**PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO**
**FACULTAD DE INGENIERÍA**
**ESCUELA DE INGENIERÍA INFORMÁTICA**

# APLICACIÓN DE TÉCNICA DE MINERÍA DE DATOS PARA UN SUPERMERCADO

**JUAN ANDRÉS JEREZ BALMACEDA**

INFORME FINAL DEL PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

VALPARAÍSO, ENERO 2015

**PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO**
**FACULTAD DE INGENIERÍA**
**ESCUELA DE INGENIERÍA INFORMÁTICA**

# APLICACIÓN DE TÉCNICA DE MINERÍA DE DATOS PARA UN SUPERMERCADO

**JUAN ANDRÉS JEREZ BALMACEDA**

PROFESOR GUIA: PAMELA HERMOSILLA
PROFESOR CO-REFERENTE: RODOLFO VILLARROEL

VALPARAÍSO, ENERO 2015

# Agradecimientos.

Soy un hombre tremendamente agradecido de la vida, en especial por lo generoso que ha sido el universo conmigo. Mi familia ha sido clave para tomar este rumbo universitario, gracias a su apoyo he podido realizar mi sueño de ser ingeniero civil en informática.

Quiero comenzar agradeciéndoles a mis dos padres (Mamá y Papá) por confiar siempre en mí y darme el mayor de los regalos, el amor incondicional, sin ellos no sería como soy. De ellos aprendí los valores de la familia, amistad y por sobre todo el ser perseverante.

Quiero agradecer a mi hermana y cuñado, con los cuales pasamos muchos momentos valiosos juntos y con los cuales compartí más de 9 años. Con ellos aprendí lo que significa la libertad y el respeto por el otro.

También quiero agradecer a mi polola por su apoyo, las inmensas alegrías que hemos vivido juntos y su motivación para terminar este proceso universitario.

A mis amigos les quiero agradecer por haberme dados tantos momentos alegres, distendidos y de aprendizaje. Cada uno aportó en lo que es este proceso completo. Finalmente agradecer a mi profesora de proyecto título que muchas veces me orientó y me corrigió con mucha paciencia y exigencia.

# Dedicatorias.

Quiero dedicarle este éxito, a las personas que más quiero y más respeto, que son a mis dos padres. A mi padre porque desde chico que me motivó a seguir por el camino de la ingeniería, por espíritu de superación y por ese apoyo incondicional.

A mi madre por darme ese amor incondicional cada vez que lo requerí, por su espíritu de orden, y por esa inteligencia emocional que me ayudó en distintas situaciones. A los dos les dedico esto así como Uds. me dieron la vida.

Finalmente quiero dedicarle este triunfo a mi querida y amada compañera de vida, Francisca, ella fue una de las que me motivó a finalizar esta etapa.

# Índice

# Resumen

La disminución en el precio de las TIC's (Tecnologías de la información) tanto en el software y hardware, han tenido como resultado un aumento en la información en el ámbito empresarial. Esta posee una gran relevancia a la hora de tomar de decisiones. Paralelamente existen datos que poseen un potencial no descubierto que pueden seguir generando nuevo conocimiento.

Existen variadas técnicas para generar conocimiento, sin embargo la minería de datos ha sido pionera, ya que esta propone un conjunto de técnicas que generan patrones previamente desconocidos y que generan conocimiento útil para distintos ámbitos. En específico para este trabajo se trabaja con las reglas de asociación que permiten formar relaciones de tipo implicancia.



Palabras Clave: Minería de datos, Reglas de asociación, Metodologías de minería de datos, base de datos, Aplicación de técnicas de minería de datos.



# Abstract

The decrease in the price of ICT (Information Technology ) both software and hardware, have resulted in an increase in business information . This has great relevance when taking decisions. Parallel data exist that possess an undiscovered potential that can continue to generate new knowledge.

There are various techniques to generate knowledge, though data mining has been a pioneer, because this proposes a set of techniques that generate previously unknown patterns and generate useful knowledge for different fields. Specifically for this job working with association rules that allow such form relationships of implication.



Keywords: Data Mining, Association Rules, Data Mining Methodologies, Database, Application of data mining techniques.

# Lista de Figuras

# Lista de Tablas

# 1 Introducción

Dentro de los últimos años, se ha desarrollado una tendencia de crecimiento exponencial al momento de recolectar (datos) y generar información, causado principalmente por la disminución del precio en los dispositivos de almacenamiento y el aumento en el poder de procesamiento de las máquinas. Sin embargo, una de los desafíos actuales es generar información y posterior conocimiento desde un gran almacén de datos. Además dentro de estas enormes masas de datos, existe una gran cantidad de información "oculta", de gran importancia estratégica, la cual no siempre se puede acceder por las técnicas clásicas de recuperación de la información.

Una de las formas de acceder a esta información "oculta" es mediante a la Minería de Datos (Data Mining), que se ocupa en encontrar patrones y relaciones dentro de los datos, permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados. Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean.

Es por esto que una de las estrategias ampliamente utilizadas por los negocios exitosos son los métodos de análisis avanzados. Su finalidad en el negocio es incrementar las ganancias, maximizar la eficiencia operativa, reducir los costos y mejoran la satisfacción del cliente.

El presente informe entrega información sobre el trabajo de investigación realizado para la obtención del título profesional. El objeto de estudio fue el análisis de la información recopilada durante un periodo de tiempo. Esta fue analizada mediante un conjunto de técnicas de minería de datos, que capaces de extraer conocimiento útil, comprensible y previamente desconocida.

En este caso particular esta fue extraído desde la fuente de datos de un Retail. La finalidad de cumplir los objetivos planteados, donde lo recabado fue herramienta útil para el mejoramiento de dicha empresa.

# 2 Descripción del Problema

Una de las problemáticas de la empresa es el crecimiento exponencial de la data en volumen y en variedad en los últimos años. Gran parte de esta, es histórica, es decir, representa transacciones o situaciones que se han producido. Esta enorme cantidad de datos generan además ciertos problemas asociados tales como:

- ✓ Sobrecarga de información
- ✓ Exceso de información genérica
- ✓ Ausencia de información personalizada y/o Relevante para los distintos perfiles que existen en el negocio.
- ✓ Falta de retroalimentación oportuna para la mejora del negocio

Se trabajó con una empresa del sector Retail, específicamente un supermercado de la región. Esta maneja cantidades grandes de información sobre sus productos. Por petición del administrador del local se protegió la identidad de dicha organización.

Ellos poseen algunos sistemas de información capaces de generar información, sin embargo el sistema tiene ciertas limitaciones, como por ejemplo este, no es capaz de encontrar relaciones entre productos.

Una de las oportunidades propuestas fue trabajar en esta arista, es decir buscar relaciones del tipo de implicancia entre los distintos tipos de producto con el fin de incrementar las ventas. Existen variados ejemplos de estos vínculos, que son más bien evidentes, tal como lo sin el del pan y la mantequilla. Sin embargo, gracias a estudios posteriores se logró encontrar relaciones entre productos que aparentemente no cumplían con reglas claras. Es el conocido caso de los pañales y la cerveza, en donde el público que los adquiría tenia ciertas características y cumplía ciertos patrones de comportamiento. La finalidad del descubrimiento de este vínculo es, principalmente apoyar la toma de decisiones de los ejecutivos en relación a la organización completa o parcial del supermercado. Esta distribución ayudara promover venta, con simples estrategias tales como, ubicar productos relacionados en un mismo pasillo, crear descuentos, promociones o packs. Todo con el fin de aumentar la probabilidad que los clientes compren dichas configuraciones de productos.

Para el trabajo de investigación se propuso tomar toda la data de un periodo de tiempo, la cual será contenida en los almacenes de datos y procesada para cumplir el objetivo planteado.

# 3 Objetivos

## 3.1 Objetivo General el proyecto

Aplicación de técnicas de minería de datos para el área de marketing de un Supermercado con el fin de apoyar la toma de decisiones.

## 3.2 Objetivos Específicos

1. Determinar metodología de minería de datos a utilizar
2. Designar la metodología de Data Warehouse para el almacén de datos.
3. Seleccionar la herramienta informática a utilizar para la solución propuesta.
4. Identificar relaciones existentes entre productos utilizando la técnica de minería de datos escogida.
5. Analizar los resultados obtenidos.

# 4 Marco Teórico

A continuación se describirán los conceptos necesarios para resolver la problemática existente. Para ello se realizó una etapa de investigación que permitió definir las bases teóricas y de acuerdo ello, se eligió la mejor alternativa. Esta indagación cumple el propósito de analizar el estado del arte completo hasta el momento. Los conceptos serán presentados en desde lo más general a lo particular en orden y relevancia.

## 4.1 Inteligencia de Negocios

Para entender la Extracción de información, es primordial, entender el concepto de BI. Este se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos.

La Inteligencia del Negocio (BI) puede representar herramientas y sistemas que juegan un papel clave en el proceso estratégico de la planificación de una compañía. Estos sistemas permiten reunir, almacenar, y analizarlos datos corporativos siendo una importante ayuda en la toma de decisiones.

En un amplio sentido la inteligencia de negocios es el puente entre los sistemas de información y los procesos relevantes para el negocio, en los cuales se manejan volúmenes de datos importantes y poseen un potencial para ser trabajados.

Se pueden definir diferentes componentes presentes en la BI, los cuales se explicaran a continuación:

- ✓ **Multidimensionalidad:** Poseer varias dimensiones en vez de un valor, dependiendo de los ejes definidos. Una herramienta de BI debe de ser capaz de reunir información dispersa en toda la empresa e incluso en diferentes fuentes para así proporcionar a los departamentos la accesibilidad, poder y flexibilidad que necesitan para analizar la información
- ✓ **Agentes:** Los agentes son programas que piensan. Ellos pueden realizar tareas a un nivel muy básico sin necesidad de intervención humana.
- ✓ **Data Warehouse:** Es una colección de datos que contiene datos necesarios o útiles para una organización. Coloca información de todas las áreas funcionales de la organización en manos de quien toma las decisiones. También proporciona herramientas para búsqueda y análisis.
- ✓ **Data Mining:** Pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no muy evidentes.

Se profundizara en los dos últimos componentes (Data Warehouse y Datamining) ya que son relativos a la investigación y solución. A continuación se profundizara en el concepto de Data Warehouse.

## 4.2 Data Warehouse

Un almacén de datos (Data Warehouse) es una colección de datos que contiene datos útiles para una organización. Este debe entregar la información correcta, a la gente indicada, en el momento óptimo y en el formato adecuado. Los datos almacenados poseen ciertas características nombradas a continuación:

- ✓ Están orientado a un determinado ámbito (empresa, organización, etc.)
- ✓ Integrados
- ✓ No volátiles
- ✓ Variable en el tiempo

Esto ayuda a la toma de decisiones en la entidad en la que se utiliza. Es una estructura de datos donde la información contenida está diseñada para favorecer el análisis y la divulgación eficiente de datos.

Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sea necesario. Dichas unidades se denominan Data Marts.

Existen varios caminos o metodologías para construir un DW, estas serán explicadas con mayor detalle.

### 4.2.1 Metodologías para la creación de Data Warehouse

Para aplicar inteligencia de negocios es importante la elección de una buena metodología de construcción del almacén de datos. Existen variadas, sin embargo abordaron las más conocidas, que son el Ralph Kimball y el Bill Inmon. Para entender las diferencias entre ambos enfoques, es necesario en tener claro la diferencia entre Data Warehouse y Data Mart.

Un Data Warehouse proporciona una visión global, común e integrada de los datos de la organización, El Data Mart a diferencia es un subconjunto de los datos del Data Warehouse con que tiene como objetivo responder a un determinado análisis, función o necesidad y con una población de usuarios específica.

Teniendo en cuenta esto, se realizó un resumen de los aspectos más importantes de cada una de las metodologías, las cuales serán vistas a continuación:

#### 4.2.1.1 Metodología de Bill Inmon.

Bill Inmon ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis (sería el CIF o Corporate Information Factory). Insiste además en que ha de tener las siguientes características:

- **Orientado a temas**.- Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- **Integrado**.- La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- **No volátil**.- La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- **Variante en el tiempo**.- Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

La información ha de estar a los máximos niveles de detalle. Los DW departamentales o Data Marts son tratados como subconjuntos de este DW corporativo, que son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir de este DW Central.

El enfoque Inmon también se referencia normalmente como Top-down. Los datos son extraídos de los sistemas operacionales por los procesos ETL y cargados en las áreas de stage, donde son validados y consolidados en el DW corporativo, donde además existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DW. Una vez realizado este proceso, los procesos de refresco de los Data Mart departamentales obtienen la información de él, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

La metodología para la construcción de un sistema de este tipo es la habitual para construir un sistema de información, utilizando las herramientas habituales (esquema Entidad Relación, DIS (Data Item Sets, etc). Para el tratamiento de los cambios en los datos, usa la Continue and Discrete Dimension Management (inserta fechas en los datos para determinar su validez para las Continue Dimension o bien mediante el concepto de snapshot o foto para las Discrete Dimension).

Al tener este enfoque global, es más difícil de desarrollar en un proyecto sencillo (pues estamos intentando abordar el “todo”, a partir del cual luego iremos al “detalle”).

### 4.2.2 Metodología de Ralph Kimball.

El Data Warehouse es un conglomerado de todos los Data Marts dentro de una empresa, siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional (no normalizado), que incluye, como ya vimos, las dimensiones de análisis y sus atributos, su organización jerárquica, así como los diferentes hechos de negocio que se quieren analizar**.** Por un lado tenemos tablas para las representar las dimensiones y por otro lado tablas para los hechos (las facts tables). Los diferentes Data Marts están conectados entre sí por la llamada **bus structure**, que contiene los elementos anteriormente citados a través de las dimensiones conformadas (que permiten que los usuarios puedan realizar querys conjuntos sobre los diferentes Data Marts, pues este bus contiene los elementos en común que los comunican).

Este enfoque también se referencia como **Bottom-up**, pues al final el Data Warehouse Corporativo no es más que la unión de los diferentes Data Marts, que están estructurados de una forma común a través de la bus structure. Esta característica le hace más flexible y sencillo de implementar, pues podemos construir un Data Marts como primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL extraen la información de los sistemas operacionales y los procesan igualmente en el área stage, realizando posteriormente el llenado de cada uno de los Data Marts de una forma individual, aunque siempre respetando la estandarización de las dimensiones (dimensiones conformadas).

La metodología para la construcción del Dw incluye las 4 fases que son:

- ✓ Selección del proceso de negocio
- ✓ Definición de la granularidad de la información
- ✓ Elección de las dimensiones de análisis
- ✓ Identificación de los hechos o métricas**.**

## 4.3 Minería de datos

Es un paso dentro de todo el proceso de KDD. Su objetivo es descubrir tendencias, situaciones anómalas y/o interesantes, patrones y secuencias en los datos. Esta intenta obtener patrones o modelos a partir de los datos recopilados. Aunque desde un punto de vista académico el término data mining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos, en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta.

Generalmente la aplicación de técnicas de data Mining en grandes bases de datos persiguen los siguientes resultados:

1. Clasificación
2. Regresión
3. Resumen
4. Análisis de Secuencias
5. Descubrimiento de reglas de asociación

Estos resultados condicionaran la elección posterior de la técnica de minería de datos que será explicada más adelante. A continuación se explicara el proceso típico de minera de datos:

### 4.3.1 Pasos típicos del proceso de Data Mining (DM)

Un proceso típico de minería de datos consta de los siguientes pasos:

1. **Selección del conjunto de datos,** Definición de las variables objetivo , como a las variables independientes, como posiblemente al muestreo de los registros disponibles.
2. **Análisis de las propiedades de los datos**, Acá se construyen los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
3. **Transformación del conjunto de datos de entrada,** Se le conoce como pre-procesamiento de los datos.
4. **Seleccionar y aplicar la técnica de minería de datos,** Se construye el modelo predictivo, de clasificación o segmentación.
5. **Extracción de conocimiento,** Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.
6. **Interpretación y evaluación de datos,** Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Se profundizara aún más en la DM para interpretar de mejor manera la solución. Se comenzara desde las metodologías, pasando por las principales técnicas de minería de datos hasta finalmente los modelos de minería de datos. Estos serán expuestos a continuación:

## 4.3.2 Metodologías de minería de datos

Las metodologías permiten llevar a cabo el proceso de mimería de datos en forma sistemática y no trivial. Ayudan a entender el proceso de descubrimiento de conocimiento y proveen un guía para la planificación y ejecución de los proyectos.

Algunos modelos conocidos como metodologías son en realidad modelo de proceso, es decir un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece que hacer y la metodología especifica cómo hacerlo. Una metodología no solo define fases de un proceso sino también tareas que decidan realizarse y como llevar a cabo las mismas.

Se investigaron 3 metodologías más un modelo de proceso, los cuales serán explicadas a continuación:

### 4.3.2.1 Metodología SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) fue creada especialmente para trabajar con el software de la empresa SAS, y se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando.

### 4.3.2.2 Metodología CRISP-DM

CRISP- DM fue creada por el grupo de empresas SPSS, NCR y Daimler Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos. Divide el proceso en seis fases o entregables:

- ✓ Comprensión del negocio
- ✓ Comprensión de los datos
- ✓ Preparación de los datos
- ✓ Modelado
- ✓ Evaluación
- ✓ Implantación.

La sucesión de fases, no es necesariamente rígida. Cada fase se descompone en varias tareas generales de segundo nivel. CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo.

### 4.3.2.3 Metodología Catalyst

La metodología Catalyst, conocida como P3TQ (Product, Place, Price, Time, Quantity), es una metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información.

El Modelo de Negocio (MII), proporciona una guía de pasos para identificar un problema (o la oportunidad del mismo) y los requerimientos reales de la organización. El foco que propone la metodología Catalyst en su Modelo de Negocio sobre la cadena de valor organizacional, hizo que sea difundida en la comunidad científica como metodología "P3TQ", aunque ésta no sea su denominación original.

### 4.3.2.4 Modelo de proceso KDD

El concepto KDD se define como la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita, donde se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones, para conseguirlo hará falta técnicas de aprendizaje, estadística y bases de datos. Más que una metodología (Como) KDD está pensado como un modelo de proceso (Que) que incluye la minería de datos como un paso dentro de la ejecución completa. Sin embargo igual se incluyó dentro del análisis, ya que sigue siendo una manera eficaz de extraer conocimiento. A continuación se explicara el proceso de KDD.

#### 4.3.2.4.1 El proceso de KDD

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos. Este proceso involucra varios pasos los cuales son:

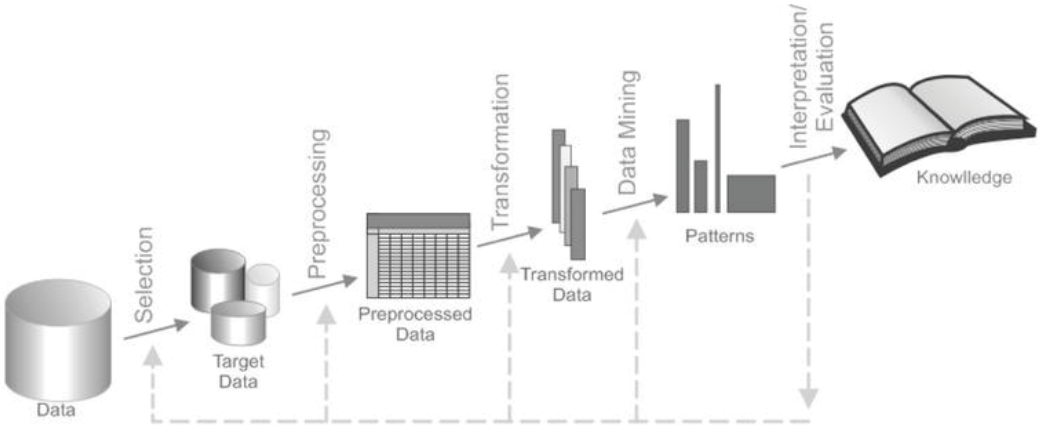


Figura 1: Proceso de KDD

- ✓ **Determinar las fuentes de información**: que pueden ser útiles y dónde conseguirlas.
- ✓ **Diseñar el esquema de un almacén de datos (Data Warehouse):** que consiga unificar de manera operativa toda la información recogida.
- ✓ **Selección, limpieza y transformación de los datos que se van a analizar:** la selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).La limpieza y pre-procesamiento de datos se logra diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario), etc.
- ✓ **Implantación del almacén de datos:** que permita la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
- ✓ **Seleccionar y aplicar el método de minería de datos apropiado:** esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc.
- ✓ **Evaluación, interpretación, transformación y representación de los patrones extraídos:** Interpretar los resultados y posiblemente regresar a los pasos anteriores. Esto puede involucrar repetir el proceso, quizás con otros datos, otros algoritmos, otras metas y otras estrategias.
- ✓ **Difusión y uso del nuevo conocimiento:** Incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) lo cual puede incluir resolver conflictos potenciales con el conocimiento existente. El conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. En este sentido, KDD implica un proceso interactivo e iterativo involucrando la aplicación de varios algoritmos de minería de datos.

#### 4.3.2.5 Técnicas de minería de datos

Como ya es conocido, las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados. Las técnicas más representativas son:

- ✓ **Redes Neuronales:** Son un paradigma de aprendizaje automático. inspirado en la forma en que funciona el funcionamiento de los animales.
- ✓ **Regresión lineal:** Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.
- ✓ **Árboles de decisión:** Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la IA, dada una base de datos se construyen estos diagramas de construcciones lógicas.

- ✓ **Modelos Estadísticos:** Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.
- ✓ **Agrupación o Clustering:** Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia
- ✓ **Reglas de Asociación**: Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Para el caso específico se desarrollara de manera extendida las reglas de asociación, ya que estas ayudan a encontrar relaciones de tipo implicancia en una transacción a la hora de hacer una compra.

#### 4.3.2.5.1 Reglas de asociación

La minería de Reglas de Asociación es una técnica importante en la Minería de Datos y consiste en encontrar las asociaciones interesantes en forma de relaciones de implicación entre los valores de los atributos de los objetos de bases de datos transaccionales, relacionales o Data Warehouse.

Numerosos y recientes estudios avalan su actualidad e importancia y aplicación en áreas como:

- ✓ El mercadeo
- ✓ Medicina seguridad en redes
- ✓ Análisis de información de ventas
  - o Diseño de catálogos
  - o Distribución de mercancías en tiendas
  - o Segmentación de clientes en base a patrones compra

Una regla de asociación es una expresión de la forma X ⇒ Z donde X y Z son conjuntos de elementos.

Formalmente se define como:

$I = \{i_1, i_2, \ldots, i_n\}$ => Un conjunto de n atributos binarios llamados ítems.

$D = \{t_1, t_2, \ldots, t_m\}$ => Un conjunto de transacciones almacenadas en una base de datos.

Cada transacción en $D$ tiene un **ID** (identificador) único y contiene un subconjunto de ítems de. Una **regla** se define como una implicación de la forma:

$$X \Rightarrow Y$$

Dónde:

$X, Y \subseteq I$ y $X \cap Y = \emptyset$

Los conjuntos de ítems $X$ y $Y$ se denominan respectivamente antecedente y consecuente.

- Soporte, s, es la probabilidad de que una transacción contenga {X , Y }
- Confianza, c, es la probabilidad condicional de que una transacción que contenga {X} también contenga {Y}.

#### 4.3.2.5.2 Evaluación de las Reglas

En minería de datos con reglas de asociación en BD transaccionales evaluamos las reglas de acuerdo al soporte y la confianza de las mismas.

En reglas de asociación, la cobertura se llama soporte y la precisión se llama confianza.

Se pueden leer como:

$$\text{Soporte}(X \Rightarrow Z) = P(X \cup Z)$$

$$\text{Confianza}(X \Rightarrow Z) = P(Z|X) = \frac{\text{soporte}(X \cup Z)}{\text{Soporte}(X)}$$

En realidad estamos interesados únicamente en reglas que tienen mucho soporte (soporte ≥ sop min y confianza ≥ conf min), por lo que buscamos (independientemente de que lado aparezcan), pares atributo-valor que cubran una gran cantidad de instancias.

A estos, se les llama ítem-sets y a cada par atributo-valor ítem. Un ejemplo típico de reglas de asociación es el análisis de la canasta de mercado.

Básicamente, encontrar asociaciones entre los productos de los clientes, las cuales pueden impactar a las estrategias de mercado. Ya que tenemos todos los conjuntos, los transformamos en reglas con la confianza mínima requerida. Algunos ítems producen más de una regla y otros no producen ninguna.

Otra criterio de evaluación útil para las reglas de asociación está el Lift o ascensor que se define como la relación del soporte observada a la esperada si X e Y eran independientes. Estas pueden definirse matemáticamente de la siguiente manera.

$$\text{Lift}(X \Rightarrow Z) = \frac{\text{soporte}(X \cup Y)}{\text{soporte}(X)\ x\ \text{soporte}(Y)}$$

Se investigaron Varios algoritmos, sin embargo se explicaran solo dos, por ser atingentes a la investigación, Estos son, algoritmos de reglas de asociación y el A priori, ambos serán explicados a continuación:

##### 4.3.2.5.2.1 Algoritmos de reglas de asociación

Una regla de asociación es una regla que implica ciertas relaciones de asociación entre un conjunto de objetos en una base de datos. De un conjunto de transacciones, donde cada transacción es un conjunto de literales (llamados ítems), una regla de asociación es una expresión de la forma X Y, donde X y Y son conjuntos de datos. El significado intuitivo de tal regla es que las transacciones de la base de datos que contiene la X tiende a contener Y.

Un ejemplo de una regla de asociación es: “el 30% de transacciones que contienen cerveza también contenga pañales; 2% de todas las transacciones contenga ambos de estos ítems. Aquí 30% es llamado la confianza de la regla, y 2% el apoyo de la regla. El problema es encontrar todas las reglas de asociación que satisfaga un mínimo de usuarios específicos.

#### 4.3.2.5.2.2 Algoritmo Apriori

Un algoritmo de regla de asociación Apriori se ha desarrollado para reglas de minería de datos para grandes transacciones sobre bases de datos por el equipo de IBM Quest.

Este algoritmo divide el problema de las reglas de asociación en dos partes:

1. Primero se debe encontrar todas las combinaciones de ítems que tienen soporte de transacción. Esas combinaciones se denominan frecuencia de conjunto de ítems.
2. Utilizar las frecuencias de los conjuntos de ítems para generar las reglas que se desean. La idea general es que si, por ejemplo, ABCD y AB son las frecuencias conjuntos de ítems, entonces podemos determinar si la regla AB CD se mantiene al computar el radio R = Minería de datos una herramienta para la toma de decisiones Capítulo 360soporte (ABCD) /soporte (AB). La regla se mantiene sólo si R >= mínimo de confianza. La regla tendrá el mínimo de soporte porque ABCD tiene más frecuencia. El algoritmo de Apriori usado en la búsqueda para encontrar la frecuencia de todos los conjuntos de ítems se describe a continuación:

## 4.4 Herramientas

Una herramienta en cualquier ámbito es un objeto elaborado con el fin de facilitar la realización de una tarea que requiere de una aplicación correcta. Para la minería de datos estas nos ayudan a aplicar de manera correcta la metodología y técnica de minería de datos. En esta sección se describirán las herramientas escogidas para la solución.

### 4.4.1 Herramientas para la Minería de Datos.

Existen variadas herramientas en el mercado que permiten aplicar diferentes técnicas de minería de datos para el descubrimiento de patrones novedosos. Estas pueden ser categorizadas en Librerías y en Herramientas propiamente tal.

Las librerías de Minería de datos son un conjunto de métodos que implementan funcionalidades y utilidades básicas como el acceso a datos, modelos de redes neuronales, métodos bayesianos, exportación de resultado. Las librerías se encargan principalmente de facilitar el desarrollo de las tareas de Minería de Datos que son más complejas, como el diseño de experimentos. El problema de las librerías, es que es precisa la comprensión de conocimientos de programación.

La Librería investigada fue:

- ✓ **WEKA (Waikato environment for knowledge analysis):** Es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de Waikato en Nueva Zelanda. Sus principales características son:
    - Acceso a los datos desde un archivo en formato ARFF(es un archivo de texto plano organizado en filas y columnas)
    - Pre-procesado de datos (selección, transformación de atributos)
    - Modelos de Aprendizaje (reglas de asociación, modelos de agrupamiento, modelos combinados)
    - Visualización del entorno

Unas herramientas se caracterizan propiamente por centrarse en un único modelo (redes neuronales, modelos estadísticos, etc.) Su principal ventaja es que no es necesario poseer grandes conocimientos de programación. La herramienta investigada fue:

- ✓ **KNIME:** es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. KNIME está desarrollado en la plataforma eclipse y programado principalmente en java. Está concebido con una herramienta gráfica y dispone de una serie de nodos (Que encapsulan distintos algoritmos) y flechas (Que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva. Los nodos implementan distintos tipos de acciones que pueden ejecutarse sobre una tabla de datos:

- Manipulación de filas y columnas, como muestreos, transformación y agrupaciones.
- Visualización(Histogramas)
- Creación de modelos estadísticos y de minería de datos, como arboles de decisión, máquina de soporte de vectores y regresiones
- Aplicación de dichos modelos sobre conjunto de nuevos datos
- Por último y lo más importante al ser una herramienta código libre posibilita su extensión mediante la creación de nuevos nodo que implementen algoritmos a la medida del usuario. Además existe la posibilidad de utilizar de manera directa y transparente a WEKA y /o de incorporar de manera sencilla código desarrollado de R Python. KNINE es una herramienta con licencia OPENGL, lo que permite descargarla y utilizarla de manera gratuita.

Una vez concluida su descripción, se justificará la elección para el desafío planteado.

# 5 Descripción de la solución

Finalmente después del marco teórico las cuales asentaron las bases teóricas sólidas para una buena elección de la solución a la problemática. En esta fase se eligió:

- La herramienta para la minería de datos.
- La Metodología de :
    - ✓ Minería de datos
    - ✓ Construcción del Data Warehouse

## 5.1 Herramientas de DM escogidas.

Para la elección de la herramienta se tomaron varios criterios en consideración a la hora de tomar la decisión. Los criterios considerados fueron:

- ✓ Facilidad de uso
- ✓ Con los algoritmos necesarios
- ✓ Que permita la integración de diferentes módulos
- ✓ Flexible
- ✓ Limpieza y transformación de los datos
- ✓ Entrega de resultados de manera gráfica

Siguiendo la pauta establecida las herramientas que cumplen con lo nombrado son KNINE y el módulo WEKA. Se eligen básicamente por su alto reconocimiento y usabilidad. Además WEKA posee la capacidad de trabajar con reglas de asociación que es lo referente al caso. Otro de los criterios cumplidos fue que KNIME es una muy buena herramienta de manipulación de los datos, ya que permite, limpiarlos, extraerlos, cargarlos desde una base de datos, leerlos, manejarlos de columnas (cambiar filas y mover columnas).

Posee una extensión especial de WEKA que permite aplicar los algoritmos necesarios. Ambas, fueron escogidas ya que la combinación de ellas puede dar los resultados esperados. En resumen, KNIME será utilizada para limpiar los datos, enviarlos a la base de datos y estructurarlos para aplicar los algoritmos y WEKA será utilizado para aplicar algoritmos que den resultados gráficos.

## 5.2 Metodología escogida para la construcción de Data Warehouse

Si bien antes del diseño de un almacén de datos, primero hay que mirar a sus objetivos de negocio, a corto plazo y largo plazo. Además se debe analizar las fuentes de datos de cantidad y calidad. Por último, evaluar su nivel de recursos, plazos y presupuesto. Esto le ayuda a llegar a qué método adoptar Inmon de Kimball o de o una combinación de ambos.

En este caso particular, se trata de los datos y la información de solo un departamento, Es decir, lo que surja de este estudio, afectara positivamente al área ventas. Por otro lado los recursos destinados a esta investigación son los puestos a disposición por el estudiante. En relación con el tiempo, se hace relevante destacar que es un estudio pensado en el corto plazo.

Tomando en consideración todo lo anterior, como la optimización local es lo suficientemente bueno y la atención se centra en la victoria rápida, es aconsejable ir por el enfoque de Kimball. Además este tipo de metodología bottom-up permite que, partiendo de cero, podamos empezar a obtener información útil en cuestión de días y después de los prototipos iniciales.

## 5.3 Metodología de minería de datos propuesta

Para llegar a una determinación de que metodología utilizar, se realizara un análisis comparativo de las metodologías, evaluando distintos criterios y cuales se adecuan al proyecto específico. Se definirán 4 criterios para la elección final, estas se mencionan a continuación:

1. **Escenarios y puntos de partida considerados para el proyecto**: Según el punto de partida del proceso, es posible clasificarlos en:

    a. Escenarios donde se aborda desde la minería de datos una situación organizacional (un problema o una oportunidad), buscando patrones y relaciones que puedan colaborar con la misma.
    b. Escenarios donde el proyecto comienza con un conjunto de datos y el objetivo es explorarlos para encontrar relaciones interesantes que puedan ser útiles en el dominio de aplicación.

2. **Estructura de fases del proceso**: Fases necesarias para llevar a cabo un proyecto de minería de datos, de las cuales se pueden identificar análisis, modelamiento de los datos (selección y preparación de los datos), comprensión del problema, etc.

3. **Nivel de detalle en las tareas de cada fase:** Se evalúa el grado de profundidad con el que se describen las actividades y tareas en cada una de las fases del proceso. Algunos modelos describen sólo las fases generales, mientras que otros establecen las tareas específicas a llevar a cabo en cada una de ellas.
4. **Actividades para la gestión del proyecto**: Las actividades que se llevan a cabo dentro de cada categoría pueden ser de planificación o bien de control. Las actividades de planificación incluyen la identificación de las tareas a realizar en el proyecto, estimación de la duración de las mismas, estimación de los recursos afectados y la definición del curso de acción. Las actividades de control tienen por objetivo el monitoreo del estado actual del proyecto para su comparación con lo planificado.

Tomando en consideración los puntos expuestos anteriormente, podemos hacer un análisis esperado desde los puntos hablados anteriormente. Estos serán expuestos de la forma que fueron presentados. A continuación el análisis:

1. **Escenarios y puntos de partida considerados para el proyecto**. Entre los cuatro modelos analizados, sólo SEMMA inicia el proyecto de minería a partir del conjunto de datos (la primera fase es el muestreo de los datos). CRISP-DM y KDD (en su versión completa de nueve pasos) comienzan con un análisis del negocio y del problema organizacional. Catalyst es la metodología más completa en este aspecto, ya que considera cinco escenarios posibles como punto de partida, entre los cuales se encuentra el inicio desde un problema u oportunidad de negocio.

2. **Estructura de fases del proceso**.

   KDD, CRISP-DM y Catalyst contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería. SEMMA excluye esta actividad del modelo. En todos los modelos se contempla la selección y preparación de los datos. Esta situación se repite para la fase de modelado, donde se aplican las técnicas de minería para obtener los nuevos patrones. La fase de evaluación de los patrones obtenidos está presente también en todas las metodologías. En SEMMA, la evaluación e interpretación de estos patrones se realiza sobre el desempeño del modelo, mientras que en las otras metodologías la evaluación se realiza en función de la utilidad que se aporta al dominio de aplicación o problema organizacional. La implementación de los resultados obtenidos es una fase que no está incluida en el modelo SEMMA. En CRISP-DM, se propone además una planificación para el control futuro y un análisis de cierre del proyecto (análisis postmortem). El análisis postmortem consiste en encontrar información objetiva acerca de la trayectoria de un proyecto, con la finalidad de poder hacer una evaluación abierta del equipo de trabajo, de las decisiones tomadas a lo largo del mismo, de las tecnologías empleadas y sus consecuencias, con el objetivo de incorporar lo aprendido en proyectos futuros.

3. **Nivel de detalle en las tareas de cada fase.**

   Los modelos KDD y SEMMA proponen sólo los pasos generales del proyecto de minería de datos, sin especificar puntualmente las tareas que deben llevarse a cabo en cada una de sus fases. En cambio, los modelos CRISP-DM y Catalyst, especifican con mayor detalle las actividades del proceso, aunque Catalyst señala además "cómo" realizarlas. KDD y SEMMA se acercan más a un modelo de proceso que a una metodología, ya que sólo definen las fases generales. En proyectos donde se desee aplicar los mismos, cada organización deberá establecer las tareas y las actividades que implementará en cada etapa. Si bien los modelos CRISP-DM y Catalyst no llegan a especificar con un alto nivel de detalle cómo realizar todas las tareas, podrían ser considerados una metodología ya que describen y puntualizan las actividades específicas a realizar en cada fase del proceso.

4. **Actividades para la gestión del proyecto**

Podemos observar que tanto la metodología CRISP-DM como la metodología Catalyst proponen actividades de planificación para las distintas áreas de la gestión del proyecto, pero no explicitan tareas de control y monitoreo. KDD y SEMMA no incluyen actividades de gestión del proyecto.

## 5.4 Base de datos

Es de vital importancia tener la información almacenada y que ésta se encuentre me manera ordenada y consistente. Justamente en este proyecto la principal fuente de información son los archivos texto plano, los cuales contienen grandes cantidades de registros. Dada la complejidad de la solución se hace importante el uso de este recurso.

Para la elección de la base de datos se tomaron en cuenta las principales características de los motores de base de datos más utilizados y con ellas se delibero cual era la más adecuada para utilizar.

Se investigaron los motores de base de datos Mysql, postgresql y SQL server, de los cuales se tomó en consideración uno que fuera de fácil uso, fiable y gratuito. Se eligió el motor Mysql ya que esta optimizado para consultas sencillas lo que ofrece un rendimiento óptimo para la solución. Además este puede ser instalado de forma gratuita lo que ofrece un ahorro en los costos. Por último provee de una fiabilidad e integridad de los datos lo que permitirá asegurar su consistencia.

## 5.5 Algoritmos.

Posteriormente a la limpieza, transformación y carga de los datos, se aplican los algoritmos de minería de datos. Sin embargo para esto se deben seleccionar los algoritmos adecuados para el caso.

Entre los tipos de algoritmos de Data Mining están los clasificadores Bayesianos, funciones de clasificación, algoritmos perezosos, meta clasificadora, reglas de asociación, reglas de clasificación, arboles, algoritmos de búsqueda, clustering y otros. Todos tienen distintas finalidades por lo mismo se escogió el que fuera más apropiado para la investigación.

El algoritmo escogido es Apriori, que permite generar las reglas de asociación y describir hechos que ocurren en común dentro de un determinado conjunto de datos y algoritmos. Básicamente el criterio que se tomó para elegir el algoritmo a utilizar es la compatibilidad con el software escogido y la información disponible.

# 6 Aplicación al caso de estudio.

Para encontrar las relaciones existentes entre los productos del supermercado fue necesario llevar a cabo el proceso de extracción de la información. Primero se comenzará explicando el método de pre-procesamiento al archivo entregado al comienzo.

Posteriormente se detallará el proceso a seguir para la construcción del repositorio de extracción explicando con anterioridad y el proceso ETL que se realizó en él.

Para finalizar ésta, se realizará la aplicación del algoritmo A-priori y se analizarán los resultados obtenidos.

## 6.1 Acerca de los datos

Este archivo fue entregado por parte del profesional informático, con una cantidad de 75.000 registros, para el periodo 2014 y con un formato establecido, en donde la mayoría de los datos son obligatorios (datos seleccionados y provistos por él). Sin embargo existe un campo que no es obligatorio el cual es el número de contacto telefónico, ya que los clientes muchas veces no tienen o derechamente no lo quieren dar. Por lo mismo este provee de datos incompletos. A pesar que ahora no es de utilidad para el estudio, podrían ser utilizados como medio para futuras investigaciones.

| Row ID | IdBoleta | Rut | Nombre | Telefono | Fecha | Genero | Cantidad | CodPro... | Nomb.P... | Precio | Categoria | Marca |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Row0 | 47228990 | 10843630-1 | JANINA DEL CARMEN ARANCIBIA SANC... | 107345577 | 10-2-14 10:04 | Mujer | 1 | 97647758 | Lomo Liso | 58455 | Carne de vac... | Super Cerdo |
| Row1 | 47506530 | 7169161-6 | EUGENIA ISABEL CARVAJAL CABALLERO | ? | 11-2-14 10:04 | Mujer | 2 | 93531017 | Super Cerdo... | 58968 | Carne de cerdo | Super Cerdo |
| Row2 | 32205760 | 10948618-3 | CLAUDIA PATRICIA MARIN FUENTES | 12282501 | 12-2-14 10:04 | Mujer | 4 | 29910027 | Vinos Santa ... | 82313 | Vino | Santa Rita |
| Row3 | 23264157 | 9996502-9 | MORAINA AMANDA RODRIGUEZ BRAVO | ? | 13-2-14 10:04 | Mujer | 0 | 68669302 | Pan Integral | 67102 | Pan | Bimbo |
| Row4 | 11093286 | 23918959-8 | KRYSMA ANTONELLA HERRERA NOVOA | ? | 14-2-14 10:04 | Mujer | 2 | 15718978 | Tomate a gr... | 28125 | Ensaladas | ? |
| Row5 | 48347531 | 8628291-7 | CECILIA DE LAS NIEVES ARRIAGADA NE... | ? | 15-2-14 10:04 | Mujer | 2 | 87522126 | Jugos Watts | 36431 | Jugos | Watts |
| Row6 | 33283991 | 7388601-5 | ALICIA DEL CARMEN BUSTOS VALDIVIESO | 59428509 | 16-2-14 10:04 | Mujer | 0 | 109619360 | Sopa super ... | 31529 | Sopas | Cari¤o |
| Row7 | 9832870 | 8157557-6 | EGLA SILVIA TAPIA RAMIREZ | 67264245 | 17-2-14 10:04 | Mujer | 3 | 30447310 | Mantequilla ... | 5615 | Mantequilla | Colun |
| Row8 | 77558771 | 7278621-1 | ALFONSO LISARDO GERA MARTÒNEZ RI... | 34523319 | 18-2-14 10:04 | Hombre | 3 | 67930234 | Queso Chanco | 36463 | Queso | Chanco |
| Row9 | 19691469 | 4409712-5 | SILVIA HUERTA FERMANDOIS | ? | 19-2-14 10:04 | Mujer | 5 | 90556411 | Tomate a gr... | 72475 | Ensaladas | ? |
| Row10 | 717029 | 5266541-8 | CLAUDIO CARTAGENA GALVEZ | 94659475 | 20-2-14 10:04 | Hombre | 2 | 51425555 | Super Cerdo... | 14216 | Carne de cerdo | Super Cerdo |
| Row11 | 80205384 | 3674812-5 | ANELBA TORRES TOLOSA | 9217951 | 21-2-14 10:04 | Mujer | 2 | 15471679 | Super Cerdo... | 62738 | Carne de pollo | Super Pollo |
| Row12 | 87731643 | 7737966-5 | GLORIA GUDELIA ROJAS CABELLO | 58677019 | 22-2-14 10:04 | Mujer | 1 | 62862049 | Lechuga | 69115 | Ensaladas | ? |
| Row13 | 72997848 | 4380665-3 | ELBA DEL ROSARIO VEGA | ? | 23-2-14 10:04 | Mujer | 4 | 23743957 | Jugos Watts | 27751 | Jugos | Watts |
| Row14 | 40349128 | 7555324-2 | Rosa Albertina Cisterna Galarce | 21718681 | 24-2-14 10:04 | Mujer | 3 | 38987531 | Sopa super ... | 80411 | Sopas | Cari¤o |
| Row15 | 44353091 | 11731435-9 | Yeny Leon Allendes | 66680338 | 25-2-14 10:04 | Mujer | 6 | 82579072 | Mantequilla ... | 524 | Mantequilla | Colun |
| Row16 | 21771122 | 6213579-4 | LORENZA DEL CARMEN AGUILERA AGUIL... | 31108513 | 26-2-14 10:04 | Mujer | 4 | 19664616 | Queso Chanco | 72280 | Queso | Chanco |

Figura 2: Registros con campos vacíos

## 6.2 Proceso de extracción

El proceso de extracción corresponde a la selección del archivo y su posterior carga del mismo a una tabla de datos relacional, en este caso Mysql, para realizar posteriormente una transformación y carga de estos en el Data Warehouse.

Para realizar la extracción desde la fuente de origen y obtener cada uno de los campo se utilizó la herramienta Knime, esta herramienta posee opciones para poder limpiar los registros, y luego enviarlos a una base de datos Mysql.

Una vez almacenados, se llena una tabla definida con anterioridad la cual servirá para trabajar sobre ella. Finalmente se realizara el proceso de transformación y carga de ellos en el data Warehouse. La figura siguiente muestra el archivo ya procesado.

| Row ID | Col0 | Col1 | Col2 | Col3 | Col4 | Col5 | Col6 | Col7 | Col8 | Col9 | Col10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Row0 | 47228990 | 10843630-1 | JANINA DEL CARMEN ARANCIBIA SANC... | 107345577 | Mujer | 1 | 97647758 | Lomo Liso | 58455 | Carne de vac... | Super Cerdo |
| Row2 | 32205760 | 10948618-3 | CLAUDIA PATRICIA MARIN FUENTES | 12282501 | Mujer | 4 | 29910027 | Vinos Santa ... | 82313 | Vino | Santa Rita |
| Row6 | 33283991 | 7388601-5 | ALICIA DEL CARMEN BUSTOS VALDIVIESO | 59428509 | Mujer | 0 | 109619360 | Sopa super ... | 31529 | Sopas | Cari×o |
| Row7 | 9832870 | 8157557-6 | EGLA SILVIA TAPIA RAMIREZ | 67264245 | Mujer | 3 | 30447310 | Martequilla ... | 5615 | Mantequilla | Colun |
| Row8 | 77558771 | 7278621-1 | ALFONSO LISARDO GERA MARTÖNEZ RI... | 34523319 | Hombre | 3 | 67930234 | Queso Chanco | 36463 | Queso | Chanco |
| Row10 | 717029 | 5266541-8 | CLAUDIO CARTAGENA GALVEZ | 94659475 | Hombre | 2 | 51425555 | Super Cerdo... | 14216 | Carne de cerdo | Super Cerdo |
| Row11 | 80205384 | 3674812-5 | ANELBA TORRES TOLOSA | 9217951 | Mujer | 2 | 15471679 | Super Cerdo... | 62738 | Carne de pollo | Super Pollo |
| Row14 | 40349128 | 7555324-2 | Rosa Albertina Cisterna Galarce | 21718681 | Mujer | 3 | 38987531 | Sopa super ... | 80411 | Sopas | Cari×o |
| Row15 | 44353091 | 11731435-9 | Yeny Leon Allendes | 66680338 | Mujer | 6 | 82579072 | Martequilla ... | 524 | Mantequilla | Colun |
| Row16 | 21771122 | 6213579-4 | LORENZA DEL CARMEN AGUILERA AGUIL... | 31108513 | Mujer | 4 | 19664616 | Queso Chanco | 72280 | Queso | Chanco |
| Row18 | 33686125 | 20181717-k | CAMILA DEL PILAR SAAVEDRA BRICE¥O | 32525506 | Mujer | 5 | 50658513 | Super Cerdo... | 105668 | Carne de cerdo | Super Cerdo |
| Row19 | 12300397 | 23086244-3 | JAIME PATRICIO VALDIVIA PARRA | 10752605 | Hombre | 4 | 35128299 | Vino blanco ... | 6538 | Vino | Santa Rita |

Figura 3: Extracto de tabla de archivo trabajado

## 6.3 Repositorio de extracción

Para tener los datos que se han extraídos de manera ordenada es necesario contar con un repositorio de extracción. Esto es útil por muchos motivos, porque además se pueden aplicar algoritmos como lo son las reglas de asociación, A continuación se muestran el repositorio propuesto.

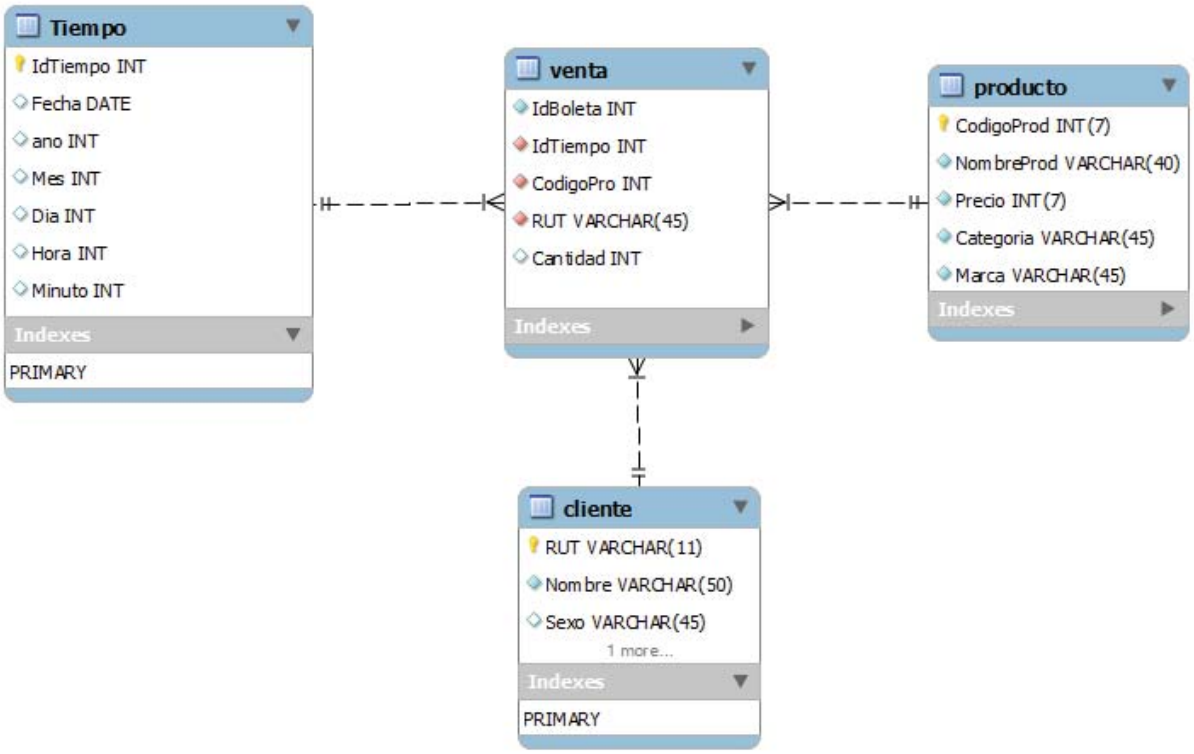


Figura 4: Repositorio de reposición propuesto

## 6.4 Proceso de Transformación y carga

A continuación se detalla la implementación del proceso de transformación y carga del archivo desde la tabla donde se encuentra alojado hacia el repositorio de extracción. Luego de la limpieza en el paso anterior se puede comenzar a cargar todos datos al repositorio de extracción antes modelado. Este proceso es conocido como proceso de carga.

Ya teniendo el repositorio cargado con toda la información, es posible tener un registro ordenado y seguro del archivo. Además ahora es posible hacer consultas de base de datos y aplicar algoritmos de data Mining.

### 6.4.1 Reglas de asociación y Algoritmo A priori

A continuación se presenta el formato en que se encuentran almacenados el Id de la boleta junto con el producto. Este formato debería ser transformado para crear los vectores de comportamiento de los usuarios, los cuales serán utilizados para realizar el estudio de las distintas preferencias que poseen mediante reglas de asociación utilizando el algoritmo A priori. En la imagen se puede apreciar el ID de la compra junto con la secuencia de compra de productos comprados.

| Row ID | transac... | produit |
|---|---|---|
| Row0 | 1 | B |
| Row1 | 1 | E |
| Row2 | 1 | H |
| Row3 | 2 | A |
| Row4 | 2 | B |
| Row5 | 2 | E |
| Row6 | 2 | F |
| Row7 | 3 | B |
| Row8 | 3 | C |
| Row9 | 3 | F |
| Row10 | 3 | H |
| Row11 | 4 | A |
| Row12 | 4 | D |
| Row13 | 4 | F |
| Row14 | 4 | G |
| Row15 | 5 | B |

Figura 5: Formado transaccional de los datos para minar

Del repositorio se extrajeron los identificadores de las boletas, IdBoletas, para traspasarlo a un formato de archivo que nos permitiera poder crear las asociaciones. Es por eso que para poder manejar las reglas de asociación se creó la matriz transaccional con la cantidad de objetos, para dicha tarea. Por medio del módulo Pivoting hizo posible esta conversión.

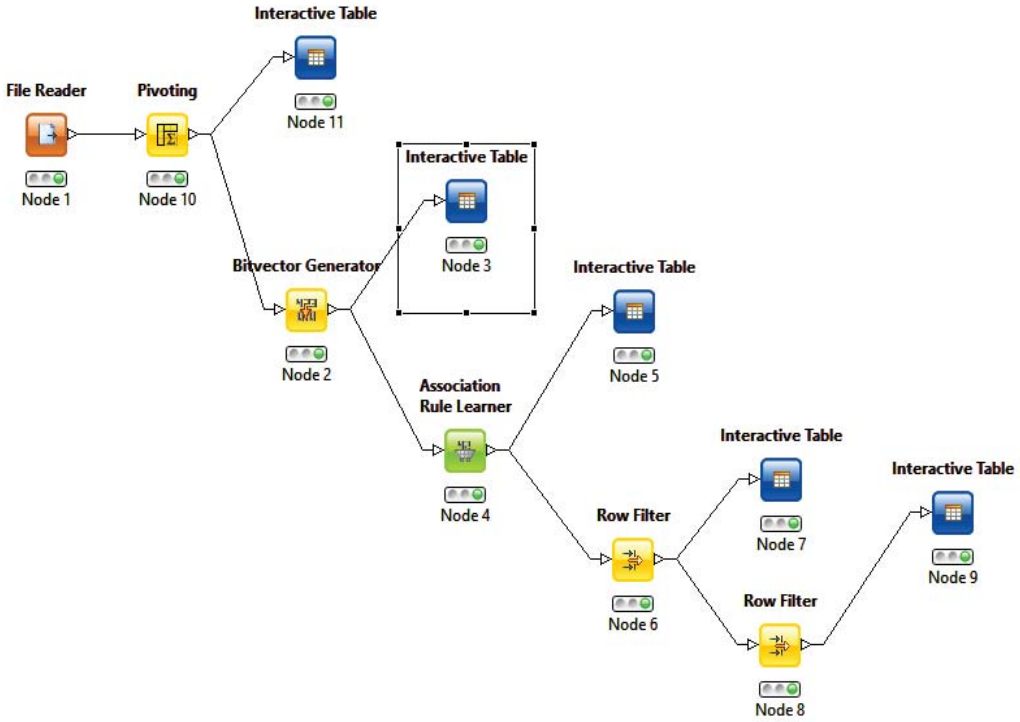


Figura 6: Módulos Knime para la conversión del archivo plano

El primer módulo es el que lee el archivo que será transformado (Por el segundo módulo llamado "**Pivoting"**) en una matriz transaccional para luego ser interpretada por los siguientes módulos. En imagen siguiente se muestra la matriz transaccional para el caso que representa cada transacción del producto como 0 o 1, en donde el uno es representativo de su existencia en la transacción.

| Row ID | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 1 | 1 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 1 | 1 | 0 | 0 |
| 13 | 0 | 0 | 0 | 1 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 |

Figura 7: Matriz transaccional para las reglas de asociación

Luego cada registro de la matriz es pasado a un vector hexadecimal que será leído por el modulo "**Association Rule Learner**" o reglas de asociación. A este módulo se le definió la confianza y el soporte requerido para las reglas de asociación, es decir en este caso particular asigno una confianza de 90% y un soporte mínimo de 0.01. Finalmente se definieron dos filtros (Módulos llamados "**Row Filter"),** el primero para definir el lift que mínimo para las reglas y el segundo para configurar el consecuente.

Para aplicar el algoritmo Apriori se utilizó el nodo de Association Rule learned, que es proporcionado por una extensión de knime, acá se utilizaron distintos nodos para cargar la matriz y adecuar los datos contenidos en ella para realizar la correcta aplicación del algoritmo.

A continuación se muestra la figura 12. Detalladamente los parámetros que fueron utilizados para configurar l nodo del algoritmo.

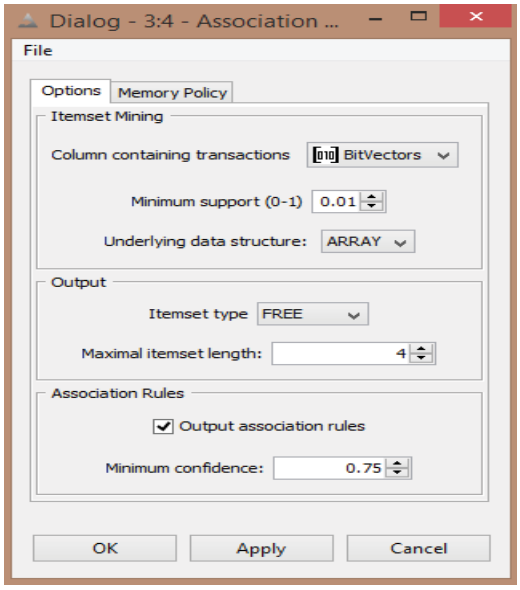


Figura 8: Configuración del Asocciation rule learned

Los valores de soporte mínimo es 0.01 y de confianza es de 0,75 los que sustentaran que las reglas encontradas tengan una relevancia para el estudio.

# 7 Resultados Obtenidos

Finalmente luego de explicar los pasos para obtener lo deseado, se abordarán y discutirán los resultados obtenidos de acuerdo a distintos criterios de separación.

Desde un comienzo se parte con un conjunto de ítems finito que genera una gran cantidad de reglas de asociación que se repiten, sin embargo el primer filtro conceptual y el que le da cierta validez inicial a cada reglas son la confianza y el soporte. En este caso en particular se utilizó una confianza del 0,75% y un soporte de 1%, en la primera prueba. Luego se definió con un 90% de confianza. Sin embargo más adelante se establecerán más criterios de separación. A continuación en la Figura 13 se puede apreciar que se generaron 77 reglas (con confianza 0.75) de asociación de las cuales algunas fueron esperadas y otras no tanto. Un valor esperado teniendo en cuenta que eran 75.000 registros.

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule47 | 0.026 | 0.92 | 13.531 | Mayonesa | <--- | [Cordero,Cereal] |
| rule48 | 0.026 | 0.921 | 13.496 | Cordero | <--- | [Mayonesa,Cereal] |
| rule49 | 0.026 | 0.921 | 13.603 | Cereal | <--- | [Cordero,Mayonesa] |
| rule50 | 0.026 | 0.923 | 13.638 | Queso lamin... | <--- | [Mayonesa,Cereal] |
| rule51 | 0.026 | 0.926 | 13.618 | Mayonesa | <--- | [Queso laminado,Cereal] |
| rule52 | 0.026 | 0.922 | 13.618 | Cereal | <--- | [Queso laminado,Mayonesa] |
| rule53 | 0.026 | 0.925 | 13.667 | Queso lamin... | <--- | [Cordero,Cereal] |
| rule54 | 0.026 | 0.929 | 13.611 | Cordero | <--- | [Queso laminado,Cereal] |
| rule55 | 0.026 | 0.919 | 13.569 | Cereal | <--- | [Queso laminado,Cordero] |
| rule56 | 0.026 | 0.926 | 13.687 | Queso lamin... | <--- | [Cordero,Mayonesa] |
| rule57 | 0.026 | 0.92 | 13.518 | Mayonesa | <--- | [Queso laminado,Cordero] |
| rule58 | 0.026 | 0.925 | 13.559 | Cordero | <--- | [Queso laminado,Mayonesa] |
| rule59 | 0.027 | 0.94 | 12.743 | Cerveza | <--- | [Ajo,kiwi] |
| rule60 | 0.027 | 0.937 | 12.095 | kiwi | <--- | [Ajo,Cerveza] |
| rule61 | 0.027 | 0.923 | 8.452 | Vino blanco | <--- | [Queso,Papas prefritas] |
| rule62 | 0.027 | 0.922 | 12.154 | Queso | <--- | [Vino blanco,Papas prefritas] |
| rule63 | 0.028 | 0.995 | 9.113 | Vino blanco | <--- | [Pescado,Salsa blanca,Papas duqueza] |
| rule64 | 0.028 | 0.993 | 12.879 | Pescado | <--- | [Vino blanco,Salsa blanca,Papas duqueza] |
| rule65 | 0.028 | 0.994 | 12.878 | Salsa blanca | <--- | [Pescado,Vino blanco,Papas duqueza] |
| rule66 | 0.028 | 0.814 | 7.926 | Papas duqu... | <--- | [Pescado,Vino blanco,Salsa blanca] |
| rule67 | 0.028 | 0.907 | 11.768 | Pescado | <--- | [Salsa blanca,Papas duqueza] |
| rule68 | 0.028 | 0.904 | 11.716 | Salsa blanca | <--- | [Pescado,Papas duqueza] |
| rule69 | 0.028 | 0.765 | 7.451 | Papas duqu... | <--- | [Pescado,Salsa blanca] |
| rule70 | 0.028 | 0.905 | 8.29 | Vino blanco | <--- | [Pescado,Papas duqueza] |
| rule71 | 0.028 | 0.89 | 11.546 | Pescado | <--- | [Vino blanco,Papas duqueza] |
| rule72 | 0.028 | 0.754 | 7.344 | Papas duqu... | <--- | [Pescado,Vino blanco] |
| rule73 | 0.028 | 0.909 | 8.327 | Vino blanco | <--- | [Salsa blanca,Papas duqueza] |
| rule74 | 0.028 | 0.891 | 11.545 | Salsa blanca | <--- | [Vino blanco,Papas duqueza] |
| rule75 | 0.028 | 0.758 | 7.38 | Papas duqu... | <--- | [Vino blanco,Salsa blanca] |
| rule76 | 0.033 | 0.928 | 11.186 | Ensaladas | <--- | [Mani,Papas duqueza] |
| rule77 | 0.033 | 0.937 | 11.153 | Mani | <--- | [Ensaladas,Papas duqueza] |

Figura 9: Reglas de asociación sin filtro

Sin embargo para efectos prácticos se definieron otros filtros, como ya se mencionó anteriormente para mejorar y hacer más confiable las reglas de asociación. El primero definido fue el lift (definida la relación del soporte observada a la esperada si X e Y eran independientes). Este se quedó con un valor entre 10 y 1.8 para incluir todos los valores. Los resultados pueden verse en la figura 14.

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule61 | 0.027 | 0.923 | 8.452 | Vino blanco | <--- | [Queso,Papas prefritas] |
| rule63 | 0.028 | 0.995 | 9.113 | Vino blanco | <--- | [Pescado,Salsa blanca,Papas duqueza] |
| rule66 | 0.028 | 0.814 | 7.926 | Papas duqu... | <--- | [Pescado,Vino blanco,Salsa blanca] |
| rule69 | 0.028 | 0.765 | 7.451 | Papas duqu... | <--- | [Pescado,Salsa blanca] |
| rule70 | 0.028 | 0.905 | 8.29 | Vino blanco | <--- | [Pescado,Papas duqueza] |
| rule72 | 0.028 | 0.754 | 7.344 | Papas duqu... | <--- | [Pescado,Vino blanco] |
| rule73 | 0.028 | 0.909 | 8.327 | Vino blanco | <--- | [Salsa blanca,Papas duqueza] |
| rule75 | 0.028 | 0.758 | 7.38 | Papas duqu... | <--- | [Vino blanco,Salsa blanca] |
| rule78 | 0.033 | 0.755 | 7.349 | Papas duqu... | <--- | [Ensaladas,Mani] |
| rule82 | 0.034 | 0.936 | 8.567 | Vino blanco | <--- | [Pescado,Salsa blanca] |
| rule85 | 0.041 | 0.774 | 9.43 | Fruta en co... | <--- | [Salmon,Vino] |

Figura 10: Reglas de asociación con lift entre 1,8 y 10

Luego y como tercer parámetro, se usó el consecuente. En este caso se utilizó el predecesor el Vino Blanco para gerenar las reglas. Producto de esta actividad quedaron armadas 5 asociaciones.

Una de las debilidades de Knime es que sólo se pueden relacionar un consecuente en cada regla, esto conlleva que se pierdan algunas más interesante con varios parámetros a la vez. Sin embargo se generaron tres posibles combinaciones individuales que permitieron suplir esta falencia. Dentro de los concecuentes propuestos por administracion estaban Vino Blanco, Carne de cordero y Mayonesa. Estos filtros se solicitaron debido a la fecha que estaba en curso, se requería de asociar de manera tal que favoreciera las ventas. En la siguiente tabla, se muestran los vínculos encontrados utilizando el Vino Blanco.

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule61 | 0.027 | 0.923 | 8.452 | Vino blanco | <--- | [Queso,Papas prefritas] |
| rule63 | 0.028 | 0.995 | 9.113 | Vino blanco | <--- | [Pescado,Salsa blanca,Papas duqueza] |
| rule70 | 0.028 | 0.905 | 8.29 | Vino blanco | <--- | [Pescado,Papas duqueza] |
| rule73 | 0.028 | 0.909 | 8.327 | Vino blanco | <--- | [Salsa blanca,Papas duqueza] |
| rule82 | 0.034 | 0.936 | 8.567 | Vino blanco | <--- | [Pescado,Salsa blanca] |

Figura 11 : Reglas de asociación encontradas con Vino Blanco como consecuente.

Como se puede observar en la tabla, y gracias a esta técnica se revelaron 4 reglas en donde el vino blanco estaba como como consecuente, con un nivel de confianza sobre el 90 % y con un lift del sobre el 1.8. Al ser el lift una medida para comprobar si la relación fue producto del azar, se puede decir que existe un 90% de probabilidad de que las reglas encontradas tendrán como consecuente el vino blanco, es decir si existe pescado, salsa blanca en la canasta existe un 93.6% de probabilidad de que exista una botella de vino blanco. Es por esto que fue declarado como regla fuerte.

Para el cordero también se evidenció reglas interesantes. Para un nivel de confianza por sobre el 90% y un lift entre el 1.8 y 10, se encontraron 10 reglas de asociación interesantes. Al igual que el vino blanco existe una probabilidad de que al seleccionar dichos ítem sets, exista en la canasta cordero. Por ejemplo, en la canasta queso laminado y mayonesa se puede decir con un 92.5% probabilidad que si ambos son están en la canasta, entonces se llevará cordero.

A continuación se muestran los resultados.

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule6 | 0.021 | 0.998 | 14.626 | Cordero | <--- | [Queso laminado,Mayonesa,vod... |
| rule11 | 0.021 | 0.99 | 14.514 | Cordero | <--- | [Mayonesa,Cereal,vodka] |
| rule16 | 0.021 | 0.995 | 14.579 | Cordero | <--- | [Queso laminado,Cereal,vodka] |
| rule20 | 0.021 | 0.908 | 13.299 | Cordero | <--- | [Cereal,vodka] |
| rule25 | 0.021 | 0.904 | 13.24 | Cordero | <--- | [Mayonesa,vodka] |
| rule27 | 0.021 | 0.901 | 13.211 | Cordero | <--- | [Queso laminado,vodka] |
| rule35 | 0.026 | 0.995 | 14.586 | Cordero | <--- | [Queso laminado,Mayonesa,Cer... |
| rule38 | 0.026 | 0.921 | 13.496 | Cordero | <--- | [Mayonesa,Cereal] |
| rule44 | 0.026 | 0.929 | 13.611 | Cordero | <--- | [Queso laminado,Cereal] |
| rule48 | 0.026 | 0.925 | 13.559 | Cordero | <--- | [Queso laminado,Mayonesa] |

Figura 12: Reglas de asociación encontradas con Cordero como consecuente.

Siguiendo la misma idea, se encontraron 9 reglas de asociación para la mayonesa como consecuente, en donde nuevamente el nivel de confianza es de 90% en conjunto con el lift que se configuró entre 1.8 y 10.

Los resultados se muestran en la siguiente tabla.

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule5 | 0.021 | 0.994 | 14.606 | Mayonesa | <--- | [Queso laminado,Cordero,vod... |
| rule8 | 0.021 | 0.994 | 14.616 | Mayonesa | <--- | [Queso laminado,Cereal,vodka] |
| rule10 | 0.021 | 0.997 | 14.653 | Mayonesa | <--- | [Cordero,Cereal,vodka] |
| rule24 | 0.021 | 0.906 | 13.313 | Mayonesa | <--- | [Cordero,vodka] |
| rule28 | 0.021 | 0.913 | 13.426 | Mayonesa | <--- | [Cereal,vodka] |
| rule34 | 0.026 | 0.993 | 14.594 | Mayonesa | <--- | [Queso laminado,Cordero,Cer... |
| rule37 | 0.026 | 0.92 | 13.531 | Mayonesa | <--- | [Cordero,Cereal] |
| rule41 | 0.026 | 0.926 | 13.618 | Mayonesa | <--- | [Queso laminado,Cereal] |
| rule47 | 0.026 | 0.92 | 13.518 | Mayonesa | <--- | [Queso laminado,Cordero] |

Figura 13: Reglas de asociación encontradas con Mayonesa como consecuente.

# 8 Conclusión

El objetivo principal del trabajo de título fue, el resolver alguna problemática propuesta utilizando alguna metodología preferente. Para ello primero se definieron las bases teóricas que permitieron tener un estado del arte y con eso proponer una solución.

Una de las características fundamentales de esta investigación es que a partir del marco teórico, se utilizaron ciertas métricas para generar conocimiento medible y cuantificable, que permitió el uso de esta información útil para el Retail.

Dentro de los objetivos logrados, cabe destacar que el hecho de identificar relaciones ocultas y útiles, es un nicho relevante y fructífero para diferentes negocios y no sólo el del Retail.

Para finalizar se puede decir que este trabajo puede ser el punto de partida para próximas investigaciones, que puedan complementar, como base para otras áreas de investigación, o para superar las falencias encontradas en la herramienta de minería de datos.

# 9 Referencias


[1] Hipp, J., Guntzer, U., y Nakhaeizadeh, G., Algorithms for Association Rule Mining: A General Survey and Comparison, SIGKDD Explorations, 2 (1), 5864, 2000.

[2] Cano, J., Herrera, F., Lozano, M, Extracción de modelos predictivos e interpretables en conjuntos de datos de tamaño grande mediante la selección de conjuntos de entrenamiento, TAMIDA2005, pp.145-152, ISBN: 84-9732-449-8.

*[3] Westphal, C.,Blacton, T.,Data Mining Solutions, Methodos and Tools for Solving Real-Work Problems.*

[4] Witten, H.,Frank, E., Practical Machine Learning Tools and Techniques with Java Implementations.

[5]grawal, R., Imielinski, T. y Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data, 207-216.

[6] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. y Verkamo, A. I. (1996). Fast Discovery of Association Rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 307-328). AAAI/MIT Press.

[7]Bigus, J.P. (1996). Data mining with neural networks: solving business problems from application development to decision support. New York: McGraw-Hill.

[8] Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence.

[9] Apte, C., and Hong, S. J. 1996. Predicting Equity Returns from Securities Data with Minimal Rule Generation. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press.

[10] Cheeseman, P. 1990. On Finding the Most Probable Model. In Computational Models of Scientific Discovery and Theory Formation, eds. J. Shrager and P. Langley, 73–95. San Francisco, Calif.: Morgan Kaufmann.

[11] Dasarathy, B. V. 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.Washington, D.C.: IEEE Computer Society

[12] Mannila, H.; Toivonen, H.; and Verkamo, A. I. 1995. Discovering Frequent Episodes in Sequences. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 210–215.

[13] Howso, C. Successful Business Intelligence, Secrest to Making BI a Killer App, ISBN: 0-07-159614-3.

[14] Vercellis C. Business Intelligence: Data Mining and Optimization for Decision Making. 2009. ISBN: 978-0-470-51138-1

[15] Rivadera G. La metodología de Kimball para el diseño de almacenes de datos. 2010

[16] Azevedo, A., Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADIS 2008. Britos, P. (2008). Procesos de Explotación de Información.

[17] Camaro H., Silva M. Two paths in seach of patterns through Data Mining: SEMMA and CRISP. 2010.

[18] SAS Enterprise Miner: "SEMMA". 2008. http://tinyurl.com/ semmaSAS

[19] Chapman, P., Clinton, J., Keber, R., et al.. "CRISP-DM 1.0 Step by step BI guide". Edited by SPSS. 2000. http://tinyurl.com/ crispDM

[20] Berrios G., Guia Metodologica para la definición y desarrollo de un Data Warehouse. Nicaragua 2003.

[21] Inmon, W.H. Building the Data Warehouse (Third Edition), New York: John Wiley & Sons, (2002).

[22] Classification of power quality considering voltage Sags in distribution systems using KDD Process, Anderson Roges Teixeira Góes, Maria Teresinha Arns Steiner and Rodrigo Antonio Peniche, (2014).

[23] Minig Association Rules between Sets of Items in Large Databases, Rakesh Agrawal, Tomasz Imielinski, Arun Swami, San Jose.

# 10 Anexos.

A continuación se explicarán los pasos de la solución:

## 10.1 Pasos de la solución

A continuación se presentará de manera conceptual el diseño que se desea aplicar para poder analizar el comportamiento de los usuarios, en este diseño será la base para posterior implementación del mismo. Así se definirán los componentes y características de la solución propuesta.

Más adelante se detallará también el proceso ETL (Extracción, transformación y carga) de los archivos texto plano, que en este caso fueron obtenidos de una fuente de internet. Los que serán almacenados en el repositorio de extracción que permitirá la implementación de algoritmos sobre estos.

### 10.1.1 Limpieza de Datos.

Es muy importante tener en cuenta que los archivos texto plano almacenan mucha información que puede contener campos en blanco. Por lo mismo para evitar datos erróneos (Valores en blanco o fuera de rango) se hace una limpieza de los mismos. Además tiene como finalidad reducir el tamaño del archivo para así agilizar el proceso de transformación y carga en el repositorio de extracción.

| IdBolteta | Rut | Nombre | Telefono | Fecha | Sexo | Canidad | CodProd | Nomb. Prod | Precio | Categoria/Producto | Marca |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3443 | 23013594-0 | agustina minett ibañez silva | | 20/10/2014 15:50:09 | M | 2 | 678 | Super Cerdo, Lomo Vetado | 3990 | Carne de cerdo | Super cerdo |
| 3443 | 20182807-4 | katalina godoy pizarro | 91246648 | 20/09/2014 15:50:09 | M | 3 | 679 | Vino blanco torreon de paredes | 3490 | Vinos | Torreon de paredes |
| 3544 | 13537830-5 | wilson alberto ahumada araya | 870307 | 20/08/2014 15:50:09 | H | 4 | 680 | Pan bimbo | 1450 | Pan | Bimbo |
| 4334 | 4449549-k | guillermo horacio diaz | 98917177 | 22/01/2014 15:50:09 | H | 5 | 681 | Cerveza cristal litro | 1100 | Cerveza | Cristal |
| 3544 | 17161231-4 | bernardita vanessa saez fernandez | 033/2715227 | 04/10/2014 15:50:09 | M | 6 | 682 | Jugo watts litro | 560 | Jugo | Watts |
| 5654 | 8758454-2 | jose luis vicencio vera | | 20/10/2014 15:50:09 | H | 2 | 683 | Carne de truto de pavo picada | 1300 | Carne de pavo | Sopraval |

Tabla 1: Representación de los datos en bruto

En la figura se puede ver como es el archivo texto plano en el cual se puede apreciar los campos contenidos.

En la tabla se pueden observar los siguientes datos que contienen el archivo:

- ✓ **IdBoleta:** Identificador de la boleta en una transacción o compra esta puede repetirse ya que en una boleta puede haber más de un producto.
- ✓ **Rut:** Identificador de un cliente del supermercado
- ✓ **Nombre:** Nombre del cliente
- ✓ **Telefono:** Telefono de contacto.
- ✓ **Sexo:** Genero del cliente
- ✓ **Nombre del producto:** Nombre del producto a la venta.
- ✓ **Precio:** Valor del producto.
- ✓ **Categoría:** Nombre Genérico para para el producto el cual permite una rápida búsqueda y clasificación. Por ejemplo el producto, SUPER CERDO LOMO Vetado, su Categoría es Carne de cerdo.
- ✓ **Marca:** Empresa que produce el producto.
- ✓ **Cantidad:** Productos que se lleva en la transacción

#### 10.1.1.1 Vectores de comportamiento del usuario

A partir de las distintas boletas es decir transacciones de compra, se puede definir un vector que contiene los objetos que solicito cuando se realizó la compra. Estos están representados de forma binaria es decir con valor 1 si está presente y 0 si es que está ausente.

La forma del vector es la siguiente:

$V = (P_1, P_2 P_3 P_4 P_5 \dots P_n)$

A partir de los vectores de comportamiento de los usuarios se pueden construir las matrices de tracciones, en las cuales se pueden aplicar una seria de algoritmos de Data mining que utilizan técnicas de aprendizaje no supervisado tales como reglas de asociación. Los algoritmos específicos que se utilizaran serán los explicados más adelante en este proyecto. A modo de ejemplo se presenta la matriz de vistas de productos comprados.

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Figura 14: Matriz transaccional de productos

#### 10.1.1.2 Modelo del Data Warehouse

La pieza fundamental de un sistema Business Intelligence es el DWH porque todos los listados y análisis que se hagan se harán a partir de esta única base de datos Beneficios. Anteriormente se escogió el modelo Ralph Kimball que nos permitirá facilitar algunas tareas. Este modelo está estructurado en jerarquías, que es una agrupación de dimensiones o campos.

Las tablas fueron diseñadas con un alto nivel de detalle para que cualquier consulta posterior pueda resolverse con cualquier campo disponible. Sin embargo esto también conlleva cargar muchos datos, lo que requiere tiempo de carga y espacio en el disco. Como se trata de un proyecto pequeño es conveniente realizarlo de esta manera.

Como nos interesa que las consultas generadas sean sencillas (con pocas tablas y pocas relaciones), El modelo en estrella es perfecto para conseguir este objetivo. Además, desaparece el problema que generan las diferentes jerarquías en que se pueden agrupar los productos. El modelo creado será presentado a continuación:

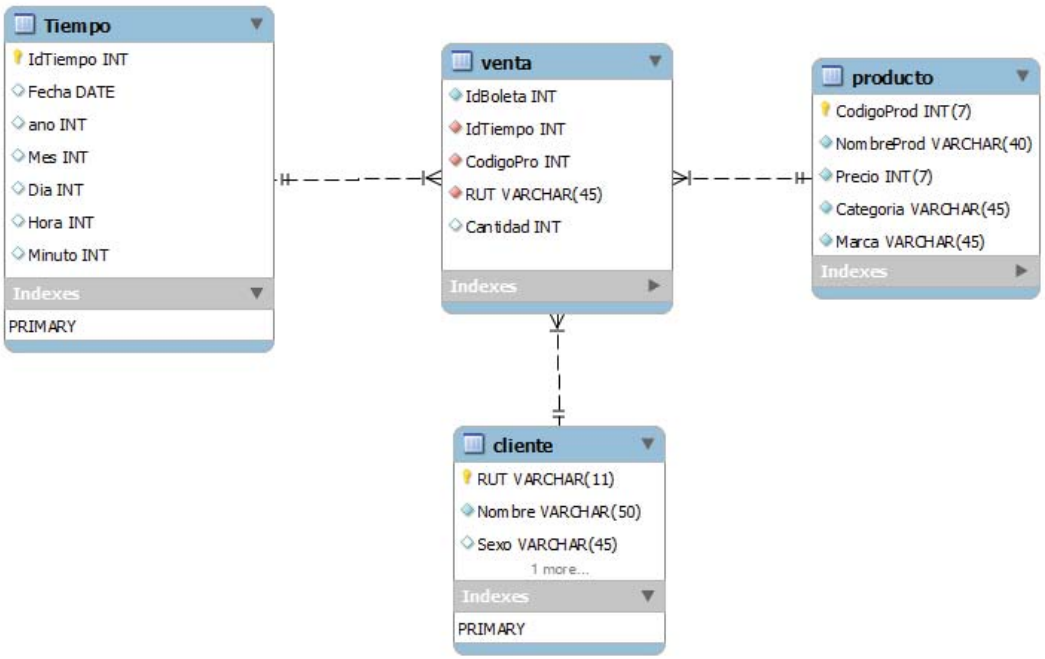


Figura 15: Modelo de DW

Ahora se pasará a explicar con mayor detalle cada dimensión en particular.

- Dimensión Cliente: Registra la información de los clientes que realizaron compras en el supermercado.

| Columna | Descripción |
|---|---|
| RUT | Clave primaria de la tabla |
| Nombre | Nombre del cliente |
| Sexo | Sexo del cliente |

- Dimensión tiempo: Registran el tiempo exacto donde se realizó una compra de productos.

| Columna | Descripción |
|---|---|
| IdTiempo | Corresponde a la clave primaria de la tabla y la clave foránea de la tabla ventas. |
| Fecha | Corresponde la fecha en donde el cliente hace la compra del producto. |
| Ano | Corresponde el año en donde el cliente hace la compra del producto. |
| Mes | Corresponde al mes en donde el cliente hace la compra del producto. |

| | |
|---|---|
| Día | Corresponde el día semana en donde el cliente hace la compra del producto. |
| Hora | Corresponde la hora en donde el cliente hace la compra del producto. |
| Minuto | Corresponde los minutos en donde el cliente hace la compra del producto. |

- Dimensión Producto: Registran los productos comprados por los clientes.

| Columna | Descripción |
|---|---|
| CodigoProd | Corresponde al identificador del producto y la clave foránea de la tabla ventas. |
| NombreProd | Corresponde al nombre del producto |
| Precio | Corresponde al precio del producto |
| Categoría | Corresponde a la familia o categoría del producto |
| Marca | Corresponde a la marca de la familia |

- Dimensión Venta: Corresponde a la tabla de hechos la cual une a las demás tablas. Esta posee el indicador de la cantidad.

| Columna | Descripción |
|---|---|
| IdBoleta | Clave primaria de una venta y corresponde a el código de la boleta. |
| IdTiempo | Clave foránea que une a la tabla Venta con la tabla tiempo |
| CodigoProd | Clave foránea que une a la tabla Venta con la tabla Productos |
| Rut | Clave foránea que une a la tabla Venta con la tabla Clientes |
| Cantidad | Cantidad de productos comprados en esa compra. |

Este repositorio de extracción tiene como fin almacenar la información importante proveniente del archivo texto plano en una estructura adecuada que permita aplicar DM. Además, en este repositorio se almacenan los vectores característicos que serán las entradas de los algoritmos de DM. Desde los cuales se podrán analizar el comportamiento de los compradores.