87. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_EquipoConsolidado=EC03 patron_prom=P2 151 ==> OK=OK 114 conf:(0.75)

88. ID_Camino=C04 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 patron_prom=P2 151 ==> OK=OK 114 conf:(0.75)

89. ID_Camino=C04 Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 patron_prom=P2 151 ==> OK=OK 114 conf:(0.75)

90. ID_EquipoConsolidado=EC03 294 ==> OK=OK 221 conf:(0.75)

91. Tiempo_Desde_Comienzo=T5800 ID_EquipoConsolidado=EC03 294 ==> OK=OK 221 conf:(0.75)

92. ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 294 ==> OK=OK 221 conf:(0.75)

93. ID_EquipoConsolidado=EC03 patron_prom=P2 294 ==> OK=OK 221 conf:(0.75)

94. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 294 ==> OK=OK 221 conf:(0.75)

95. Tiempo_Desde_Comienzo=T5800 ID_EquipoConsolidado=EC03 patron_prom=P2 294 ==> OK=OK 221 conf:(0.75)

96. ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 patron_prom=P2 294 ==> OK=OK 221 conf:(0.75)

97. Tiempo_Desde_Comienzo=T5800 ID_TareasConsolidadas=TC02 ID_EquipoConsolidado=EC03 patron_prom=P2 294 ==> OK=OK 221 conf:(0.75)