ANÁLISIS PARA PREDICCIÓN DE VENTAS UTILIZANDO MINERÍA DE DATOS EN ALMACENES DE VENTAS DE GRANDES SUPERFICIES.

JOSÉ ANTONIO GARCÍA BERMÚDEZ
ÁNGELA MARÍA ACEVEDO RAMIREZ

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERIAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PEREIRA
2010

ANÁLISIS PARA PREDICCIÓN DE VENTAS UTILIZANDO MINERÍA DE DATOS EN ALMACENES DE VENTAS DE GRANDES SUPERFICIES.

JOSÉ ANTONIO GARCÍA BERMÚDEZ
ÁNGELA MARÍA ACEVEDO RAMIREZ

TRABAJO DE GRADO PARA OPTAR AL TÍTULO DE INGENIERO DE SISTEMAS Y COMPUTACIÓN

DIRECTOR
JORGE IVÁN RIOS PATIÑO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERIAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PEREIRA
2010

NOTA DE ACEPTACIÓN

_____
_____
_____
_____
_____
_____

_____
FIRMA DEL PRESIDENTE DEL JURADO

_____
FIRMA DEL JURADO

_____
FIRMA DEL JURADO

PEREIRA, 2010.

TABLA DE CONTENIDO.

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág

LISTA DE IMÁGENES

Pág

LISTA DE ANEXOS

INTRODUCCIÓN

MARCO PROBLÉMICO

Desde hace algunas décadas, el hombre se ha visto en la necesidad de administrar sus actividades la mayoría de éstas comerciales, como lo es el uso que le da al dinero tanto en el hogar como a nivel empresarial, por lo tanto le es necesario almacenar un historial de algunos o la mayoría de sus actividades comerciales, lo que lo obliga a llevar de manera ordenada el cómo y en qué ha gastado su dinero, hasta el punto de ser necesario contar con una persona que se dedique a administrar, almacenar y vigilar dichas actividades a nivel empresarial.

Con el paso del tiempo se ha visto que para dar una adecuada administración de todas esas actividades y con el fin de evitar muchos conflictos, en la mayoría de los lugares como por ejemplo los hospitales se realizan historial de visitas, entradas y salidas de pacientes; en las estaciones de policía se registran con hora y fecha exactas de los hechos sucedidos; en almacenes grandes se registran las transacciones en facturas con fecha de compra y en algunos casos con nombre del cajero, entre otros ejemplos; por lo que se comienza a formar una generación masiva de datos los cuales llevan a la creación de almacenes o bodegas de datos, algunos con un crecimiento tan exagerado que hasta para las consultas realizadas por lenguajes como SQL es imposible lograr resultados eficientes.

A nivel comercial se puede observar que las empresas logran la recolección de grandes volúmenes de información acerca de su actividad, tales como compras, ventas, inventarios, entre otros, de los cuales algunos de estos datos serán usados y otros se acumularán hasta inclusive llegar a perderse por falta de actualidad o cambio en las políticas de manejo de datos. Para darle un poco de utilidad a esta información se han aplicado diversos modelos y técnicas especialmente desarrollados en el campo de la *Minería de Datos* por medio de los cuales es posible describir el movimiento de los inventarios así como encontrar posibles relaciones que se puedan dar entre determinados productos.

1. JUSTIFICACIÓN

La recolección y almacenamiento de datos ha sido una de las tareas más comunes en todo tipo de empresas, puesto que se hace necesario contar con un histórico de los movimientos comerciales que realizan en una organización para poder llegar a controlar dichos movimientos, pero hasta ahora en muchas de estas empresas este control se lleva de una manera muy arcaica o con métodos que no son muy efectivos, por lo que los resultados pueden no ser los esperados.

Sabiendo que la *Minería de Datos* se fundamenta en la búsqueda de patrones dentro de grandes bases de datos, utilizando diversos métodos tanto de estadística como de inteligencia artificial, haciendo uso de recursos informáticos y tecnológicos, en el presente proyecto se busca aprovechar los beneficios de la misma con el fin de extraer información e inclusive conocimiento oculto en los datos con el fin de apoyar la toma de decisiones en una organización.

Cuando a los datos previamente recolectados y almacenados se les da un trato adecuado, es posible aplicar sobre éstos diversas metodologías, entre las cuales se encuentran las técnicas de *Minería de Datos*, de tal forma que éstas permitan conocer el comportamiento de los inventarios o las posibles relaciones que se presenten entre dos o más productos. La aplicación de este modelo ayuda de tal forma que se puede encontrar a partir de los datos información que hasta el momento había sido desconocida, además de que dicha información obtenida ayuda en la toma de decisiones o al desarrollo de algún proceso a nivel empresarial.

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Utilizar la *Minería de Datos* haciendo uso de la plataforma *RapidMiner*, para aplicar un modelo de predicción de ventas sobre un conjunto de datos seleccionados de una gran superficie de venta, con el fin de encontrar relaciones entre dos o más productos.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar la gestión necesaria para obtener una base de datos que permita la aplicación de alguna de las técnicas de Minería de Datos.
- Determinar la técnica a utilizar sobre la base de datos de acuerdo a los datos que ésta contenga.
- Realizar las funciones de manipulación, selección y procesamiento de los datos.
- Seleccionar una técnica de *Minería de Datos* que permita descubrir asociaciones entre dos o más productos.
- Aplicar una técnica de *Minería de Datos* que permita descubrir asociaciones o correlaciones entre productos que se venden juntos.
- Validar la técnica, comprobando que ésta se ajusta apropiadamente a los requerimientos del problema planteado.
- Dar una breve explicación de los resultados obtenidos y el por qué de los mismos.

3. MARCO REFERENCIAL

3.1 TÉCNICAS DE ALMACENAMIENTO DE DATOS

La mayoría de las decisiones que se toman en una empresa u organización, se hace con base en la información obtenida de los datos que se tienen almacenados sobre la actividad de dicha organización; generalmente éste almacenamiento está consolidado como una base de datos en primera instancia, la cual puede ser transformada en una bodega de datos para facilitar la toma de decisiones, pero inclusive y mejor aún se puede hacer una búsqueda inteligente de patrones o tendencias con una nueva técnica llamada *Minería de Datos*.

Para no dejar estos conceptos tan difusos, a continuación se hará una descripción de lo que son y cómo se están utilizando en éste campo de crecimiento masivo de los datos.

3.1.1 BASES DE DATOS

Una base de datos, es el almacenamiento organizado de datos que tienen una dependencia y que han sido recolectados y explotados por una organización o empresa en particular, dicha explotación se hace con programas creados para la manipulación de la misma.

El término base de datos fue escuchado por primera vez en 1.963 en un simposio realizado en Estados Unidos California, donde se definió como “un conjunto de información relacionada que se encuentra agrupada ó estructurada”[1]; aunque realmente no se puede afirmar que sea un conjunto de información ya que en realidad son datos, los cuales como tal no son información a menos que se les intervenga para que lo sean, pero por ser la primera vez que se utilizó, se puede decir que es aceptable tal definición.

Para dar una información más detallada sobre las bases de datos es bueno hacer énfasis en cómo surgieron y para qué han sido utilizadas, se comenzará por hacer saber que éstas no han sido como hoy se presentan ya procesadas en computador, sino que esta información se tenía almacenada en papel como en archivos o bibliotecas por ejemplo; en las cuales se encuentra información y está organizada de acuerdo a unas métricas utilizadas por los archivistas/bibliotecarios; en las empresas, los datos que almacenaban eran demasiados y a la hora de

[1]CAMPBELL, Mary. base IV Guía de Autoenseñanza. España. Editorial McGraw Hill – Interamericana. 1990. pp110/111,121/122,161,169, 179-191/192. (4 Mar 2009)

necesitar hacer uso de estos podría resultar engorroso dependiendo del orden con el que los tuviesen guardados; quizás es a raíz de este problema en cuanto a tiempo, eficiencia y espacio que surgen las bases de datos en formato digital, las cuales solucionan estos problemas puesto que es mejor el tiempo de respuesta de una máquina que de una persona, los datos son los que se requieren y se ahorra espacio y dinero en papelería; además tienen ciertas características como reducir la redundancia de datos, evitar inconsistencias, se pueden crear restricciones de seguridad como acceso a ciertas personas, acceso concurrente por parte de múltiples usuarios, respaldo y recuperación, y su gestión en cuanto al almacenamiento es mejor.

Otra de las definiciones para bases de datos la da el autor Daniel Cohen[2] quien dice que:

"Se define una base de datos como una serie de datos organizados y relacionados entre sí, los cuales son recolectados y explotados por los sistemas de información de una empresa o negocio en particular.**".**

Desde el punto de vista informático, la base de datos es un sistema formado por un conjunto de datos almacenados en discos que permiten el acceso directo a ellos y un conjunto de programas que manipulen ese conjunto de datos.

Cada base de datos se compone de una o más tablas que guarda un conjunto de datos. Cada tabla tiene una o más columnas y filas. Las columnas guardan una parte de la información sobre cada elemento que se requiera guardar en la tabla, cada fila de la

3.1.1.1 CARACTERÍSTICAS

Entre las principales características de los sistemas de base de datos se pueden mencionar:

- Independencia lógica y física de los datos.
- Redundancia mínima.
- Acceso concurrente por parte de múltiples usuarios.
- Integridad de los datos.

[2] Cohen Karen Daniel. (1996). Sistemas de información para la toma de decisiones. México. McGraw-Hill. 243p. (4 Mar 2009)

- Consultas complejas optimizadas.
- Seguridad de acceso y auditoria.
- Respaldo y recuperación.
- Acceso a través de lenguajes de programación estándar.

Como se puede ver, las características que muestran las bases de datos dan el respaldo para que se le pueda aplicar en este caso la *Minería de Datos* y pueda arrojar el resultado que se espera.

3.1.1.2 VENTAJAS DE LAS BASES DE DATOS

Utilizar bases de datos en este proyecto es primordial, ya que por medio de los datos que proporcionarán, se hará una conexión directa y efectiva para aplicar la *Minería de Datos* y su correspondiente herramienta que en éste caso es *RapidMiner*, por tal motivo a continuación se mostrará qué ventajas tiene el uso de éstas para este proyecto.

El autor Daniel Cohen muestra las grandes ventajas[3] que dan confiabilidad para tomar la base datos como base en la utilización de *Minería de Datos*:

- **Control Sobre la Redundancia de Datos:**

Los sistemas de ficheros almacenan varias copias de los mismos datos en ficheros distintos. Esto hace que se desperdicie espacio de almacenamiento, además de provocar la falta de consistencia de datos.

En los sistemas de bases de datos todos los registros están integrados, por lo que no se almacenan varias copias de los mismos datos.

- **Consistencia de Datos:**

Eliminando o controlando las redundancias de datos se reduce en gran medida el riesgo de que haya inconsistencias. Si un dato está almacenado una sola vez, cualquier actualización se debe realizar sólo una vez, y está disponible para todos los usuarios inmediatamente. Si un dato está duplicado y el sistema conoce esta redundancia, el propio sistema puede encargarse de garantizar que todas las copias se mantienen consistentes.

[3] Ibídem

- **Datos Compartidos:**

En los sistemas de ficheros, los ficheros pertenecen a las personas o a los departamentos que los utilizan. Pero en los sistemas de bases de datos, la base de datos pertenece a la empresa y puede ser compartida por todos los usuarios que estén autorizados.

- **Mantenimiento de Estándares:**

Gracias a la integración es más fácil respetar los estándares necesarios, tanto los establecidos a nivel de la empresa como los nacionales e internacionales. Estos estándares pueden establecerse sobre el formato de los datos para facilitar su intercambio, pueden ser de documentación, procedimientos de actualización y también reglas de acceso.

- **Mejora en la Integridad de Datos:**

La integridad de la base de datos se refiere a la validez y la consistencia de los datos almacenados. Normalmente, la integridad se expresa mediante restricciones o reglas que no se pueden violar. Estas restricciones se pueden aplicar tanto a los datos, como a sus relaciones, y es el Sistema de Gestión de Base de Datos (SGBD) quien se debe encargar de mantenerlas.

- **Mejora en la Seguridad:**

La seguridad de la base de datos es la protección de la base de datos frente a usuarios no autorizados. Sin unas buenas medidas de seguridad, la integración de datos en los sistemas de bases de datos hace que éstos sean más vulnerables que en los sistemas de ficheros.

- **Mejora en la Accesibilidad a los Datos**

Muchos SGBD proporcionan lenguajes de consultas o generadores de informes que permiten al usuario hacer cualquier tipo de consulta sobre los datos, sin que sea necesario que un programador escriba una aplicación que realice tal tarea.

- **Mejora en la Productividad**

El SGBD proporciona muchas de las funciones estándar que el programador necesita escribir en un sistema de ficheros. A nivel básico, éste proporciona todas las rutinas de manejo de ficheros típicas de los programas de aplicación.

El hecho de disponer de estas funciones permite al programador centrarse mejor en la función específica requerida por los usuarios, sin tener que preocuparse de los detalles de implementación de bajo nivel.

- **Mejora en el Mantenimiento:**

En los sistemas de ficheros, las descripciones de los datos se encuentran inmersas en los programas de aplicación que los manejan.

Esto hace que los programas sean dependientes de los datos, de modo que un cambio en su estructura, o un cambio en el modo en que se almacena en disco, requiere cambios importantes en los programas cuyos datos se ven afectados.

Sin embargo, los SGBD separan las descripciones de los datos de las aplicaciones. Esto es lo que se conoce como independencia de datos, gracias a la cual se simplifica el mantenimiento de las aplicaciones que acceden a la base de datos.

- **Aumento de la Concurrencia:**

En algunos sistemas de ficheros, si hay varios usuarios que pueden acceder simultáneamente a un mismo fichero, es posible que el acceso interfiera entre ellos de modo que se pierda información o se pierda la integridad. La mayoría de los SGBD gestionan el acceso concurrente a la base de datos y garantizan que no ocurran problemas de este tipo.

- **Mejora en los Servicios de Copias de Seguridad:**

Muchos sistemas de ficheros dejan que sea el usuario quien proporcione las medidas necesarias para proteger los datos ante fallos en el sistema o en las aplicaciones. Los usuarios tienen que hacer copias de seguridad cada día, y si se produce algún fallo, utilizar estas copias para restaurarlos.

En este caso, todo el trabajo realizado sobre los datos desde que se hizo la última copia de seguridad se pierde y se tiene que volver a realizar. Sin embargo, los SGBD actuales funcionan de modo que se minimiza la cantidad de trabajo perdido cuando se produce un fallo.

Teniendo en cuenta la explicación dada anteriormente ante las ventajas de las bases de datos se puede asegurar que se puede emplear con confiabilidad; aunque esta herramienta se muestra con tan buenas posibilidades de trabajo, también tiene sus desventajas, las cuales se mostrarán a continuación.

3.1.1.3 DESVENTAJAS DE LAS BASES DE DATOS

- **Complejidad**

Los SGBD se componen de un conjunto de programas y funcionalidades que debido a la cantidad de operaciones y la capacidad de cómputo de los mismos, se convierte en un producto de complejo funcionamiento y es por eso que para el correcto desempeño de dichas funcionalidades se exige la aplicación de procedimientos altamente especializados, por lo cual las personas encargadas de su mantenimiento requieren de conocimientos altamente especializados y específicos.

- **Deducción de Información Específica**

Los SGBD actuales carecen de funcionalidades que permitan definir Reglas Deductivas y Activas que permitan modelar directamente los datos para la deducción, inferencia y obtención de información precisa derivada de dichos datos.

- **Vulnerable a los Fallos**

El hecho de que todo esté centralizado en el SGBD hace que el sistema sea más vulnerable ante los fallos que puedan producirse. Es por ello que deben tenerse copias de seguridad (Backup).

3.1.1.4 TIPOS DE CAMPOS

Cada Sistema de Base de Datos posee tipos de campos que pueden ser similares o diferentes. Entre los más comunes podemos nombrar:

- **Numérico:** entre los diferentes tipos de campos numéricos podemos encontrar enteros “sin decimales” y reales “decimales”.

- **Booleanos:** poseen dos estados: Verdadero “Si” y Falso “No”.

- **Memos:** Estos campos son particularmente adecuados para dotar a cada registro de la tabla de un lugar para escribir todo tipo de comentarios. No es necesario definir su longitud, ya que la misma se maneja de manera automática, extendiéndose a medida que se le agrega información.

- **Fechas:** almacenan fechas facilitando posteriormente su explotación. Almacenar fechas de esta forma posibilita ordenar los registros por fechas o calcular los días entre una fecha y otra.

- **Alfanuméricos:** contienen cifras y letras. Presentan una longitud limitada (255 caracteres).

- **Autoincrementables:** son campos numéricos enteros que incrementan en una unidad su valor para cada registro incorporado. Su utilidad resulta: Servir de identificador ya que resultan exclusivos de un registro.

Estos campos muestran en qué áreas y de qué forma se están sometiendo las bases de datos en el desarrollo de la *Minería de Datos* y en qué manera ésta puede ser útil.
Teniendo en cuenta lo anteriormente presentado, se puede deducir que las Bases de Datos y los SGBD son ideales para el correcto almacenamiento de los datos en una organización, sin embargo dichos SGBD exigen la aplicación de procedimientos altamente complejos para la extracción de información que permanece oculta en los datos, por lo cual para lograr el objetivo planeado se aplicará el Modelado en *Minería de Datos* ya que permite la obtención de información deseada de una forma más eficiente.

3.1.2 BODEGAS DE DATOS

También conocidas como Almacenes de Datos, es un concepto relativamente nuevo, orientado al manejo de grandes volúmenes de datos, provenientes de diversas fuentes, de muy diversos tipos. Estos datos cubren largos períodos de tiempo, lo que trae consigo que se tengan diferentes esquemas de las bases de datos fuentes. La concentración de esta información está orientada a su análisis para apoyar la toma de decisiones oportunas y fundamentadas.
Una Bodega de Datos, es un conjunto de datos integrados orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales sirven de soporte en el proceso de toma de decisiones de la administración[4].

Su nombre, Bodega de Datos de ahora en adelante BGDs., se asocia con una colección de datos de gran volumen, provenientes de sistemas en operación y otras fuentes, después de aplicarles procesos de análisis, selección y transferencia de datos seleccionados. Su misión consiste en, a partir de estos

[4] Duque Méndez, Néstor Darío. Bases de Datos. Universidad Nacional de Colombia. (2005). http://www.virtual.unal.edu.co/cursos/sedes/manizales/4060029/lecciones/cap8-1.html. (9 Mar 2009).

datos y apoyado en herramientas sofisticadas de análisis, obtener información útil para el soporte a la toma de decisiones.

En síntesis una BGDs. es una gran colección de datos que recoge información de múltiples sistemas fuentes u operacionales dispersos, y cuya actividad se centra en la *Toma de Decisiones* -es decir, en el análisis de la información- en vez de en su captura. Una vez reunidos los datos de los sistemas fuentes se guardan durante mucho tiempo, lo que permite el acceso a datos históricos; así los almacenes de datos proporcionan al usuario una interfaz consolidada única para los datos, lo que hace más fácil escribir las consultas para la toma de decisiones[5].

Según lo anterior se puede observar como las bodegas de datos, están más orientadas para ayudar al usuario en la toma de decisiones, diferente a las bases de datos las cuales tienen como principal objetivo la captura y almacenamiento de datos, siendo esto lo necesario para el desarrollo de éste proyecto.

3.1.2.1 OBJETIVOS DE LAS BODEGAS DE DATOS

A continuación se muestran los objetivos[6] que las BGDs. tienen para la utilización de datos de una empresa:

- Proveer una visión única de los clientes en toda la empresa.
- Poner tanta información comercial como sea posible en manos de tantos usuarios diferentes como sea posible.
- Mejorar el tiempo de espera que insumen los informes habituales.
- Monitorear el comportamiento de los clientes.
- Predecir compras de productos.
- Mejorar la capacidad de respuesta a problemas comerciales.
- Incrementar la precisión de las mediciones.

[5] Velazco, Roberto Hernando. Almacenes de datos (Datawarehouse). (2007). http://www.rhernando.net/modules/tutorials/doc/bd/dw.html. (7 Mar 2009)
[6] Bressán Griselda E. (2003). Lic. en sistemas de información. Almacenes de datos y Minería de Datos. Trabajo monográfico de adscripción. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm. (12 Mar. 2009)

- Aumentar la productividad.
- Incrementar y distribuir las responsabilidades.

Con estos objetivos se hace más preciso el alcance y responsabilidad que las BGDS. tienen sobre los datos y la empresa para la cual son aplicadas.

3.1.2.2 UTILIDAD DE LAS BODEGAS DE DATOS

Teniendo en cuenta lo mencionado se puede decir que las BGDS. pueden ser utilizadas para:

- Manejo de relaciones de marketing.
- Análisis de rentabilidad.
- Reducción de costos.

Como se puede ver las BGDs. se utilizan para ayudar en la toma de decisiones basada en la información que muestran de manera clara.

3.1.2.3 DIFERENCIAS ENTRE BASE DE DATOS Y LAS BODEGAS DE DATOS

A continuación se presenta una serie de diferencias que muestran desde diversos puntos de vista, desde que perspectiva funciona cada almacén.

BASE DE DATOS OPERACIONAL	BODEGA DE DATOS
Datos Operacionales	Datos del Negocio para Información
Orientado a la Aplicación	Orientado al Sujeto
Actual	Actual + Histórico
Detallada	Detallada + Resumida
Cambia Continuamente	Estable

Tabla No 1. Diferencias entre Bases de Datos y Bodegas de Datos

Como se pudo observar en la tabla anterior (Tabla No 1.)[7], las BGDs., aportan información de mayor valor que las bases de datos, ya que dicha información es más completa y detallada. Aunque las BGDs. serían de gran utilidad para este trabajo, se sigue manejando el prospecto de bases de datos ya que permite trabajar directamente sobre los datos en el estado en que han sido almacenados.

3.1.2.4 CARACTERÍSTICAS DE LAS BODEGAS DE DATOS

Con lo mencionado anteriormente en el texto, las BGDs. muestran su capacidad para la integración, ejecución, agrupamiento, análisis y control de los datos, ahora se mencionará que características[8] rodean los almacenes de datos las cuales muestran que son:

- **Organizados en torno a temas:** La información se clasifica con base a los aspectos que son de interés para la empresa.

- **Integrado:** Es el aspecto más importante. La integración de datos consiste en convenciones de nombres, codificaciones consistentes, medida uniforme de variables, etc.

- **Dependiente del tiempo:** Esta dependencia aparece de tres formas:
  - ✓ La información representa los datos sobre un horizonte largo de tiempo.
  - ✓ Cada estructura clave contiene (implícita o explícitamente) un elemento de tiempo (día, semana, mes, etc.).
  - ✓ La información, una vez registrada correctamente, no puede ser actualizada.

- **No volátil:** El Almacén de Datos sólo permite cargar nuevos datos y acceder a los ya almacenados, pero no permite ni borrar ni modificar los datos.

Estas características muestran una BGDs. organizada y centrada en realizar un buen trabajo en el uso de los datos, la cual con el paso del tiempo mostrará la elaboración de un buen desempeño según sus funciones.

[7] Ibídem
[8] Ibídem

3.1.2.5 FUNCIONALIDADES DE LAS BODEGAS DE DATOS

Con lo anteriormente mencionado se hace una fusión que lleva a implementar diferentes funcionalidades que haciendo uso de las BGDs, facilita la creación y explotación de los datos almacenados en ella, y por consiguiente lo hacen más eficaz en el momento de su uso.

Las BGDs. incluyen funcionalidades como:

- Integración de bases de datos heterogéneas (relacionales, documentales, geográficas, archivos, etc.)
- Ejecución de consultas complejas no predefinidas visualizando el resultado en forma gráfica y en diferentes niveles de agrupamiento y totalización de datos.
- Agrupamiento y desagrupamiento de datos en forma interactiva.
- Análisis del problema en términos de dimensiones.
- Control de calidad de datos.

Teniendo en cuenta las funciones permiten observar las diversas y amplias temáticas que se le pueden aplicar a las BGDs. y la efectividad que sus datos puedan arrojar debido a la variedad de funciones que emplea para su desarrollo.

3.1.2.6 ARQUITECTURA DE LAS BODEGAS DE DATOS

Como se ha mencionado anteriormente, y con lo que se mostrará a continuación, se hace saber por qué al usar las BGDs. se estaría incrementando de forma innecesaria el trabajo en este proyecto, ya que el tiempo que se estaría utilizando en crear una arquitectura de BGDs., podría emplearse en otro campo del proyecto; a continuación se mostrará la arquitectura que emplea un almacén de datos.
La estructura básica de la arquitectura de las BGDs. incluye[9]:

1. **Datos Operacionales**: Origen de datos para el componente de almacenamiento físico de las BGDs.
2. **Extracción de datos**: Selección sistemática de datos operacionales usados para formar parte de las BGDs.

[9] Ibídem

3. **Transformación de datos**: Procesos para resumir y realizar cambios en los datos operacionales.

4. **Carga de datos**: Inserción de datos en las BGDs.

5. **Almacén**: Almacenamiento físico de datos de la arquitectura de las BGDs.

6. **Herramienta de acceso**: Herramientas que proveen acceso a los datos.

Como se ha visto, para generar una arquitectura de BGDs. se emplea una cantidad de datos, los cuales se deben presentar en una forma muy detallada y seleccionada, la cual requiere el empleo de una cantidad de tiempo y datos que quizá para el alcance de este proyecto, sea innecesario utilizar, por eso se sigue con la ayuda de la *Minería de Datos* accediendo a los datos directamente desde la Base de Datos.

3.2 MINERIA DE DATOS

3.2.1 INTRODUCCIÓN

A lo largo del funcionamiento de una empresa se acumulan grandes cantidades de datos que son almacenados, algunos de ellos serán usados, otros se acumularán hasta perderse por falta de actualidad o por cambios en las políticas de manejo de los mismos.
Con el desarrollo de los sistemas de cómputo que se ha incrementado considerablemente en los últimos 20 (veinte) años[10], las empresas han tenido la capacidad de almacenar grandes volúmenes de datos históricos sobre las operaciones diarias en todas las áreas de la organización, con el fin de satisfacer las necesidades propias de la empresa, pero en la mayoría de las organizaciones se presenta exceso de registros, por lo que se hace más complicado encontrar información específica y verdaderamente significativa que permita obtener conocimiento que hasta el momento permanecía oculto y el cual brinda una visión más completa y clara de la situación operacional de la empresa ayudando a mejorar la forma en la que se toman las decisiones en la organización[11].

La creciente necesidad de información ha hecho que en las empresas se diseñen sistemas de información y de apoyo a la toma de decisiones, que han tenido como objetivo primordial proveer de toda la información necesaria a los ejecutivos de alto nivel para apoyarlos en la toma de decisiones, además de que les permite tener acceso rápido y efectivo a la información compartida y crítica del negocio; sin embargo actualmente la demanda de las empresas en cuanto a la información, va mas allá de simples consultas, o reportes consolidados.

Como respuesta a dichos requerimientos, se han creado técnicas de almacenamiento y análisis de información, haciendo uso de diversas áreas de conocimiento como la estadística, inteligencia artificial, computación gráfica, bases de datos y el procesamiento masivo[12], que han servido como fundamento de nuevas técnicas de análisis como la *Minería de Datos*, que ha facilitado el proceso de extraer información de los grandes volúmenes de datos, revelando conocimiento innovador a las organizaciones, permitiendo conocer de una forma más detallada el comportamiento de variables sumamente importantes para las

[10] Molina, Luis Carlos. (2002). Data Mining: Torturando los datos hasta que confiesen. http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html. (4 Mar. 2009).
[11] Larrieta, María Isabel Ángeles y Santillán Angélica María. (2004). Minería de Datos: Concepto, características, estructura y aplicaciones. http://www.ejournal.unam.mx/rca/190/RCA19007.pdf (10 Mar.2009).
[12] Vallejos, Sofía. (2006). Minería de Datos. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf (7 Mar.2009).

empresas, como lo son inventarios, ventas, el comportamiento de los clientes, entre otros factores.

3.2.2 HISTORIA

Las tecnologías de la información han facilitado los procesos administrativos de las organizaciones, ya que mediante éstas, las empresas han podido almacenar de manera segura todos los datos referentes a las funciones que desempeñan, entre las cuales se encuentran, las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, entre otras muchas funciones que se llevan a cabo a diario en las empresas, que representan la memoria de la organización.

Una vez satisfechas éstas necesidades, surgen un nuevo grupo de requerimientos relacionados con información precisa sobre los sistemas de las organizaciones que exigen pasar a la acción inteligente sobre los datos para extraer la información oculta que representan dichos datos, y que puede servir de base para la toma de decisiones[13], éste es el objetivo de la *Minería de Datos*.

La *Minería de Datos* tiene sus raíces básicamente en dos áreas del conocimiento: la primera y más grande en la cual tiene sus cimientos, es la estadística clásica, la cual cuenta con diversos conceptos como la distribución estándar, la varianza, análisis de *clustering*[14], entre muchos otros, los cuales juegan un papel muy importante en el proceso de la misma, ya que éstos, brindan gran parte de la fundamentación bajo la cual muchos de sus modelos han sido construidos.

La segunda área de conocimiento que hace parte de la fundamentación de la *Minería de Datos* es la inteligencia artificial, ésta disciplina procura aplicar procesamiento lógico a través de algoritmos genéticos, redes neuronales, árboles de decisión, entre otros, a diversos problemas estadísticos; para poder aplicar dicho procesamiento, es necesario contar con gran capacidad de poder de cómputo lo cual no fue posible hasta comienzos de los 80's cuando los computadores empezaron a ofrecer mayor capacidad de procesamiento a precios más asequibles, permitiendo que se empezaran a generar diferentes aplicaciones de éste tipo, que en un principio tuvieron fines científicos y de investigación[15].

A pesar que las técnicas de análisis estadístico permiten conocer información que puede ser útil, no permiten identificar relaciones cualitativas entre los datos, que podrían llegar a ser bastante significativas para las empresas.

---

[13] Aranguren, Silvia Mónica y Muzachiodi, Silvia Liliana.(2003). Implicancias del Data Mining. http://www.fceco.uner.edu.ar/extinv/publicdocent/sarangur/pdf/introduccion.pdf (4 Mar.2009)

[14] Agrupación

[15] "A Brief History of Data Mining. Data mining software". (2006). http://www.data-mining-software.com/data_mining_history.htm (5 Mar. 2009)

Para poder obtener de los datos cierto tipo de información que aporte conocimiento altamente valioso para las organizaciones, se requiere disponer también de técnicas y métodos de análisis inteligente que aunque todavía no han sido perfectamente establecidos, están siendo desarrollados dentro de la inteligencia artificial con el fin de descubrir dicha información que se encuentra oculta en las bases de datos de la organizaciones.

El concepto de *Minería de Datos* fue usado por primera vez en los años sesenta, cuando los estadísticos manejaban términos como *data fishing*[16], *data mining*[17] o *data archaealogy*[18] con la idea de encontrar correlaciones entre los datos sin una hipótesis previa, en bases de datos imprecisas e inconsistentes. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de *la Minería de Datos.* A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; en 2002 ya existían más de 100 empresas en el mundo ofreciendo alrededor de 300 soluciones. En la actualidad, las listas de discusión sobre este tema, las forman investigadores de más de ochenta países[19].

Actualmente el proceso de *Minería de Datos*, al estar compuesto por varias etapas, hace el uso de diferentes disciplinas, como la visualización, la computación de alto rendimiento, la estadística, modelos matemáticos y la inteligencia artificial, los cuales le permiten obtener mejores resultados a la hora de extraer información de las bases de datos[20], al igual que existen gran variedad de aplicaciones o herramientas comerciales que además de ser muy poderosas ya que cuentan con un sinfín de utilerías que facilitan el desarrollo de un proyecto, éstas pueden complementarse entre sí para poder arrojar resultados satisfactorios[21] que entreguen información altamente significativa para la toma de decisiones en una organización.

3.2.3 DEFINICIÓN

La *Minería de Datos* es un proceso no trivial que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones, relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tiene

[16] Pesca de Datos. http://www.businesspme.com/uk/articles/technologies/13/Data-dredging,-data-fishing.html
[17] Minería de Datos. http://en.wikipedia.org/wiki/Data_mining
[18] Arqueología de Datos. http://en.wikipedia.org/wiki/Data_archaeology
[19] Molina, Luis Carlos.(2002). Data Mining: Torturando los datos hasta que confiesen. http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html. (4 Mar. 2009)
[20] Christen, Peter.(2005) A very short introduction to… Data Mining. http://datamining.anu.edu.au/talks/2005/datamining-comp2340-2005.pdf (9 Mar. 2009)
[21] Vallejos, Sofía.(2006). Minería de Datos. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf (7 Mar.2009)

una explicación que pueden descubrirse mediante diversos métodos de esta técnica[22].

Integra un conjunto de áreas como lo son la estadística y la inteligencia artificial para identificar información valiosa en grandes volúmenes de datos, con el propósito que dicha información hallada en el proceso, aporte un sesgo hacia la toma de decisiones a nivel empresarial; debido a esta característica, el proceso de *Minería de Datos* incluye dos componentes para servir de apoyo a la toma de decisiones: el componente de análisis de verificación y el de descubrimiento.

El análisis de verificación permite obtener conclusiones basadas en el comportamiento pasado, la *Minería de Datos* con enfoque en el descubrimiento ayuda a descubrir nuevas oportunidades de negocio. El análisis de verificación permite confirmar o rechazar los descubrimientos obtenidos con el nuevo enfoque. Por lo cual se puede decir que la *Minería de Datos* es un proceso que ayuda a descubrir información útil desde las bases de datos y por lo tanto es una herramienta relacionada directamente al negocio.

La *Minería de Datos* se encarga de extraer relaciones fundamentales en el comportamiento de los clientes, productos e incluso proveedores de una empresa, invirtiendo la dinámica del método científico, ya que se concentra en llenar la necesidad de descubrir el por qué, para luego predecir y pronosticar las posibles acciones a tomar con cierto factor de confianza para cada predicción.

En el proceso de *Minería de Datos*, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son como son. Luego entonces, se valida si la hipótesis inspirada por los datos en los mismos, será numéricamente significativa, pero experimentalmente inválida. De ahí que la *Minería de Datos* debe presentar un enfoque exploratorio, y no confirmador. Usarla para confirmar las hipótesis formuladas no es recomendable, pues se estaría haciendo una inferencia poco válida[23].

La *Minería de Datos* deriva patrones y tendencias que existen en los datos. Estos patrones y tendencias se pueden recopilar y definir como un modelo de la misma

Los modelos de *Minería de Datos* se pueden aplicar, entre otras muchas, a situaciones empresariales como las siguientes:

- Predicción de ventas.

[22] Larrieta, María Isabel Ángeles y Santillán Angélica María.(2004). Minería de Datos: Concepto, características, estructura y aplicaciones. http://www.ejournal.unam.mx/rca/190/RCA19007.pdf (10 Mar.2009)
[23] Bressán Griselda E. (2003). Lic. en sistemas de información. Almacenes de datos y Minería de Datos. Trabajo monográfico de adscripción. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm (12 Mar. 2009).

- Clasificación y estratificación de Clientes.
- Determinar relaciones entre productos que generalmente se venden juntos.
- Buscar secuencias en el orden en que los clientes agregan productos a una cesta de compra.

3.2.4 ANTECEDENTES DE LA MINERÍA DE DATOS

La extracción de conocimiento a partir de datos, tiene como objetivo descubrir patrones que, entre otras cosas, deben ser válidos, novedosos, interesantes y, en última instancia, comprensibles. Los seres humanos tienen una capacidad innata de ver patrones a su alrededor. Las técnicas de *Minería de Datos* han querido emular, estas capacidades de aprendizaje para deducir información significativa a partir de grandes volúmenes de datos en diversas áreas y disciplinas[24].

Desde sus inicios, la *Minería de Datos* ha sido utilizada en diversas áreas de conocimiento, en las cuales ha sido de gran ayuda en la extracción de información a partir de grandes volúmenes de datos, se han desarrollado investigaciones y proyectos en casi todas las ramas de la ciencia, como la astronomía, medicina, mercadotecnia, entre otros, obteniendo resultados satisfactorios tanto a nivel científico como empresarial.

Dichos proyectos e investigaciones han contribuido enormemente al desarrollo, evolución y especialización de los algoritmos y métodos de extracción de conocimiento, permitiendo obtener cada vez con mayor exactitud, información altamente significativa tanto para una investigación académica como para una organización o empresa.

Los algoritmos y métodos de *Minería de Datos*, han servido de apoyo en la toma de decisiones en una organización, facilitando de ésta forma la labor de administración del negocio, como también ha sido de gran ayuda en el ámbito de las ciencias, en las investigaciones científicas y académicas, permitiendo entender las causas que generan los fenómenos, así como ha sido útil para abordar los problemas desde una perspectiva apropiada para resolver conflictos desde su causa más básica.

La *Minería de Datos* ha posibilitado el desarrollo de proyectos de investigación en un menor tiempo que los métodos tradicionales, logrando alcanzar resultados altamente significativos para los grupos y organizaciones, ya que los resultados

[24] Zamarron. Sanz, Carlos et al. (2008). Aplicación de la Minería de Datos al estudio de las alteraciones respiratorias durante el sueño. http://www.sogapar.org/pneuma/pneuma6/pneuma-n-6-5c.pdf (29 Ag. 2009).

obtenidos arrojan información altamente valiosa, que ha contribuido en investigaciones tanto académicas como experimentales en un área definida de la ciencia.

A continuación se presentan algunas de las áreas en las que se han aplicado la *Minería de Datos* obteniendo resultados altamente significativos[25]:

- **Astronomía:** clasificación de cuerpos celestes.
- **Metereología:** predicción de tormentas, entre otros.
- **Medicina:** caracterización y predicción de enfermedades, probabilidad de respuesta satisfactoria a tratamiento médico.
- **Industria y manufactura:** diagnóstico de fallas.
- **Mercadotecnia:** identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, selección de sitios de tiendas, afinidad de productos, entre otros.
- **Inversión en casas de bolsa y banca:** análisis de clientes, aprobación de préstamos, determinación de montos de crédito, entre otros.
- **Gestión de Riesgos:** Las compañías de seguros y empresas de hipotecas utilizan la de *Minería de Datos* para descubrir los riesgos asociados con los clientes potenciales.
- **Detección de fraudes y comportamientos inusuales:** telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, entre otros.
- **Análisis de canasta de mercado:** para mejorar la organización de tiendas, segmentación de mercado.
- **Rating:** Determinación de niveles de audiencia de programas televisivos.

[25] Bressán Griselda E. (2003). Lic. en sistemas de información. Almacenes de datos y Minería de Datos. Trabajo monográfico de adscripción.
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm (12 Mar. 2009)

3.2.5 FASES DEL PROCESO DE MINERÍA DE DATOS

La aplicación de los algoritmos de *Minería de Datos*, requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen datos erróneos o redundantes. Por otra parte, es necesario interpretar y evaluar los resultados obtenidos.

El proceso completo consta de las siguientes fases[26]:

1. Definición de los objetivos.
2. Preparación de datos.
3. Análisis exploratorio de los datos.
4. Especificación de los métodos.
5. Análisis de los datos.
6. Evaluación de los métodos.
7. Implementación de los métodos.

3.2.5.1 Definición de los Objetivos

Esta fase del proceso de *Minería de Datos* consiste en definir los objetivos del análisis. No siempre es fácil definir el fenómeno que queremos analizar, de hecho, los objetivos del grupo u organización interesado en realizar éste proceso generalmente son claros, pero los problemas de fondo como los que se pretenden abordar a partir de la misma, pueden ser difíciles de traducir en objetivos concretos que deben ser analizados.
Una definición clara del problema y los objetivos que se persiguen son los requisitos previos para iniciar el análisis correctamente. Esta es ciertamente una de las fases más difíciles del proceso, ya que lo establecido en esta fase determina cómo se organizan las fases siguientes, por lo tanto, los objetivos deben ser claros y no debe haber lugar para dudas ni incertidumbres.

[26] Moreno. García, María et al. (2009). Aplicación de técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf (29 Ag. 2009)

3.2.5.2 Preparación de los Datos

Una vez que los objetivos del análisis han sido identificados, es primordial identificar las fuentes de información externas e internas para seleccionar el subconjunto de datos necesario de datos para el análisis. Las fuentes de datos corresponden generalmente a fuentes de datos internas de la organización interesada en el proceso de *Minería de Datos*, ya que de ésta forma los datos pueden ser más fiables para la investigación.

Estos datos también tienen la ventaja de ser el resultado de experiencias y procedimientos de la propia empresa. La fuente de datos ideal es el almacén de datos de la empresa a la que se le realiza el proceso de *Minería de Datos*. En un almacén de datos se recopilan datos históricos que no están sujetos a cambios, lo cual les da una alta fiabilidad a los datos, además de que puede resultar más sencillo extraer porciones de las bases de datos de acuerdo a áreas específicas de la organización.

Los datos utilizados para la realización del presente proyecto, son datos que no están sujetos a cambios ya que son una copia de la base de datos original y no están sujetos a sufrir ningún cambio que si podría tener la misma a lo largo del tiempo.

El primer paso fundamental antes de realizar el análisis de datos, es disponer de una adecuada selección de los mismos, lo que implica que es necesario tener una apropiada representación de los datos que generalmente se encuentran condensados en una tabla que se conoce como matriz de datos.

La matriz de datos se construye según las necesidades analíticas del problema y de los objetivos previamente establecidos. Una vez que una matriz de datos está disponible, es necesario llevar a cabo una limpieza preliminar de los datos, en otras palabras, se realiza un control de calidad de los datos disponibles, que generalmente es conocido como la limpieza de datos o preprocesado.

El preprocesado, es un proceso formal que se usa para resaltar las variables que existen en la matriz de datos, pero que no son adecuadas para el análisis que se desea realizar, el preprocesado de los datos es también un importante control sobre los elementos de las variables y la posible presencia de datos erróneos o redundantes.

Por último, es útil establecer un subconjunto o una muestra de los datos disponibles, esto se debe a que la calidad de la información recogida a partir del análisis completo de todo el conjunto de datos disponibles no siempre es mejor que la información obtenida del análisis sobre las muestras de los mismos datos. De hecho, en la *Minería de Datos*, las bases de datos que usualmente se analizan

son muy grandes, por lo tanto al usar una muestra de los datos se reduce el tiempo de análisis.

Trabajar con muestras permite comprobar la validez del modelo para el resto de los datos, siendo de ésta forma una importante herramienta de diagnóstico, además de que también reduce el riesgo de que el método aplicado presente irregularidades perdiendo su capacidad de generalizar y de diagnosticar sobre los datos.

3.2.5.3 Análisis Exploratorio de los Datos

En el Análisis exploratorio de los datos se puede destacar cualquier anomalía en los mismos, ya que se pueden presentar elementos que son diferentes del resto, estos elementos de los datos no necesariamente deben ser eliminados ya que podrían contener información que es importante para alcanzar los objetivos del análisis.
Un análisis exploratorio de los datos es esencial, ya que permite al analista predecir cuál o cuáles métodos pueden ser más adecuados para aplicar en la siguiente fase del análisis. En esta opción, hay que tener en cuenta la calidad de los datos obtenidos en la fase anterior.

El análisis exploratorio también puede sugerir la necesidad de realizar una nueva extracción de datos porque los datos anteriormente recogidos se pueden considerar insuficientes para alcanzar los objetivos establecidos.

3.2.5.4 Especificación del Método

Existen varios métodos y algoritmos que se pueden aplicar en el proceso de *Minería de Datos*, por lo que es importante tener una clasificación de los métodos existentes.
La elección del método depende del problema en estudio o el tipo de datos disponibles, el proceso de extracción de datos se rige por las aplicaciones, por esta razón, los métodos utilizados se pueden clasificar de acuerdo con el objetivo de los análisis. Se pueden distinguir tres clases principales:

- **Métodos Descriptivos**: Permiten formar grupos de datos rápidamente, también son conocidos como métodos simétricos, no supervisados o indirectos. Las observaciones son generalmente clasificadas en grupos que no son conocidos con anterioridad (análisis de conglomerados, mapas de Kohonen), los elementos de las variables pueden estar conectados entre sí de acuerdo a vínculos desconocidos de antemano (los métodos de

asociación, los modelos log-lineales, los modelos gráficos), de esta manera, todas las variables disponibles son tratados en el mismo nivel y no hay hipótesis de causalidad.

- **Métodos de Predicción**: Su objetivo es describir una o más de las variables en relación con todas las demás, son conocidos como métodos asimétricos, supervisados o directos. Se llevan a cabo mediante la búsqueda de normas de clasificación o de predicción basada en los datos, estas normas nos ayudan a predecir o clasificar el resultado futuro de una o más variables de respuesta o de destino en relación a lo que ocurre en la práctica con los motivos que la causan o bien en relación con las variables de entrada. Los principales métodos de este tipo son los desarrollados en el ámbito de la máquina de aprendizaje, tales como las redes neuronales (perceptrón de multicapa y árboles de decisión), como también lo son modelos estadísticos clásicos, como los modelos de regresión lineal y logística.

- **Métodos Locales**: Su objetivo es identificar las características particulares relacionadas con un subconjunto de la base de datos, los métodos descriptivos y métodos de predicción, son globales más que locales. Ejemplos de métodos locales son las reglas de asociación para el análisis de datos transaccionales y la determinación de observaciones anómalas u *outliers*.

La anterior clasificación de los métodos de *Minería de Datos* es exhaustiva, sobre todo desde el punto de vista funcional y cada método puede ser utilizado como tal o como una etapa en un análisis de varias etapas en el proceso de *Minería de Datos*.

3.2.5.5 El Análisis de Datos

Una vez que el o los métodos a aplicar en el proceso han sido especificados, deben traducirse en algoritmos apropiados para realizar los cálculos que ayudan a sintetizar los resultados que se necesitan de la base de datos disponible. La amplia gama de programas informáticos especializados y no especializados para la *Minería de Datos* significa que para la mayoría de aplicaciones estándar, no es necesario el desarrollo de algoritmos, sin embargo, aquellos que gestionan este proceso deben tener un buen conocimiento de los diferentes métodos, así como las soluciones de software, para que puedan adaptar el proceso a las necesidades específicas de la organización e interpretar correctamente los resultados para la toma de decisiones.

3.2.5.6 Evaluación del Método

Una vez aplicado el método, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber aplicado varios métodos mediante el uso de distintas técnicas, se deben comparar los métodos en busca de aquel que se ajuste mejor al problema. Esta es una verificación de diagnóstico importante sobre la validez del método específico que se aplica a los datos disponibles. Idealmente, los patrones descubiertos deben tener las siguientes cualidades: ser precisos, comprensibles, es decir, inteligibles e interesantes lo que implica que deben ser útiles y novedosos[27].

Es posible que ninguno de los métodos utilizados permitan alcanzar el conjunto de objetivos propuestos de manera satisfactoria, en ese caso será necesario volver y especificar un nuevo método que es más apropiado para el análisis. Al evaluar el rendimiento de un método específico, así como medidas de diagnóstico de un tipo de estadísticas, otras cosas deben ser consideradas, como lo son las restricciones de tiempo, las limitaciones de recursos, y la calidad y disponibilidad de datos.

3.2.5.7 Implementación de los Métodos

La *Minería de Datos* no es sólo un análisis de los datos, es también la integración de los resultados en el proceso de decisión de la empresa u organización. El Conocimiento del negocio, la extracción de las normas y su participación en el proceso de decisión permite pasar de la fase analítica a la producción de un motor de decisión.
Una vez que el modelo ha sido elegido y probado con un conjunto de datos, la regla de clasificación se puede aplicar a toda la población de referencia. Por ejemplo, se puede adquirir la capacidad de distinguir con anterioridad qué clientes son más rentables para la empresa, o también se puede adquirir la habilidad de calibrar las políticas comerciales diferenciadas para los distintos grupos de consumidores, lo que aumenta los beneficios de la empresa.

Después de haber visto los beneficios que se pueden obtener de la *Minería de Datos*, es fundamental implementar correctamente dicho proceso en una organización, con el fin de explotar todo su potencial.

Aunque las fases anteriores se realizan en el orden en que aparecen, el proceso es altamente iterativo, estableciéndose retroalimentación entre los mismos. Además, no todos las fases requieren el mismo esfuerzo, generalmente la fase de

[27] Zamarron. Sanz, Carlos et al. (2008). Aplicación de la Minería de Datos al estudio de las alteraciones respiratorias durante el sueño. http://www.sogapar.org/pneuma/pneuma6/pneuma-n-6-5c.pdf (29 Ag. 2009)

preparación de datos es la más dispendiosa ya que representa aproximadamente el 60 % del esfuerzo total de todo el proyecto[28].

3.2.6 ALGORITMOS DE MINERÍA DE DATOS

La aplicación automatizada de algoritmos de *Minería de Datos* permite detectar patrones en los datos eficientemente y a partir de éstos, derivar información implícita en ellos, de tal forma que se pueda comprobar qué tan útiles son las predicciones que se derivan de los datos para una organización en general.

Dichos algoritmos de *Minería de Datos* se encuentran en continua evolución y se desarrollan como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones.

A continuación se muestra la clasificación de los algoritmos en la *Minería de Datos*,[29] la cual se da en dos grandes categorías:

- **Supervisados o Predictivos:** Los algoritmos supervisados o predictivos predicen el valor de un atributo (*etiqueta*) de un conjunto de datos, conocidos otros atributos (*atributos descriptivos*). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como *aprendizaje supervisado* y se desarrolla en dos fases:

    a) Entrenamiento: Se construye un modelo usando un subconjunto de datos con una etiqueta conocida.
    b) Prueba: Se prueba del modelo sobre el resto de los datos.

- **No Supervisados o de Descubrimiento del Conocimiento:** Descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas.

---

[28] Giudici, Paolo. (2003). Applied Data Mining Stadistical Methods for Bussines and Industry. Chichester. Jhon Wiley & Sons, Inc. 364p.

[29] Moreno. García, María et al. (2009). Aplicación de técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf (29 Ag. 2009)

En la Tabla No 2., se muestran algunas de las técnicas de minería de ambas categorías[30].

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Detección de Desviaciones
Inducción Neuronal	Segmentación
Regresión	Agrupamiento ("*clustering*")
Series Temporales	Reglas de Asociación
	Patrones Secuenciales

Tabla No 2. Clasificación de las técnicas de *Minería de Datos*.

Como se puede observar en la tabla anterior (Tabla No 2.), las reglas de asociación utilizan algoritmos no supervisados o de descubrimiento de información, como lo son el algoritmo a priori, GRI (Inducción generalizada de reglas), FP Growth (crecimiento de patrones frecuentes), entre otros, los cuales se expondrán en otra sección (sección 3.3.6) más adelante en este documento. Debido a que las reglas de asociación tienen como propósito descubrir patrones y tendencias que se presentan en los datos, en el presente proyecto se empleará dicha técnica de *Minería de Datos*, ya que éstas permiten alcanzar los objetivos del presente proyecto satisfactoriamente.

3.2.7 TAREAS DE LA MINERIA DE DATOS

Los algoritmos de *Minería de Datos* realizan en general tareas de predicción de información desconocida que puede estar contenida en los datos, como también puede realizar la labor de describir de patrones de comportamiento de los datos[31].

Muchos de los problemas de tipo intelectual, económico, y de interés comercial se pueden solucionar en términos de las tareas que la *Minería de Datos* está planteada a cumplir; la siguiente lista muestra las tareas más comunes de *Minería de Datos*[32]:

[30] Ibídem.
[31] Larose, Daniel T.(2005).Discovering Knowledge in Data an Introduction to Data Mining. Hoboken, New Jersey. Jhon Wiley & Sons, Inc Publication. 222p.
[32] Berry,Michael J.A y Gordon S. Linoff.(2004).Data Mining techniques for Marketing, Sales, and Customer Relationship Management. Indianapolis. Wiley Publishing, Inc. 637p.

- Clasificación.
- Estimación.
- Predicción.
- Asociación.
- Agrupamiento o *Clustering*.
- Descripción.

A continuación se describen cada una de las tareas de la *Minería de Datos*:

3.2.7.1 Clasificación.

La clasificación, es una de las tareas más comunes de *Minería de Datos*, que parece ser un imperativo humano, ya que con el fin de comprender el mundo que nos rodea, usamos constantemente la clasificación y categorización.

Consiste en examinar las características de un elemento presente en el conjunto de datos y asignarlo a uno de los conjuntos predefinidos de clases. Los elementos que van a ser clasificados, están generalmente representados por los registros que se contienen de ese elemento en una tabla de base de datos o un archivo.

La tarea de clasificación se caracteriza por contar con una correcta definición de las clases, y de una formación de entrenamiento que consiste en ejemplos preclasificados. La tarea es construir o aplicar un modelo de algún tipo que pueda ser empleado en los datos que aún no hayan sido clasificados con el fin de clasificarlos.

Algunos ejemplos de las tareas de clasificación tanto en el ámbito empresarial como en la investigación son:

- Establecer si una determinada transacción mediante tarjeta de crédito es fraudulenta.
- Asignar a un nuevo estudiante en un tema particular con respecto a sus necesidades especiales.
- Evaluar si una solicitud de hipoteca es un buen riesgo de crédito o no.
- Diagnosticar determinadas enfermedades.
- Determinar qué números de teléfono corresponden a máquinas de fax.

Las técnicas que generalmente emplean la clasificación son: Los árboles de decisión y las técnicas del vecino más cercano. Las redes neuronales y análisis de enlaces también utilizan la clasificación en ciertos casos.

3.2.7.2 Estimación

Es la actividad donde dado unos datos de entrada, se debe estimar los valores para algunas variables continuas desconocidas, tales como ingreso, balance de una tarjeta de crédito, entre otros. La estimación es similar a la clasificación, salvo que la variable de destino es numérica y no categórica. Los modelos son construidos usando registros "completos", que proporcionan el valor de la variable de destino así como los predictores. Entonces, para las nuevas observaciones, las estimaciones del valor de la variable son realizadas en base a los valores de las variables predictoras.

El método de estimación tiene la gran ventaja de que los registros individuales pueden ser ordenados de acuerdo con rango de la estimación.

Para ver la importancia de esto, se puede suponer que una empresa de botas de esquí ha presupuestado para una distribución de 500.000 catálogos de publicidad, si el método de clasificación se utiliza y son identificados 1,5 millones de esquiadores, entonces se podría simplemente colocar el anuncio en las facturas de 500.000 personas seleccionadas al azar de esa muestra, si, por otra parte, cada titular tiene una propensión a la puntuación de esquí, se puede enviar el anuncio a los 500.000 candidatos más probables.

Algunos ejemplos de las tareas de estimación son:

- Estimación del número de niños en una familia.
- Estimación de los ingresos totales del hogar de una familia.
- La estimación del valor de toda la vida de un cliente como comprador.
- Estimación de la probabilidad de que alguien pague un crédito solicitado.

Los modelos de regresión y las redes neuronales se adaptan bien a las tareas de estimación.

3.2.7.3 Predicción

La predicción es similar a la clasificación y a la estimación, salvo que para la predicción, los registros se clasifican de acuerdo a algún comportamiento futuro, o valor futuro estimado. La predicción es una de las más importantes actividades realizadas en la *Minería de Datos*.

En una tarea de predicción, la única manera de comprobar la exactitud de la clasificación es esperar los resultados y evaluarlos. La razón principal para el tratamiento de la predicción como una actividad separada de la clasificación y la estimación es que en el modelado predictivo hay otras cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo.

Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos. Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos. Cuando este modelo se aplica a nuevas entradas de datos, el resultado es una predicción del comportamiento futuro de los mismos.

Algunos ejemplos de las tareas de predicción tanto en el ámbito empresarial como en la investigación son:

- Predecir qué clientes se retirarán dentro de los próximos seis meses.
- Predecir qué suscriptores de telefonía ordenarán un servicio de valor agregado.
- Predecir el porcentaje de aumento en las muertes de tráfico el próximo año si se aumenta el límite de velocidad.
- Predecir si una molécula particular, en el descubrimiento de fármacos dará lugar a un nuevo medicamento rentable para una empresa farmacéutica.

La mayoría de las técnicas de *Minería de Datos* son adecuadas para usar la predicción a partir de datos históricos como también de datos de entrenamiento de forma adecuada. La elección de la técnica depende de la naturaleza de los datos de entrada, el tipo de valor que se predice, y la importancia concedida a la explicabilidad de la predicción.

3.2.7.4 Asociación

La tarea de asociación en *Minería de Datos* es encontrar los atributos que deben "ir juntos", es decir aquellos atributos que se relacionan entre sí.

Más prevalente en el mundo de los negocios, donde es conocido como análisis de afinidad, o análisis de la canasta de mercado, la tarea de la asociación trata de descubrir las reglas para cuantificar la relación entre dos o más atributos. Reglas de la Asociación son de la forma "Si antecedente, a continuación, como consecuencia," junto con una medida de la confianza y apoyo relacionados con la regla.

Por ejemplo, un supermercado en particular puede encontrar que de las compras de 1000 clientes en un jueves por la noche, 200 compraron pañales, y de los 200 que compraron pañales, 50 compraron cerveza. Así pues, la regla de asociación sería "Si se compran pañales entonces se compra cerveza" con un soporte de “pañales” de 200/1000 = 20%; soporte de “pañales entonces cerveza” de 50/1000 y una confianza (“soporte pañales entonces cerveza/soporte pañales”) de 50/200 = 25%.

Algunos ejemplos de las tareas de asociación tanto en el ámbito empresarial como en la investigación son:

- Investigar la proporción de suscriptores de un plan telefónico de una compañía de celulares que responden positivamente a una oferta de un servicio de actualización.
- Examinar la proporción de niños cuyos padres les leen a sí mismos y que son buenos lectores
- La predicción de la degradación de las redes de telecomunicaciones.
- Encontrar qué artículos en un supermercado se compran juntos y qué artículos no se compran juntos.
- Determinar la proporción de casos en que un nuevo fármaco se presenta efectos secundarios peligrosos.

Existen varios algoritmos para la generación de reglas de asociación, como el algoritmo a priori, el algoritmo de GRI (Inducción generalizada de reglas), entre otros.

3.2.7.5 Agrupamiento o *Clustering*

El *Clustering* se refiere a la agrupación de los registros, las observaciones, o los casos en las clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, y diferentes a los registros de las otras categorías.

El agrupamiento o *Clustering* difiere de la clasificación en que no hay ninguna variable de destino para la agrupación o bien no existen tipos predefinidos ni modelos de clasificación. La tarea de la agrupación no es tratar de clasificar, calcular o predecir el valor de una variable de destino. En cambio, los algoritmos de agrupamiento buscan segmentar el conjunto de datos en subgrupos o grupos relativamente homogéneos, donde la similitud de los registros dentro de la agrupación se maximiza y la similitud con los registros fuera del clúster se minimiza.

Algunos ejemplos de las tareas de agrupamiento o *clustering* tanto en el ámbito empresarial como en la investigación son:

- Marketing de destino de un producto de nicho para un negocio de pequeña capitalización que no tiene un gran presupuesto publicitario.
- Para efectos de auditoría contable, se pueden segmentar los comportamientos financieros en las categorías de benignos y sospechosos.
- Como una herramienta para reducir, cuando el conjunto de datos tiene cientos de atributos
- Expresión de genes por agrupación, en donde grandes cantidades de genes pueden mostrar un comportamiento similar

La agrupación o *Clustering* se hace a menudo como un preludio a alguna otro modelo o tarea de *Minería de Datos*. Por ejemplo, la agrupación podría ser el primer paso en un esfuerzo de la segmentación del mercado: en lugar de tratar de llegar a todos los clientes de una empresa, se puede crear una norma como "a qué tipo de promoción no responden mejor los clientes", primero dividir la base de clientes en grupos o personas con hábitos de compra similar, y luego preguntar qué tipo de promoción funciona mejor para cada grupo.

3.2.7.6 Descripción

A veces el propósito de aplicar la *Minería de Datos* es simplemente para describir lo que está ocurriendo en una base de datos compleja de una manera que aumente nuestra comprensión de las personas, productos o procesos que produjeron los datos en primer lugar. Una suficientemente buena descripción de un comportamiento, usualmente sugiere una explicación para ello también.

Por lo menos, una buena descripción sugiere dónde empezar a buscar una explicación. La famosa brecha de género en la política estadounidense es un ejemplo de cómo una simple descripción, "las mujeres apoyan a los demócratas en mayor número que los hombres," puede provocar gran interés y estudio por parte de periodistas, sociólogos, economistas y científicos políticos, por no hablar de los candidatos a cargos públicos[33].

Los modelos de *Minería de Datos* debe ser lo más transparente posible, es decir, los resultados del modelo deben describir patrones claros que se puedan explicar e interpretar intuitivamente. Algunos métodos de *Minería de Datos* son más apropiados que otros a la interpretación transparente. Por ejemplo, árboles de decisión para dar una explicación intuitiva y entendible de sus resultados. Por otra parte, las redes neuronales son relativamente opacas a los no especialistas, debido a la no linealidad y la complejidad del modelo en otras palabras porque funcionan similar a una caja negra.

Los árboles de decisiones son una herramienta poderosa para describir a los clientes (o cualquier otra cosa) de una empresa con respecto a un objetivo o resultado particular. Las reglas de la asociación y la agrupación o *clustering* también pueden ser utilizados para construir los perfiles.

3.2.8 TÉCNICAS DE MINERÍA DE DATOS

Dado que la *Minería de Datos* es un campo muy interdisciplinar, existe un conjunto de tareas que cumplen con sus propósitos y que pueden ser utilizadas en áreas de aplicación específicos, diversas técnicas de *Minería de Datos* se utilizan para llevar a cabo las tareas de la misma, estas técnicas consisten en algoritmos específicos que pueden ser utilizados para cada función.

[33] Ibidem.

Dentro de las principales técnicas de *Minería de Datos* se encuentran[34]:

- Técnicas de inferencia estadística.
- Visualización.
- Razonamiento basado en memoria.
- Detección de conglomerados.
- Análisis de vínculos.
- Árboles de decisión.
- Redes neuronales.
- Algoritmos genéticos.

3.2.8.1 Técnicas de Inferencia Estadística

Las técnicas y métodos estadísticos del razonamiento han sido utilizados durante varias décadas, siendo éstos los únicos medios para analizar los datos en el pasado. Numerosos paquetes estadísticos están ahora disponibles para calcular promedios, sumas, y diferentes distribuciones para distintas aplicaciones. Más recientemente, las técnicas estadísticas del razonamiento están jugando un papel importante en la *Minería de Datos*.

Cabe destacar que la *Minería de Datos* no sustituye la estadística "clásica", sino que la complementa. Así pues, la estadística juega un importante papel en el análisis de los datos, e incluso también en el aprendizaje automático. Debido a esto, no se puede estudiar la *Minería de Datos* sin conocimientos previos de estadística[35].

3.2.8.2 Visualización

Las tecnologías de visualización son apropiadas para identificar patrones ocultos en un conjunto de datos, usualmente son usadas al comienzo de un proceso de *Minería de Datos* para determinar la calidad del conjunto de datos.

---

[34] Ponniah, Pauraj. (2001). Data Warehousing Fundamentals a Comprehensive Guide for IT Professionals. New York. Jhon Wiley & Sons, Inc Publication.516p.

[35] *Marcano, Yelitza. Talavera, Rosalba (2007)* Minería de Datos como soporte a la toma de decisiones empresariales *Universidad del Zulia* http://www.serbi.luz.edu.ve/scielo.php?pid=S1012-15872007004000008&script=sci_arttext (29 Ag. 2009).

Los modelos de visualización pueden ser bidimensionales, tridimensionales o incluso multidimensionales, se han desarrollado varias herramientas de visualización para integrarse con las bases de datos ofreciendo una visualización de forma interactiva a la *Minería de Datos*.

3.2.8.3 Razonamiento Basado en Memoria

Esta técnica mantiene un conjunto de datos de los registros conocidos, el algoritmo conoce las características de los registros de este conjunto de datos de formación, y cuando llega un nuevo registro de datos para la evaluación, el algoritmo encuentra registros del conjunto de datos de formación que sean similares a los nuevos, de tal forma que a partir de las características de los registros conocidos realiza la predicción y clasificación para los nuevos, por lo tanto el razonamiento basado en memoria se considera una técnica directa de *Minería de Datos* que utiliza instancias conocidas como modelo para realizar predicciones sobre instancias desconocidas.
Cuando es incluido un nuevo conjunto de datos a la herramienta de Minería, en primer lugar ella calcula la "distancia" entre estos datos y los registros contenidos en el conjunto de datos de formación, la función de la distancia de la herramienta realiza el cálculo, los resultados determinan qué registros de datos en el conjunto de datos de formación califican para ser considerados como vecinos para el registro de datos entrantes, a continuación, el algoritmo utiliza una función de combinación para combinar los resultados de las distintas funciones de distancia para obtener la respuesta final.

La función de la distancia y la función de combinación son componentes clave de la memoria técnica basada en el razonamiento.

3.2.8.4 Detección de Conglomerados

Consiste en aplicar modelos que encuentran registros de datos que son similares a otros. La detección de conglomerados es inherentemente indirecta, puesto que la meta es encontrar similitudes en los datos previamente desconocidas.

El agrupamiento o la detección de conglomerados es una de las primeras técnicas de *Minería de Datos*, y se caracteriza porque mediante ella es posible descubrir conocimiento a partir de un algoritmo no dirigido, o bien, aprendizaje no supervisado.

En la técnica de detección de conglomerados, no se realiza una búsqueda de datos preclasificados y no se hace distinción entre las variables independientes y

dependientes de los registros, en ésta técnica se realizan búsquedas en los datos por medio de un algoritmo para la detección de grupos o conjuntos de elementos de datos que son similares entre sí, ya que se espera que los clientes similares o productos similares tiendan a comportarse de la misma manera.

Cuando se hayan formado los grupos de datos similares se puede escoger un grupo y hacer algo útil con él. Cuando el algoritmo de minería produce un conglomerado, se debe entender lo que exactamente representa ese grupo, ya que sólo entonces es posible hacer algo útil con él.

Si sólo hay dos o tres variables o dimensiones en los registros de datos, es bastante fácil de detectar los grupos, incluso cuando se trata con muchos registros, pero si se trata de 500 variables de 100.000 registros, se necesita contar con una herramienta especial para procesar dichos datos.

3.2.8.5 Análisis de Vínculos

Esta técnica es muy útil para identificar las relaciones entre registros aplicando modelos basados en descubrimiento de patrones presentes en los datos. Dependiendo de los tipos de descubrimiento de conocimiento, las técnicas de análisis de vínculos tienen tres tipos de aplicaciones: *descubrimiento de asociaciones, descubrimiento de patrones secuenciales, y descubrimiento de secuencias de tiempo similares*. A continuación se Analizan brevemente cada una de estas aplicaciones:

- **Descubrimiento de Asociaciones***:* Las asociaciones son las afinidades entre los elementos, los algoritmos de descubrimiento de asociaciones encuentran sistemática y eficientemente combinaciones donde la presencia de un elemento sugiere la presencia de otro. Al aplicar estos algoritmos para las operaciones de compras en un supermercado, se descubren las afinidades entre los productos que pueden ser adquiridos juntos, las reglas de Asociación representan tales afinidades entre los datos. Los patrones de soporte y de confiabilidad indican la fuerza de la asociación, las reglas con altos valores de soporte y confiabilidad son más válidas, relevantes y útiles para un grupo u organización.

- **Descubrimiento de Patrones Secuenciales:** Como su nombre lo indica, estos algoritmos se encargan de descubrir patrones en una serie de registros, donde un grupo de elementos sigue otro grupo específico. El tiempo desempeña un papel importante para el descubrimiento de dichos patrones, ya que al seleccionar los registros para el análisis, se debe tener la fecha y el tiempo como elementos de datos para permitir el descubrimiento de patrones secuenciales.

- **Descubrimiento de Secuencias de Tiempo Similares:** Esta técnica depende de la disponibilidad de las secuencias de tiempo, en la técnica anterior, los resultados indican eventos secuenciales en el tiempo, esta técnica, sin embargo, encuentra una secuencia de acontecimientos y luego viene con otras secuencias similares de acontecimientos.

3.2.8.6 Árboles de Decisión

Esta técnica se aplica a la clasificación y predicción. Los árboles de decisión son ampliamente usados y pueden ser fácilmente explicados basándose en el criterio usado para dividir los datos en las extremidades del árbol. Los árboles de decisión son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos.

Las técnicas basadas en árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos, los valores faltantes y los datos incongruentes que se puedan presentar en el conjunto de datos. Son bastante eficientes y obtienen resultados para clasificación, los métodos obtenidos se pueden expresar como conjuntos de reglas. Uno de los inconvenientes de los árboles de decisión es su limitada expresividad y que son inestables ante variaciones de la muestra.

3.2.8.7 Redes Neuronales

Las redes neuronales simulan el cerebro humano mediante el aprendizaje de un conjunto de datos de formación y la aplicación del aprendizaje para generalizar los patrones para la clasificación y predicción. Estos algoritmos son eficaces cuando los datos carecen de un patrón aparente.

Las redes neuronales consisten en modelos predecibles, no lineales que aprenden a través del entrenamiento, generalizando los patrones que se encuentran en él, para clasificarlos y hacer pronósticos con ellos. Una vez la red neuronal ha sido entrenada, puede trabajar con gran cantidad de datos en una fracción del tiempo gastado por un humano. Las redes neuronales son ampliamente usadas para detectar actividades fraudulentas.

Su ventaja principal es que, cuando están bien ajustadas, obtienen precisiones muy altas. Además son muy expresivas y permiten capturar modelos no lineales. Entre sus inconvenientes se suelen nombrar su sensibilidad a valores anómalos, aunque son robustos frente a pocas incongruencias que se puedan presentar en

los datos y a los atributos no significativos, dentro de sus inconvenientes también se encuentra el hecho de que necesitan muchos ejemplos para el aprendizaje y son relativamente lentas, además de que en la mayoría de los casos son bastante incomprensibles.

3.2.8.8 Algoritmos Genéticos

En cierto modo, los algoritmos genéticos tienen algo en común con las redes neuronales ya que ésta técnica también tiene su base en la biología. Los algoritmos genéticos aplican los mecanismos de la genética y de la selección natural para buscar conjuntos óptimos de parámetros que describan una función de predicción.

Esta técnica utiliza un proceso muy iterativo de selección, cruzado, y de mutación de operadores, evolucionando las sucesivas generaciones de modelos. En cada iteración, cada modelo compite con todos los otros modelos por la herencia de los rasgos de los anteriores hasta que sólo el modelo más predictivo sobrevive.

3.2.9 HERRAMIENTAS DE MINERÍA DE DATOS

Actualmente existen aplicaciones o herramientas de *Minería de Datos* muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

Para elegir una herramienta de *Minería de Datos* adecuada según las necesidades de cada proyecto, se debe tener en cuenta generalmente los siguientes aspectos[36]:

- **Acceso a Datos**: La herramienta de *Minería de Datos* debe ser capaz de acceder a fuentes de datos tales como el almacén de datos y llevar rápidamente en los conjuntos de datos necesarios para su entorno. En muchas ocasiones, es posible que tenga datos de otras fuentes para aumentar los datos extraídos del almacén de datos. La herramienta debe ser capaz de leer otras fuentes de datos y formatos de entrada.
- **Selección de Datos**: Para la selección y extracción de los datos para el proceso de minería, la herramienta debe ser capaz de realizar sus

[36] Ibidem.

operaciones de acuerdo con una variedad de criterios de selección que debe incluir la capacidad de filtrado de datos no deseados y la obtención de nuevos elementos de datos de los ya existentes.

- **Sensibilidad a la Calidad de los Datos**: Debido a su importancia, la herramienta de *Minería de Datos* debe ser sensible a la calidad de los datos que se utilizan para realizar el proceso, de tal forma que la herramienta debe ser capaz de reconocer los datos faltantes o incompletos y compensar el problema; la herramienta también debe ser capaz de producir informes de error.

- **Visualización de Datos**: Técnicas de *Minería de Datos* que utilizan grandes volúmenes de datos que en la mayoría de los casos producen una amplia gama de resultados, la incapacidad para mostrar los resultados de forma gráfica y esquemática disminuye el valor de la herramienta, es recomendable seleccionar una herramienta con buenas capacidades de visualización de datos.

- **Extensibilidad**: La arquitectura de la herramienta debe ser capaz de integrarse con la administración de almacenamiento de datos y otras funciones como la extracción de datos y gestión de metadatos.

- **Rendimiento**: La herramienta debe proporcionar un rendimiento constante, independientemente de la cantidad de datos que se extraen, el algoritmo específico que se aplica, el número de variables especificadas, y el nivel de precisión exigido.

- **Escalabilidad**: La *Minería de Datos* tiene que trabajar con grandes volúmenes de datos para descubrir patrones significativos y útiles y las relaciones. Por lo tanto, es necesario contar con una herramienta que cuente con esta característica.

- **Apertura**: La apertura se refiere a ser capaz de integrarse con el medio ambiente y otros tipos de herramientas. Existen herramientas que tienen la capacidad de conectarse a aplicaciones externas donde los usuarios pueden acceder a los algoritmos de *Minería de Datos* desde otras aplicaciones. La herramienta debe ser capaz de compartir la producción con herramientas de escritorio tales como presentaciones gráficas, hojas de cálculo y utilidades de base de datos. La característica de la apertura debe incluir también la disponibilidad de la herramienta en las plataformas de servidor principal.

Algunas de las herramientas que se utilizan en el proceso de *Minería de Datos* contienen diferentes funcionalidades como lo son: *las herramientas estadísticas, herramientas núcleo de Minería de Datos, herramientas de consulta y*

herramientas de visualización de datos, a continuación se presenta una breve descripción de cada una de ellas:

- **Herramientas Estadísticos**: proporcionan análisis exploratorio para analistas expertos o estadísticos. Al igual que con la programación tradicional, el usuario necesita entender el lenguaje, ya que se necesita una sintaxis específica para poder extraer los datos de interés. Adicionalmente, los usuarios necesitan saber cómo condicionar los datos, así como saber estructurar y administrar las consultas.

- **Herramientas Núcleo de *Minería de Datos***: suministran análisis exploratorio a analistas expertos y estadísticos usando un formato gráfico y amigable con el usuario. No se necesita entender un lenguaje o sintaxis. Este tipo de herramientas descubren patrones escondidos, tendencias, relaciones e indicadores predicativos. A pesar de ser muy gráfico el ambiente de trabajo, se requiere saber cómo condicionar los datos y como estructurar y manejar las consultas. Una herramientas núcleo de *Minería de Datos*, puede proporcionar la capacidad de usar múltiples técnicas, como por ejemplo detección de conglomerados, árboles de decisión y redes neuronales; queda a criterio del minero escoger la técnicas que mejor se ajusten a las situaciones del negocio.

- **Herramientas de Consulta**: proporcionan el acceso a los datos detallados, también pueden ser usadas para extraer datos. Estas herramientas son directas, es decir se requiere que el usuario tenga un muy buen conocimiento de lo que está buscando. Son útiles para desarrollar una idea particular o para probar o invalidar una hipótesis.

- **Herramientas de Visualización de Datos**: muestran los datos gráficamente para mejorar su comprensión, permiten entender grandes cantidades de datos, y datos con complejas relaciones, usando generalmente cubos que muestran jerarquías de dimensiones, aunque no son exactamente herramientas de minería, asisten al minero a visualizar los factores más predictivos en cierta situación, también ayudan a comunicar los resultados que arrojan algoritmos complejos de minería como por ejemplo los de agrupación o *clustering*, a personas que no tienen conocimientos previos en estadística.

Para la realización de este proyecto, se utilizó la herramienta llamada *RapidMiner*, la cual se dará a conocer a continuación.

3.2.10 RAPIDMINER

RapidMiner es un entorno para el aprendizaje automático y para procesos de *Minería de Datos*, bajo el concepto de operador modular permite el diseño de cadenas de operadores complejos anidados para un gran número de problemas de aprendizaje.
El manejo de los datos es transparente para los usuarios, ya no tienen que hacer frente con el formato de los datos reales o de los datos desde diferentes puntos de vista, el núcleo de *RapidMiner* se ocupa de las transformaciones necesarias.

Hoy en día, *RapidMiner* es el líder mundial en soluciones de *Minería de Datos* con código abierto y es ampliamente utilizado por los investigadores y las empresas[37].

RapidMiner introduce nuevos conceptos de manejo transparente de datos facilita el proceso de configuración para usuarios finales, además de que cuenta con interfaces claras y una especie de lenguaje de script basado en XML lo que la convierte en un entorno de desarrollo integrado para la *Minería de Datos* y el aprendizaje automático.

3.2.10.1 RapidMiner como Herramienta de Minería de Datos

RapidMiner utiliza XML (Extensible Markup Language), un lenguaje ampliamente utilizado que es muy adecuado para describir objetos estructurados que se utilizan para describir los árboles de operaciones modelados durante el proceso de descubrimiento de conocimiento. XML se ha convertido en un formato estándar para el intercambio de datos, además, este formato es fácilmente interpretado por humanos y máquinas.

Todos los procesos de *RapidMiner* son descritos fácilmente en un formato XML y es posible ver esta descripción como un lenguaje de secuencias de comandos para el aprendizaje automático y para el proceso de *Minería de Datos*.
La interfaz gráfica de usuario y el lenguaje de secuencias de comandos XML, convierte a *RapidMiner* en un entorno de desarrollo y en un intérprete para el aprendizaje automático y la *Minería de Datos*; además, el proceso de

[37] Mierswa, I. and Wurst, M. and Klinkenberg, R. and Scholz, M. and Euler, T. (2006) Yale (now: RapidMiner): Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006).

configuración de archivos XML define un estándar para el formato de intercambio para los procesos de *Minería de Datos*.

3.2.10.2 Diversas maneras de utilizar RapidMiner

RapidMiner se puede iniciar en línea, si se suministra el proceso de configuración como un archivo XML; alternativamente, la interfaz gráfica de usuario de ésta se puede utilizar para el diseño de la descripción XML de los árboles de operación, de forma interactiva para controlar e inspeccionar las los procesos en ejecución, y seguimiento continuo de la visualización del proceso de resultados; se pueden utilizar *Break points*[38] para comprobar los resultados intermedios y los datos de el flujo entre los operadores.

También puede usar *RapidMiner* desde un programa, borrar interfaces definidas de una forma fácil aplicando operadores individuales, operadores de cadenas, o árboles de operación sobre los datos de entrada. Una versión de línea de comandos y una API Java permite el uso de *RapidMiner* desde un programa sin usar la interfaz gráfica de usuario. Además está escrito completamente en Java, que se ejecuta en cualquier plataforma y/o sistema operativo.

3.2.10.3 Manipulación transparente de los datos

RapidMiner apoya el proceso flexible de acuerdo a que permite buscar el mejor esquema de aprendizaje y de preprocesamiento de los datos y de las tareas de aprendizaje que estén a la mano de acuerdo con el problema planteado.

Además logra un manejo transparente de los datos mediante el apoyo a varios tipos de fuentes de datos, ocultando la transformación y partición interna de éstos del usuario. Debido al concepto modular de operador a menudo sólo un operador tiene que ser reemplazado para evaluar su desempeño, mientras que el resto del proceso de diseño sigue siendo el mismo. Esta es una característica importante tanto para la investigación científica como para la optimización de las aplicaciones del mundo real.

[38] Punto de interrupción.

3.2.10.4 Operadores integrados para la Minería de Datos

RapidMiner proporciona más de 400 Operadores incluyendo los siguientes[39]:

- **Algoritmos de aprendizaje máquina:** Cuenta con un gran número de esquemas de aprendizaje para la regresión y las tareas de clasificación, incluidas las máquinas de soporte vectorial (SVM), árboles de decisión, varios algoritmos de minería de reglas de asociación y de agrupación.

- **Los operadores de Weka:** Cuenta con todas las operaciones de Weka como esquemas de aprendizaje y evaluadores de atributos del ambiente de aprendizaje de Weka también están disponibles y pueden ser utilizados como todos los demás operadores *RapidMiner*.

- **Operadores de preprocesamiento de datos:** Cuenta con varios operadores como son los de discretización, función de filtrado, la reposición de valores faltantes y de valores infinitos, la normalización, la eliminación de características inútiles, toma de muestras, la reducción de dimensionalidad, entre otras.

- **Operadores de característica:** Cuenta con algoritmos de selección como la selección hacia adelante y hacia atrás, la eliminación, y varios algoritmos genéticos, los operadores para la extracción de características de series de tiempo, característica de ponderación, la pertinencia característica, y la generación de nuevas características.

- **Meta operadores:** Cuenta con operadores de optimización para el diseño de procesos, por ejemplo, operadores de iteraciones o varios esquemas de optimización de parámetros.

- **De evaluación de la ejecución:** Cuenta con la validación cruzada y otros esquemas de evaluación, criterios de rendimiento para clasificaciones y regresiones, operadores para la optimización de parámetros, operadores cerrados u operadores de cadenas.

- **Visualización:** Cuenta con operadores para el registro y la presentación de los resultados. Crear en línea gráficos en 2D y 3D de los datos, modelos aprendidos y los resultados de otros procesos.

[39] Mierswa, I. and Wurst, M. and Klinkenberg, R. and Scholz, M. and Euler, T. (2006) Yale (now: RapidMiner): Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006).

- **Entradas y Salidas:** Cuenta con operadores flexibles para datos de entrada y salida, soporte para varios formatos de archivos incluyendo arff, C4.5, CSV, bibtex, dBase, y lee directamente de bases de datos.

La herramienta *RapidMiner* cuenta con grandes ventajas que han permitido la realización del proceso de *Minería de Datos* en varias ocasiones, en éste caso en particular ha sido de gran apoyo para la obtención de algunos de los objetivos propuestos en el presente proyecto, así como ha facilitado todo el proceso gracias a la robustez de la herramienta, resaltando también la facilidad de uso y transparencia en el momento de manipular los datos.

3.3 REGLAS DE ASOCIACION

3.3.1 INTRODUCCIÓN

Encontrar relaciones entre los ítems en una base de datos no es una tarea trivial, principalmente si se requiere extraer dichas relaciones de una base de datos con centenas de atributos y docenas de miles o millones de registros.

La minería de reglas de asociación permite encontrar, en un tiempo relativamente corto, las asociaciones más frecuentes en una base de datos. Para eso, es válido el principio de que, si un conjunto de ítems no es frecuente, cualquier combinación de ítems que incluya este conjunto también será infrecuente.

El objetivo de las reglas de asociación es encontrar asociaciones entre los ítems referenciados en bases de datos transaccionales, relacionales o en bodegas de datos.

Las reglas de asociación tienen diversas aplicaciones como el análisis de la canasta de mercado, marketing cruzado con correo, diseño de catálogos, segmentación de clientes respecto a las compras y el soporte para la toma de decisiones.

Actualmente en varios entornos tanto empresariales como comerciales es necesario contar con herramientas que permitan obtener conocimiento útil que brinde soporte a la toma de decisiones, para ello se necesita de un proceso que utiliza una serie de técnicas para el procesamiento de los datos, como lo es la *Minería de Datos*, que permite llevar a cabo un proceso de descubrimiento de información automático.

Para la implementación de las reglas de asociación existen una variedad de algoritmos como los son A Priori, DHP, Partition, FP-Growth y Eclat (Ver sección 3.3.6), que pueden ser aplicados con el fin de encontrar asociaciones entre un conjunto de datos, para seleccionar el más adecuado, se definen una serie de criterios, entre los que se encuentran: accesos a la base de datos, costo computacional, tiempo de ejecución y rendimiento, los cuales se analizan para cada algoritmo con el fin de realizar la selección de aquel que más se ajusta a los requerimientos del presente proyecto[40].

[40] Naranjo, Roberto. Sierra, Luz Marina. (2009). Herramienta de software para el análisis de canasta de mercado sin selección de candidatos. http://www.scielo.org.co/pdf/iei/v29n1/v29n1a08.pdf. (29 de Ag. 2009)

3.3.2 DEFINICION DE REGLAS DE ASOCIACIÓN

Dada una colección de conjuntos de ítems, las reglas de asociación describen como varias combinaciones de ítems están apareciendo juntas en los mismos registros.

Una típica aplicación de reglas de asociación está dentro del análisis de datos llamado análisis de la canasta de mercado, donde el objetivo es encontrar regularidades en los comportamientos de los clientes dentro de los términos de combinación de productos que son comprados muchas veces en conjunto.

En un dominio de análisis de canasta de mercado, los ítems representan los productos en el estante de un supermercado. Estos podrían ser ítems como cerveza, papas, leche, pan, pañales, entre otros. Los registros dentro de una base de datos de canasta corresponden entonces al contenido de un cesto de compras: para cada tipo de producto en una canasta, el correspondiente ítem está dentro del registro. Si un cliente compra leche y pañales, entonces existirá un correspondiente registro {leche, pañales} en la base de datos.

La cantidad o el precio de los ítems no es considerado en esta técnica, solamente la información binaria de que si un producto fue comprado o no. Nótese que el número de diferentes ítems puede ser del orden de miles, donde compras típicas solamente contienen en su mayoría docenas de ítems.

3.3.3 ANÁLISIS DE LA CANASTA DE MERCADO

Hoy en día para las empresas se ha convertido en una necesidad y oportunidad el conocer la información y analizarla en pro de tomar decisiones que en el momento apropiado apoyen su gestión y supervivencia en la actual y competitiva economía. De ahí surge la necesidad de incorporar en su dinámica herramientas informáticas que permitan procesar y obtener de los volúmenes de información almacenados los elementos suficientes para tomar decisiones.

Es necesario tener clara y precisa comprensión de que para una empresa tomar una decisión sin el conocimiento profundo de la información implica la posibilidad de errar en la toma de decisiones, dado que conlleva el costo requerido para poner en marcha un plan que busque la fidelidad de los clientes o capturar nuevos, o cautivar a un nuevo nicho de mercado. Según sea el fin que se pretende alcanzar con la toma de la decisión, si este no se logra se habrá perdido el esfuerzo de dicha estrategia.

Al analizar los datos para apoyar sus decisiones de marketing, dado que ese análisis deberá proveer respuestas a preguntas como: ¿qué se debe hacer para entender cómo compran los clientes?. Para responder esta pregunta, es necesario partir del análisis de canasta de mercado, ya que una canasta de mercado típica contiene los datos de la compra de productos de un cliente, en qué cantidad cada uno, y en qué época lo hace. Por lo tanto, es necesario descubrir patrones interesantes ocultos, no triviales, y de interés para las empresas alrededor de los mismos, lo cual es el objetivo principal de la *Minería de Datos*.

Las herramientas de *Minería de Datos* predicen futuras tendencias y comportamientos, permitiendo tomar decisiones conducidas por un conocimiento obtenido a partir de la información; para conseguirlo, hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales.

3.3.4 REGLAS DE ASOCIACIÓN EN LA TRANSACCIÓN DE NEGOCIOS

Actualmente, con la masiva cantidad de datos que las organizaciones recolectan en sus procesos de negocio el descubrimiento de asociaciones interesantes en los registros de transacciones puede ayudar para la toma de decisiones en los procesos de marketing.

En el ejemplo típico para reglas de asociación, “el análisis de canasta de mercado”, supóngase que un granjero local ha puesto un stand de verduras y está ofreciendo los siguientes artículos: **{espárragos, fríjoles, brócoli, maíz, pimientas verdes, calabazas, tomates}**, a este conjunto de artículos lo denotaremos I, y en la Tabla No 3. se mostrarán los artículos comprados.

No	Contenido Canasta
1	Brócoli, pimienta, maíz
2	Espárragos, calabaza, maíz
3	Maíz, tomates, fríjoles, calabazas
4	Pimienta, maíz, tomates, fríjoles
5	Fríjoles, espárragos, fríjoles, tomates
6	Calabaza, espárragos, fríjoles, tomates
7	Tomates, maíz,
8	Brócoli, tomates, pimienta
9	Calabaza, espárragos, fríjoles
10	Fríjoles, maíz
11	Pimienta, brócoli, fríjoles, calabaza

12	Espárragos, fríjoles, calabaza
13	Calabaza, maíz, espárragos, fríjoles
14	Maíz, pimienta, tomates, fríjoles, brócoli

Tabla No 3. Artículos comprados por los clientes

En el conjunto D de transacciones representadas en la Tabla No 3., cada transacción (T) en D representa un conjunto de artículos contenidos en I. Suponga que se tiene un conjunto particular de artículos A (e. g., fríjoles y calabazas), y otro conjunto de artículos B (e.g., espárragos). Luego una regla de asociación toma la forma de (A =>B), donde el antecedente A y el consecuente B son subconjuntos propios de I, y A y B son mutuamente excluyentes.

Existen dos medidas asociadas a una regla de asociación: soporte y confianza, que le dan validez a la misma. El soporte para una regla de asociación particular A=>B es la proporción de transacciones en D que contienen A y B[41].

$$Soporte = P(A \cap B) = \frac{No.TransaccionesConAyB}{No.TotalTransacciones}$$

Figura No 1. Soporte de A ⇒B

La confianza C de la regla de asociación A=>B es una medida de exactitud de la regla, determinada por el porcentaje de transacciones en D que contienen A y B[42].

$$Confianza = P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Figura No 2. Confianza de A ⇒B

El analista puede preferir reglas que tengan alto soporte o alta confianza, o usualmente ambas. Las reglas fuertes son las que reúnen o superan ciertos soportes mínimos y criterios de confianza. Por ejemplo, un analista interesado en encontrar qué artículos del supermercado se compran juntos, puede establecer un nivel de soporte mínimo de 20% y un nivel de confianza mínimo del 70%.

[41] Larose, Daniel T.(2005).Discovering Knowledge in Data an Introduction to Data Mining. Hoboken, New Jersey. Jhon Wiley & Sons, Inc Publication. 222p.
[42] Ibidem.

Por otro lado, en detección de fraude o de terrorismo, se necesitaría reducir el nivel de soporte mínimo a 1% o menos, ya que comparativamente pocas transacciones son fraudulentas o relacionadas con terrorismo.

Un *itemset* es un conjunto de artículos contenidos en I, y un k-itemset es un itemset que contiene k artículos; por ejemplo, **{fríjoles, calabazas}** es un 2-itemset, y **{brócoli, pimienta verde, maíz}** es un 3-itemset, cada uno de los estantes de vegetales puestos en I. La frecuencia Φ del conjunto de artículos (itemset) es simplemente el número de transacciones que contienen el conjunto de artículos particular. Un conjunto de artículos frecuente es aquel que ocurre al menos un cierto mínimo número de veces, teniendo una frecuencia de conjunto de artículos, por ejemplo: suponiendo que Φ = 4, los conjuntos de artículos que ocurren más de cuatro veces se dice que son frecuentes; denotamos el conjunto de k-itemsets como Fk.

Las reglas de asociación para minería de grandes bases de datos son procesos de dos pasos:

1. Encontrar todos los conjuntos de artículos frecuentes, es decir, aquellos con frecuencia ≥ Φ.

2. Del conjunto de artículos frecuentes, generar reglas de asociación que satisfagan condiciones mínimas de soporte y confianza.

Se observa que la técnica de reglas de asociación es la que más se adecúa para el desarrollo del presente proyecto, ya que se quiere descubrir las relaciones existentes entre los productos ofrecidos analizando el comportamiento de las transacciones.

Además esta técnica sugiere una búsqueda por toda la base de datos, realizando una clasificación en cada barrido, por lo tanto no hay límite establecido para la cantidad de datos que puede manejar, busca las características presentes en las transacciones realizadas, las cuales pueden tener atributos de diferentes tipos, por lo tanto no es necesario hacer una conversión a un tipo de datos específico.

La capacidad predictiva de la técnica depende de las medidas establecidas de confianza y soporte, ya que esta técnica se basa en el conteo de ocurrencias posibles entre las combinaciones de ítems en la tabla de transacciones, y posee gran escalabilidad ya que realiza un barrido por la base de datos, por lo que puede operar sin mayores problemas con un número grande de datos.

3.3.5 BÚSQUEDA DE ÍTEMSETS FRECUENTES

La identificación de itemsets frecuentes es computacionalmente costosa ya que requiere considerar todas las combinaciones de los distintos ítems, resultando en una búsqueda exponencial. La imagen No 1., muestra el lattice de espacio de búsqueda resultante de E= {a, b, c, d}. Para la búsqueda de los ítemsets frecuentes se emplean dos formas comunes de búsqueda en árbol: primero, a lo ancho (Imagen No 2.) (BFS, por sus siglas en inglés), y segundo (Imagen No 3.), en profundidad (DFS, por sus siglas en inglés), sobre árboles similares (Imagen No 3.).

Estos algoritmos trabajan por lo general de la siguiente manera: buscan un conjunto Ck de k-itemsets con alta probabilidad de ser frecuentes, llamémosles en lo sucesivo k-itemsets candidatos, comenzando por k=1. Se comprueba cuáles son frecuentes y a partir de estos se genera nuevamente un conjunto de candidatos de tamaño k+1, C_{k+1}. Este proceso se repite hasta que no se pueda generar un nuevo conjunto candidato. Tal estrategia garantiza que sean visitados todos los *itemsets* frecuentes, al mismo tiempo que se reduce el número de itemsets infrecuentes visitados.

Con la estrategia BFS el valor del soporte de los (k-1) itemsets son determinados antes de contar el soporte de todos los k-itemsets, ello le permite utilizar la propiedad arriba enunciada. Con la estrategia DFS no son conocidos todos los (k-1) itemsets, pero sí los necesarios (k-1) itemsets cuando se generan cada uno de los k-itemsets, pues trabaja recursivamente descendiendo por el árbol y siguiendo la estructura del segundo (Imagen No 3.).

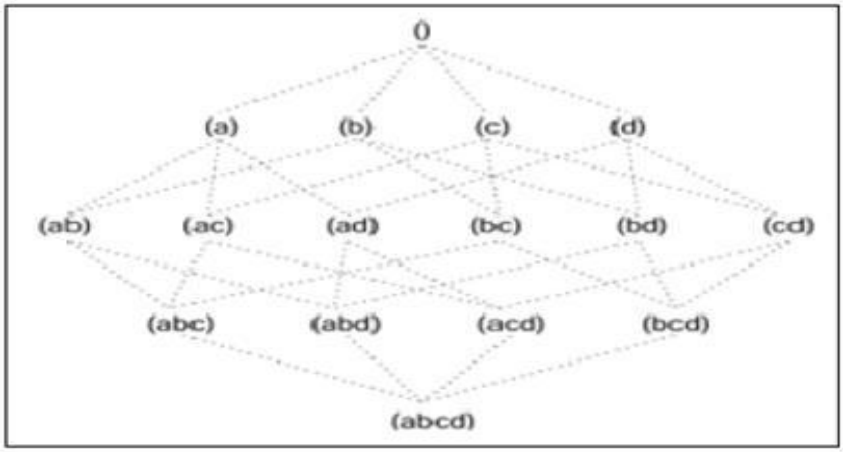


Imagen No 1. Lattice del espacio de búsqueda[43]

[43] Naranjo, Roberto. Sierra, Luz Marina. (2009). Herramienta de software para el análisis de canasta de mercado sin selección de candidatos. http://www.scielo.org.co/pdf/iei/v29n1/v29n1a08.pdf. (29 de Ag. 2009)

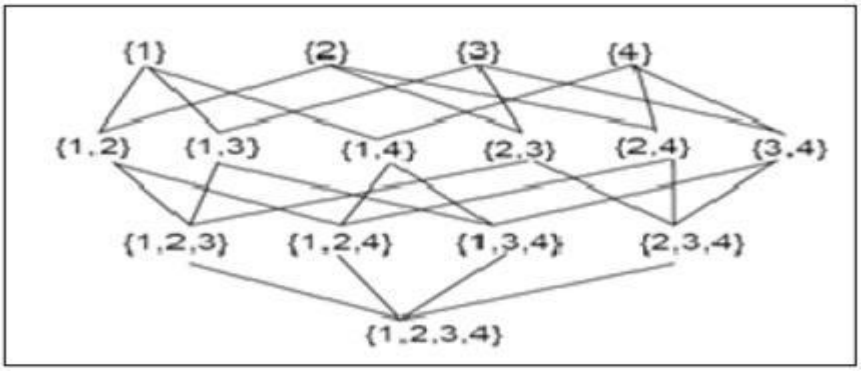


Imagen No 2. Árbol utilizado en la estrategia BFS

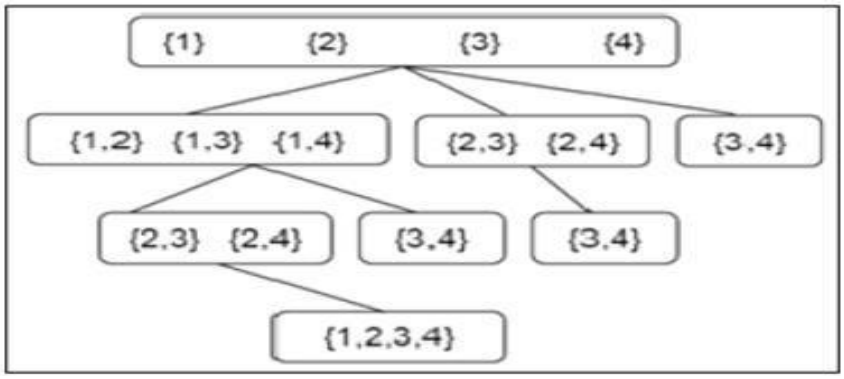


Imagen No 3. Árbol usado en los algoritmos con estrategia DFS[44]

Para contar el soporte de todos los conjuntos de ítems también se emplean por lo general dos mecanismos:

1. Determinar el valor del soporte contando directamente sus ocurrencias en la base de datos.

2. Determinar el soporte empleando la intersección entre conjuntos. Un tid es un identificador único de una transacción.

[44] Ibidem.

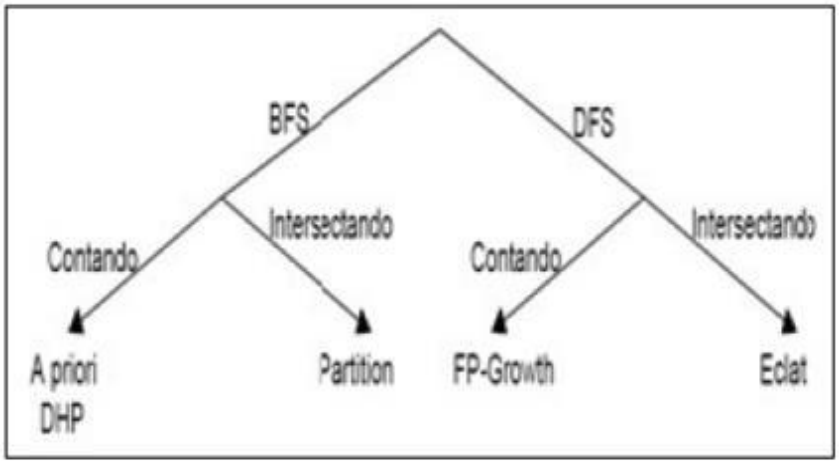


Imagen No 4. Algoritmos para el cálculo de itemsets frecuentes[45]

3.3.6 ALGORITMOS DE REGLAS DE ASOCIACIÓN

Para la técnica de reglas de asociación existen una serie de algoritmos tales como: *A priori, DHP, Partition, FP-Growth y Eclat*, de los cuales se seleccionó el más adecuado teniendo en cuenta los siguientes criterios[46]:

Accesos a la base de datos: es importante que los algoritmos minimicen el recorrido por la base o bodega de datos, pues el número de reglas crece exponencialmente con el de ítems considerados, lo cual afecta el rendimiento del algoritmo cuando se accede constantemente a la base o bodega de datos.

Costo computacional*:* es importante que el algoritmo no realice un gran número de operaciones.

Tiempo de ejecución*:* se desea que el tiempo utilizado para la generación de reglas sea razonable.

Rendimiento*:* es importante que el algoritmo realice las operaciones y procesos de forma eficiente.

Se revisó cada uno de los algoritmos mencionados con los criterios definidos:

[45] Ibidem.

[46] Naranjo, Roberto. Sierra, Luz Marina. (2009). Herramienta de software para el análisis de canasta de mercado sin selección de candidatos. http://www.scielo.org.co/pdf/iei/v29n1/v29n1a08.pdf. (29 de Ag. 209)

3.3.6.1 Algoritmo A Priori

Este algoritmo busca primero todos los conjuntos frecuentes unitarios (contando sus ocurrencias directamente en la base de datos), se mezclan estos para formar los conjuntos de ítems candidatos de dos elementos y seleccionan entre ellos los frecuentes. Considerando la propiedad de los conjuntos de ítems frecuentes, se vuelve a mezclar estos últimos y se seleccionan los frecuentes (hasta el momento ya han sido generados todos los conjuntos de ítems frecuentes de tres o menos elementos). Así sucesivamente se repite el proceso hasta que en una iteración no se obtengan conjuntos frecuentes[47].

La evaluación de éste algoritmo es la siguiente:

Accesos en la base de datos*:* este algoritmo busca todos los conjuntos frecuentes unitarios contando sus ocurrencias directamente en la base de datos, por lo tanto se realizan varias pasadas en dicha base.

Costo computacional: el conteo de soporte de los candidatos es costoso debido a que el número de subconjuntos frecuentes en cada candidato es cada vez mayor y los niveles en el árbol hash de candidatos se incrementa.

Tiempo de ejecución*:* hay que hacer tantos recorridos como sea necesario para encontrar todos los ítems frecuentes, por lo que no solo es costosa la solución en memoria, sino además en tiempo.

Rendimiento: este algoritmo tiene algunas mejoras para el rendimiento, entre ellas está la de reducir el número de ítems que contienen subconjuntos infrecuentes, aunque posteriormente al mezclar pares de conjuntos frecuentes con k-2 elementos iguales hay que verificar si todos los subconjuntos de k-1 elementos pertenecen al conjunto de itemsets frecuentes, con lo cual mejora el rendimiento.

3.3.6.2 Algoritmo DHP (Direct Hashing Pruning: Poda y Hashing Directa)

Este algoritmo emplea una técnica de hash para eliminar los conjuntos de ítems innecesarios para la generación del próximo conjunto de ítems candidatos. Cada (k+1)- *itemse*t es añadido a una tabla hash en un valor *hash* dependiente de las ocurrencias en la base de datos de los conjuntos candidatos de k elementos que lo formaron, o sea, dependiente del soporte de los conjuntos candidatos de k elementos. Estas ocurrencias son contadas explorando en las transacciones de la base de datos. Si el soporte asociado a un valor hash es menor que el soporte

[47] Agrawal, Rakesh et al. (1993).Mining Association Rules between Sets of Items in Large Database. http://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf.(26 Oct. 2009)

mínimo entonces todos los conjuntos de ítems de k+1 elementos con este valor hash no serán incluidos entre los candidatos de k+1 elementos en la próxima pasada.

La evaluación de éste algoritmo es la siguiente:

Accesos en la base de datos: este algoritmo emplea una tabla hash con un valor hash dependiente de las ocurrencias en la base de datos de los conjuntos candidatos, por lo cual se requiere hacer varias accesos en la base de datos.

Costo computacional*:* se emplea una tabla hash para reducir el número de candidatos; el espacio de memoria empleado por la tabla compite con el necesitado por al árbol hash, de ahí que, con tablas hash muy grandes (para reducir la cantidad de falsos positivos) la memoria se vuelve insuficiente.

Tiempo de ejecución*:* este algoritmo requiere de varias pasadas a la base de datos para su funcionamiento, en cada pasada cuenta el soporte de cada ítem y coloca en la tabla hash los conjuntos de K - ítems de acuerdo al valor del soporte de cada uno de dichos conjuntos, lo que representa un costo en tiempo.

Rendimiento*:* en este algoritmo el número de candidatos que tienen igual valor hash está directamente relacionado con el tamaño de la tabla, por tanto el espacio de memoria empleado por la tabla compite con el necesitado por al árbol hash y la memoria se vuelve insuficiente, afectando el rendimiento del algoritmo.

3.3.6.3 Algoritmo Partition

Este algoritmo propone fraccionar la base de datos en tantas partes como fueren necesarias para que todas las transacciones en cada partición estén en la memoria.

En contraste con los vistos hasta el momento, este algoritmo recorre la base de datos sólo dos veces. En la primera, cada partición es minada independientemente para encontrar los conjuntos de ítems frecuentes en la partición y luego se mezclan estos para generar el total de los conjuntos de ítems candidatos. Muchos de estos pueden ser falsos positivos, pero ninguno falso negativo (notemos que si existen m particiones, para que un itemset tenga soporte *s* debe poseer un soporte no menor que *s/m* al menos en una de las m particiones, los conjuntos candidatos serán por tanto los que cumplan esta condición). En la segunda pasada se cuenta la ocurrencia de cada candidato, aquellos cuyo soporte es mayor que el mínimo soporte especificado se retienen como conjuntos frecuentes.

Este algoritmo emplea el mecanismo de intersección entre conjuntos para determinar su soporte, en este caso cada ítem en una partición mantiene la lista de los identificadores de las transacciones que contienen a dicho ítem.

La evaluación de éste algoritmo es la siguiente:

Accesos en la base de datos*:* sólo requiere dos pasadas a través de la base de datos, para el cálculo de los ítems frecuentes.

Costo computacional*:* es relativamente más eficiente que el A priori pero tiene dos problemas: el costo en memoria es mayor, pues requiere almacenar para cada ítem el conjunto de transacciones que lo contiene; y además el cálculo del soporte de un candidato obtenido por la unión de dos conjuntos frecuentes obliga a intersectar los dos conjuntos.

Tiempo de ejecución*:* este algoritmo mantiene la base de datos en memoria y evita las operaciones de E/S en disco, divide la base de datos en tantas partes como sean necesarias para que todas las transacciones queden en la memoria, al reducir las operaciones de entrada/salida disminuye el tiempo de ejecución.

Rendimiento: este algoritmo, al igual que el A priori, mejora el rendimiento al reducir el número de ítems que contienen subconjuntos infrecuentes, aunque posteriormente al mezclar pares de conjuntos frecuentes con k-2 elementos iguales hay que verificar si todos los subconjuntos de k-1 elementos pertenecen al conjunto de itemsets frecuentes.

3.3.6.4 Algoritmo ECLAT

Los algoritmos de tipo Eclat reducen la cantidad de operaciones de E/S, aunque esta vez atravesando la base de datos sólo una vez.

Se basan en realizar un agrupamiento (*clustering*) entre los ítems para aproximarse al conjunto de ítems frecuentes maximales y luego emplean algoritmos eficientes para generar los ítems frecuentes contenidos en cada grupo. Para el agrupamiento proponen dos métodos que son empleados después de descubrir los conjuntos frecuentes de dos elementos: el primero, por clases de equivalencia: esta técnica agrupa los itemsets que tienen el primer ítem igual. El segundo, por la búsqueda de cliques maximales: se genera un grafo de equivalencia cuyos nodos son los ítems, y los arcos conectan los ítems de los 2-itemsets frecuentes, se agrupan los ítems por aquellos que forman cliques maximales.

La evaluación de éste algoritmo es la siguiente:

Accesos en la base de datos: este algoritmo reduce la cantidad de operaciones de entrada/salida atravesando la base de datos sólo una vez.

Costo computacional*:* es más eficiente que el A priori, sin embargo presenta el mismo problema que el Partition: el costo en memoria es mayor, pues requiere almacenar para cada ítem el conjunto de transacciones que lo contiene; y además el cálculo del soporte de un candidato obtenido por la unión de dos conjuntos frecuentes obliga a intersectar los dos conjuntos.

Tiempo de ejecución: este algoritmo se basa en realizar un agrupamiento (*clustering*) entre los ítems, lo que infiere en el tiempo de ejecución.

Rendimiento: el realizar tareas de agrupamiento requiere de pasos adicionales en su funcionamiento, sin embargo computacionalmente es más eficiente que el A priori.

Los algoritmos mencionados se basan en la estrategia del algoritmo A priori, por lo tanto todos ellos presentan generación de candidatos para seleccionar las reglas de asociación. En el A priori, cuando la base de datos presenta gran cantidad de ítems frecuentes, grandes patrones, o mínimas medidas de soporte, el algoritmo presenta los siguientes problemas:

- Es costoso manejar una gran cantidad de conjuntos candidatos. Por ejemplo, para describir patrones de 100 ítems tal como {a1, a2, …a100}, es necesario crear cerca de 1030 candidatos, que representa un alto costo computacional sin importar la técnica aplicada.

- Es tedioso repetir este proceso para comparar los candidatos en búsqueda de concordancia en la base de datos, especialmente aquellos patrones considerados como largos.

Es por esto que para solucionar el problema de generación de candidatos se plantea el siguiente algoritmo el cual no requiere generación de candidatos y mejora el rendimiento de esta técnica.

3.3.6.5 FP-Growth (Frequent Pattern Growth: Crecimiento de Patrones Frecuentes)

Este algoritmo está basado en una representación de árbol de prefijos de una base de datos de transacciones llamada Frequente Pattern Tree.

La idea básica del algoritmo FP-Growth puede ser descrita como un esquema de eliminación recursiva: en un primer paso de preprocesamiento se borran todos los ítems de las transacciones que no son frecuentes individualmente o no aparecen en el mínimo soporte de transacciones, luego se seleccionan todas las transacciones que contienen al menos un ítem frecuente, se realiza esto de manera recursiva hasta obtener una base de datos reducida. Al retorno, se remueven los ítems procesados de la base datos de transacciones en la memoria y se empieza otra vez, y así con el siguiente ítem frecuente. Los ítems en cada transacción son almacenados y luego se ordena descendentemente su frecuencia en la base de datos.

Después de que se han borrado todos los ítems infrecuentes de la base de datos de transacciones, se pasa al árbol FP. Un árbol FP es básicamente de prefijos para las transacciones, esto es: cada camino representa el grupo de transacciones que comparten el mismo prefijo, cada nodo corresponde a un ítem. Todos los nodos que referencian al mismo ítem son referenciados juntos en una lista, de modo que todas las transacciones que contienen un ítem específico pueden encontrarse fácilmente y contarse al atravesar la lista. Esta lista puede ser accesada a través de la cabeza, lo cual también expone el número total de ocurrencias del ítem en la base de datos.

La evaluación de éste algoritmo es la siguiente:

Accesos en la base de datos: este algoritmo no requiere de la generación de candidatos, por lo tanto, precisa de pocos accesos a la base de datos.

Costo computacional: el algoritmo está basado en una representación de árbol de prefijos de una base de datos de transacciones, por lo tanto necesita de la creación de un árbol de prefijos; sin embargo, la creación de dicho árbol no requiere de un costo computacional elevado.

Tiempo de ejecución: este algoritmo busca patrones frecuentes con una corta búsqueda recursiva de prefijos, lo que en tiempo de ejecución es muy superior al del A priori, ya que no requiere de constantes accesos a la base de datos.

Rendimiento: puede generar un árbol FP-Tree de una base de datos proyectada si el árbol inicial no se puede alojar completamente en la memoria principal, lo que le permite adecuarse a los recursos disponibles.

De acuerdo a lo anterior, se decide aplicar el algoritmo FPGrowth ya que tiene ventajas operacionales sobre los otros al no necesitar de la generación de ítems candidatos y ser computacionalmente más rápido.

Entre las razones por la que se seleccionó este algoritmo es debido a que requiere de pocos accesos a la base o bodega de datos. Este algoritmo está basado en una representación de árbol de prefijos de una base de datos de transacciones; sin embargo, la creación de dicho árbol no requiere de un costo computacional elevado. El algoritmo busca patrones frecuentes con una corta búsqueda recursiva de prefijos, lo que en tiempo de ejecución es muy superior al A priori, ya que no requiere constantes accesos a la base o bodega de datos.

ALGORITMOS	Accesos en la Base de Datos	Costo Computacional	Tiempo de Ejecución	Rendimiento
A Priori	Requiere de varios accesos a la base de datos.	Es costoso debido a que el número de subconjuntos frecuentes.	Consume mucho tiempo	Tiene algunas mejoras en el rendimiento.
DHP	Requiere de varios accesos a la base de datos.	La memoria puede volverse insuficiente	Los accesos a la BD pueden llegar a disminuir el tiempo.	La memoria se puede volver insuficiente, afectando el rendimiento del algoritmo.
Partition	Requiere de dos accesos a la BD.	Tiene un elevado costo en memoria.	La reducción de las operaciones de E/S disminuye el tiempo de ejecución.	Tiene algunas mejoras en el rendimiento.
ECLAT	Requiere de un acceso en la BD.	Tiene un elevado costo en memoria.	Realiza un agrupamiento que infiere en el tiempo de ejecución.	Computacionalmente es más eficiente que el A priori.
FP-Growth	Requiere de pocos accesos a la BD.	No requiere de un costo computacional elevado.	En tiempo de ejecución es muy superior al del A priori.	Se adecua fácilmente a los recursos disponibles.

Tabla No 4. Criterios de Evaluación de Algoritmos.

4. INFORME DE RESULTADOS

Trabajar con grandes superficies de ventas, parte del interés de aprovechar los grandes volúmenes de datos, que generalmente son desperdiciados en la mayoría de las empresas, y que haciendo buen uso de ellos se puede adquirir información pertinente y precisa que puede cambiar y predecir en gran sentido las ventas, o los productos que la empresa está comercializando.

Para hacer realidad esta propuesta, se planteó como objetivo general aplicar la *Minería de Datos* haciendo uso de la herramienta ***RapidMiner***, para utilizar una técnica para predicción de ventas sobre un conjunto de datos seleccionados, y poder con esto descubrir asociaciones entre dos o más productos.

Partiendo del objetivo general, se tuvieron en cuenta los siguientes objetivos específicos:

- Realizar la gestión necesaria para obtener una base de datos que permita la aplicación de alguna de las técnicas de Minería de Datos.
- Determinar la técnica a utilizar sobre la base de datos de acuerdo a los datos que ésta contenga.
- Realizar las funciones de manipulación, selección y procesamiento de los datos.
- Seleccionar una técnica de *Minería de Datos* que permita descubrir asociaciones entre dos o más productos.
- Aplicar una técnica de *Minería de Datos* que permita descubrir asociaciones o correlaciones entre productos que se venden juntos.
- Validar la técnica, comprobando que ésta se ajusta apropiadamente a los requerimientos del problema planteado.
- Dar una breve explicación de los resultados obtenidos y el por qué de los mismos.

4.1 OBTENCIÓN DE LOS DATOS.

Teniendo en cuenta el primer objetivo, se pasó a hacer una exposición sobre lo que se deseaba hacer con los datos de la empresa a los dueños de la superficie si

era posible o en su defecto al encargado del mismo, se presento rechazo por parte de algunos debido a la desconfianza del manejo que se le pudiera dar a los datos; cuando se contactó la persona que facilitó los datos, éste, por el poco conocimiento sobre el área, solo estableció prestar los datos y que se emplearan de la forma más adecuada, haciendo énfasis en la importancia de mantener la confidencialidad de los mismos y que fueran utilizados para fines académicos únicamente.

4.2 FUNCIONES DE MANIPULACIÓN, SELECCIÓN Y PROCESAMIENTO DE LOS DATOS.

4.2.1 MANIPULACIÓN DE LOS DATOS

Originalmente, la base de datos se encontraba en un archivo de texto plano teniendo la siguiente estructura (Imagen No.5):

Imagen No 5. Base de datos original.

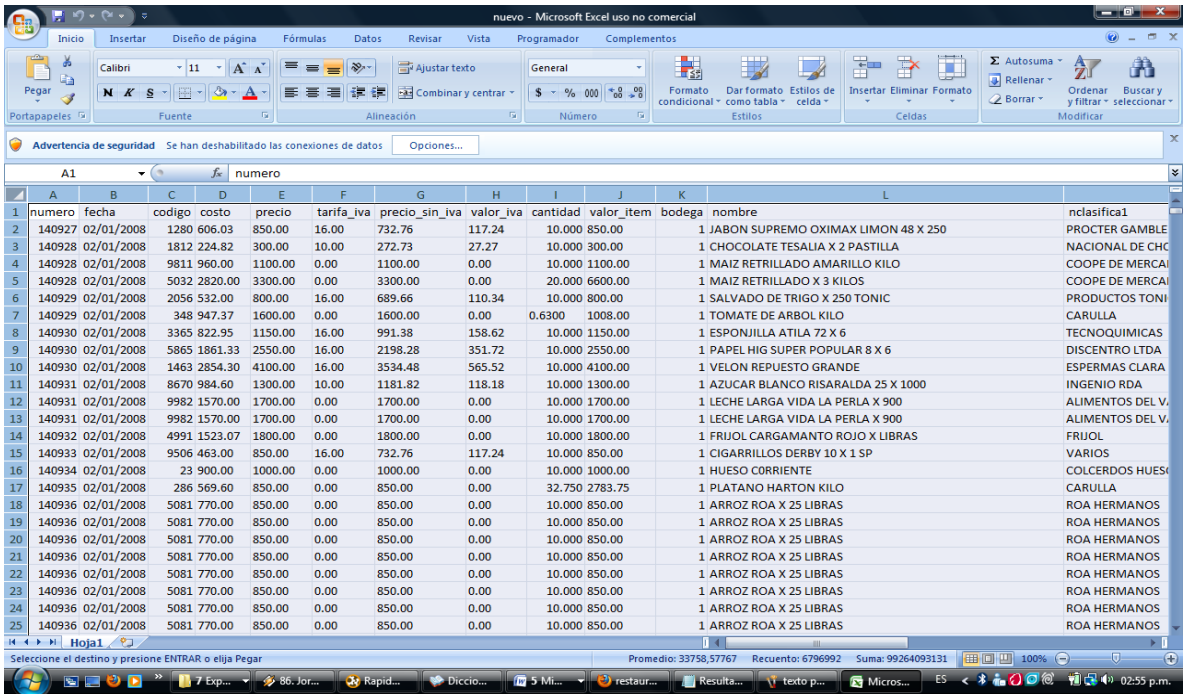
Este archivo visto de ésta manera (Imagen No.5) no es técnicamente manipulable a la hora de hacer filtrados y análisis pertinente, ya que por una parte los procesadores de texto plano son ineficientes para hacer búsquedas y agrupaciones de datos específicos, y se hace necesario poder organizarlos de tal manera que se puedan identificar los datos inconsistentes que pudieran entorpecer la investigación, tales como valores nulos, incoherencias de las transacciones; como se comenta en la teoría vista anteriormente. Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas.

4.2.2 SELECCIÓN DE LOS DATOS.

Consiste no solamente en eliminar los datos no válidos, sino también seleccionar los que permitan encontrar correlaciones ocultas entre ellos, identificar los orígenes de los que son más precisos, y determinar qué variables son las más adecuadas para usarse en el análisis. Normalmente se trabaja con un conjunto de datos muy grande y no se puede comprobar cada transacción.

4.2.3 PROCESAMIENTO DE LOS DATOS.

Teniendo en cuenta los inconvenientes que se pueden tener con la utilización del archivo en texto plano se decide migrarlo a Excel, donde es más viable hacer los respectivos análisis y estudios de los datos. Después de la migración de los datos a Excel, estos tomaron el siguiente aspecto:



numero	fecha	codigo	costo	precio	tarifa_iva	precio_sin_iva	valor_iva	cantidad	valor_item	bodega	nombre	nclasifica1
140927	02/01/2008	1280	606.03	850.00	16.00	732.76	117.24	10.000	850.00	1	JABON SUPREMO OXIMAX LIMON 48 X 250	PROCTER GAMBLE
140928	02/01/2008	1812	224.82	300.00	10.00	272.73	27.27	10.000	300.00	1	CHOCOLATE TESALIA X 2 PASTILLA	NACIONAL DE CHO
140928	02/01/2008	9811	960.00	1100.00	0.00	1100.00	0.00	10.000	1100.00	1	MAIZ RETRILLADO AMARILLO KILO	COOPE DE MERCAI
140928	02/01/2008	5032	2820.00	3300.00	0.00	3300.00	0.00	20.000	6600.00	1	MAIZ RETRILLADO X 3 KILOS	COOPE DE MERCAI
140929	02/01/2008	2056	532.00	800.00	16.00	689.66	110.34	10.000	800.00	1	SALVADO DE TRIGO X 250 TONIC	PRODUCTOS TONI
140929	02/01/2008	348	947.37	1600.00	0.00	1600.00	0.00	0.6300	1008.00	1	TOMATE DE ARBOL KILO	CARULLA
140930	02/01/2008	3365	822.95	1150.00	16.00	991.38	158.62	10.000	1150.00	1	ESPONJILLA ATILA 72 X 6	TECNOQUIMICAS
140930	02/01/2008	5865	1861.33	2550.00	16.00	2198.28	351.72	10.000	2550.00	1	PAPEL HIG SUPER POPULAR 8 X 6	DISCENTRO LTDA
140930	02/01/2008	1463	2854.30	4100.00	16.00	3534.48	565.52	10.000	4100.00	1	VELON REPUESTO GRANDE	ESPERMAS CLARA
140931	02/01/2008	8670	984.60	1300.00	10.00	1181.82	118.18	10.000	1300.00	1	AZUCAR BLANCO RISARALDA 25 X 1000	INGENIO RDA
140931	02/01/2008	9982	1570.00	1700.00	0.00	1700.00	0.00	10.000	1700.00	1	LECHE LARGA VIDA LA PERLA X 900	ALIMENTOS DEL V
140931	02/01/2008	9982	1570.00	1700.00	0.00	1700.00	0.00	10.000	1700.00	1	LECHE LARGA VIDA LA PERLA X 900	ALIMENTOS DEL V
140932	02/01/2008	4991	1523.07	1800.00	0.00	1800.00	0.00	10.000	1800.00	1	FRIJOL CARGAMANTO ROJO X LIBRAS	FRIJOL
140933	02/01/2008	9506	463.00	850.00	16.00	732.76	117.24	10.000	850.00	1	CIGARRILLOS DERBY 10 X 1 SP	VARIOS
140934	02/01/2008	23	900.00	1000.00	0.00	1000.00	0.00	10.000	1000.00	1	HUESO CORRIENTE	COLCERDOS HUES
140935	02/01/2008	286	569.60	850.00	0.00	850.00	0.00	32.750	2783.75	1	PLATANO HARTON KILO	CARULLA
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS
140936	02/01/2008	5081	770.00	850.00	0.00	850.00	0.00	10.000	850.00	1	ARROZ ROA X 25 LIBRAS	ROA HERMANOS

Imagen No 6. Base de datos en Excel.

Ya con esta migración se logró obtener una mejor estructura (Imagen No. 6) de los datos, donde se puede apreciar cada atributo con sus respectivos valores. Estando en este formato se comenzó por analizar los datos que tenían incoherencias, lo cual se logró haciendo un ordenamiento por filas principalmente

en las filas de atributo ***nombre***, puesto que los atributos claves para esta investigación fueron, ***número*** el cual representa el número de la factura, ***código*** el cual representa el código del producto y ***nombre*** el cual representa el nombre del producto.

Una vez realizada dicha estructuración, se obtuvieron los siguientes resultados (Imagen No 7.). (Cabe resaltar que estos resultados son sacados de la última y definitiva base de datos de la cual se ampliará más adelante):

numero	fecha	codigo	costo	precio	tarifa_iva	precio_sin_iva	valor_iva	cantidad	valor_item	bodega	nombre	nclasifica1	clasifica1	nclasifica2	clasifica2
142408	06/01/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
144056	11/01/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
149143	26/01/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
150793	30/01/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
158881	20/02/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
163390	01/03/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
165059	06/03/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
180029	16/04/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
186014	06/05/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
195493	16/06/2008	993	3151.00	4250.00	16.00	3663.79	586.21	10.000	4250.00	1	.....	QUALA	280	CREMAS DENTALES	180
140992	02/01/2008	5430	948.27	1450.00	16.00	1250.00	200.00	10.000	1450.00	1	\N	\N	\N	\N	\N
141191	02/01/2008	4911	1270.00	1700.00	10.00	1545.45	154.55	10.000	1700.00	1	\N	\N	\N	\N	\N
141278	02/01/2008	5430	948.27	1450.00	16.00	1250.00	200.00	10.000	1450.00	1	\N	\N	\N	\N	\N
141313	02/01/2008	7540	3017.00	4100.00	16.00	3534.48	565.52	10.000	4100.00	1	\N	\N	\N	\N	\N
141313	02/01/2008	7540	3017.00	4100.00	16.00	3534.48	565.52	10.000	4100.00	1	\N	\N	\N	\N	\N
141332	03/01/2008	3074	11121.00	13800.00	16.00	11896.55	1903.45	10.000	13800.00	1	\N	\N	\N	\N	\N
141345	03/01/2008	3074	11121.00	13800.00	16.00	11896.55	1903.45	10.000	13800.00	1	\N	\N	\N	\N	\N
141533	03/01/2008	5271	1758.20	2350.00	16.00	2025.86	324.14	10.000	2350.00	1	\N	\N	\N	\N	\N
141674	04/01/2008	7540	3017.00	4100.00	16.00	3534.48	565.52	10.000	4100.00	1	\N	\N	\N	\N	\N
141833	04/01/2008	3292	6281.40	8250.00	16.00	7112.07	1137.93	10.000	8250.00	1	\N	\N	\N	\N	\N
141836	04/01/2008	2814	5118.07	7150.00	16.00	6163.79	986.21	10.000	7150.00	1	\N	\N	\N	\N	\N
141855	04/01/2008	7540	3017.00	4100.00	16.00	3534.48	565.52	10.000	4100.00	1	\N	\N	\N	\N	\N
142992	08/01/2008	4026	2490.00	2900.00	0.00	2900.00	0.00	10.000	2900.00	1	\N	\N	\N	\N	\N
143131	08/01/2008	5271	1758.20	2350.00	16.00	2025.86	324.14	10.000	2350.00	1	\N	\N	\N	\N	\N
143131	08/01/2008	5271	1758.20	2350.00	16.00	2025.86	324.14	10.000	2350.00	1	\N	\N	\N	\N	\N
143144	08/01/2008	4026	2490.00	2900.00	0.00	2900.00	0.00	10.000	2900.00	1	\N	\N	\N	\N	\N
143173	08/01/2008	3074	11121.00	13800.00	16.00	11896.55	1903.45	10.000	13800.00	1	\N	\N	\N	\N	\N
143388	09/01/2008	5271	1758.20	2350.00	16.00	2025.86	324.14	10.000	2350.00	1	\N	\N	\N	\N	\N
143434	09/01/2008	4026	2490.00	2900.00	0.00	2900.00	0.00	10.000	2900.00	1	\N	\N	\N	\N	\N
143452	09/01/2008	9146	2884.00	3600.00	10.00	3272.73	327.27	10.000	3600.00	1	\N	\N	\N	\N	\N
144280	12/01/2008	5271	1758.20	2350.00	16.00	2025.86	324.14	10.000	2350.00	1	\N	\N	\N	\N	\N
144364	12/01/2008	9429	1575.00	2200.00	16.00	1896.55	303.45	10.000	2200.00	1	\N	\N	\N	\N	\N
144508	12/01/2008	3292	6281.40	8250.00	16.00	7112.07	1137.93	10.000	8250.00	1	\N	\N	\N	\N	\N
145660	15/01/2008	2950	6033.00	8350.00	16.00	7198.28	1151.72	10.000	8350.00	1	\N	\N	\N	\N	\N
146586	18/01/2008	4244	5410.34	7500.00	16.00	6465.52	1034.48	10.000	7500.00	1	\N	\N	\N	\N	\N
148022	23/01/2008	5430	948.27	1450.00	16.00	1250.00	200.00	10.000	1450.00	1	\N	\N	\N	\N	\N
148213	23/01/2008	5430	948.27	1450.00	16.00	1250.00	200.00	10.000	1450.00	1	\N	\N	\N	\N	\N
149243	26/01/2008	5430	948.27	1450.00	16.00	1250.00	200.00	20.000	2900.00	1	\N	\N	\N	\N	\N
150018	28/01/2008	8094	6880.72	8750.00	0.00	8750.00	0.00	10.000	8750.00	1	\N	\N	\N	\N	\N
151371	01/02/2008	4698	5687.68	7800.00	16.00	6724.14	1075.86	10.000	7800.00	1	\N	\N	\N	\N	\N
152765	04/02/2008	3292	6281.40	8250.00	16.00	7112.07	1137.93	10.000	8250.00	1	\N	\N	\N	\N	\N
152982	05/02/2008	521	2917.60	3950.00	16.00	3405.17	544.83	10.000	3950.00	1	\N	\N	\N	\N	\N
153735	07/02/2008	4698	5687.68	7800.00	16.00	6724.14	1075.86	10.000	7800.00	1	\N	\N	\N	\N	\N
155255	11/02/2008	3292	6281.40	8250.00	16.00	7112.07	1137.93	10.000	8250.00	1	\N	\N	\N	\N	\N
155842	12/02/2008	7239	660.00	950.00	16.00	818.97	131.03	10.000	950.00	1	\N	\N	\N	\N	\N
155842	12/02/2008	7239	660.00	950.00	16.00	818.97	131.03	10.000	950.00	1	\N	\N	\N	\N	\N
155842	12/02/2008	7239	660.00	950.00	16.00	818.97	131.03	10.000	950.00	1	\N	\N	\N	\N	\N
157054	15/02/2008	3292	6281.40	8250.00	16.00	7112.07	1137.93	10.000	8250.00	1	\N	\N	\N	\N	\N
159406	21/02/2008	4698	5687.68	7800.00	16.00	6724.14	1075.86	10.000	7800.00	1	\N	\N	\N	\N	\N

Imagen No 7. Incoherencias en la Base de Datos.

Como se puede observar en la Imagen No 7., se tienen incoherencias en cuanto al *nombre del producto*, nombre de la clasificación 1, código de la clasificación 1, nombre de la clasificación 2 y código de la clasificación 2; de donde en total se eliminaron los siguientes artículos (tabla No 5.):

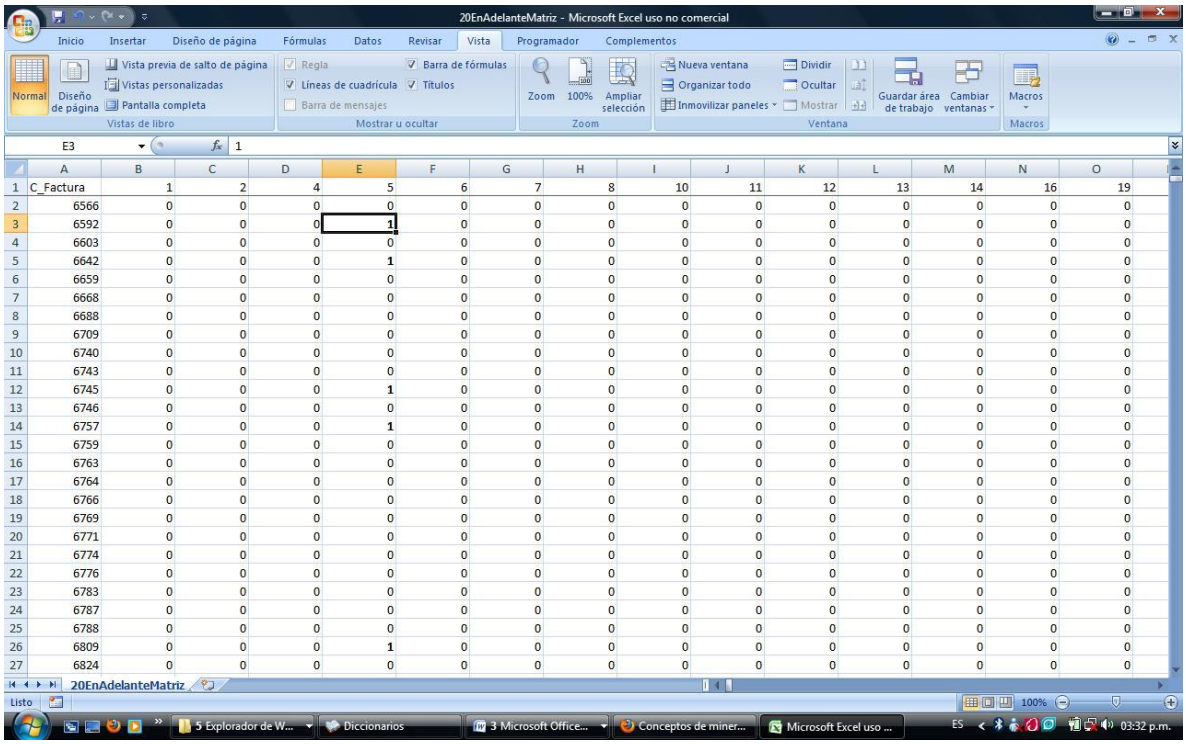
Códigos sin nombre (\N ó …. = 17):

codigo	nombre	nclasifica1	clasifica1	nclasifica2	clasifica2
993	.....	QUALA	280	CREMAS DENTALES	180
5430	\N	\N	\N	\N	\N
4911	\N	\N	\N	\N	\N
7540	\N	\N	\N	\N	\N
3074	\N	\N	\N	\N	\N
5271	\N	\N	\N	\N	\N
3292	\N	\N	\N	\N	\N
2814	\N	\N	\N	\N	\N
4026	\N	\N	\N	\N	\N
9146	\N	\N	\N	\N	\N
9429	\N	\N	\N	\N	\N
2950	\N	\N	\N	\N	\N
4244	\N	\N	\N	\N	\N
8094	\N	\N	\N	\N	\N
4698	\N	\N	\N	\N	\N
521	\N	\N	\N	\N	\N
7239	\N	\N	\N	\N	\N

Tabla No 5. Códigos de producto eliminados por incoherencias.

Después de eliminar los datos innecesarios, se pasó a seleccionar los atributos precisos para la creación de la matriz transaccional, la cual es uno de los principales pasos para poder aplicar el algoritmo ***FP-Growth*** de **Reglas de asociación** con el fin de obtener relaciones entre ítems.

Para dicha matriz los atributos seleccionados en este caso son el *número de factura* y el *código de producto*, ya que con solo ellos, se pueden describir las *compras* hechas por cada factura; la matriz que se creó tiene como columnas los códigos de productos y como filas los números de factura, y cada celda que confronta un número de factura con un código de producto tiene un valor binario (0 ó 1) el cual representa si se compró dicho producto en dicha factura, siendo 1 comprado y 0 no comprado. Una vista del resultado de la matriz es la siguiente (Imagen No.8):



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	C_Factura	1	2	4	5	6	7	8	10	11	12	13	14	16	19
2	6566	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	6592	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	6603	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	6642	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	6659	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	6668	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	6688	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	6709	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	6740	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	6743	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	6745	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	6746	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	6757	0	0	0	1	0	0	0	0	0	0	0	0	0	0
15	6759	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	6763	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	6764	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	6766	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	6769	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	6771	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	6774	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	6776	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	6783	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	6787	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	6788	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	6809	0	0	0	1	0	0	0	0	0	0	0	0	0	0
27	6824	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Imagen No 8. Ejemplo de matriz utilizada.

Pero para lograr el llenado de ésta matriz (Imagen No 8.) se tuvo que hacer un programa, aprovechando la capacidad que trae Excel de crear y ejecutar programas desarrollados en Visual Basic.

Una vez obtenida la matriz completamente diligenciada, lo cual es un proceso un poco largo dependiendo del tamaño de la matriz, se almacenó el archivo como un CSV (Comma Separated Values: Valores Separados por Comas), porque es uno de los formatos soportados por la herramienta ***RapidMiner*** y que es generado por Excel.

A pesar de tener tantas fuentes que hablaban acerca del proceso de limpieza, en ninguna se expresaba de una manera explícita cómo era el procedimiento o cuáles eran los operadores apropiados para dicha limpieza, por tanto el llegar a la matriz, como se mencionó anteriormente, se realizo implementando los conocimientos adquiridos en el campo de la ingeniería, más directamente en la programación.

De la misma manera, cuáles son los algoritmos previos a utilizar para hacer el análisis de la canasta de mercado fue una decisión basada en los varios documentos consultados referentes a *Minería de Datos*, de los cuales el mejor algoritmo que busca los ítems frecuentes en una base de datos es el ***FP-Growth*** *(*Ver *FP-Growth* pág. 69 *)*, y dicha mejoría se basa en una combinación de *Costo computacional, Tiempo de ejecución y Rendimiento*; de dichos ítems se sacan las

Reglas de Asociación necesarias de acuerdo a ciertos criterios que cada algoritmo utiliza para hacer el descarte de cada ítem y regla respectivamente, dichos criterios son el ***soporte*** y la ***confianza***.

Estos criterios son dependientes del usuario, es decir el *minero* es quien decide cual va a ser el soporte y la confianza para sacar sus propios resultados de acuerdo al valor que les dé. Pero para esta herramienta, el tener la base de datos en forma de una matriz transaccional, no era suficiente ya que el algoritmo ***FP-Growth*** solo maneja atributos nominales, y la matriz como estaba, tenía valores numéricos; para solucionar este problema se acudió a un foro (una muestra de lo que fue el foro que se abrió en la página de la herramienta se presentará como Anexo No 1 al final del trabajo) que la empresa desarrolladora de la herramienta ***RapidMiner*** tiene a disposición de los usuarios, allí luego de hacer una navegación profunda y de contar con varios colaboradores, se conoció la existencia de un operador llamado ***Numerical2Nominal*** el cual evitaría el error que hasta el momento presentaba la herramienta al hacer los intentos y éste fue el siguiente (Imagen No.9):

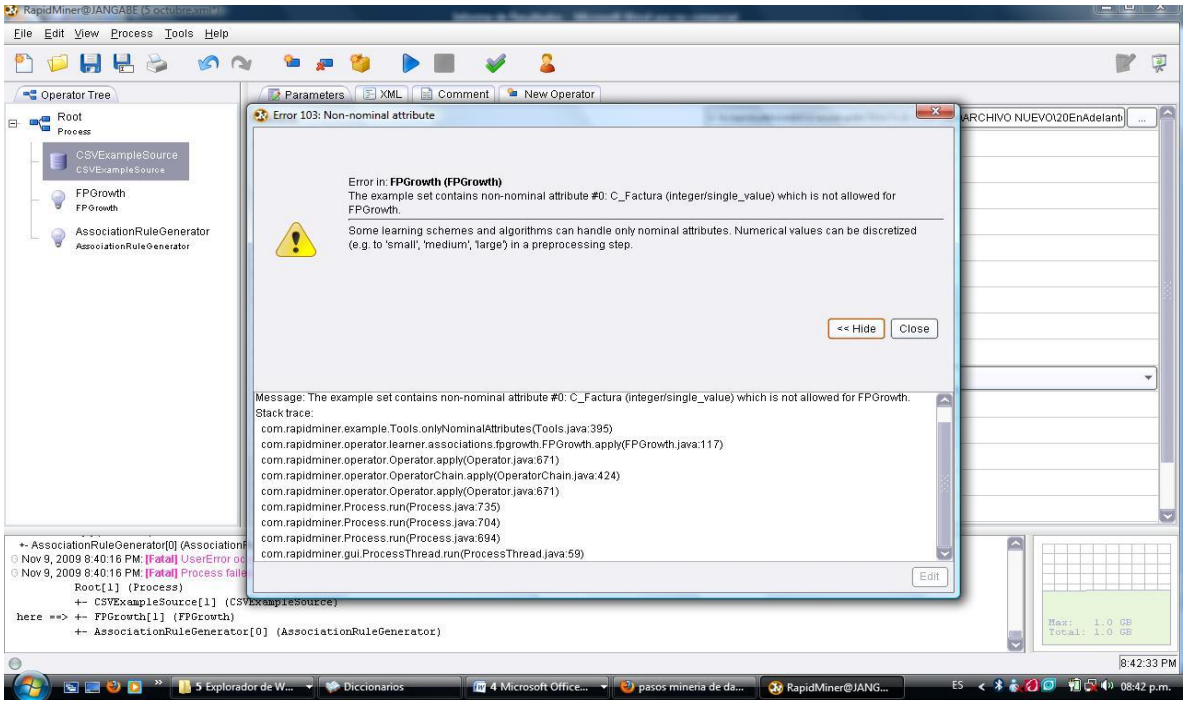


Imagen No 9. Error presentado por la herramienta.

Ya el operador *Numerical2Nominal* solucionó el anterior problema y permitió seguir adelante con la investigación.

Al principio, y debido a la restricción de Excel (office 2003), solo se pudo contar con una base de datos con un total de 65.535 registros (filas), lo cual obviamente dio pie a algunas dudas en cuanto al tamaño, ya que casualmente ese es el tamaño máximo permitido por Excel.

Los primeros intentos con estos datos no arrojaron ninguna regla, lo cual generó la duda si la herramienta era confiable o no, para lo cual se logró conseguir un archivo para ser usado como prueba y que se confrontó con otra herramienta llamada ***XLMiner***, la cual no es una herramienta libre pero que tiene una versión de prueba con ciertas restricciones en el tamaño de los datos y que para este fin fue apropiada, además se hizo una prueba manual; los resultados de ambas herramientas y la prueba manual fueron exactamente iguales, teniendo como resultado lo siguiente (Ver imágenes 10, 11 y 12):

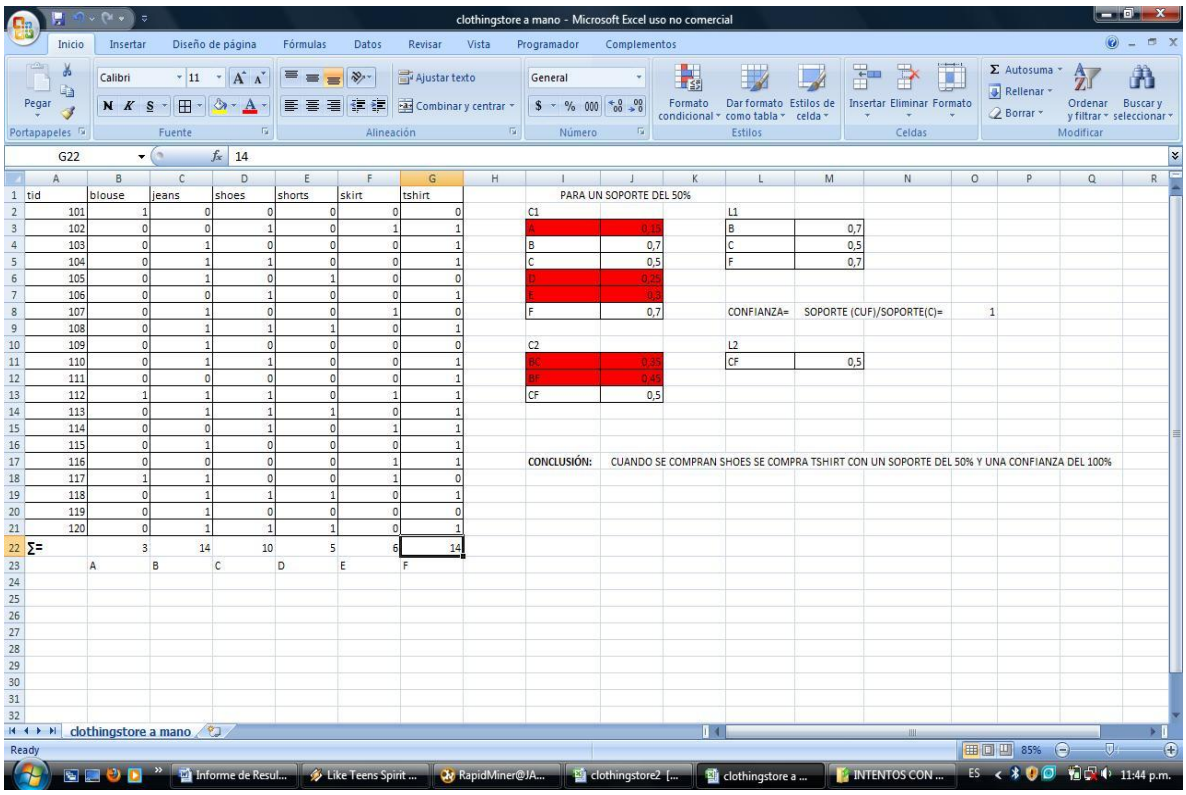


Imagen No 10. Prueba Manual (Soporte = 50% y Confianza = 100%).

XLMiner : Association Rules

Data	
Input Data	clothingstore2!$A$1:$F$21
Data Format	Binary Matrix
Minimum Support	10
Minimum Confidence %	100
# Rules	1
Overall Time (secs)	1

Place the cursor on a cell in the rules table to read a rule.
Use up / down arrow keys to browse through the rules.

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support (a)	Support (c)	Support (a U c)	Lift Ratio
1	100	shoes=>	tshirt	10	14	10	1,42857

Imagen No 11. Prueba con XLMiner (Soporte = 50% y Confianza = 100%).

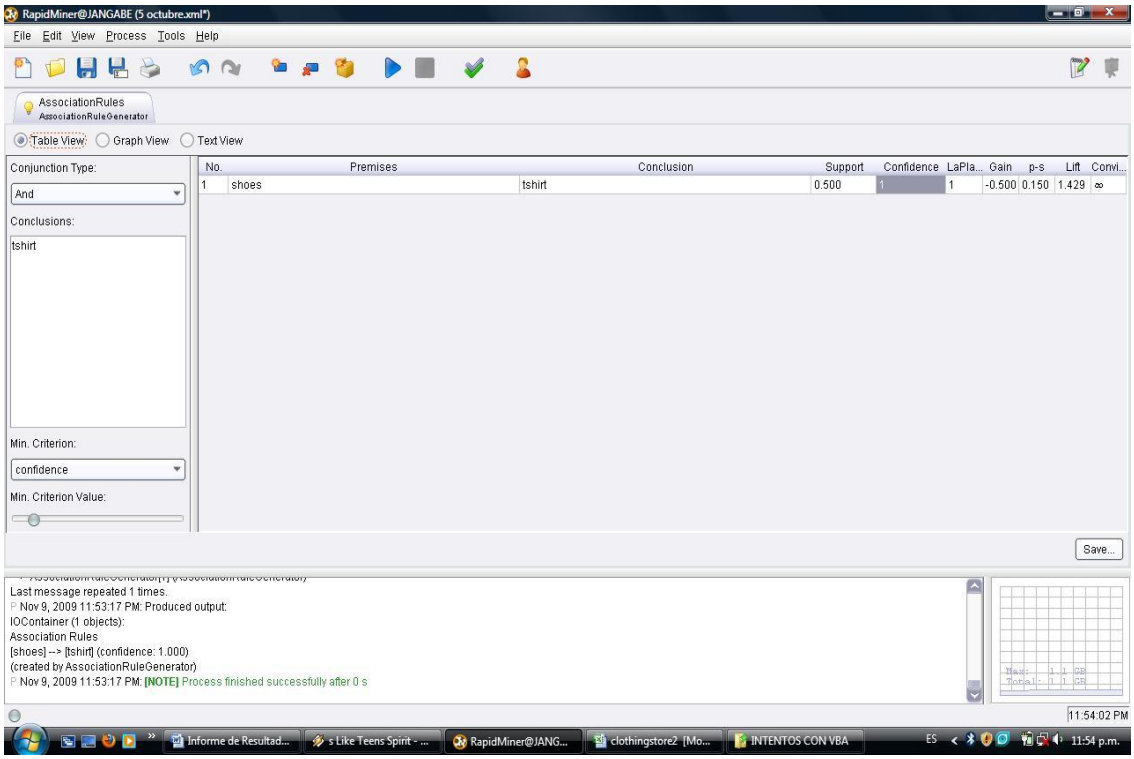


Imagen No 12. Prueba con *RapidMiner* (Soporte = 50% y Confianza = 100%).

Como se pudo observar en las anteriores imágenes (10, 11 y 12), que a su vez representan en cada una de ellas las diferentes herramientas para su sustento, muestran que el resultado en las tres es el mismo, teniendo en cuenta igual nivel de ***soporte*** y ***confianza***; esta prueba se convirtió en la prueba piloto que da credibilidad a los resultados arrojados por la herramienta ***RapidMiner***.

4.3 SELECCIÓN DEL MODELO.

De acuerdo a los puntos anteriores, el modelo seleccionado fue el siguiente:

- Las columnas seleccionadas para el análisis de la base de datos las cuales fueron: *Número de Factura* y *Código de Producto*.
- Se seleccionaron los algoritmos ***FP-Growth*** y ***AssociationRules*** para obtener resultados.
- Se tienen en cuenta el soporte y la confianza como parámetros que indican a los algoritmos cómo procesar los datos.

4.4 APLICACIÓN DE LA TÉCNICA

Retomando parte de lo mencionado anteriormente, la aplicación de la técnica se hizo pero con unos resultados insatisfactorios, por lo cual fue necesario primero que todo hacer la verificación de la confiabilidad de la herramienta, luego se procedió a hacer un análisis para tratar de tener idea del por qué no se obtenían resultados hasta el momento. Dicho análisis consistió en hacer un migrado de la base de datos a ***Postgres*** *(Gestor de bases de datos)*. Una vez tenida la base de datos allí se procedió a hacer las siguientes consultas (por consola) con sus respectivos resultados:

```
postgres=# select * from itemsPorFactura order by articulosFactura desc,
numeroFactura asc;
```

Figura No 3. Consulta SQL para conocer cuántos artículos se compraron en cada factura

Con esta consulta (Figura No.3) se puede apreciar cuantos artículos se compraron en cada factura de acuerdo al siguiente resultado (Tabla No 6):

Número factura	Artículo factura
209350	67
173594	53
184030	51
187885	46
206042	45
179672	41
205078	41
186965	39
216576	36
176186	35
195836	35
216265	35
184320	34
179526	33
174371	32
175449	32
184728	32
188420	32
190022	32
190755	32
192328	32
215845	32
--Más--	

Tabla No 6. Total artículos en cada factura

Viendo el anterior resultado se puede observar que entre los datos con los que se contaban el máximo de productos comprados fue de 67; ahora con el resultado de la siguiente consulta (Figura No 4.) se puede apreciar en cuántas facturas se compro dicho número de productos:

```
postgres=# select articulosFactura, count(*) from itemsPorFactura group by articulosFactura;
```

Figura No 4. Consulta SQL para conocer cuántas veces se compró la misma cantidad de productos

El resultado de la anterior consulta arrojó los siguientes resultados (Tabla No.7):

artículosFactura	Count
67	1
53	1
51	1
46	1
45	1
41	2
39	1
36	1
35	3
34	1
33	1
32	8
30	4
29	8
28	2
27	8
26	12
25	12
24	15
23	13
22	22
21	16
20	23
19	41
18	40
17	62
16	78
15	79
14	90
13	103
12	154
11	158
10	204
9	248
8	266
7	419
6	587
5	846
4	1233
3	2229
2	4742
1	12154

Tabla No 7. Número de veces en la venta de una misma cantidad de producto

Observando este resultado se puede apreciar que la compra de muchos productos al mismo tiempo es muy baja, contrario a las compras unitarias, las cuales son de 12.154 de un total de 23.890 como se puede apreciar gracias a la siguiente consulta (Figura No.5):

postgres=# select count(distinct numero) from "DatosLa29";
Total = 23.890

Figura No 5. Total de facturas en la base de datos

Y lo cual muestra que esa información hace obvio el fallo en los resultados, debido a que esas 12.154 facturas de compras unitarias, representan más de un 50% del

total de las facturas, haciendo menos probable el hallazgo de ítems que se venden juntos y con reiteración. A estas alturas y con la base de datos que se tenía, se creía que solo había un total de 1.481 productos gracias a la consulta (Figura No.6):

postgres=# select count(distinct codigo) from "DatosLa29";
Total = 1.481

Figura No.6 Total de artículos en la base de datos.

Para conocer cuantas veces se compró un determinado artículo se ejecutó la siguiente consulta (Figura No.7):

```
postgres=# select codigo as codigo, count(codigo) as cantidad from
"DatosLa29" group by codigo order by cantidad desc;
```

Figura No 7. Cuantas veces se compró determinado artículo en total.

Los resultados para dicha consulta fueron (Tabla No.10):

Código	Cantidad
4995	1987
6391	1645
4550	1641
6049	1486
3241	1356
9523	1346
298	1100
903	840
4810	811
54	808
5021	705
9820	662
4994	631
4822	616
4828	597
4717	592
8847	547
1307	519
4992	518
4119	447
3978	418
1697	417

Tabla No 10. Cuantas veces se compró determinado artículo en total.

Luego de haber realizado varios análisis sobre la base de datos, se tuvo la posibilidad de trabajar con el Excel del office 2007; se le hizo la migración a esta herramienta, donde se encontró que la base de datos original no era nada comparado a lo que se había encontrado al comienzo, ya que los 65.535 registros se convirtieron ahora en 424.811 registros, teniendo una diferencia enorme entre ambas bases de datos.

Al tener esta nueva base de datos se procedió a aplicar las mismas consultas en *Postgres*, para lo cual por ser una base de datos mucho más grande, se tuvo que hacer un llenado de la tabla en *Postgres* por partes, realizando en cada parte una inserción en la base de datos de a 10.000 registros a la vez, tomando cierto tiempo en hacerlo; los resultados de las consultas anteriormente aplicadas a la base de datos inicial arrojaron los siguientes resultados:

postgres=# select count(distinct codigo) from "DatosLa29";
Total = 5.768

Figura No 8. Total de artículos en la base de datos

(Se pasó de tener 1.481 productos a 5.768.)

postgres=# select count(distinct numero) from "DatosLa29";
Total = 76.571

Figura No 9. Total de facturas en la base de datos.

(Se pasó de tener 23.890 facturas a 76.571.)

postgres=# select * from itemsPorFactura2 order by articulosFactura desc, numeroFactura asc;

Figura No 10. Consulta SQL para determinar cuántos artículos se compraron en determinada factura.

El resultado obtenido de la anterior consulta (Figura No.10) se muestra en la siguiente tabla (Tabla No 9.):

Numerofactura	Articulosfactura
173594	181
186965	175
191136	170
184030	167
205078	163
187885	160
6757	154
179158	149
206042	147
162717	145
195272	145
6766	133
209350	133
145297	131
162925	131
6995	127
209177	127
184728	126
151308	125
211880	125
162864	122
170805	120
185575	120
197501	119
6849	118
147440	118
144069	115
176330	115
163185	114
190581	114
214407	113
214501	113
195836	112
148372	111
166674	110
184320	109
190755	109
204854	108
159544	107
162386	105
216235	105
176186	104
192328	104
174322	103
177931	102
165683	101
182566	101
142209	100
150590	100
155567	99
179421	99
160568	98
188420	98
191380	98
214395	98
163262	97
165532	97
186883	97
199242	97
152676	96
152912	96
143725	95
210676	95
151399	93
208527	93
155771	92
168902	92
178622	92
185902	92
174236	91
--Más--	

Tabla No 9. Cuantos artículos se compraron en determinada factura.

(Se pasó de un máximo de 67 productos comprados en una factura a 181)

```
postgres=# select articulosFactura, count(*) from itemsPorFactura2 group by articulosFactura;
```

Figura No 11. Consulta SQL para determinar cuántas veces se compró determinada cantidad de productos en una factura

La anterior consulta (Figura No.11) arrojó los siguiente resultados (Tabla No 10):

Articulosfactura	Count
181	1
175	1
170	1
167	1
163	1
160	1
154	1
149	1
147	1
145	2
133	2
131	2
127	2
126	1
125	2
122	1
120	2
119	1
118	2
115	2
114	2
113	2
112	1
111	1
110	1
109	2
108	1
107	1
105	2
104	2
103	1
102	1
101	2
100	2
99	2
98	4
97	4
96	2
95	2
93	2
92	4
91	4
90	3
89	2
88	5
87	4
86	1
85	2
84	10
83	4
82	10
81	5
80	7
79	8
78	4
77	8
76	15
75	10
74	13
73	15
72	13
71	18
70	17
69	13
68	21
67	21
66	19
65	16
64	27
63	17
62	35
61	21
60	20
59	32
58	28
57	36
56	28
55	35
54	43
53	39
52	45
51	37
50	42
49	41
48	51
47	53
46	44
45	54
44	53
43	63
42	76
41	77
40	76
39	78
38	86
37	90
36	110
35	88
34	100
33	117
32	137
31	129
30	119
29	138
28	154
27	161
26	160
25	176
24	193
23	195
22	221
21	275
20	254
19	283
18	304
17	376
16	434
15	498
14	579
13	671
12	765
11	959
10	1140
9	1300
8	1678
7	2229
6	3131
5	4138
4	5964
3	8603
2	13244
1	25884

Tabla No 10. Cuantas veces se compró determinada cantidad de productos en una factura.

Luego de consultar la cantidad de productos que se adquirieron en una factura, se ejecutó la siguiente consulta (Figura No.12) para determinar un producto cuántas veces se compró:

```
postgres=# select codigo as codigo, count(codigo) as cantidad from "DatosLa29" group by codigo order by cantidad desc;
```

Figura No 12. Consulta SQL para determinar cuántas veces se compró determinado artículo en específico.

La anterior consulta (Figura No 12.), arrojó los resultados que se muestran en la tabla a continuación (Tabla No 11.):

Codigo	Cantidad
5081	12656
286	4419
235	3765
6867	3742
3956	3626
7474	3446
4762	3223
4995	3183
8670	3010
4550	3010
5	2756
6049	2558
6391	2416
228	2399
6044	2213
3241	2201
8137	2194
261	2054
9523	2036
903	1993
307	1813
9854	1809
298	1795
4991	1727
4936	1591
301	1578
3952	1547
6946	1479
4912	1330
4810	1232
1944	1227
803	1216
54	1215
4920	1206
6141	1162
6045	1135
5020	1122
954	1121
5021	1120
4999	1107
6920	1095
5014	1065
4882	1056
4994	1032
1812	1008
5028	1003
4828	997
8849	966
4822	964
8847	959
9820	922
9926	918
4717	916
520	911
273	910
1307	896
4866	894
4992	888
1945	872
4867	871
262	871
265	862
308	858
4834	843
3133	833
2844	821
4119	786
1697	783
6046	782
3859	778
1854	771
9811	769
1532	765
4823	731
3666	720
272	720
1832	718

Tabla No 11. Cuantas veces se compró determinado artículo en específico.

(Se pasó de un artículo comprado máximo 1.987 veces a 12.656)

Como se puede apreciar, la variabilidad en cuanto a todas las consultas realizadas fue muy grande; pero desafortunadamente, el hecho de ser una base de datos tan inmensa, hace que el manejo de la misma sea más difícil, ya que el consumo de recursos es superior; al intentar aplicar *Minería de Datos* con la B.D completa se presentaron errores en cuanto a memoria por parte de Excel en el momento de crear la matriz; con ***RapidMiner*** se presentó el siguiente error (Imagen No 13.):

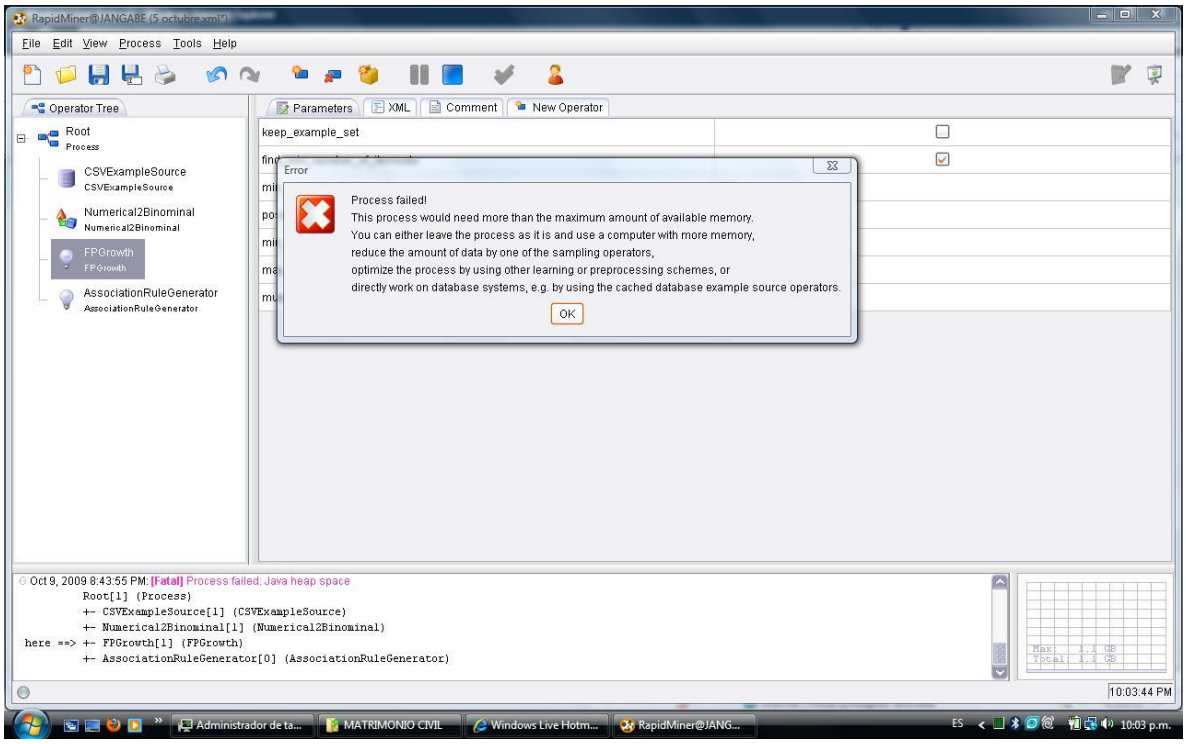


Imagen No 13. Error por desbordamiento de memoria.

Allí se aclara que el proceso necesita más memoria de la que se tiene disponible, que se puede o usar un computador con más memoria o reducir la cantidad de datos.

Haciendo uso de solo las facturas con compras mayores o iguales a 20, se redujo la cantidad que antes era de 76.573 a solo 4.392, y los productos de 5.779 a 5.258. Al tener esa disminución en los datos se pudo hacer una matriz para la cual no se requería tanto consumo de memoria o por lo menos se hacía factible para el computador con el cual se contaba, con ésta base de datos se obtuvieron los siguientes resultados aplicando el modelo anteriormente establecido:

Reglas de asociación obtenidas para un soporte mayor que 0 y menor o igual al 12%, con una confianza mayor que 70%:

REGLAS DE ASOCIACIÓN OBTENIDAS		
Soporte entre **0% y 12%** y Confianza **> 70%**		
Regla	**Soporte**	**Cofianza**
[7474, 4999] --> [4995]	**0.059**	**0.796**
[272] --> [298]	**0.055**	**0.778**
[298, 228] --> [261]	**0.057**	**0.750**
[369] --> [298]	**0.059**	**0.750**
[4994] --> [4995]	**0.093**	**0.746**
[4999] --> [4995]	**0.114**	**0.735**

Tabla No 12. Reglas de asociación obtenidas Soporte entre 0% y 12% y Confianza > 70%.

REGLAS DE ASOCIACIÓN OBTENIDAS		
Soporte entre **0**% y **12**% y Confianza > **50**%		
Regla	**Soporte**	**Confianza**
[7474, 4999] --> [4995]	**0.059**	**0.796**
[272] --> [298]	**0.055**	**0.778**
[298, 228] --> [261]	**0.057**	**0.750**
[369] --> [298]	**0.059**	**0.750**
[4994] --> [4995]	**0.093**	**0.746**
[4999] --> [4995]	**0.114**	**0.735**
[261, 228] --> [298]	0.059	0.689
[298] --> [261]	0.096	0.658
[261] --> [298]	0.096	0.653
[228] --> [261]	0.086	0.651
[273] --> [298]	0.051	0.634
[261, 298] --> [228]	0.059	0.620
[228] --> [298]	0.079	0.589
[261] --> [228]	0.086	0.588
[8847] --> [4995]	0.050	0.585
[4828] --> [4995]	0.054	0.560
[4992] --> [4995]	0.051	0.545

[4991] --> [4995]	0.085	0.544
[298] --> [228]	0.079	0.544
[6046] --> [4995]	0.052	0.535
[6141] --> [4995]	0.070	0.528
[6946] --> [4995]	0.066	0.523
[8849] --> [4995]	0.052	0.516
[4995, 4999] --> [7474]	0.059	0.514
[307] --> [298]	0.056	0.511
[6141] --> [7474]	0.068	0.509
[307] --> [261]	0.056	0.509
[7474] --> [4995]	0.173	0.509
[286] --> [298]	0.054	0.506
[4994] --> [7474]	0.063	0.503
[4882] --> [7474]	0.059	0.5
[5021] --> [4995]	0.053	0.5

Tabla No 13. Reglas de asociación obtenidas Soporte entre 0% y 12% y Confianza > 50%.

Entre las tablas anteriores (Tablas No.12 y 13), se puede apreciar que cuando la confianza es menor, el número de reglas de asociación que se obtiene se incrementa, teniendo además de las que ya se habían producido otras más. Pero puede que sea un poco confuso tal y como se muestra en dichos cuadro, por tanto a continuación se muestra la forma correcta como se deben interpretar dichos resultados, no sin antes mostrar que significa cada código:

CodigoProducto	Nombre
228	TOMATE LARGA VIDA KILO
261	CEBOLLA HUEVO BLANCA KILO
272	HABICHUELA KILO
273	LECHUGA BATAVIA KILO
286	PLATANO HARTON KILO
298	ZANAHORIA KILO
307	PAPA R-12 KILO
369	PAPA CRIOLLA KILO
4828	ESPAGUETTIS LA MUÑECA 24 X 250

4882	CHOCOLATE LUKER 100UND X 250
4991	FRIJOL CARGAMANTO ROJO X LIBRAS
4992	FRIJOL BOLON X LIBRAS
4992	ARVEJA X LIBRAS
4995	LENTEJA X LIBRAS
4999	BLANQUILLOS LA 29 25 X 500
5021	AZUCAR LA 29 15 X 3 KILOS
6046	ARROZ ROA 3UN X 5KILOS
6141	JABON REY 16 X 3
6946	GALLETAS SALTIN PENTA TACO X 24
7474	SAL REFISAL X KILO
8847	FRIJOL CARGAMANTO ROJO X KILOS
8849	FRIJOL BOLON X KILO

Tabla No 14. Códigos de producto con su respectivo nombre.

En la tabla anterior (Tabla No 14) se puede identificar cada código implícito en las tablas de reglas obtenidas (Tablas No.12 y 13) con su respectivo nombre; la interpretación de la regla es:

REGLA DE ASOCIACIÓN	INTERPRETACIÓN
[4999] → [4995] [Blanquillos la 29 25x500] → [lenteja x libras]	Cuando se compran blanquillos la 29 25x500, se compra lenteja x libras con un soporte del 11,4% y una confianza del 73,5%
[272] → [298] [Habichuela kilo]→[zanahoria kilo]	Cuando se compra habichuela kilo, se compra zanahoria kilo con un soporte del 5,5% y una confianza del 77,8%
[298, 228] → [261] [Zanahoria kilo, tomate larga vida kilo] → [cebolla de huevo blanca kilo]	Cuando se compran zanahoria kilo y tomate larga vida kilo, se compra cebolla de huevo blanca kilo con un soporte del 5,7% y una confianza del 75%
[5021]→[4995] [Azúcar la 29 15x3 kilos] → [lenteja x libras]	Cuando se compra azúcar la 29 15x3 kilos, se compra lenteja x libras con un soporte del 5,3 % y una confianza del 50%

Tabla No 15. Interpretación de las reglas de asociación.

Como se puede observar en la tabla anterior (Tabla No.15) la interpretación de cada regla resultante no es difícil, lo que se debe tener claro es que por ejemplo, para la última regla allí interpretada, se tiene que ese 5,3% de soporte hace referencia a que de una base de datos en la cual el número de transacciones es

de 162.835, en 8.630 de esas transacciones cuando se compró *azúcar la 29 15x3 kilos* también se compró *lenteja x libras*; y que la confianza sería conociendo de antemano el soporte del *azúcar la 29 15x3 kilos*, hacer la división del soporte de *azúcar* → *lenteja*, entre el soporte del *azúcar* que para obtener este resultado de 50% debe ser igual a 10,6%.

5. CONCLUSIONES.

Teniendo en cuenta el desarrollo del anterior trabajo, se concluye lo siguiente.

- La herramienta utilizada ***RapidMiner***, es una muy buena ayuda para llevar a cabo un proceso de *Minería de Datos*, por un lado porque es muy completa en cuanto a todo lo que en este campo se necesita, y por otro porque aparte de que tiene facilidades para la enseña el manejo de la herramienta más a fondo, cuenta con un foro en el que se da una atención excelente y se resuelven muchas dudas que surgen a lo largo de la experiencia de realizar un proyecto de *Minería de Datos*.

- A pesar que se pueden evidenciar resultados positivos, existe una gran variedad de software que no es libre el cual posiblemente facilite el preprocesado de los datos, herramientas como: SPSS Clementine, SPSS, SAS, entre otras.

- Este trabajo parte de la necesidad de encontrar comportamientos en los inventarios, los cuales a simple vista serían tediosos, por no decir imposible, de visualizar, tales como tendencias de compras, gustos o hábitos en las mismas, con el fin de establecer políticas para la fidelidad de los mismos y el crecimiento de las ventas, que es el fin último que se persigue.

- El objetivo del análisis fue poder encontrar patrones de comportamiento en las compras por parte de los clientes, y poder descubrir que productos son adquiridos al mismo tiempo con una frecuencia interesante para llegar a contribuir en el incremento de las ventas.

- Obtener un modelo soportado en las tecnologías de la información y las comunicaciones que simule o permita predecir el comportamiento de los clientes al hacer sus compras, propone una clara ventaja competitiva para las empresas.

- Para realizar un proceso de *Minería de Datos*, se hace necesario tener un gran conocimiento sobre las diferentes tareas que en este campo se presentan para poder tener claridad de cuál de estas es la que en realidad se acomoda a los requerimientos y necesidades del cliente.

- Se hace evidente que el proceso de *Minería de Datos*, es una tarea que no es fácil de realizar ya que se necesita hacer una gran cantidad de pruebas para llegar a obtener algún resultado, quizás no el mejor o esperado pero si

uno que puede dar cierta claridad en los movimientos que se pueden dar en grandes superficies de ventas como en este caso.

- Las reglas de asociación como técnica de *Minería de Datos* ofrecen muy buena alternativa para el análisis de canasta de mercado, permitiendo encontrar patrones de conducta para apoyar la toma de decisiones en marketing.

- Los algoritmos ***FP-Growth*** y ***AssociationRulesGenerator***, fueron utilizados para este proyecto ya que son los que más se acoplan a las necesidades que en este caso surgen de una base de datos que contiene movimientos de inventario de una gran superficie de venta, y al mismo tiempo arrojan información que es sencilla de interpretar.

- El uso del algoritmo *FP-Growth*, uno de los más rápidos y eficientes para establecer reglas de asociación, facilita la realización del proceso de selección de dichas reglas en forma rápida, ya que permite encontrarlas sin selección de candidatos.

- Como se puede observar en las Tablas No.12 y 13, se destaca que hay una variación en cuanto al número de reglas generadas en el orden de mayor número de reglas para un menor valor de la confianza; esto sucede debido a que la confianza es un valor resultante de dividir el soporte de la premisa entre el soporte de la conclusión; por tanto la variación del valor de la confianza varia el resultado de las reglas de asociación.

- El soporte al igual que la confianza es un factor determinante a la hora de obtener resultados, ya que es el primer filtro que se hace de los datos antes de pasarlo por el filtro que se realiza con la confianza.

- Teniendo en cuenta la gran cantidad de productos que maneja el supermercado, se presentó una particular situación, la cual consistió en que el mayor número de transacciones presentadas fue de compras unitarias y éstas a su vez fueron de productos distintos en la mayoría de los casos, haciendo que el proceso sea menos efectivo y que las reglas de asociación generadas no sean suficientes o por lo menos el soporte y la confianza con la que resultan, es bajo, dando la impresión de que el supermercado maneja demasiados productos para la cantidad de ventas que realiza, y al haber tantos productos con el mismo fin pero de diferentes marcas, se presta para que cada persona que compra por decir mantequilla, no compre la misma sino que cada quien compre de cada diferente marca disponible, o sabor, o característica que hace que una mantequilla se diferencie de otra y por tanto tenga su propia identificación.

- Los procesos ejecutados en la aplicación de *Minería de Datos* requieren de unas capacidades computacionales un poco costosas, eso sí, dependiendo del tamaño de la base de datos que se utilice para esto, y de la habilidad que se tenga para el manejo de dicha base de datos combinado con el conocimiento sobre el tema de minería.

Como sugerencias se tiene lo siguiente:

1. Se debe tener claro el tipo de datos al cual se le va a aplicar la *Minería de Datos*, ya que de esto depende que se elijan los correctos algoritmos para encontrar los mejores resultados.

2. Se recomienda aplicar la técnica utilizada en este proyecto, realizando un manejo diferente de los datos, aplicando taxonomías de acuerdo a las diferentes categorizaciones en las que se puede incluir cada ítem, un ejemplo sería crear la categoría mantequillas, donde se incluyan todas las mantequillas manejadas por la superficie de venta independiente de la marca, el sabor, el proveedor, entre otros, con el fin de obtener mejores resultados.

BIBLIOGRAFÍA

BIBLIOGRAFÍA REFERENCIADA

LIBROS

- Campell, Mary. base IV Guía de Autoenseñanza. España. Editorial McGraw Hill – Interamericana. 1990. pp110/111,121/122,16,169, 179-191/192.

- Cohen Karen Daniel. (1996). Sistemas de información para la toma de decisiones. México. McGraw-Hill. 243p.

- Larose, Daniel T. (2005). Discovering Knowledge in Data an Introduction to Data Mining. Hoboken, New Jersey. Jhon Wiley & Sons, Inc Publication. 222p.

- Berry,Michael J.A y Gordon S. Linoff.(2004). Data Mining techniques for Marketing, Sales, and Customer Relationship Management. Indianapolis. Wiley Publishing, Inc. 637p.

- Giudici, Paolo. (2003). Applied Data Mining Stadistical Methods for Bussines and Industry. Chichester. Jhon Wiley & Sons, Inc. 364p.

- Ponniah, Pauraj. (2001). Data Warehousing Fundamentals a Comprehensive Guide for IT Professionals. New York. Jhon Wiley & Sons, Inc Publication. 516p.

PÁGINAS WEB

- Agrawal, Rakesh et al. (1993). Mining Association Rules between Sets of Items in Large Database. http://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf. (26 Oct.2009).

- Aranguren, Silvia Mónica y Muzachiodi, Silvia Liliana. (2003). Implicancias del Data Mining. http://www.fceco.uner.edu.ar/extinv/publicdocent/sarangur/pdf/introduccion.pdf (4 Mar.2009).

- "A Brief History of Data Mining. Data mining software". (2006). http://www.data-mining-software.com/data_mining_history.htm (5 Mar. 2009).

- Bressán Griselda E. (2003). Lic. en sistemas de información. Almacenes de datos y *Minería de Datos*. Trabajo monográfico de adscripción. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosBressan.htm. (12 Mar. 2009).

- Christen, Peter. (2005) A very short introduction to… Data Mining. http://datamining.anu.edu.au/talks/2005/datamining-comp2340-2005.pdf (9 Mar. 2009).

- Chávez García, Carlos ¿Qué son las bases de datos? (2007). http://www.maestrosdelweb.com/principiantes/%C2%BFque-son-las-bases-de-datos/.(4 Mar 2009).

- Duque Méndez, Néstor Darío. Bases de Datos. Universidad Nacional de Colombia.(2005). http://www.virtual.unal.edu.co/cursos/sedes/manizales/4060029/lecciones/cap8-1.html. (9 Mar 2009).

- Larrieta, María Isabel Ángeles y Santillán Angélica María. (2004). *Minería de Datos*: Concepto, características, estructura y aplicaciones. http://www.ejournal.unam.mx/rca/190/RCA19007.pdf (10 Mar.2009).

- Marcano, Yelitza. Talavera, Rosalba. (2007). *Minería de Datos* como soporte a la toma de decisiones empresariales Universidad del Zulia http://www.serbi.luz.edu.ve/scielo.php?pid=S1012-15872007004000008&script=sci_arttext (29 Ag. 2009).

- Mierswa, I. and Wurst, M. and Klinkenberg, R. and Scholz, M. and Euler, T. (2006) Yale (now: *RapidMiner*): Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006). http://ufpr.dl.sourceforge.net/project/yale/1.%20RapidMiner/4.4/rapidminer-4.4-tutorial.pdf (1 Dic. 2009)

- Molina, Luis Carlos. (2002). Data Mining: Torturando los datos hasta que confiesen. http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html. (4 Mar. 2009).

- Molina Félix, Luis Carlos y Béjar, Javier. (1999). INTEGRACIÓN DE REGLAS DE ASOCIACIÓN Y DE CLASIFICACIÓN. http://www.lsi.upc.es/~lcmolina/SC/html/paper/cars-lmf.pdf (9 Ag. 2009).

- Moreno. García, María et al. (2009). Aplicación de técnicas de *Minería de Datos* en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf (29 Ag. 2009).

- Naranjo, Roberto. Sierra, Luz Marina. (2009). Herramienta de software para el análisis de canasta de mercado sin selección de candidatos. http://www.scielo.org.co/pdf/iei/v29n1/v29n1a08.pdf. (29 de Ag. 2009).

- Reyes, Jose Fernando y García, Rodolfo. (2005) EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS. http://ingenierias.uanl.mx/26/pdfs/26_**el proceso**.pdf. (30 Ag. 2009).

- Vallejos, Sofía. (2006). *Minería de Datos*. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf (7 Mar.2009).

- Velazco, Roberto Hernando. Almacenes de datos (Datawarehouse). (2007). http://www.rhernando.net/modules/tutorials/doc/bd/dw.html. (7 Mar 2009).

- Zamarron. Sanz, Carlos et al. (2008). Aplicación de la *Minería de Datos* al estudio de las alteraciones respiratorias durante el sueño. http://www.sogapar.org/pneuma/pneuma6/pneuma-n-6-5c.pdf (29 Ag. 2009).

BIBLIOGRAFÍA CONSULTADA

LIBROS

- Berry, Michael y Linoff, Gordon. (2004). Data Mining Techniques for Marketing, Sales and Customer Relationship Management. Indianapolis. Wiley Publishing Inc. 643p.

- Kimball, Ralph y Ross Margy. (2002). The Data Warehouse Toolkit Olap Solutions. Indianapolis. Wiley Publising, Inc. 421p.

- Non, Ye. (2003).The Handbook of Data Mining. New York. Laurence Erlbaum Associates Publishers. 689p.

- Tang, Zhao Hui y Mac Lennon, Jamie. (2005). Data Mining With SQL Server 2005. Indianapolis. Wiley Publishing, Inc. 460p.

- Voges, Kevin y Pope Nigel. (2006). Bussines Aplication and Computational Intelligence. London. Idea Group Publishing.481p.

- Witten, Ian y Frank, Eibe. (2005). Data Mining Practical Machine Learning Tools and Techniques. San Francisco. ELSEVIER. 525p.

ANEXOS.

Anexo 1.

Rapid-I Forum

RapidMiner => Getting Started => Topic started by: jangabe on August 31, 2009, 07:01:56 PM

Title: **How can I start???**
Post by: **jangabe** on **August 31, 2009, 07:01:56 PM**

Hello, first than everything I don't speak english so good, so please sorry if I make mistakes. Hope you could understand me.
My problem is that I want to apply association rules to a DB of a supermarket, but this data have to be cleaned (data preprocessing) and I don't know how to make this.
So, for an idea the DB has the follow items:

numero(Bill_Number) fecha(date) codigo(ProductCode) costo(Cost) precio(Price) tarifa_iva(Value-added Tax Rate) ->

precio_sin_iva(Price_whitout_Value-added tax) valor_iva(Value_Value-added tax) cantidad(Amount) valor_item(Value_item) ->

bodega(storeRoom) nombre(ProductName) nclasifica1(ClassificationNane) clasifica1(ClassificationCode) ->

nclasifica2(ClassificationName2) clasifica2(ClassificationCode2)

The data for this are:

216932 31/08/2008 4756 2531.66 3200.00 16.00 2758.62 441.38 10.000 3200.00 1 MARGARINA PRACTIS X 400 GR LLOREDA DISTRIBUCIONES 70 MARGARINAS 90

Please if you can help me about what must I to apply for clean the data or has it of the correct form by apply fp growt then association rules.
Thank you by the attention.

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **September 01, 2009, 09:14:29 AM**

Hi,
FP-Growth needs a transaction format. For each possible item it has to specify if its contained in one transaction or not (true/false). Hence you have to make your dataset consisting of binominal attributes.
You could use the Nominal2Binominal Operator, but as his name states it only transforms nominal attributes. Numerical ones must be converted to nominal beforehand using for example a discretization operator.

Greetings,
Sebastian

Title: **Re: How can I start???**
Post by: **jangabe** on **September 25, 2009, 10:44:09 PM**

Hi Sebastian,
I have been trying to do a matrix with the data that I have, only using the bill_number and the product_code, having the bill_number like column and the code_product like row, but I would like to know which is the maximum of data that RM support, because I have a matrix of 56126(columns) X 1481(rows) and I want to know if it's possible working whit this or I have to reduce the matrix, or I have to change some parameter on the RM configuration for accept this data, I say this because after I run the tool I had the follow log file:

P Sep 25, 2009 4:23:24 PM: Initialising process setup
P Sep 25, 2009 4:23:24 PM: [NOTE] No filename given for result file, using stdout for logging results!
P Sep 25, 2009 4:23:24 PM: Checking properties...
P Sep 25, 2009 4:23:24 PM: Properties are ok.
P Sep 25, 2009 4:23:24 PM: Checking process setup...
P Sep 25, 2009 4:23:24 PM: Inner operators are ok.
P Sep 25, 2009 4:23:24 PM: Checking i/o classes...
P Sep 25, 2009 4:23:24 PM: i/o classes are ok. Process output: AssociationRules.
P Sep 25, 2009 4:23:24 PM: Process ok.
P Sep 25, 2009 4:23:24 PM: Process initialised
P Sep 25, 2009 4:23:24 PM: [NOTE] Process starts
P Sep 25, 2009 4:23:24 PM: Process:
  Root[1] (Process)
  +- CSVExampleSource[1] (CSVExampleSource)
  +- Numerical2Binominal[1] (Numerical2Binominal)
  +- FPGrowth[1] (FPGrowth)
  +- AssociationRuleGenerator[1] (AssociationRuleGenerator)
P Sep 25, 2009 4:24:47 PM: [NOTE] Numerical2Binominal: Breakpoint reached
G Sep 25, 2009 4:25:04 PM: [Warning] Cannot plot all data points, using only a sample of 5000 rows. You can increase the number of values in the properties dialog from the tools menu, the property name is 'rapidminer.gui.plotter.rows.maximum'
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'Scatter Matrix': Data table must have between 0 and 50 columns, was 1482.
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'Survey': Data table must have between 0 and 100 columns, was 1482.
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'Andrews Curves': Data table must have between 0 and 1000 columns, was 1482.
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'Quartile Color Matrix': Data table must have between 0 and 100 columns, was 1482.
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'RadViz': Data table must have between 0 and 1000 columns, was 1482.
G Sep 25, 2009 4:25:04 PM: [NOTE] Cannot use plotter 'Surface 3D': Data table must have between 0 and 50 rows, was 5000.
P Sep 25, 2009 4:37:48 PM: [NOTE] FPGrowth: Breakpoint reached
G Sep 25, 2009 4:38:08 PM: [Warning] Cannot plot all data points, using only a sample of 5000 rows. You can increase the number of values in the properties dialog from the tools menu, the property name is 'rapidminer.gui.plotter.rows.maximum'
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'Scatter Matrix': Data table must have between 0 and 50 columns, was 1482.
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'Survey': Data table must have between 0 and 100 columns, was 1482.
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'Andrews Curves': Data table must have between 0 and 1000 columns, was 1482.
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'Quartile Color Matrix': Data table must have between 0 and 100 columns, was 1482.
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'RadViz': Data table must have between 0 and 1000 columns, was 1482.
G Sep 25, 2009 4:38:08 PM: [NOTE] Cannot use plotter 'Surface 3D': Data table must have between 0 and 50 rows, was 5000.
P Sep 25, 2009 4:40:48 PM: [NOTE] AssociationRuleGenerator: Breakpoint reached
P Sep 25, 2009 4:42:10 PM: Process:
  Root[1] (Process)
  +- CSVExampleSource[1] (CSVExampleSource)
  +- Numerical2Binominal[1] (Numerical2Binominal)
  +- FPGrowth[1] (FPGrowth)
  +- AssociationRuleGenerator[1] (AssociationRuleGenerator)
P Sep 25, 2009 4:42:10 PM: Produced output:
IOContainer (1 objects):
Association Rules
[4995] --> [5430] (confidence: 1.000)
[4550] --> [5430] (confidence: 1.000)
[6049] --> [5430] (confidence: 1.000)
[9523] --> [5430] (confidence: 1.000)
(created by AssociationRuleGenerator)
P Sep 25, 2009 4:42:10 PM: [NOTE] Process finished successfully after 18:45

Or if you can help me about what is happening because the association rules obtained are practically no one, because the 3 obtained no one is good for me.
Thank you by the attention.

Title: **Re: How can I start???**
Post by: **steffen** on **September 26, 2009, 10:30:27 AM**

Hello jangabe

Dont worry, RapidMiner can handle this amount of data. The log-file says that some of the plotters are configured to handle only a certain amount of points. You can change that behaviour in the settings (located in the menu bar). Note that these default settings have been made to decrease computation time of the plotters.

@association-rules: You have to convert the data into transaction format as Sebastian said. Then the association rules will be more useful. If you do not understand what he means, go and grab yourself a data mining book where the procedure is explained. This is recommended anyway so you can judge the quality and the behavior of the outcome.

regards,

Steffen

Title: **Re: How can I start???**
Post by: **jangabe** on **September 29, 2009, 12:15:33 AM**

Good afternoon,
Thanks for answer my doubt, and with relation to the transaction format i made a little program that filled the matrix of that form 0s and 1s, but i made what you do with the 'rapidminer.gui.plotter.rows.maximum' and i dont know why this dont gives a good results if the amount of data is big.
Can you tell me if am i doing something wrong, maybe because i dont use name else code that represent that name.
If you want write me and i gives you a file sample...

Thank you by the attention.

Title: **Re: How can I start???**
Post by: **jangabe** on **September 29, 2009, 12:40:49 AM**

Error in: CSVExampleSource (CSVExampleSource) Could not read file 'C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes 65000 NombreProd.csv': Number of columns in line 1 was unexpected, was: 1482, expected: 244. The given file could not be read. Please make sure that the file exists and that the RapidMiner process has sufficient privileges.

In another attempt i obtained that message mistake, where can I configure it for 1482 columns??

Title: **Re: How can I start???**
Post by: **steffen** on **September 29, 2009, 05:02:59 PM**

Hello again

Quote
Thanks for answer my doubt, and with relation to the transaction format i made a little program that filled the matrix of that form 0s and 1s, but i made what you do with the 'rapidminer.gui.plotter.rows.maximum' and i dont know why this dont gives a good results if the amount of data is big.
Can you tell me if am i doing something wrong, maybe because i dont use name else code that represent that name.
If you want write me and i gives you a file sample...

Okay, this is what I understood: You think that the parameter "rapidminer.gui.plotter.rows.maximum" is limiting the number of rows used in learning (association rules) , but that is wrong. This parameter only affects the number of rows used for plotting.

Regarding:
Quote
Error in: CSVExampleSource (CSVExampleSource) Could not read file 'C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes 65000 NombreProd.csv': Number of columns in line 1 was unexpected, was: 1482, expected: 244. The given file could not be read. Please make sure that the file exists and that the RapidMiner process has sufficient privileges.

This means that your csv - files is somehow messed up. There are many reasons for this, but my first idea is that the operator infers the number of columns from the first line where are not all column names are specified. If could post the process (operator tab, copy and paste the xml -code to this forum), this would be helpful.

regards,

Steffen

Title: **Re: How can I start???**
Post by: **jangabe** on **September 30, 2009, 12:15:20 AM**

Hello and thanks again.
About the parameter I made a new change in the "rapidminer.gui.attributeeditor.rowlimit" but the result is the same.
About the another theme, I dont understand what do you want to say with 'file is somehow messed up'?. Like i told you, in the first row is the product_code and and no name, but the mistake was because i thought that changing the product_code by the real product_name it was to give the hoped results; at once, the Xml tha is generated is the next:

```
<operator name="Root" class="Process" expanded="yes">
   <operator name="CSVExampleSource" class="CSVExampleSource">
      <parameter key="filename"   value="C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes 65000 Lleno.csv"/>
      <parameter key="label_name"   value="C_Factura"/>
   </operator>
   <operator name="Numerical2Binominal" class="Numerical2Binominal">
   </operator>
   <operator name="FPGrowth" class="FPGrowth">
      <parameter key="min_support"   value="0.4"/>
   </operator>
   <operator name="AssociationRuleGenerator" class="AssociationRuleGenerator">
      <parameter key="min_confidence"   value="0.4"/>
   </operator>
</operator>
```

The min_support and min_confidence has been tested with another values and the results are the same:

[4995] --> [5430] (confidence: 1.000)
[4550] --> [5430] (confidence: 1.000)
[6049] --> [5430] (confidence: 1.000)
[9523] --> [5430] (confidence: 1.000)

And realy I am surprised because the amount of data is big for only obtain that rules, or is it possible?
How can I to know that maybe that result is the only one that I'm goign to obtain???
Thank you by the attention.

Title: **Re: How can I start???**
Post by: **steffen** on **September 30, 2009, 02:13:06 PM**

Hello jangabe

First of all: I have never flamed anyone for his / her skills in english (I am not a native speaker, too) , but frankly, your sentences are giving me a headache. Please try to form shorter sentences.

@your problem: I cannot figure out from your last posts what your data looks like **NOW**. Please post the first rows of the csv-file your are loading ("C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes 65000

Lleno.csv") (including first line).

then we will see ... I have a vague idea what could be wrong ...

regards,

Steffen

Title: **Re: How can I start???**
Post by: **jangabe** on **October 02, 2009, 05:22:27 AM**

Hello, and sorry for my bad english...

Well, the first csv-file row is the next:

C_Factura 4 22 23 36 37 39 41 46 54 61 64.....

140992 0 0 0 0 0 0 0 0 0 0 0.....
141191 0 0 0 0 0 0 0 0 0 0 0.....
141278 0 0 0 0 0 0 0 0 0 0 0.....
. .
. .
. .
. .
. .

I didn't put more because are 1481 product_code and 56126 bill_code, but basically that's the form; the product_codes are orderly A-Z as the same way the bill_codes.

Thank you by the attention.

Title: **Re: How can I start???**
Post by: **steffen** on **October 02, 2009, 08:22:56 AM**

No problem

Now we are getting somewhere ...

I copied the data into a text-file and played a little bit. The CSVExampleSource-Operator (as you have posted it) causes a single-column-attribute. I suggest to try this setup:

Code:
```
<operator name="Root" class="Process" expanded="yes">
   <operator name="SimpleExampleSource" class="SimpleExampleSource">
      <parameter key="filename"            value="/home/steffen/Desktop/check.txt"/>
      <parameter key="read_attribute_names"       value="true"/>
      <parameter key="id_column"         value="1"/>
   </operator>
   <operator name="Numerical2Binominal" class="Numerical2Binominal">
   </operator>
   <operator name="FPGrowth" class="FPGrowth">
      <parameter key="min_support"      value="0.4"/>
   </operator>
   <operator name="AssociationRuleGenerator" class="AssociationRuleGenerator">
      <parameter key="min_confidence" value="0.4"/>
   </operator>
</operator
```

Another tip: If you doubleclick on an operator (or rightlick -> Breakpoint after) you can set (guess) a breakpoint. This allows you to see the result of the selected operator. If you now click "resume" (the arrow) at the top task bar, the process continues. This is extremely helpful when it comes to process debugging.

oh and I think, that "thank you **for** the / your attention" is correct ;)

kind regards,

Steffen

Title: **Re: How can I start???**
Post by: **jangabe** on **October 02, 2009, 11:13:33 PM**

Hello,
I made exactly what you wrote and I obtained the same result...
The Xml code was:

```
<operator name="Root" class="Process" expanded="yes">
    <operator name="SimpleExampleSource" class="SimpleExampleSource">
        <parameter key="filename"    value="C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes
65000 Lleno.csv"/>
        <parameter key="read_attribute_names"    value="true"/>
        <parameter key="id_column"    value="1"/>
    </operator>
    <operator name="Numerical2Binominal" class="Numerical2Binominal" breakpoints="before">
    </operator>
    <operator name="FPGrowth" class="FPGrowth" breakpoints="before">
        <parameter key="min_support"    value="0.4"/>
    </operator>
    <operator name="AssociationRuleGenerator" class="AssociationRuleGenerator" breakpoints="before">
        <parameter key="min_confidence"    value="0.4"/>
    </operator>
</operator>
```

So, will be that those are the only association rules on that DB??
I want to tell you that if you can give me your e-mail for To send you the complete file and you cuoul to make tests??

Thank you for your attention.

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **October 05, 2009, 08:38:55 AM**

Hi,
although I didn't read the hole thread here, just a little remark: Did you lower the min_support parameter of fp-growth and the min_confidence parameter of the rules generator? Might be, there are no more rules until the support and confidence gets really low...

Greetings,
Sebastian

Title: **Re: How can I start???**
Post by: **jangabe** on **October 05, 2009, 06:27:34 PM**

Hi,
This is the Xml-code on the last try:

```
<operator name="Root" class="Process" expanded="yes">
    <operator name="CSVExampleSource" class="CSVExampleSource">
        <parameter key="filename"    value="C:\Users\JANGABE\Desktop\DATOS PRUEBA CLASIFICADOS 1 mes
65000 Lleno.csv"/>
        <parameter key="id_column"    value="1"/>
    </operator>
    <operator name="Numerical2Binominal" class="Numerical2Binominal">
    </operator>
    <operator name="FPGrowth" class="FPGrowth">
        <parameter key="min_support"    value="0.1"/>
    </operator>
```

```
   <operator name="AssociationRuleGenerator" class="AssociationRuleGenerator">
      <parameter key="min_confidence"   value="0.1"/>
   </operator>
</operator>
```

Like you see the min_support and the min_confidence are in their lower value; and the results are the same:

Association Rules
[4995] --> [5430] (confidence: 1.000)
[4550] --> [5430] (confidence: 1.000)
[6049] --> [5430] (confidence: 1.000)
[9523] --> [5430] (confidence: 1.000)

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **October 06, 2009, 08:37:06 AM**

Yes,
I see. Obviously there isn't any lower value between 0 and 1 than 0.1...

Title: **Re: How can I start???**
Post by: **jangabe** on **October 08, 2009, 12:28:28 AM**

Hi,
I have downloaded the latest RM version.
Curiosly, I did the test with this version applying the same parameters and I didn't obtain any association rule.
That's strange 'cause before the result was:

Association Rules
[4995] --> [5430] (confidence: 1.000)
[4550] --> [5430] (confidence: 1.000)
[6049] --> [5430] (confidence: 1.000)
[9523] --> [5430] (confidence: 1.000)

So, what happend in this case??

Thank you for your attention.

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **October 08, 2009, 08:44:57 AM**

Hi,
sorry, but I'm not a magician. I cannot even guess whats causing your problems, because I don't have your data, I do not even have your process. You don't even said, which is your rapid miner version, and since there are two most recent versions, 4.6 and 5.0beta, I can only guess. And last but not least, if you don't follow my advice to lower your support and confidence threshold (BELOW 0.1! For example just take 0.0000001 for testing if it works anyway.) I cannot help you at all.

Greetings,
  Sebastian

Title: **Re: How can I start???**
Post by: **jangabe** on **October 08, 2009, 05:44:29 PM**

Hi,
From this link you can download the file tha contain my data:
http://www.4shared.com/file/137659584/22782223/DATOS_PRUEBA_CLASIFICADOS_1_mes_65000_Lleno.html

The version of RM is the 4.6, and I tried with confidence and support in 0.0000001 and I obtained 32 association rules

Something strange happened me, I always try with a new process, and don't obtain any result, just now i did it again but this time with a saved process and amazingly the result were immediately(always take around 18- 20

minutes) and positive(between 32 an 20 varying the confidence and the support), but made a new try again and the result was NO RULES FOUND. What would happen??

Thanks you for your attention.

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **October 09, 2009, 08:11:27 AM**

Ok, thank you very much for this. No I have the hope to be able to reproduce your problems. I will check this as soon as I can, but I doubt I will find the time before next week.

Greetings,
Sebastian

Title: **Re: How can I start???**
Post by: **jangabe** on **October 09, 2009, 04:01:05 PM**

Hi,
Thanks for your help and I want ask you something:
Why do i have to reduce the min_confidence and the min_support for obtain so many rules??
Is there some relation between the min_support, the min_confidence and the amount of data??

Thanks you very much...:)
Best Regards.

Title: **Re: How can I start???**
Post by: **Sebastian Land** on **October 12, 2009, 08:59:28 AM**

Hi,
the both thresholds specify in how many examples this item set or this rule have to occur before it is called frequent. So if you have 1000 examples and a support of 0.1 then the item set must be contained in 100 examples, otherwise the set is discarded.
The level of support needed for gaining some rules depends on your data. You will always find rules, but with lesser support, these rules are more worthless, because they are less general, describing only a small number (or only one) of transactions.

Greetings,
Sebastian
