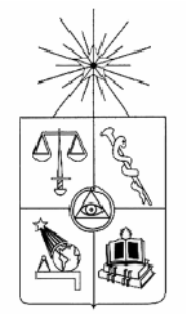


UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

DESARROLLO DE UN SISTEMA DE APOYO A LA TOMA DE DECISIONES PARA EL MANEJO DE PRODUCTOS Y TIENDAS EN UNA CADENA DE RETAIL A PARTIR DE DATOS TRANSACCIONAL DE VENTAS Y CARCTERÍSTICAS DE TIENDAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUTRIAL

JORGE ALEJANDRO GAETE VILLEGAS

PROFESOR GUÍA
JUAN VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN
VICTOR REBOLLEDO LORCA
ROBERT CERCOS BROWNELL

SANTIAGO DE CHILE
OCTUBRE 2009

Índice de contenidos

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR: JORGE GAETE VILLEGAS
FECHA: 06/10/09
PROF. GUIA: SR. JUAN VELASQUEZ

DESARROLLO DE UN SISTEMA DE APOYO A LA TOMA DE DECISIONES PARA EL MANEJO DE PRODUCTOS Y TIENDAS EN UNA CADENA DE RETAIL A PARTIR DE DATOS TRANSACCIONAL DE VENTAS Y CARCTERÍSTICAS DE TIENDAS

El presente trabajo propone la construcción de un sistema de apoyo a la toma de decisiones para la realización de agrupaciones de tiendas y predicción de demanda de productos en una cadena de *retail*.

El proyecto fue realizado en la gerencia de procesos comerciales de Easy S.A. y consideró la información de venta diaria de los dos últimos años para la sección 13, "Corral de herramientas", en todas las tiendas de Easy S.A. en el país.

El trabajo se enmarca dentro de la necesidad de extraer conocimiento desde grandes bases de datos, de manera oportuna, eficiente en términos de recursos humanos y utilizando la mayor cantidad de información posible, tanto para el seguimiento como para el planteamiento de nuevas acciones comerciales, en particular la evolución de datos históricos, agrupación de tiendas según criterios comerciales y la determinación de demanda por productos.

Como solución se proponen la creación de un sistema de apoyo a la toma de decisiones compuesto por un *data mart* para el almacenamiento de los datos y la información relevante para el área de Procesos Comerciales, la generación de herramientas para la recolección, tratamiento y almacenamiento de estos datos, la realización de agrupaciones de tiendas, pronóstico de demanda por productos y la visualización de reportes.

La solución se desarrolló utilizando la metodología KDD para el diseño de los procesos de selección, extracción, limpieza y carga de datos hacia el *data mart.* Para la obtención de agrupaciones de tiendas se utilizaron los algoritmos de s*elf organizing feature maps y Kmeans* y para la realización de predicciones de demanda de productos se utilizaron los algoritmos de redes neuronales artificiales y medias móviles.

Como resultado se encuentran agrupacione de tiendas a partir de sus características físicas de utilización de espacio en sala y variables socioeconómicas de las comunas en las que se encuentran ubicadas, se realizan predicciones de demanda validadas en datos históricos y se implementa un prototipo funcional del sistema de apoyo a la toma de decisiones diseñado.

La presente memoria contiene los lineamientos para la implementación de un sistema de apoyo a la toma de decisiones basado en herramientas *open source*, aplicable a cualquier problema que involucre grandes cúmulos de datos y sobre los cuales sea necesario la aplicación de herramientas para la extracción de información.

Agradecimientos

A mis padres por lo que quiero ser,

A mi familia por lo que soy,

A mis amigos por apoyarme en ser,

A mis profesores por creer en lo que soy.

En especial a mis padres, Josefina y Jorge por el apoyo incondicional y por el amor que siempre me han dado.

A Juan Velásquez por la paciencia y la guía en este proceso, a Robert Cercos y Víctor Rebolledo por las revisiones de mi trabajo y a Tomás Zavala y Claudio Larrea por su ayuda desde Easy y su siempre buena disposición.

A mis amigos, Jorge, Sergio, Matías, Bárbara, Ignacio y Sergio por los momentos y los años que compartimos dentro y fuera de la Universidad.

A mis hermanos Josefina y Benjamín por compartir conmigo su inocencia y su sabiduría y por dejarme ser parte de sus vidas.

1. INTRODUCCIÓN

1.1 Antecedentes generales

El sector del *retail*, o de ventas minoristas, se ha desarrollado en los últimos años como una de las áreas de la economía chilena más potentes y emergentes.

A nivel Sudamericano la importancia de los *retailers* chilenos es creciente, con dos agentes como protagonistas de la industria a nivel regional [1] y con incipientes incursiones en expansiones a nivel sudamericano[2] para el año 2009 y 2010.

La importancia de esta industria en Chile se aprecia en el 21%[3] que representan sus ventas en el PIB nacional y en la tendencia de constante crecimiento en los últimos años[1].

Dentro de este mercado se encuentra una variedad de formatos de tiendas, de entre las que sobresalen[1, 4]: supermercados (26%), "*home improvement*" (19%) y tiendas por departamentos (15%).

En particular el formato de *Home Improvement*, que tuvo ventas avaluadas en MMUS$ 5800 al año 2006[1] tiene a 3 actores como los más relevantes en cuanto a participación de mercado[1,5]: Homecenter (23%), Easy (7%) y MTS (8%).

El contexto actual de crisis financiera ha afectado de manera importante a la industria del *retail*, la que explicó por sí sola más de la mitad de la caída del 47% en la inversión chilena en el exterior durante el tercer trimestre del año 2008[6].

La industria del *retail* se presenta como una industria altamente competitiva, con una alta penetración de mercado en el país[1] y con gran cantidad de clientes muy diversos tanto en hábitos de compra como en necesidades e intereses.

Es entonces que aparece la necesidad de diferenciación como vital para cada uno de los actores del *retail*, de manera de aparecer como una solución real y única a las necesidades de sus clientes.

Para esto, el manejo de los datos sobre ventas de productos y comportamiento de clientes se torna central para lograr tanto los objetivos de posicionamiento como de efectividad en el manejo de inventarios y rentabilidad del metro cuadrado de espacio en sala.

---

[1] Ver Anexo A "Presentación del retail en Chile"
[2] Censcosud y el grupo Falabella con operaciones en Perú, Colombia, Argentina y Brasil.
[3] 33,5 MMUS$ el año 2007
[4] Porcentajes expresados como parte del total de ventas de la industria
[5] Evaluado en MMUS$ al año 2006.
[6] CCS (Cámara de comercio de Santiago), 2008

1.2 Presentación de la empresa

Easy S.A. es una empresa del rubro de tiendas para el mejoramiento del hogar, perteneciente al grupo Cencosud, importante *retailer* chileno.

Cencosud es un grupo con operaciones en la industria del *retail* que cuenta con operaciones en supermercados, hipermercados, *homecenters*, tiendas por departamento, centros comerciales, desarrollo inmobiliario y servicios financieros. Adicionalmente desarrolla líneas de negocios que complementan sus operaciones principales, estas son: corretaje de seguros, agencia de viajes y centros de entretención familiar.

Dentro de este esquema de negocio, Easy S.A. pertenece a la división de "Home Improvement"[7], y cuenta con 25 locales en Chile, 46 locales en Argentina y 1 local en Colombia.

Easy S.A. nace en el año 1993 con la inauguración de su primera tienda en Argentina. Su expansión internacional continúa al año siguiente con la apertura de su primera tienda en Chile, en el mal Alto Las condes. El siguiente hito en su expansión internacional se dio el año 2008, con la inauguración de la primera tienda en Colombia.

Al año 2008 la empresa sigue su crecimiento con la apertura de 4 nuevas tiendas y la expansión de sus canales de ventas inaugurando la venta por internet, a través de la página www.easy.cl.

En cuanto al estado financiero (1), para el año 2008 la empresa presento ingresos por explotación por $278.788.679.000 y una utilidad operacional de $4.610.622.000.

1.3 Planteamiento del problema y justificación

El creciente nivel de competitividad y la complejidad operacional presente en el área del retail, hacen necesario la implementación de nuevas herramientas que generen ventajas competitivas en el mercado.

En el caso de Easy S.A. se hace indispensable la implementación de mecanismos que le permitan hacer uso eficiente tanto de sus recursos humanos como de sus datos, de manera de lograr coherencia entre las acciones comerciales y los objetivos estratégicos, dentro de los que destacan:

- Posicionamiento efectivo: Para ser congruentes con la imagen de marketing deseada y lograr un posicionamiento efectivo es necesario mantener un mix de productos acorde y actualizado para el público objetivo.

- Rentabilizar el espacio en sala: Un importante indicador es la rentabilidad por metro cuadrado, la forma de optimizar este indicador es actualizando el mix de

[7] "Mejoramiento del hogar" traducido al Castellano

productos de manera constante[8], eliminando los productos que muestren bajo rendimiento y remplazándolos por otros con posibilidades de buen rendimiento o rendimiento comprobado.

- Disminuir inventarios: La mala planificación de demanda genera en algunas ocasiones sobreestimación y en otras subestimación de la demanda generando un mal manejo de los inventarios, lo que acarrea costos de almacenamiento por un lado y la subutilización del espacio por otra.

- Utilización eficiente y eficaz de RRHH[9]: Focalizar la utilización de RRHH en tareas que generen valor agregado.
  Se estima que el costo en RRHH utilizados en realizar tareas relacionadas con la catalogación de productos, pronóstico de demanda y agrupación de tiendas de manera manual asciende a los $4.950.000[8]. Cabe mencionar que estos RRHH no tienen dentro de sus responsabilidades estas tareas.

Como se describe anteriormente el proceso de catalogación, pronóstico de demanda y agrupación de tiendas se hace indispensable en el cumplimiento de las metas comerciales de Easy, tanto en marketing como en el área comercial y financiero.

El actual proceso de manejo de información y realización de pronóstico de demanda, agrupación de tiendas y catalogación presenta las siguientes características que constituyen problema al dificultar la obtención de los objetivos antes mencionados:

- Utilización ineficiente e ineficaz de RRHH: Sobre ocupación de su tiempo: Dada la carga de trabajo, no alcanzan a realizar un trabajo detallado.
- Dificultades y demoras en la obtención de datos.

Se puede agregar además la lentitud del proceso que solo permite realizar el proceso de catalogación dos veces por año, y con una calidad deficiente: existen tiendas que de 20.000 productos catalogados, tiene 10.000 que ya no se venden en Easy[10], lo que recarga el sistema computacional y ocupa espacio físico en tiendas.

Un último corresponde al de la calidad de datos, los que presentan alta volatilidad y tienen integrados problemas como quiebres de Stock, lo que los hacen poco confiables como para ser utilizados en trabajos de de *clustering* predicción de demanda.

Para enfrentar estos problemas, se propone la creación de un sistema de apoyo a la toma de decisiones que:

- Estandarice la recopilación, tratmiento y carga de datos para su psterior uso.
- Permita la aplicación de metodologías de data *mining* para *clustering* y pronósticos de demanda.

[8] Idealmente cada 3 meses, Fuente Claudio Larrea, Category Maneger, departamento de procesos comerciales.
[9] Recursos Humanos
[10] Tomas Zavala, Subgerente de Precios y Surtidos. Para mayor detalle sobre el cálculo referirse a anexo H “Entrevistas”, punto H.2

- Facilite la visualización de datos históricos como de reportes con información relevante a la gerencia de Procesos Comerciales.

1.4 Objetivos

1.3.1 Objetivo General

Construir un Datamart orientado a la determinación de agrupaciones de tiendas y al pronóstico de demanda de productos.

1.3.2 Objetivos Específicos

Respecto al Datamart

1. Estandarización de los procesos de extracción, transformación y carga de datos.
2. Generación de una base de datos con la información obtenida y con una estructura orientada al minado de datos.
3. Generación de indicadores orientados a responder las preguntas de *clusterizacion*.

Respecto a la *clusterización*

1. Generación de una metodología de automatización del proceso de agrupación de tiendas.
2. Incluir la experiencia del experto en negocios.
3. Permitir la fácil identificación de relaciones entre tiendas.
4. Permitir la corroboración del cumplimiento de objetivos comerciales en las tiendas.
5. Disminuir los tiempos empleados en este proceso.

Respecto al pronóstico de demanda

1. Generación de una metodología de automatización del proceso de predicción de demanda en tiendas.
2. Disminuir los tiempos empleados en este proceso.

1.5 Hipótesis

Dada la situación antes expuesta se plantea la siguiente hipótesis:

"La existencia de un sistema computacional de manejo y consolidación de datos, además de la utilización de algoritmos de *clustering* y predicción de demanda, permitirán la disminución de inventarios, facilitará y hará efectivo el proceso de catalogación de productos a tiendas y disminuirá el costo (al menos en horas hombre) de los procesos de catalogación y pronostico de demanda."

1.6 Alcances

El trabajo se centrará en la creación conceptual del sistema de mecanización de procesos y datos para la obtención de los objetivos planteados además de la creación de un prototipo funcional de las propuestas realizadas.

La descripción del prototipo se hará en función de los tiempos de ejecución e implementación y los análisis de los procesos de minado de información serán en términos de las variables tiendas, productos, características socioeconómicas y ventas.

El trabajo comprenderá 23 tiendas de Easy S.A. y los productos involucrados son los pertenecientes a la sección 13, “corral de herramients” lo que corresponde a 129 SKU.

1.7 Contribuciones

El trabajo realizado representa una primera aproximación a la automatización del trabajo de category managment realizado en Easy S.A. entregando herramientas para la utilización de los datos y planteando desafíos futuros para el desarrollo de herramientas y perfeccionamiento de la presentada.

2. MARCO CONCEPTUAL

En el siguiente capítulo se presentan diferentes tópicos relacionados con el desarrollo de la presente memoria, de manera de facilitar su lectura y comprensión.

2.1 Extracción de Conocimiento en Bases de Datos

En su concepción más general, una base de datos (BD) es: "Un conjunto de datos almacenados para su posterior uso"[11]. Bajo este punto de vista, tanto una biblioteca como una discoteca son bases de datos, por lo que se entenderá en particular para el desarrollo del presente trabajo una Base de datos como un conjunto de datos transaccionales almacenados digitalmente para su posterior utilización.

En la actualidad el proceso de acumulación de datos en bases de datos es cada vez más fácil gracias a dos factores principalmente: 1) La automatización de las transacciones con la aparición y el extendido uso de tecnología como códigos de barras, RFID[12] o a través de la web y 2) La disminución en el costo del hardware necesario para almacenar estos datos (2) (3).

La situación anterior ha generado cúmulos de datos de tal magnitud que las capacidades humanas de análisis e interpretación se han visto largamente sobrepasadas, con lo que surge la necesidad de nuevas técnicas y teorías computacionales para hacer frente al reto de analizar, interpretar y extraer información desde grandes volúmenes de datos (2) (3) (4).

2.1.1 Proceso KDD

El proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD[13]) se define como la "extracción no trivial de patrones en los datos válidos, novedosos, potencialmente útiles y entendibles a partir de una base de datos" (4) y aparece como solución a la necesidad de nuevas técnicas y metodologías para el trabajo con grandes bases de datos.

Este proceso KDD se hace cargo y da solución, conceptualmente, a un número de problemas que surgen del constante crecimiento de las BD, tales como (4) (5):

1. Manejo de gran cantidad de datos.
2. Integración de datos de distintas fuentes (en distinto formato).
3. Manejo de datos dañados, errados o faltantes.
4. Encontrar patrones donde resulta muy complejo hacer hipótesis a priori dado el tamaño de la muestra.

La figura 1, esquematiza el proceso de KDD, proceso iterativo e interactivo entre etapas y con el usuario.

[11] es.wikipedia.org/wiki/Base_de_datos
[12] Siglas en ingles para identificación por radio frecuencia
[13] Por sus siglas en Inglés "Knowledge Discovery in Databases"

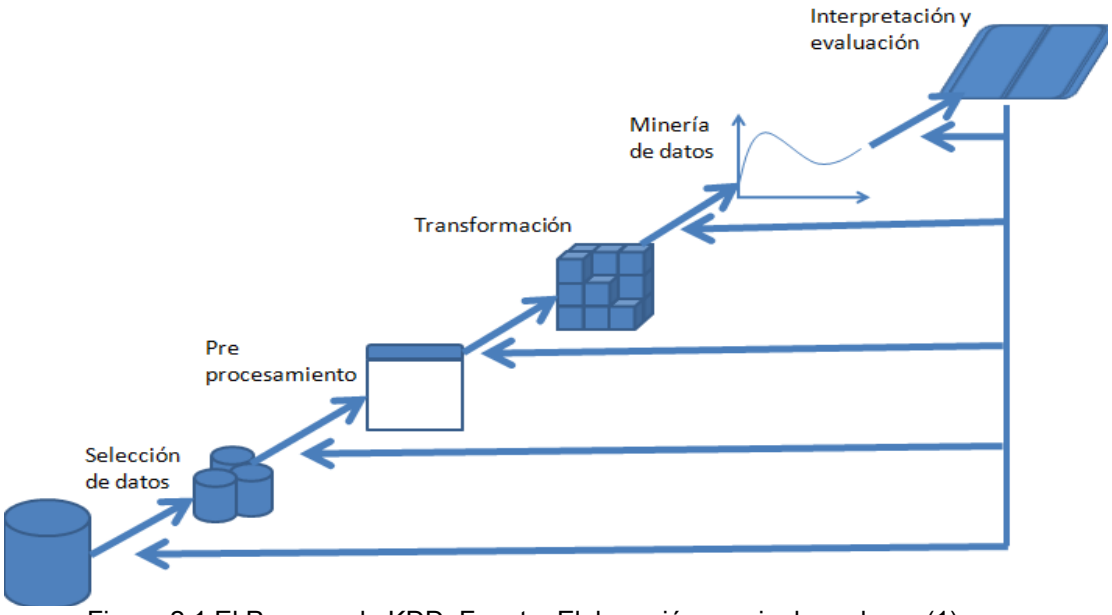


Figura 2.1 El Proceso de KDD. Fuente: Elaboración propia, basado en (1)

Las etapas constitutivas del proceso KDD se explican en la tabla 2.1.

Tabla 2.1: Descripción del proceso KDD

Etapa	Descripción	Entrada al proceso	Salida del proceso
Selección de datos	Determinación de Objetivos	BD completa	Objetivos del proyecto y selección de datos
Pre procesamiento	Limpieza de datos: Detección y manejo de outliers e información faltante	Datos objetivos para el proceso	Datos limpios y con formato común
Transformación	Reducción de datos en datos relevantes al proyecto	Datos limpios y con formato común	BD estructurada según el modelo lógico de datos.
Minería de datos	Aplicación de algoritmos para encontrar patrones en los datos	BD estructurada según modelo	Patrones en los datos
Interpretación y evaluación	Validación de los patrones encontrados por el experto en el negocio y usuario	Patrones en los datos	Patrones evaluados, entendidos y validados
Generación de conocimiento	Documentación del conocimiento adquirido e incorporado al proceso KDD	Patrones validados	Conocimiento acorde con los objetivos del proyecto

Fuente: Elaboración propia, basado en (2) y (3)

De esta forma, el proceso KDD presenta una metodología que permite encontrar información y conocimientos relevantes de manera estructurada, asegurando la calidad de los datos a utilizar, foco sólo en los datos relevantes al problema, clasificación de los métodos para el descubrimiento de patrones y alineamiento de estos con los objetivos del proyecto y finalmente el entendimiento de patrones y su validación en la práctica.

2.2 Data Warehousing y el proceso KDD

Un *data warehouse* es un repositorio de datos que busca responder a necesidades concretas de información del negocio, a tiempo para poder ser útiles a las necesidades comerciales y operativas de la empresa.

La relación entre el proceso el proceso KDD y el data warehouse se tiene de forma natural, ya que la estructura y construcción de un data warehouse entrega herramientas concretas para la implementación de los procesos propuestos en el modelo KDD.

Como se observa en la figura 2.2, la estructura y los procesos incluidos en la formación de un *data warehouse* responden a la necesidad que el proceso KDD presenta en su implementación, específicamente en las etapas de pre-procesamiento y transformación de datos, ofreciendo una representación de estos con sentido para el usuario[14], limpios y consolidados[15] para la aplicación de algoritmos de data mining.

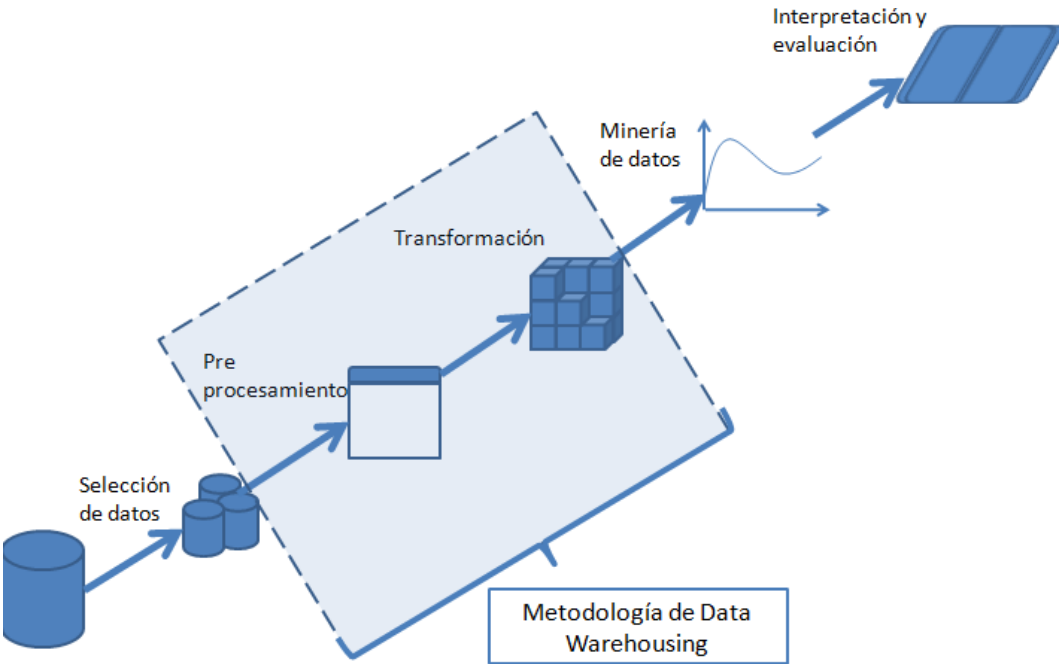


Figura 2.2: D*ata warehouse* y el proceso KDD. Fuente: Elaboración propia, basado en (6)

2.2.2 Características principales de un *data warehouse*

Un data warehouse es el resultado de una arquitectura de datos y procesos, y no una nueva tecnología, la que entrega finalmente como resultado: "Una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis" (6).

La definición anterior se puede complementar con la de Bill Inmon quien caracteriza esta “copia de los datos transaccionales” como (8):

[14] Modelamiento Multidimensiona, referirse al punto 2.2.3.1
[15] Proceso de ETL, referirse al punto 2.2.3.2

Orientado a objetivos: La relación que se establece entre los datos en la base de datos responde a la relación que los datos tienen entre sí en el mundo real, en el contexto del problema.
Integrado: Se reúne la información de todos los sistemas operacionales de la empresa. Los cuales integran un solo repositorio, de forma consistente.
No volátil: Los datos, una vez que entran al *data warehouse*, se transforman en datos de "solo lectura" es decir ya no serán modificado ni eliminados.
Variante en el tiempo: Los cambios en los datos a través del tiempo quedan registrados, de esta forma que los cambios se reflejan en el repositorio y son accesibles.

Si bien un *data warehouse* es solo un repositorio de datos, en su creación están implícitos una serie de procesos destinados a generar datos de calidad.

Estos procesos están bien documentados y son una razón de peso para la implementación de un *data warehouse*, ya que permiten hacerse cargo de los problemas asociados a la construcción de un repositorio de datos que resulte útil:

- Extracción, Transformación y Carga (ETL) de cantidades masivas de datos asegurando la integridad de los datos y la mecanización de este proceso.
- Formar un repositorio de datos con una arquitectura tecnológica y lógica que permita responder a las necesidades del negocio.

Existen una serie de elementos que interactúan en la construcción de un *data warehouse* y que caracterizan a esta arquitectura de información, a continuación se describen algunas de las más relevantes.

2.2.2.1 Proceso ETL

El proceso de extracción transformación y carga de datos (ETL) es la pieza inicial en la construcción del *data warehouse,* crítico en el éxito del mismo (4) (9). Constituye el rescate, procesamiento y limpieza de los datos especificados según requerimiento desde los sistemas operacionales hacia el servidor del *data warehouse*.

El proceso de ETL se realiza en un área aislada del *data warehouse*, accesible sólo por los programas constituyentes del proceso de ETL, denominada *data staging area* (DSA).

El DSA es el lugar de trabajo en la construcción *data warehouse*, y presenta ventajas como el permitir la modificación de datos y tablas sin afectar a los usuarios finales ni a la base de datos multidimensional, además de representar un respaldo de la información (5) (9).

Se estima que este proceso es el más caro en la implementación del *data warehouse*, implicando entre el 60% y un 80% del costo total (10). Esto debido a que es un proceso intensivo en mano de obra y diseño.

Los datos generalmente vienen con errores (tanto de completitud como de de integridad) además de no encontrarse en formatos compatibles por lo que en esta etapa se incurre en las tareas de:

1. Ajuste de Formatos.
2. Cálculos Intermedios.
3. Ajustes de Consistencia.
4. Ajustes de corrección de errores.

Figura 2.3: Proceso ETL. Fuente: Elaboración propia

Extracción: Lectura de datos provenientes desde distintas fuentes, para la carga de estos en el DSA.

Transformación: Este proceso se lleva a cabo en el DSA, y busca realizar modificaciones en los datos para corregir posibles errores en ellos y entregar el formato necesario para la carga de estos al data warehouse.

Carga: Corresponde a la carga final de los datos transformados desde el DSA hacia el *data warehouse*. Se debe considerar el diseño de la base de datos para lograr la arquitectura multidimensional.

2.2.2.2 Modelamiento Multidimensional

Los sistemas de apoyo a la toma de decisiones tienen por objeto entregar información oportuna y precisa para analizar y asistir en la realización de acciones atingentes a una realidad observada.

Dado esto, los sistemas de apoyo a la toma de decisiones necesitan de consultas complejas en las que los datos son manipulados para entregar información (5).

Ante estos requerimientos en el manejo de los datos, al modelamiento multidimensional surge como una opción eficiente de almacenamiento y organización de datos, ya que este tipo de modelamiento propone una representación que emula la visión que el usuario final tiene del negocio, quien ve los datos como la resultante de la interacción de diversas variables.

Por ejemplo, el sólo valor del total de ventas no dice mucho al tomador de decisiones, la información realmente interesante es cuál fue el resultado de las ventas de un producto en particular en una tienda para un período específico.

Conceptualmente el modelamiento multidimensional se materializa en el hipercubo de información, que como se muestra en la figura 2.5, representa el concepto multidimensional descrito en el párrafo anterior.

En lugar de tablas bidimensionales para el almacenamiento de información, el hipercubo propone un ordenamiento de los datos o variables en torno a dimensiones. Una dimensión es entendida como un conjunto de datos categóricos que determinan una variable, por ejemplo producto, cliente, tienda, etc.

De esta forma, cada celda es la resultante de la interacción de las distintas dimensiones, por lo que la información contenida en la celda esta contextualizada por las dimensiones que la determinan.

Además cada miembro de una dimensión esta agrupada en una o varias jerarquías, las que facilitan la agregación y desagregación de los datos dependientes de cada miembro de la dimensión.

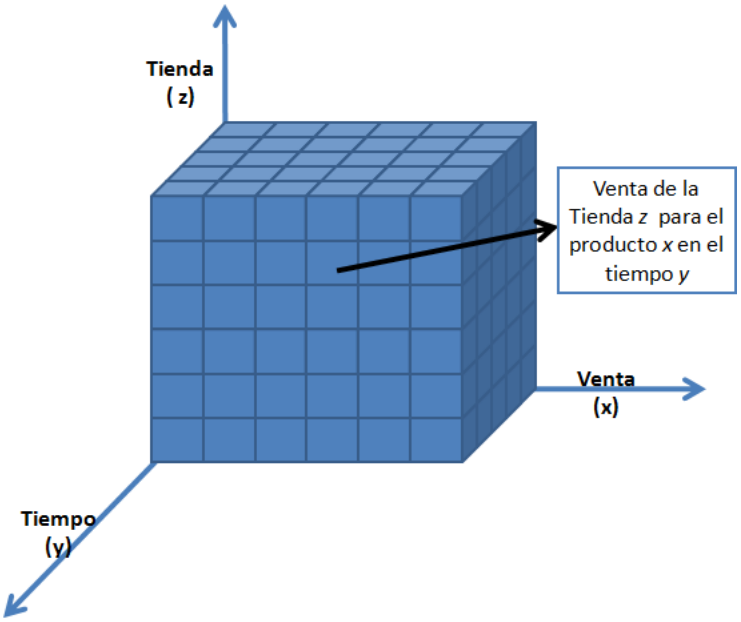


Figura 2.4: Representación de un hipercubo de información. Fuente: Elaboración propia

La realización computacional del hipercubo de información se puede obtener mediante dos técnicas: 1) el cubo de información y 2) El modelo estrella (12).

El cubo de información es la representación directa del hipercubo descrito en la figura 2.5. y debe implementarse con un motor de bases de datos multidimensional, el que a diferencia de los modelos relacionales que estructuran los datos según relaciones y tablas bidimensionales, organiza los datos en base a arreglos multidimensionales (8). Por ejemplo un hipercubo con tres dimensiones se representara de la siguiente manera:

$$\text{cubo: arreglo } (1 \dots \alpha_1, 1 \dots \alpha_2, 1 \dots \alpha_3) \qquad (2.1)$$

Con α_i el numero de atributos de la dimensión i. Entonces el acceso al dato de la venta del producto x, en la tienda z para el período y, es simplemente el acceso a la celda con coordenadas (x,y,z) en el arreglo “cubo”.

Por otra parte el modelo estrella es una representación del hipercubo de información mediante la utilización del modelo entidad relación de datos y bases de datos relacionales.

Bajo esta arquitectura de diseño, cada dimensión se encuentra representada por una tabla bidimensional, y el hipercubo es una tabla, llamada “fact-table”, en la que cada fila es una medida tomada para la intersección de todas las dimensiones que la definen.

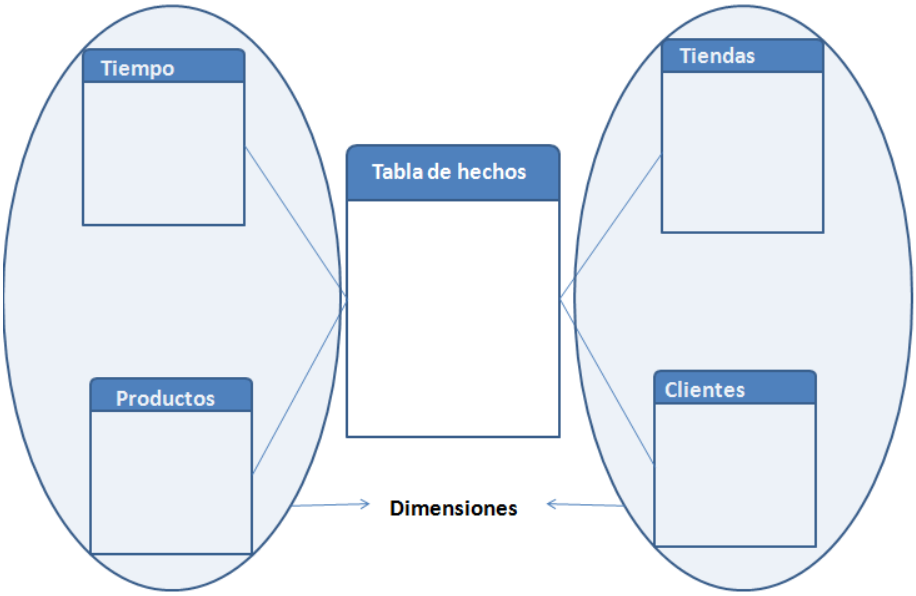


Figura 2.5: Representación del modelo estrella. Fuente: Elaboración propia

El modelo estrella es la técnica más utilizada en la implementación del modelamiento multidimensional de datos (10). Esto debido a las complicaciones que conlleva la implementación de un sistema de manejo multidimensional de bases de datos donde ya existen sistemas de manejo de bases de datos relacionales.

2.2.3 *Data warehouse* y Data Mart

Un proyecto de *data warehouse* es un proceso altamente complejo y largo, entre 12 y 18 meses (9), que aborda de forma holística a la empresa, por lo que su implementación es un asunto muy sensible.

Una solución es la implementación parcelada del *data warehouse*, mediante la puesta en marcha de módulos más pequeños, los *datamarts*.

Un *datamart* es un módulo más pequeño y potencialmente constituyente de un *data warehouse*; el concepto es definido por Ralph Kimball (6) al definir alternativamente un *data warehouse* como la unión de todos los *datamarts* de una entidad.

Un *datamart* presenta las mismas características que un data warehouse, es decir, orientado a objetos, integrado, no volátil y variante en el tiempo y solo se distingue de un *data warehouse* en el alcance de sus objetivos, ya que un *datamart* representa una

solución a una línea de negocio simple o departamento particular, mientras un *data warehouse* incluye a la empresa en su totalidad.

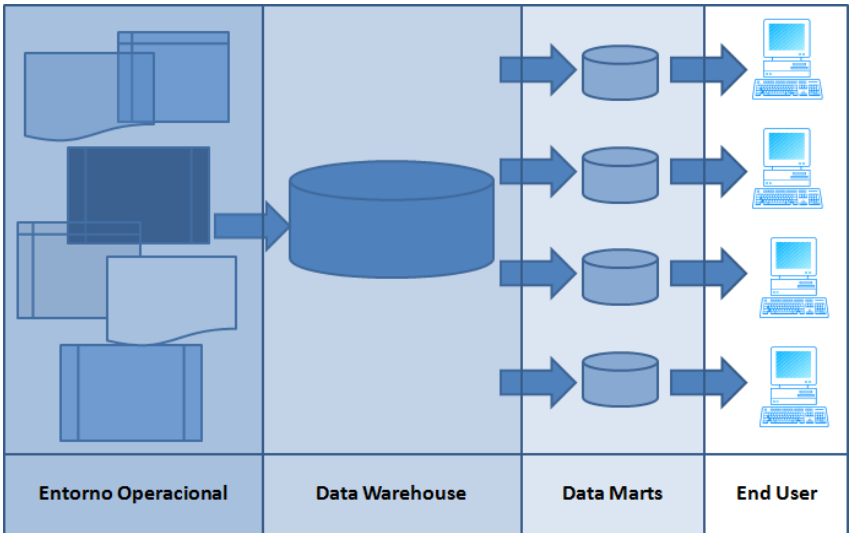


Figura 2.6: *Data warehouse* y Data marts en el entorno operacional. Basado en (5)

A partir de esta definición se desprenden dos enfoques en la implementación de un *data warehouse*, el enfoque *top-down* y el *bottom-up*.

El enfoque *top-down*, planteado por Bill Inmon (8), propone la construcción de un *data warehouse*, que incluya la información de todos los departamentos de la empresa y considere a los potenciales usuarios del sistema, para luego crear *datamarts* orientados a áreas específicas del negocio.

El enfoque bottom-up, planteado por Ralph Kimball (6), aboga por la construcción de *data marts* para cada área de negocio, orientados siempre en la futura conexión entre ellos para la construcción final de un solo *data warehouse* para la empresa.

En la presente memoria, se utilizará el enfoque “bottom – up” de implementación de un *data warehouse*, realizando un *datamart* para el área de operaciones de un *retailer*, orientado a realizar predicciones de consumo de productos y agrupaciones de tiendas de modo optimizar las acciones operacionales.

2.3 Minería de datos

Puede definirse como “El Proceso iterativo de extracción de patrones ocultos desde grandes BD, mediante la utilización de tecnologías de Inteligencia artificial o herramientas estadísticas” (13) o “El proceso de análisis secundario de grandes BD en búsqueda de relaciones insospechadas que signifiquen interés y valor para el dueño de la BD” (14).

De estas definiciones se extraen las dos principales áreas del conocimiento que entran en el terreno de la minería de datos (3): 1) La Inteligencia artificial (específicamente el Machine Learning) y 2) La estadística.

Machine Learning (13): “Rama de la Inteligencia artificial que se encarga del diseño y aplicación de algoritmos de aprendizaje”.

Estadística (15):”Metodología para la extracción de información desde los datos y expresar la cantidad de incerteza en las decisiones que tomamos”.

Es bajo estas áreas de estudio que se enfocan los principales problemas abordados por la minería de datos hoy en día, de entre los que podemos mencionar (3):

1. Búsqueda de lo inesperado por descripción de la realidad multivariante: A diferencia de la estadística clásica el mayor poder de procesamiento de datos hace que el principio de parsimonia ya no sea necesario, es decir, mientras más variables mejor. Lo anterior permite la inclusión de no linealidad en la descripción de observaciones. Para esto es fundamental el uso de grandes BD, pues el muestreo aleatorio alcanza para describir los sucesos regulares, pero no para predecir lo irregular.
2. Búsqueda de Asociaciones: Búsqueda de relaciones entre suceso. Por ejemplo: ¿Están relacionados la compra de pañales con la de cerveza en un supermercado?[16], ¿Determinados sucesos ocurren simultáneamente más de lo que se esperaría si fuesen independientes?
3. Definición de tipologías: Detección de grupos con comportamientos similares según alguna variable preponderante o criterio.
4. Predicciones: Estimación del valor de una variable en función de datos anteriores u otros datos disponibles que la puedan afectar.

2.3.1 Herramientas para la minería de datos

Para abordar los problemas antes mencionados existen múltiples modelos algorítmicos. Estos modelos pueden agruparse en dos grandes grupos: Modelos de aprendizaje supervisado y modelos de aprendizaje no supervisado.

Los modelos de aprendizaje supervisado, también conocidos como modelos predictivos, son utilizados en el desarrollo de problemas en los que el resultado final es conocido, y su objetivo es la predicción de un valor objetivo.

Dentro de los problemas abordados con modelos supervisados de minería de datos se pueden contar (3) (16) (17) (12) (18):

- Clasificación de observaciones: Se refiere a la identificación de observaciones dentro de clases definidas en función de ciertas variables y a la generación de modelos que permitan la clasificación de nuevas observaciones dentro de las clases observadas a partir de información histórica.

[16] Es un clásico ejemplo en la minería de datos, la respuesta es sí en una cadena de retail llamada Osco se encontró que entre las 5 y 7 pm los días domingos las compras de cerveza y pañales se encontraba altamente relacionada la explicación es que los padres que eran mandados a comprar pañales aprovechaban de comprar cerveza también.

Esta técnica es utilizada en clasificación de clientes, análisis crediticio, análisis de riesgo, etc.

- Predicciones: Se refiere a la determinación del valor de una variable, a través de un modelo congruente con la información pasada de esa observación.
  Esta técnica es utilizada en problemas de predicción de ventas para producto, condiciones climáticas, etc.

Los modelos de aprendizaje no supervisados, también conocidos como modelos predictivos, se utilizan con un objetivo exploratorio de los datos, cuando no se tiene idea del resultado final al que se llegará y la idea es extraer directamente de los datos relaciones y patrones que entreguen información para su posterior uso.

Dentro de los problemas abordados con modelos no supervisados de minería de datos se pueden contar (3) (16) (17) (12) (18):

- Clustering: Se refiere al problema de identificar grupos entre variables de un set de observaciones, esto sin un conocimiento previo de los grupos a obtener ni de las variables que resultarán relevantes.
  Estos modelos son utilizados en identificación de comportamientos de clientes, identificación de observaciones con comportamientos similares, etc.
- Reglas de asociación: Se refiere a la identificación de causalidades en el ocurrencia de eventos y busca identificar la probabilidad ocurrencia de observaciones dada la ocurrencia de un fenómeno no relacionado necesariamente a priori.
  Estos modelos son utilizados en la identificación de canastas de compra, diseños de catálogos, etc.

- **Técnicas aplicadas a la minería de datos.**

Existen diversas técnicas utilizadas para la realización de los modelos descritos anteriormente. A continuación mencionan algunas de las técnicas utilizadas.

- Clasificación de observaciones: Árboles de decisión, redes bayesianas, *Support vector machines*, etc.
- Predicciones: Redes neuronales artificiales, métodos univariantes de pronóstico, métodos multivariantes de pronóstico, etc.
- Clustering: K-means, K-medoids, *Self Organizing feature Maps*, Análisis discrimante, etc.
- Reglas de asociación: Redes bayesianas, arboles de decisión, etc.

En el desarrollo del presente trabajo se utilizarán los modelos de pronóstico de variables y clustering, en particular Rede Neuronales Artificiales y la metodología de Box Jenkins para predicciones y las técnicas de K-Means y Self Organizing Feature Maps para el clustering, las que serán explicadas en mayor detalle en los puntos 2.3.1 y 2.3.2.

2.3.2 Herramientas de minería de datos aplicadas al pronóstico de variables

La predicción se presenta como una necesidad de primera orden dada la constante evolución del mundo. La predicción se puede entender como la ciencia o el arte de estimar el valor de una variable de manera anticipada a su realización.

En el contexto científico esta tarea se realiza mediante la utilización de los datos históricos, utilizando como hipótesis el que estos datos son el resultado de la interrelación de los factores causales (sean estos conocidos o no) y que esta relación debiera mantenerse en el futuro.

Factores a considerar en el pronóstico son (16): 1) El objetivo del pronóstico, 2) el grado de exactitud deseado, 3) el intervalo de tiempo para el cual se quieren realizar pronósticos y 4) la calidad y disponibilidad de los datos históricos.

Los pronósticos son realizados mediante la utilización de series de tiempos, las que se entienden como un conjunto de observaciones ordenadas en el tiempo y que describen la evolución de un evento en particular.

En el punto 2.3.2.1 se presentará la técnica de redes neuronales artificiales correspondiente al *machine learning*, aplicada al pronóstico de series de tiempo, en el punto 2.3.2.2 se explicará la metodología estadística de Box Jenkins para el pronóstico de series de tiempo y en el punto 2.3.2.3 se presentará la metodología de medias móviles para el pronóstico de series de tiempo.

2.3.2.1 Redes Neuronales Artificiales aplicadas al pronóstico de series de tiempo

Como se explicó en el punto 2.3.1, una red neuronal artificial es una arquitectura que busca replicar el funcionamiento de una red neuronal biológica, como la cerebral, mediante la construcción de una red de neuronas organizada en capas (3).

La figura 2.7 ejemplifica la estructuración de la red, en ella se distinguen tres capas de neuronas:

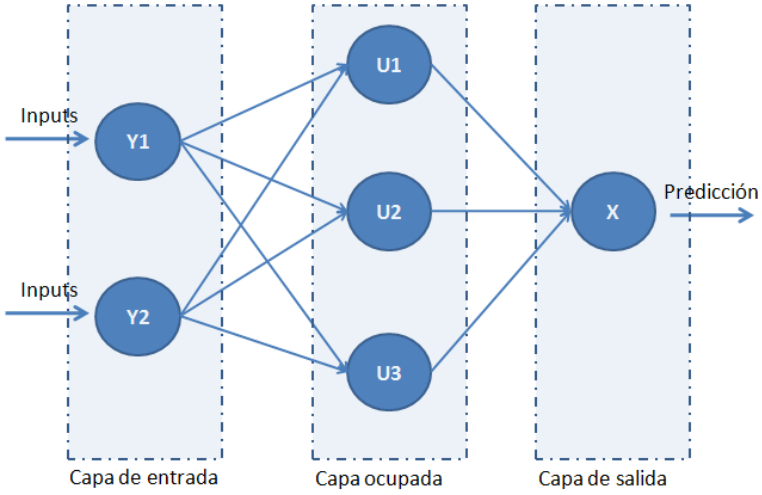


Figura 2.7: arquitectura de un red neuronal artificial. Fuente: (14)

1) Neuronas de la capa de entrada: Actúan como receptoras de los estímulos externos, propagando estos hacia la capa oculta de neuronas.
2) Neuronas de la capa oculta: Realizan un procesamiento no lineal de los estímulos recibidos. Estos patrones modificados son luego entregados a la capa de salida
3) Neuronas de la capa de salida: Resumen los impulsos recibidos desde la capa oculta de neuronas entregando finalmente las respuestas de la red.

Propagación de las variables de entrada

Cada nodo en una red neuronal artificial es una estructura de procesamiento de información que implementa una función que se encarga de computar un valor para la propagación de la información que le llega desde la red hacia los nodos con los que se encuentra conectado.

Considérese una red neuronal artificial de *C* capas, n_c neuronas en la capa *C* y *wijc* el peso de la conexión entre la neurona i de la capa *C-1* y la *j* de la capa *C*. El valor de activación de la neurona *i* de la capa *C* será (20):

$$a_i^c = f\left(\sum_{1}^{j=n_{c-1}} \left(w_{ij}^{c-1} \times a_j^{c-1} + b^c\right)\right) \quad para\ i \in (1, n_c) \qquad (2.2)$$

Donde la función f es llamada de función de activación y se utiliza para controlar la salida de valores demasiado grandes que puedan retrasar la convergencia del método. Esta función varía de problema en problema y no existe una metodología estricta para determinarla, sin embargo existen funciones que son comúnmente utilizadas, las que son presentadas en la tabla 2.2.

Tabla 2.2: Descripción de algunas funciones de transferencia

Nombre	Función	Característica
Limitador Fuerte	$a = 0\ n < 0$ $a = 1\ n \geq 0$	Respuesta binaria 0 o 1 ante cualquier estímulo
Limitador Fuerte Simétrico	$a = -1\ n < 0$ $a = +1\ n \geq 0$	Respuesta binaria -1 o 1 ante cualquier estímulo
Lineal Positiva	$a = 0\ n < 0$ $a = n\ 0 \leq n$	Respuesta lineal positiva para cualquier estímulo
Lineal	$a = n$	Devuelve el valor que entra
Lineal Saturado	$a = 0\ n < 0$ $a = n\ 0 \leq n \leq 1$ $a = 1\ n > 1$	Respuesta lineal entre 0 y 1, 0 para valores negativos y 1 para valores sobre 1

Fuente: Elaboración propia, basado en (9) (15) (14)

Algoritmo de retro programación del error

El mecanismo de aprendizaje supervisado de la red neuronal artificial es el de "*Back Propagation*". Se implementa mediante el cálculo del error de la predicción con respecto al valor deseado, este error determina una regla de corrección, la que se aplica modificando los pesos de interconexión entre neuronas (W). Se denomina de retro programación porque la modificación nace al final de la red, en la capa de salida y luego se transmite hacia el resto de la red, hacia atrás.

Luego el problema que resuelve la red es el de minimización del error de predicción (12) (20):

$$Min_W E(W, B) \qquad (2.3)$$

Donde W es la matriz de pesos de la red y B el vector de umbrales (thresholds) de las capas.

El error de la red neuronal artificial se define como el promedio del error cometido en cada predicción realizada, es decir, como:

$$E = \frac{1}{N}\sum_{n=1}^{N} e(n) \qquad (2.4)$$

Donde e(n) es el error en la predicción de la n-ésima observación. El cual se puede expresar como:

$$e(n) = h((\overline{y_n} - y_n)^2) \qquad (2.5)$$

Donde es la predicción de la red para la observación n, es el valor deseado y h() es una función del error cuadrático.

Luego el nuevo valor de los pesos W de la red se define como:

$$W_t = W_{t-1} - \mu\frac{dE}{dW_{t-1}} \quad t \in (1, T^*) \quad (2.6)$$

Con Wt el valor de la matriz de pesos en la iteración t y μ la tasa de aprendizaje, que determina la magnitud del movimiento hacia el mínimo en la superficie de error y T* la ultima iteración donde se cumpla la condición de detención de la iteración.

Esta condición es que la diferencia entre el valor de $W_t - W_{t-1} \approx 0$

2.3.2.2 Metodología de Box Jenkins para el pronóstico de series de tiempo

En la figura 2.8 se presenta una clasificación de los métodos estadísticos para el pronóstico de series de tiempo.

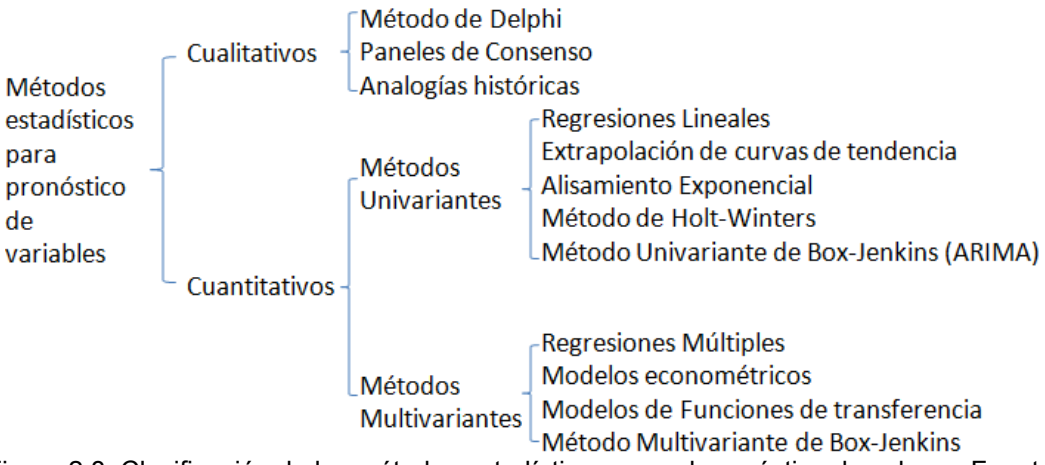


Figura 2.8: Clasificación de los métodos estadísticos para el pronóstico de valores. Fuente: (13)

Dentro de las metodologías estadísticas para el pronóstico de series de tiempo la metodología de Box Jenkins se sitúa como un método cuantitativo univariante[17] de pronóstico de series de tiempo. Es decir utiliza valores pasados de la variable a pronosticar para realiza pronósticos.

Esta metodología permite la búsqueda del mejor modelo ARIMA para una serie de tiempo dada.

La elección de esta metodología se justifica pues: 1) Al ser un método univariante, facilita la aplicación a un gran número de series, esto porque resulta más sencillo ajustar un modelo con solo un regresor que uno con más de uno y 2) existe evidencia de que para el pronósticos de corto y mediano plazo el método ARIMA se muestra superior al resto de los métodos estadísticos univariante.

La metodología de Box-Jenkins fue desarrollada en los años 70 por George Box y Gwilym Jenkins y propone una serie de procesos para la determinación de un modelo de predicción univariante para una serie de tiempo particular.

Si bien esta metodología fue propuesta hace cerca de 35 años, no ha sido sino hasta hoy, cuando los adelantos en tecnología han permitido que su uso se extienda. Numerosos son los trabajos donde se puede comprobar la efectividad del método (2) (16).

Esta metodología propone un trabajo iterativo en la búsqueda de un modelo que sea bien comportado para el pronóstico de la serie. Esto es, que loserrores entre la predicción y los datos reales sean arbitrariamente reducidos y se distribuyen de manera aleatoria e independiente.

La estructura de desarrollo de la metodología se explica a continuación, en la figura 2.9.

[17] Ver Anexo B “Estadística”

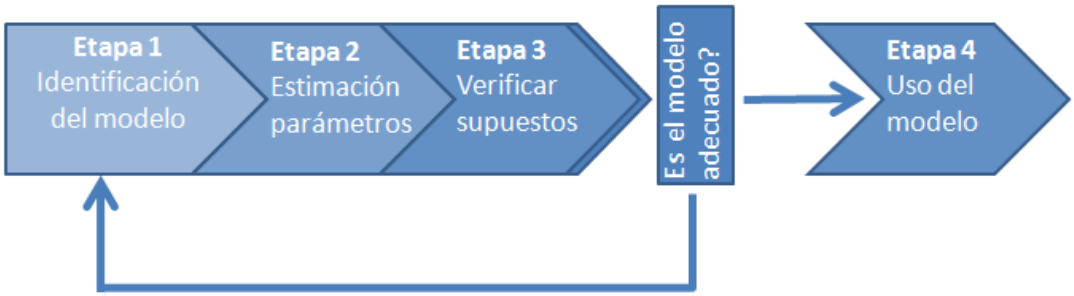


Figura 2.9: Esquema de la metodología de Box-Jenkins. Fuente: (15)

Las series con las que se trabaja son series estacionarias en covarianza lo que involucra tres conceptos (22) (23):

1. Oscilan en torno a un nivel constante.
2. Presentan baja volatilidad de los datos en torno a este nivel.
3. Los patrones de correlación de movimientos de la serie con su pasado no dependen del momento en el que se hagan las observaciones.

Luego de encontrado un modelo factible, se corrobora que este:

1. Presente errores normales, independientes e idénticamente distribuidos.
2. Que los regresores del modelo sean estadísticamente significativos.

Una vez cumplidos estos supuestos, se llega a la presentación de un modelo de ARIMA válido para la predicción de la serie de tiempo.

Modelos utilizados en la metodología de Box Jenkins (16) (23)

La metodología de Box.Jenkins propone la utilización de 2 modelos matemáticos de predicción: Modelo Auto Regresivo (AR) y Modelo de promedios móviles (MA), los que son mezclados para la generación de los modelos más complejos: Modelo Auto regresivo de promedios móviles (ARMA), Modelos Auto regresivos integrado (ARIMA) y Modelos Auto regresivos integrados con estacionalidad (SARIMA).

Los modelos autoregresivos proponen que la evolución de la serie es función de los valores pasados (rezagos) de la misma serie, por otra parte los modelos de medias móviles postulan la existencia de una relación entre el valor presente del valor de la serie con los valores de error cometidos en la predicción para períodos pasados[18].

Con los modelos AR y MA se genera un modelo ARMA, que es una combinación lineal de ambos modelos, AR y MA, incorporando tanto valores de rezagos de la serie como de errores en la predicción para períodos pasados.

A continuación se describe este modelo.

[18]Para mayor detalle en los modelos AR y MA referirse al anexo B “Estadística”

a) Modelo Autorregresivo de Promedios móviles (ARMA): En este caso el modelo propuesto es una combinación lineal de los modelos AR y MA.

$$y_t = \phi_1 * y_{t-1} + \phi_2 * y_{t-2} + \cdots + \phi_p * y_{t-p} + \varepsilon_t - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - \cdots - \theta_q * \varepsilon_{t-q} \qquad (2.7)$$

Donde:

- p el número de observaciones pasadas (rezagos utilizados)
- ϕ_i un conjunto de regresores de las observaciones pasadas
- y_t el valor del pronóstico
- q el numero de errores pasados
- Θ_i un conjunto de regresores
- ε_t el error en la predicción para la observación t

O mediante el uso del operador rezago "B":

$$\Phi_p(B) * y_t = \varepsilon_t * \Theta_q(B) \qquad (2.8)$$

Como se menciono inicialmente, estos modelos son bien comportados[19] para series estacionarias. Una serie puede no ser estacionaria por tres motivos (16) (23):

1. Varianza no constante
2. Existencia de tendencia
3. Existencia de estacionalidad

SI la serie no es estacionaria se le aplican transformaciones destinadas a convertirla en una serie estacionaria en covarianza[20].

La aplicación de la transformación de Box-Cox puede lograr series con varianza constante y la sucesiva diferenciación de los valores de la serie puede eliminar la tendencia y estacionalidades presentes.

Para suplir las dificultades de series que presenten estacionalidades y/o tendencias existen los métodos ARIMA y SARIMA.

Ambos métodos incluyen un termino de diferencia "d", el que internaliza la diferencias necesarias para quitar la tendencia (ARIMA) y la estacionalidad (SARIMA).

Diferenciar una serie es el proceso de transformación de una serie en otra mediante la resta consecutiva de sus datos.

[19] Punto 2.3.2.2 para la definición de bien comportados
[20] Por ejemplo transformaciones logarítmicas

b) Modelo Autoregresivos integrado (ARIMA): Este modelo resuelve problemas de tendencia que hagan no estacionaria[21] a una serie en particular.

La metodología consiste en diferenciar la serie hasta eliminar este elemento de tendencia.

El modelo ARIMA introduce este término de diferenciabilidad, "d", que representa el número de veces que la serie debe diferenciarse para eliminar el efecto de tendencia; mientras integra los términos p y q correspondientes a las modelos AR(p) y MA(q).

Finalmente el modelo ARIMA(p,d,q) se expresa de la siguiente forma reducida:

$$\Phi_p(B) * (1-B)^d * y_t = \varepsilon_t * \Theta_q(B) \qquad (2.9)$$

c) Modelo Autoregresivos integrado con estacionalidad (SARIMA): Este modelo resuelve problemas de estacionalidad que hagan no estacionaria a una serie[22].

La metodología consiste en diferenciar la serie hasta eliminar este elemento de estacionalidad.

El modelo SARIMA introduce este término de diferenciabilidad estacional, "D", que representa el número de veces que la serie debe diferenciarse para eliminar el efecto de estacionalidad; mientras integra los términos p y q correspondientes a las modelos AR(p) y MA(q).

Finalmente el modelo SARIMA(P,D,Q) se expresa de la siguiente forma reducida:

$$\Phi_P(B^s) * (1-B^s)^D * y_t = \varepsilon_t * \Theta_Q(B^s) \qquad (2.10)$$

Finalmente la forma generalizada del Modelo ARIMA, que incluye el término de estacionalidad y tendencia que expresado como ARIMA(p,d,q)XSARIMA(P,D,Q):

$$\Phi_P(B^s) * (1-B^s)^d \Phi_p(B) * (1-B)^d * y_t = \varepsilon_t * \Theta_q(B) * \Theta_Q(B^s) \qquad (2.11)$$

Donde:

- $\Phi_P(B^s)$ Modelo AR con estacionalidad.
- $\Phi_p(B)$ Modelo AR con tendencia.
- $\Theta_q(B)$ Modelo MA con tendencia.
- $\Theta_Q(B^s)$ Modelo MA con estacionalidad.
- y_t el valor del pronóstico.

[21] La tendencia produce que el promedio de la serie no sea constante
[22] La estacionalidad produce que el promedio de la serie no sea constante

- p el número de observaciones pasadas (rezagos utilizados).
- q el número de errores pasados.
- P el número de observaciones pasadas del modelo con estacionalidad.
- Q el número de errores pasados del modelo con estacionalidad.
- ε_t el error en la predicción para la observación t.
- S el período de estacionalidad.
- d orden de la diferencia en el modelo con tendencia.
- D orden de la diferencia en el modelo con estacionalidad.

2.3.2.3 Metodología de medias móviles para la predicción de series de tiempo

Esta es una de las metodologías básicas en la predicción de series de tiempo.

Dada una serie de tiempo, el valor del período *t+1* corresponderá al promedio de los *n* valores anteriores. Luego, el valor de la serie para el período *t+2* corresponderá al promedio de los *n* valores anteriores, incluyendo a *t+1*.

2.3.2.4 Validación de resultados del pronóstico

La validación de los resultados obtenidos por los dos métodos de predicción son comparados en base a un número de indicadores de desviación de los valores esperados, estos indicadores son (21) (20):

a) Error Absoluto medio (MAE): Mide el error absoluta en la predicción. Esto considerando el promedio del valor absoluto entre el valor predicho y el valor real

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \overline{y_i}| \quad (2.12)$$

Donde:

- $\overline{y_n}$ es la predicción
- y_t el valor observado
- N es el número de observaciones

b) Error Cuadrático medio (MSE): Mide el error al cuadrado entre el valor pronosticado y el observado.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2 \quad (2.13)$$

Donde:

- $\overline{y_n}$ es la predicción
- y_t el valor observado
- N es el número de observaciones

En la presente memoria estos indicadores serán modificados levemente para hacer sentido en el contexto del negocio.

Dado que para la gerencia comercial y de procesos de negocios los valores de pronóstico a corto plazo revisten mayor importancia que los valores pronosticados a mediano plazo, los errores en pronóstico realizados para observaciones a corto plazo serán penalizados.

Esta penalización se introducirá multiplicando los errores cometidos por un ponderador entre 0 y 1.

De esta forma los errores cometidos en pronósticos cercanos en el tiempo tendrán un ponderador mayor que aquellos errores cometidos en el pronóstico de observaciones lejanas en el tiempo[23].

2.3.3 Herramientas de minería de datos aplicadas al clustering

"Clustering" es un término utilizado para caracterizar a un conjunto de técnicas que busca agrupar conjuntos de observaciones en subconjuntos o "clusters", de forma que las observaciones dentro de un cluster sean lo más parecido entre si y a la vez lo más disímiles a las observaciones fuera del cluster al que pertenecen.

Estas técnicas son de gran utilidad pues reflejan conductas observadas en grupos de observaciones con variables características.

Independientemente del método técnico utilizado para la realización de la clusterización existen pasos estándares, necesarios para llevar a cabo la realización de un análisis de conglomerados, estos pasos son (18):

1. Seleccionar variables de segmentación: Esto es, definir que bajo que características se agruparan o diferenciarán las observaciones. Estas variables deberán estar alineadas con el objetivo de la segmentación.

2. Seleccionar técnica a aplicar para la segmentación: La elección del método y la tecnología para la realización del clústering. Existen principalmente dos tipos de clustering: jerárquicos y no jerárquicos.

3. Seleccionar medidas de similitud: Corresponde al criterio para establecer la similitud o "cercanía" de un par de observaciones. Existen diversos métodos para fijar este criterio y las variables a considerar para la elección de uno u otro son: Tipo de dato y objetivo del clustering.

4. Decidir número de grupos: No existe una metodología específica para decidir esto, sin embargo dentro de los criterios utilizados para obtener el número de clusters se encuentran: Consideraciones estratégicas o comerciales, Tamaños

[23] El valor de estos ponderadores será obtenido mediante experimentos en el capítulo 5

de los grupos a formar, Detección de puntos de cambio significativos en la diferenciación entre clusters.

5. Interpretar y evaluar el perfil de los conglomerados: Una vez realizada la clusterización es necesario corroborar que los resultados obtenidos entregan la información requerida, si este paso no se logra se tendrá que volver a realizar el proceso.

Las técnicas de clustering se pueden dividir en dos grupos (17) (18):1) técnicas jerárquicas y 2) técnicas no jerárquicas.

a) Técnicas de clustering jerárquicas: Consiste en la generación de agrupaciones mediante agregaciones o desagregaciones entre categorías de observaciones. De esta forma en el nivel jerárquico más alto, cada observación es una clase[24] y en el nivel jerárquico más bajo solo existe un cluster[25] que contiene a todas las observaciones.

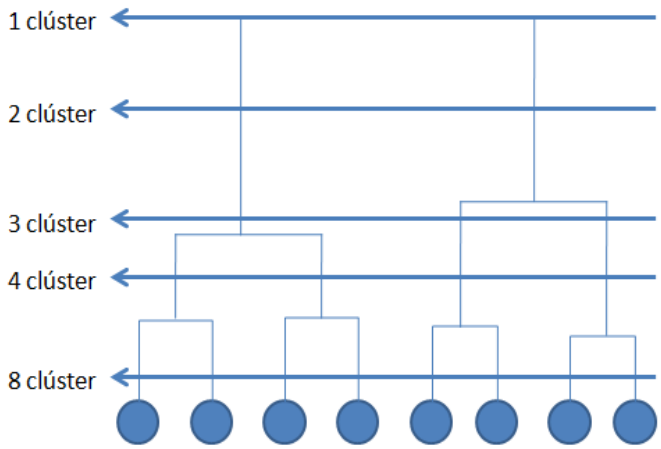


Figura 2.10: Dendograma del método de clustering jerárquico. Fuente: (12)

Existen dos metodologías de clustering jerárquico: aglomerativos y divisivos.

Los métodos aglomerativos realizan agrupaciones uniendo las clases más próximas entre sí del nivel jerárquico anterior. Por otra parte en los métodos divisivos un conjunto de clusters aparece de la división de cluster del nivel jerárquico inmediatamente siguiente.

b) Técnicas de clustering no jerárquicas: Se asigna cada observación al cluster con el que presente mayor similitud. La similitud de una observación con un cluster se realiza mediante el cálculo de la distancia entre la observación y un elemento representativo del cluster. Una vez que se han asignado todos las observaciones se vuelve a calcular el elemento representante del *cluster* y se itera, volviendo a asignar las observaciones al *cluster* con el que presenten mayor similitud.

[24] Existen tantos *clusters* como observaciones
[25] Las observaciones correspondes a un solo *cluster*

Se pueden distinguir dos tipos de técnicas de *clustering* no jerárquico: métodos discretos y los métodos difusos.

Los métodos discretos asignan una asignación única de pertenencia de una observación a un *cluster*, es decir, probabilidad de pertenecía a un *cluster* 1. Los métodos difusos por su parte entregan una probabilidad de pertenencia a un *cluster*, típicamente no igual a 1.

A modo de resumen se presenta la figura 2.11, la que ilustra la clasificación de los distintos métodos de *clustering*.

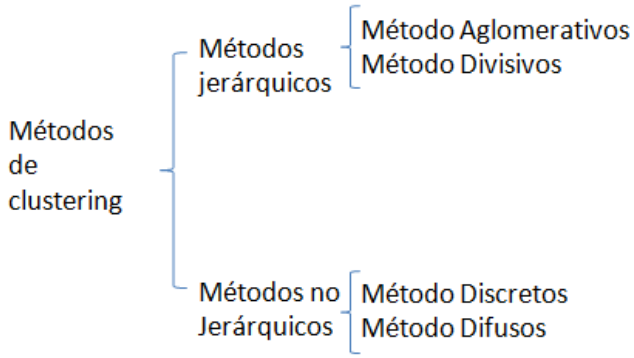


Figura 2.11: Clasificación de los métodos de *clustering.* Fuente: (13)

Cabe mencionar que desde el punto de vista computacional, la principal diferencia entre los métodos no jerárquicos y jerárquicos de *clustering* es que los primeros, permiten trabajar con una mayor cantidad de datos, debido a que no es necesario guardar la información de una iteración a otra, mientras que en los métodos jerárquicos es necesario almacenar la información de distancias entre cada observación, además del valor de *cluster* asignado para cada nivel jerárquico (18).

Medidas de similitud

Como ha sido mencionando en el punto anterior, toda metodología de *clustering* utiliza una medida de similitud o disimilitud entre observaciones.

Existe una amplia cantidad de métricas para medir distancias y no se puede decir que exista una mejor que otra, sólo alguna más adecuada que otras para cada caso ya que la medida deberá estar alineada con los objetivos de la segmentación.

A continuación se presentan algunos ejemplos típicos de medidas utilizadas para la realización de clustering.

a) Medidas de similitud continuas (17) (18) (24): Medida de distancia entre observaciones con atributos numéricos. Considérese 2 observaciones, *"i"* y *"j"*, cada uno con N atributos.

1. Distancia Euclidiana: Es la distancia común que separa a un par de datos en un plano euclideano y es determinada por el teorema de Pitágoras.

$$d(i,j) = \sqrt{\sum_{n=1}^{N} \left(x_i^n - x_j^n\right)^2} \qquad (2.14)$$

2. Distancia Manhattan: La distancia entre dos puntos es la suma del valor absoluto de la diferencia entre sus coordenadas.

$$d(i,j) = \sqrt{\sum_{n=1}^{N} \left|x_i^n - x_j^n\right|} \qquad (2.15)$$

En este caso, la distancia es medida como la suma de la proyección de la distancia euclideana sobre cada dimensión.

3. Distancia Chevychev: La distancia entre dos puntos se representa por la máxima diferencia observada entre sus coordenadas.

$$d(i,j) = \max_n \left|x_i^n - x_j^n\right| \qquad (2.16)$$

b) Medidas de similitud discontinua (17) (24): Medida de distancia entre observaciones con atributos nominales no continuos (por ejemplo: verde =1, rojo=2, azul=3).

1. Distancia Jaccard Es utilizada para medir disimilitud entre sets de datos. Considérese 2 sets de datos, A y B.

La distancia entre A y B será el cociente entre la intersección entre A y B y la unión de A y B.

$$d(i,j) = \frac{|A \cup B| - |A \cap B|}{A \cup B} \qquad (2.17)$$

2. Distancia Simple: Mide que tan iguales son dos vectores. Considérese 2 vectores de observaciones, "i" y "j", cada uno con N atributos con valores en nominales no continuos entre 0 y 1.

$$d(i,j) = \frac{a + c}{a + b + c + d} \qquad (2.18)$$

Donde a = # atributos con valor 1 tanto en "i" como en "j".
b = # atributos con valor 1 en "i" y 0 en "j".
c = # atributos con valor 0 en "i" y 1 en "j".
d = # atributos con valor 0 tanto en "i" como en "j".

2.3.3.1 K-means

Es un método no jerárquico de *clusterización* que utiliza como input, N observaciones con A atributos cada uno y un número inicial de *clusters* K entregados por el usuario. El resultado es una solución (*C*) que especifica la asignación de cada observación a solo uno de los K *clusters* .

El esquema general del algoritmo es el siguiente (17) (24) (18):

1. Se tienen N observaciones y K *clusters*.
2. Iteración 0: Se eligen arbitrariamente centroides para estos K *clusters*.
3. Se asigna cada observación al *cluster* con cuyo centroide presente mayor similitud.
4. Asignados todas las observaciones a algún *cluste*r, se calculan los nuevos centroides de cada *cluster*.
5. Se calcula la variación en distancia del nuevo centroide encontrado con el centroide de la iteración anterior.
6. Si la variación no es significativa: Se detiene el proceso.
   Si se ha alcanzado el número máximo de iteraciones: Se detiene el proceso.
7. Si no se cumple ninguno de los criterios anteriores se vuelve al punto 2.

Dada una medida de distancia entre un par de puntos:

$$\mathrm{d}(\mathrm{x_i}\,,\mathrm{x_i}) = \mathrm{f}((\mathrm{x_i} - \mathrm{x_i})^2) \qquad (2.19)$$

Se mide el grado de dispersión de los puntos en un cluster como la suma de la diferencia entre un punto del cluster y el centroide del mismo:

$$\mathrm{D(k)} = \sum_{\mathrm{i=2}}^{\mathrm{N_k}} \mathrm{d}(\mathrm{x_i}\,,\mathrm{x_{\acute{k}}}) \qquad \text{para el cluster k} \qquad (2.20)$$

Donde representa al centroide del cluster k dada la solución C.

Se define una dispersión intra-grupos como la agregación de las dispersiones observadas en un *cluster*, dada una solución de *clusterización* C:

$$DIG(C) = \sum_{k=1}^{K} D(k) \qquad (2.21)$$

2.3.3.2 Self organizing feature maps (17) (18) (24) (25)

El SOM[26] es un proceso de cuantización, en donde se busca la representación gráfica de observaciones altamente dimensionales en un plano bidimensional.

[26] Por sus siglas en inglés, (S)elf (O)rganizing (F)eatures (M)aps

Para esto se cuenta con una malla bidimensional de N neuronas, organizadas rectangular o hexagonalmente, donde cada neurona es una representación vectorial, el vector prototipo (Xn), en el espacio de dimensión determinado por las observaciones de entrenamiento cuyas componentes son los pesos sinápticos de la red.

Las M observaciones son presentadas una a una a la red de forma de, mediante una métrica, determinar el vector prototipo más cercano. La neurona más parecida a la observación se denomina la "Best-Matching Unit" (BMU) y se determina como:

$$\mathrm{BMU_i} = \underset{\bar{x}}{\operatorname{argmin}} \|x_i - x_c\| \quad \forall\, i \in (1, M) \quad (2.22)$$

Con $\bar{x}$ el conjunto de todos los vectores prototipos.

Las neuronas vecinas de la BMU son luego activadas para aprender el ejemplo i. Esto se realiza mediante la siguiente regla de actualización de los pesos de la red es la siguiente:

$$x_n(t+1) = x_n(t) + \alpha(t) h_{ci}(t)[x_i - x_n(t)] \quad (2.23)$$

Donde $h_{ci}(t)$ es una función que determina una vecindad en torne a la BMU "c" y el vector prototipo "i".

Dependiendo de la concepción que se tenga de vecindad la topología de de la grilla bidimensional cambia, pudiéndose tener (12):

1. Topologías abiertas: La topología del espacio bidimensional de neuronas se mantiene como un plano bidimensional.
2. Topologías cilíndricas: La topología del espacio es tal que los vecino de las neuronas ubicadas en el borde izquierdo son las neuronas del borde derecho.
3. Topologías toroidales: Se mantiene la topología cilíndrica per so agrega que las vecinas de las neuronas del borde superior son las neuronas del borde inferior.

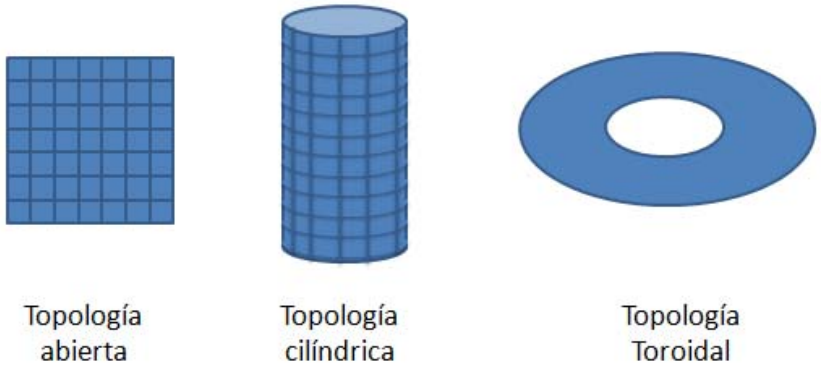


Figura 2.12: visualización de las topologías dada la definición de vecindad. Fuente: (7)

2.3.3.3 Validación de resultados

Los resultados obtenidos mediante procedimiento de *clustering* deben cumplir con dos requisitos (18): 1) ser consistentes con otras metodologías de clustering y 2) ser validadas por el usuario de la información a inferir.

Por un lado la consistencia en los resultados deben ser validados con otros métodos dados las variaciones que las condiciones iniciales y el tratamiento de los datos tienen en la solución final. Por ejemplo, K-means tiene gran dependencia en la solución inicial entregada y la SOFM presenta problemas con el manejo de los datos en los bordes del mapa.

Luego si los resultados por ambos métodos son consistentes, la solución será robusta.

Por otro lado, la información debe ser validad por el usuario final, ya que el procedimiento de clusterización no incluye el objetivo del proyecto, por lo que la solución que entregue puede no ser útil.

Finalmente, existe la posibilidad de que más de una solución presente validez, en cuyo caso se propone la utilización de un indicador de rankeo de las soluciones.

Considerando que el objetivo de la *clusterinzación* es la obtención de grupos compactos y diferenciados unos de otro, se propone la utilización de un segundo criterio de selección de *clusters* (26)*,* presentado en la ecuación 2.24, el que minimiza el cociente entre la varianza intragrupo y la varianza intergrupo.

$$\frac{Varianza\ intragrupo}{Distancia\ intergrupo} \qquad (2.24)$$

Entendiéndose como varianza intragrupo la suma sobre todos los clusters de la varianza observada en cada cluster y la distancia intergrupo como la mínima distancia existente entre dos observaciones de clusters distintos.

De esta forma se busca por un lado privilegiar clusters compactos y a la vez distantes entre ellos y por otro lado tener un criterio de selección entre soluciones válidas que permita la automatización de la elección.

2.3.4 Los datos en el proceso de Mining

Se puede entender como dato[27]:

1. m. Antecedente necesario para llegar al conocimiento exacto de algo o para deducir las consecuencias legítimas de un hecho.
2. m. Documento, testimonio, fundamento.
3. m. Inform. Información dispuesta de manera adecuada para su tratamiento por un ordenador."

[27] www.rae.es

De lo que se infiere que un dato es un la pieza fundamental para la adquisición de información.

En la siguiente sección se revisa la importancia de los datos en el proceso de minería de datos y las principales fuentes de error.

2.3.4.1 Caracterización de problemas en los datos

Desde el punto de vista de la metodología KDD[28], los datos (y sus problemas) son tratados en el proceso de transformación y Carga de datos.

Más en particular, se pueden definir dos contextos en los que estos problemas ocurren (27) y, desde esta perspectiva describir los posibles errores y sus orígenes en el proceso de extracción, transformación y carga de los datos.

Por un lado están los problemas ocurridos en el proceso de ETL, indepedientemente de la intervención de un usuario, y por otro, los ocurridos por la intervención de algún usuario.

a) Problemas ocurridos en el proceso: Problemas de diseño del proceso de ETL, pueden llevar a la pérdida o distorsión de la información.

   Estos errores se deben a errores en el análisis y diseño del proceso y se pueden mencionar como ejemplos: problemas de compatibilidad de formatos, lecturas nulas de datos, mala elección de funciones para la transformación de datos, fallas en el diseño de la base de datos, etc.

b) Problemas de origen operacional (27) (12): Problemas ocurridos por la intervención humana tanto en la captura como en la transcripción de datos al sistema.

   - Ruido en los datos: Se refiere a la modificación de los valores originales.
     Se pueden mencionar como ejemplo: distorsión de la voz que se escucha al hablar por teléfono.
   - Outliers: Datos que se alejan mucho de sus pares y resultan extraños al contexto del resto de los datos.
     Se pueden mencionar como ejemplos: Errores de medición, errores de digitación de la información.
   - Valores faltantes: Valores que no aparecen en el registro. Se pueden mencioar como ejemplos: datos no recogidos,
   - Atributos no son aplicables a todos los datos: Como ejemplo se pueden mencionar: tipeo de un número en un campo en el que se espera una letra
   - Datos duplicados: Valores que aparecen más de una vez y resulta complicado reducirlos, en muchas ocasiones el costo de perder datos es mayor que el de mantener los datos redundantemente. Como ejemplo se puede mencionar la misma persona con dos mails distintos.

[28] Figura 2.1

2.4 Herramientas tecnológicas

Tanto para la implementación de las técnicas de data mining como de ETL y la creación de las bases de datos, existen variadas herramientas tecnológicas.

En lo siguiente se nombrarán algunas opciones existentes con sus ventajas y desventajas.

2.4.1 Minería de datos

A continuación se describen las dos herramientas open source más utilizadas: Weka y R proyect.

Ambas ponen a disposición paquetes preprogramados con herramientas estadísticas y de machine learning, con interfaces amigables y fáciles de usar.

Weka[29]: Software open source escrito en java para la realización de *data mining*.

Presenta una gran variedad de técnicas de *machine learnig*, dentro de sus ventajas se cuenta el que implementa herramientas de pre procesamiento de datos.

Dentro de sus desventajas se cuenta que no implementa metodologías estadísticas, presenta peor soporte gráfico que R y la actualización de los sus paquetes es más lenta que en R, por lo que no tiene implementadas las últimas técnicas de *machine learnig*.

R proyect[30]: Es un paquete estadístico y de data mining, consistente de un lenguaje iinterpretado de programación, escrito en C++, muy simple y una interfaz gráfica para la realización de graficos.

Su gran ventaja es que está diseñado para la programación, por lo que existe una comunidad a su alrededor que genera una cantidad de paquetes de expansión y documentación.

Sin embargo, la mayor diferencia entre Weka y R es que R se encuentra enfocado a la programación de rutinas, mientras que Weka se enfoca a la utilización directa de las herramientas de minería de datos por parte del usuario final.

Este último punto, el del enfoque en la programación, inclina la decisión para la utilización de R en la presente memoria.

[29] http://www.cs.waikato.ac.nz/ml/weka/
[30] http://www.r-project.org/

2.4.2 Proceso ETL

A continuación se especifican distintas soluciones al problema de extracción, transformación y carga de datos.

Estas asisten el problema mediante una interfaz gráfica que permite realizar los procesos de extracción, transformación y carga de datos, de manera visual y más fácil, ayudando al diseño del proceso y facilitando la escritura y corrección de los elementos de programación involucrados en el proceso ETL.

Soluciones Open Source: Estas soluciones se encuentran gratuitamente en la red, dentro de las que podemos mencionar Kettle[31], Clover[32] ETL.

Mientras Kettle utiliza javascript para la realización de los cálculos Clover diversos lenguajes, como java, pearl. Ambos tienen la capacidad de correr con procesos paralelos además de tener ambientes gráficos bastante parecidos.

Soluciones Comerciales: Estos software se encuentran disponible previo pago de una licencia por su uso. Algunos ejemplos son Sunosis[33] ETL, Websphere[34]:

Sunopsis ETL es el software de ETL de Oracle, mientras Wesphere es el de IBM. El pago necesario para su utilización los hace inaccesibles para el presente trabajo.

Luego de los datos presentados se descartan las soluciones comerciales dado que el siguiente trabajo trata de la realización de un prototipo. Dado que no se encuentran diferencias sustanciales entre Clover y Kettle, se escoge Kettle dado el grado de familiaridad que tiene el autor con esta herramienta.

2.5 Levantamiento de la Situación

La complejidad de un proyecto de intervención tecnológica en una organización comienza, por la determinación de la problemática a intervenir, definir los alcances del trabajo, métricas para establecer objetivos claros. Todo esto antes de comenzar con el desarrollo de la solución y la realización de experimentos.

Para la realización de este trabajo previo existen un número de metodologías que guían en el levantamiento de los requisitos y entregables así como en el diagnóstico de la situación a intervenir

2.5.1 Áreas a determinar en el levantamiento de la situación

Se distinguen una serie de procesos necesarios para la localización y correcta ubicación del proyecto tecnológico en la empresa. De manera de minimizar el impacto

[31] http://kettle.pentaho.org/
[32] http://www.cloveretl.com/
[33] http://www.oracle.com/sunopsis/index.html
[34] http://www-01.ibm.com/software/websphere/

traumático sobre las operaciones de la organización, optimizar el tiempo de acceso a los datos y del desarrollo del proyecto y facilitar la obtención de resultados efectivos y utilizables.

Estos procesos son:

a) Obtención no ambigua de objetivos
b) Contextualización del proceso a intervenir en el marco organizacional de la empresa.
c) Identificación de los actores relevantes en el cumplimiento del objetivo.
d) Identificación de las fuentes de información, accesibilidad y calidad de los datos.
e) Caracterización de los requerimientos de mecanización.

En el desarrollo de estos procesos existen un número de herramientas metodológicas que facilitan la realización de los procesos descritos anteriormente. Los utilizados para este trabajo en particular serán descritos en lo que sigue.

2.5.2 Metodología Usada (10)

Los pasos propuestos para el levantamiento de la situación son:

a) Obtención de Objetivos: La obtención de los objetivos es fundamental tanto para el consultor como para el cliente pues permite a los desarrolladores entender y aclarar roles y responsabilidades y al cliente entender "qué quiere y cómo lo quiere"

   Este proceso incluye además la identificación de las métricas a utilizar para poder decir objetivamente cuando el proyecto está terminado o no.

   La principal herramienta metodológica para el desarrollo de esta etapa son las entrevistas con el cliente.

   Previa la realización de la entrevista es necesario prepararla, es decir, identificar la información clave a obtener, seleccionar al entrevistado, agendar la entrevista con tiempo, planificarla para no durar más de 30 minutos y preparar la formulación de las preguntas.

A continuación se presenta, en la tabla 2.3, la información básica a extraer de una entrevista.

Tabla 2.3: Información a extraer de una entrevista para la obtención de Objetivos

Información a extraer
El problema en palabras del entrevistado
Como le afecta el problema
Quién y Cómo toma las decisiones
Frecuencia y toma de las decisiones
Fuentes y Formatos de la información
Calidad de la información
Qué le falta

Fuente: Elaboración propia basado en (19)

Con respecto a la formulación de las preguntas se recomienda un estilo abierto de preguntas (10), que permita al entrevistado explayarse.

b) Contextualización del proceso a intervenir en el marco organizacional de la empresa: Luego de definidos los objetivos es necesario la contextualización de la solución, de manera de identificar las entidades relacionadas con la entrega de la información y el uso de la solución, además de quienes serán los usuarios directos del proyecto.

Para esta etapa, se propone la construcción de un Diagrama de Contexto, en el cual debe estar especificado, a nivel muy agregado, las entidades involucradas en el proceso a intervenir y su intervención con el sistema a desarrollar. La figura 2.13 ejemplifica un diagrama de contexto.

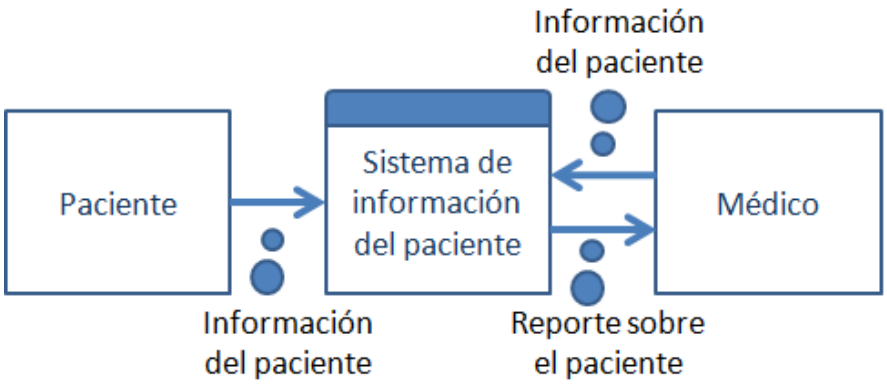


Figura 2.13 Ejemplo de diagrama de contexto. Fuente: (19)

c) Identificación de los actores relevantes en el cumplimiento del objetivo: La identificación de estos actores es relevante para la realización tanto de levantamiento de requerimientos de mecanización como obtención de datos y de implementación.

Para esto, se propone la utilización del diagrama de contexto. Este diagrama más un organigrama permite la identificación de los cargos específicos a los que el proyecto debe referirse.

d) Identificación de las fuentes de información, accesibilidad y calidad de los datos: Es indispensable para la realización de una propuesta, el saber el ambiente de manejo y mantención de los datos, así como de las fuentes, cantidad e información y la facilidad para accederlos es fundamental en la determinación de una propuesta de solución.

e) Caracterización de los requerimientos funcionales: Los requerimientos de mecanización se definen como el paso mediante el cual se pasa de un modelo lógico de la solución a la implementación física. Comprende 3 áreas de desarrollo: Especificación de requerimientos de datos, de procesos, y de distribución.

Estas áreas de desarrollo y la ubicación lógica del proceso de levantamiento de requerimientos se encuentran graficadas en la figura 2.14.

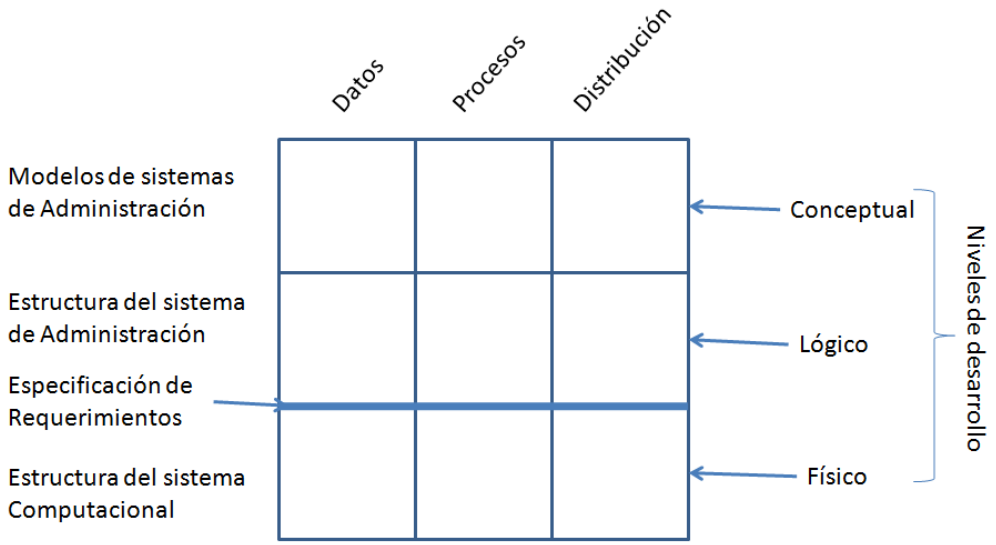


Figura 2.14: Levantamiento de requerimientos. Fuente: (19; 6)

El esquema de la figura 2.14 es un esquema, extracto de la matriz de Zachman, que presenta las etapas de construcción de un proceso.

La matriz de Zachman[35] comprende, más específicamente, una ontología y no una metodología. Es decir es una descripción de las partes constituyentes de un proceso de negocio.

Dado lo anterior, la utilización de éste esquema es funcional al ordenamiento de las partes constituyentes de cualquier proceso de negocio y no describe el *Cómo* describir sus componentes. Sin embargo, su función principal es la de guiar el desarrollo de un proyecto, asegurando la integración de todas las áreas involucradas y la relación entre cada una de ellas.

Existen otras metodologías para el levantamiento particular de cada procesos, como los diagrama de contexto, diagramas de flujo, estudios de caso, modelo entidad relación y muchos otros.

Las metodologías desarrolladas en particular para la materialización de cada etapa del diseño involucrada en la figura 2.14 será discutida y decidida en el capítulo 3: Levantamiento de la situación actual.

[35] http://www.zachmaninternational.com/index.php/the-zachman-framework

3. LEVANTAMIENTO DE LA SITUACIÓN ACTUAL

La caracterización del problema y el entorno organizacional toma suma importancia en la determinación de una solución factible y aplicable en el marco de la cultura de la empresa y las expectativas de los usuarios.

En el presente capítulo se identificará la problemática a analizar y se obtendrán los requerimientos deberá cumplir la solución al problema propuesto en el contexto de la organización.

En primer término se realizará una descripción del negocio de Easy, con el objetivo de presentar el funcionamiento y manejo de productos en la empresa. Luego, según se menciona en el capítulo 2.5, se procede a hacer el levantamiento de la situación actual para finalmente proponer los requerimientos de mecanización de datos, procesos y distribución para el sistema a proponer.

3.1 Descripción del negocio de Easy S.A.

A modo de introducción para el desarrollo del presente trabajo se describe el negocio de Easy S.A. y el manejo operacional y táctico de este.

3.1.1. Misión, visión y estrategia comercial

Como se mencionó en el capítulo 1 punto 1.3, Easy S.A. es una empresa del grupo Cencosud, con la que comparte misión y visión (1).

Misión: *"Ser el retailer más rentable y prestigioso de Latinoamérica, en base a la excelencia en nuestra calidad de servicio, el respeto a las comunidades con las que convivimos y el compromiso de nuestro equipo de colaboradores con los pilares básicos de nuestra compañía: Visión, Desafío, Emprendimiento y Perseverancia"*.

Visión:*"No solo visualizaremos el futuro, si no que trabajaremos para construirlo en grande"*.

La visión y misión antes descritas dan marco para el desarrollo de la estrategia comercial de la empresa, igualmente la visión da pie al desarrollo de herramientas que permitan la concreción de la misión de la empresa, que es donde el desarrollo de la siguiente memoria tiene sustento en la empresa.

La administración estratégica de la división de *"Home Improvement"* está a cargo de un gerente regional único, esto con el objeto de generar una propuesta de valor única, traspasar las mejores prácticas, crear sinergias en los diferentes procesos y mejorar la eficiencia operacional.

La estrategia comercial contempla la separación del negocio de mejoramiento del hogar en dos áreas: 1) Ventas de productos de materiales de construcción para

empresas y profesionales del rubro de construcción y 2) Venta asistida de artículos de reparación, equipamiento, decoración y mejoramiento del hogar.

El desarrollo del presente trabajo contemplará datos de la línea de negocios de venta asistida.

3.1.2. Manejo operacional de ventas

El área de venta asistida de artículos de reparación, equipamiento, decoración y mejoramiento del hogar se materializa a través de dos canales: 1) a través del portal de internet[36] y 2) en las tiendas de Easy S.A. a lo largo del país.

Las tiendas, son grandes galpones organizados de manera similar a las tiendas de supermercados. El espacio físico está dividido según categorías de productos, los cuales se encuentran físicamente localizados en góndolas o estanterías.

El cliente recorre la tienda con un carro buscando lo que necesita y es en esta búsqueda asistido por personal de la tienda.

Luego de identificar lo requerido, el cliente pasa por las cajas, apostadas cerca de la entrada de la tienda, y procede al pago de los productos que se desean llevar.

Existen dos mecanismos de mantención del *stock* de la tienda:1) mediante las bodegas de la propia tienda y 2) mediante el despacho desde las bodegas centrales a través de un sistema de *cross docking*[37].

Dado lo anterior es importante la estimación de la demanda, de manera de no tener sobre *stock* por un lado, ni "quiebres de *stock*"[38] por el otro.

3.1.3. Manejo táctico de tiendas y productos

En la siguiente sección se describirá el manejo y administración de las tiendas y productos a nivel estartégico y operacional.

Easy maneja un stock promedio de 35.000 productos, los que son asignados para cada una de sus 26 tiendas[39].

Existen dos tipos de organización jerárquica de productos, una jerarquía comercial y jerarquía operacional.

[36] *www.easy.cl*
[37] Esto quiere decir que existe una sola bodega central, donde llegan los pedidos generados para todas las tiendas y donde se generan los despachos especificados para cada tienda según pedido.
[38] Un quiebre de *stock* ocurre cuando se pierde una venta por no contar con la unidad para venderla.
[39] Para el detalle de las tiendas referirse al Anexo C "Datos de la empresa"

Operacionalmente los productos se encuentran organizados en secciones, rubros y subrubros, mientras que desde el punto de vista comercial la organización de los productos es según categorías.

Existen 29 secciones de productos y cada sección tiene un número de rubros y subrubros independientes, mientras que desde el punto de vista comercial los productos son agrupados en 44 categorías

Los productos son asignados a una tienda, es decir, se hacen disponibles para la venta, mediante el proceso de catalogación de productos.

Dado el gran número de productos que se manejan en Easy, la catalogación no se realiza entre cada producto y cada tienda, sino que entre agrupaciones de productos y agrupaciones de tiendas.

Para esto, los productos son previamente agrupados según características de venta y estacionalidad al igual que las tiendas son agrupadas por características físicas, para finalmente realizar las catalogaciones correspondientes.

Descripción y agrupación de las tiendas

En la tabla 3.1 se muestran las 44 categorías son existentes y el porcentaje promedio de espacio que ocupan para el conjunto de tiendas.

Tabla 3.1: Distribución del espacio en tiendas.

Categoría	% promedio espacio en salas	Categoría	% promedio espacio en salas
Accesorios de Baños	1,74%	Mat. Construcción obra Intermedia	5,30%
Adhesivos, Sellantes Y Cintas	0,57%	Muebles	7,62%
Alfombras de rollo y vinilicos	1,17%	Muebles de Baño y Cocina	3,62%
Ampolletas y Tubos	0,54%	Organización	1,43%
Automotor	1,15%	Papel Mural	0,20%
Bazar / Hogar	4,04%	Piletas	0,44%
Cerámicas	5,11%	Pinturas	4,55%
Electricidad	1,99%	Pisos de Madera	1,02%
Electrodomesticos	1,79%	Placas	4,46%
Electro-Entretenimiento	0,63%	Plantas de Jardín	4,83%
Fijaciones	1,59%	Plomería	2,10%
Grifería	0,56%	Puertas y Ventanas, Molduras	3,73%
Herramientas Manuales	1,13%	Quincallería y Herrajes	1,39%
Herramientas Manuales de Jardín	0,39%	Riego	0,53%
Herramientas y Máquinas Eléctricas	2,83%	Sanitarios	1,00%
Iluminación	3,63%	Seguridad Industrial	0,98%
Limpieza	0,69%	Semillas, Fertilizantes y Plaguicidas	0,72%
Macetería y otros de Jardín	2,51%	Techos	2,10%
Maderas	3,61%	Tiempo Libre	0,91%
Maq. Eléctricas de Jardín	1,05%	Termotanques y Calefont	0,40%
Mascotas	1,01%	Textil Hogar	5,25%
Mat. Construcción obra Gruesa	5,76%	Temporada	3,92%

Fuente: Elaboración propia, información del departamento de planning

La asignación del espacio ocupado por cada categoría en tiendas, no responde a una variable estratégica, ni a un manejo central, sino a un modelo base de tienda y a las necesidades particulares que se presenta cada tienda dadas las ventas observadas.

Las tiendas se encuentran clasificadas, por razones de manejo operativo, en tres grupos dependiendo solo del tamaño total de sus salas de venta: tiendas grandes, tiendas medianas y tiendas pequeñas.

En la tabla 3.2 se muestran estos grupos y las tiendas que los componen

Tabla 3.2: Tiendas agrupadas según clasificación

Grandes	Medianas	Pequeñas
E503	E520	E502
E504	E592	E513
E507	E524	
E508	E591	
E510	E585	
E512	E529	
E514	E781	
E517		
E518		
E521		
E525		
E534		
E522		

Fuente: Elaboración propia, en base a información de la gerencia de procesos comerciales

Cabe mencionar que esta agrupación se realiza sin ninguna otra variable de decisión que el tamaño de las tiendas, obviando información tanto de ventas como de características socioeconómicas o de utilización de espacio en las tiendas.

Manejo de productos

Los productos son agrupados según las características observadas de ventas tales como estacionalidad y demanda.

Según este criterio existen siete grupos relevantes para el desarrollo del presente trabajo: productos básicos, complementarios, complementarios tipo 1, de temporada, básicos de temporada, complementarios de temporada y excepciones.

Los productos básicos son aquellos considerados para la venta en todas las tiendas, mientras que los artículos complementarios responden a productos con demandas específicas en algunas tiendas y los artículos complementarios de tipo 1 son considerados exclusivos, aun más específicos que los artículos complementarios.

Además existe una diferencia entre productos a la venta durante todo el año y aquellos que dispuestos solo a la venta durante temporadas especiales, como artículos de piscinas o de calefacción. Estos artículos son denominados de temporada y reciben la misma agrupación que los artículos que no son de temporada.

La tabla 3.3 muestra los grupos relevantes para el trabajo realizado, indicando además a que grupos de tiendas se encuentran asignados.

Tabla 3.3: grupos de productos y su asignación a cada grupo de tienda

Grupo de Producto	Tienda catalogada
Básico	Grande, Mediana, Pequeña
Complementario uno	Grande
Complementario	Grande, Mediana
Temporada Básica	Grande, Mediana, Pequeña
Temporada Complementaria	Grande y Mediana
Temporada	Grande, Mediana y Pequeña
Excepción	A una o más tiendas según criterios comerciales

Fuente: Elaboración propia, información de la gerencia de procesos comerciales

3.2 Levantamiento de la situación

A continuación se realiza la identificación de los procesos a intervenir en Easy S.A., su situación de la situación actual y los requerimientos necesarios para realizar una propuesta de solución.

3.2.1 Obtención de Objetivos

La obtención de los objetivos se realizó mediante entrevistas con las principales áreas implicadas en la generación de requerimientos: Precios y Surtidos y Planning.

Se entrevistó a: Tomás Zavala, Subgerente de precios y surtidos, Raúl Iglesias, Head Planner y Claudio Larrea, Category manager.

La tabla 3.4 presenta los resultados de las entrevistas realizadas, obtención de las principales problemáticas identificadas por los usuarios directos de la solución a desarrollar.

Tabla 3.4. Resumen de las problemáticas levantadas y quienes las presentan

Problemática	Tomás Zavala	Claudio Larrea	Raúl Iglesias
Velocidad en realización de Pronóstico de demanda	X		X
Calidad en el pronóstico			X
Velocidad en asignación del Mix de productos por tienda	X	X	
Velocidad en el acceso a la información		X	
Seguimiento de las acciones tomadas	X		

Fuente: Elaboración propia en base a entrevistas[40]

Con esta información se obtiene el siguiente Objetivo General desde el punto de vista de los usuarios:

“Realización de pronósticos de demandas y determinación del Mix de productos, permitiendo el seguimiento oportuno de los resultados obtenidos y los reales”

[40] Ver Anexo H “Entrevistas”

Cabe mencionar que este objetivo esta alineado con el objetivo propuesto en el capítulo 1, mediante la creación de un sistema de apoyo a la toma de decisiones que entregue reportes sobre la información de ventas y que permita la aplicación de herramientas de minado de datos para la obtención de información relevante para el tomador de decisiones.

3.2.2 Contextualización del proceso a intervenir en el marco organizacional de la empresa

Para la concreción del objetivo general recién presentado es necesario el desarrollo de un sistema que interactúe entre las fuentes de datos y los usuarios que requieran la información.

En la presente sección se presentará la ubicación lógica de este sistema en el proceso en el que se encuentra inmerso, indicando además a los actores que se convertirán en sus clientes.

La figura 3.1 muestra la ubicación del sistema, contextualizándolo en el marco organizacional de gerencias de Easy S.A.. Se especifican además los principales departamentos clientes del sistema.

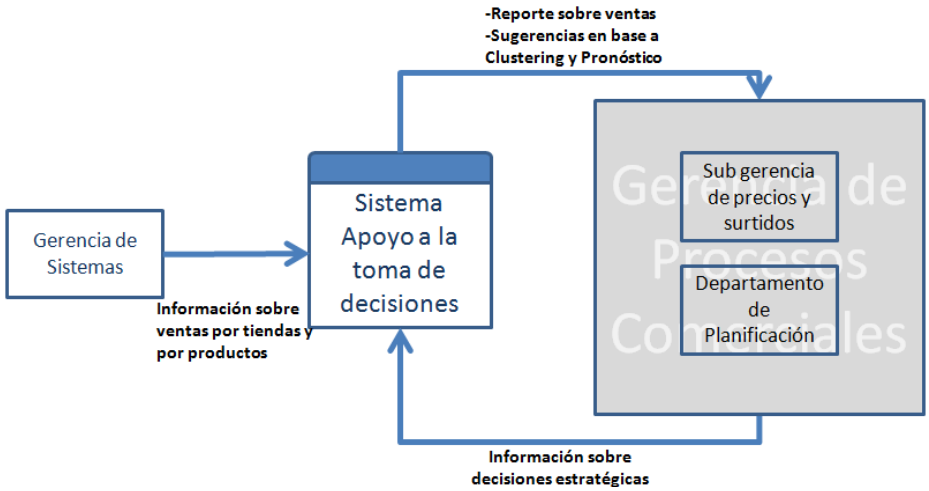


Figura 3.1: El sistema en el entorno organizacional. Fuente: Elaboración propia

Como se observa en la figura 3.1 el sistema a diseñar necesitará estar en contacto con 2 distintas gerencias de la organización, la gerencia de sistemas y la gerencia de procesos comerciales.

La gerencia usuaria será la gerencia de procesos comerciales, en particular los Departamentos de Planificación y Precios y Surtidos.quienes requerirán de la la información de ventas proporcionada por la gerencia de sistemas.

Tanto el departamento de planificación como el de precios y surtidos requerirá reportes sobre el comportamiento de ventas para la realización del seguimiento de las acciones comerciales tomadas y deberán a su vez, corroborar o rechazar las

recomendaciones que el sistema de apoyo pueda realizar tanto en pronóstico de demandas como en agrupación de tiendas.

3.2.3 Identificación de los actores relevantes en la obtención de los objetivos

Mediante la utilización del organigrama institucional, se busca identificar a los usuarios directos del sistema y a los validadores del mismo

De esta forma, se busca asegurar la identificación de los clientes y las fuentes de información.

A continuación, se presenta el organigrama institucional de Easy en la gerencia de procesos comerciales.

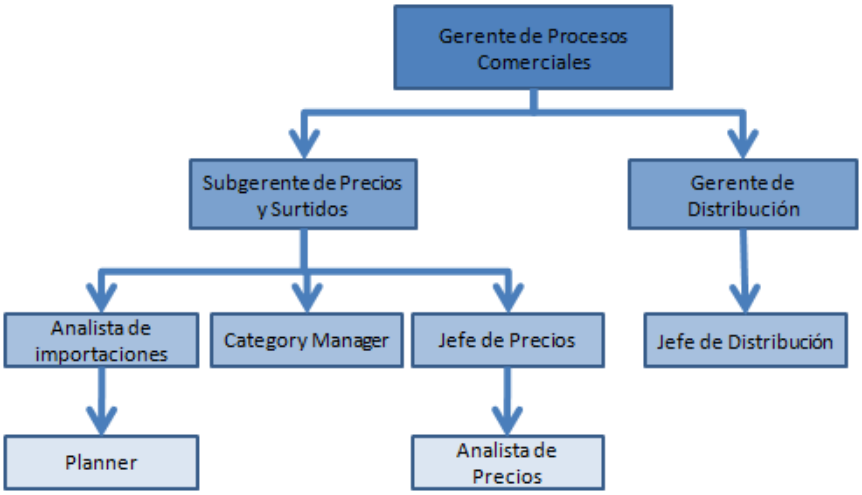


Figura 3.2: Organigrama de Easy S.A., Gerencia de Procesos Comerciales. Fuente: Elaboración propia, basado en información del departamento de procesos comerciales

Se observa que el área relevante para la concreción del proyecto es la subgerencia de precios y surtido, donde los usuarios directos del sistema serán el Analista de importaciones y el Category Manager.

Ambos necesitan la proyección de demanda para la realización de sus tareas así como de un constante monitoreo del desempeño de sus acciones comerciales.

Se identifica como un apoyo clave al Subgerente de Precios y Surtidos dado el carácter de confidencial de cierta información.

En el área de sistemas, el encargado de la entrega de información relevante es el Jefe de control de gestión del BI, siendo BI la denominación actual del sistema de visualización de los datos relevantes para la realización de los procesos comerciales[41].

[41] Referirse al punto 3.6.1 “Descripción de las fuentes de información”

3.2.4 Métricas en la medición de Objetivos

Para la determinación del éxito del proyecto, se ha consensuado con el cliente la siguiente lista de indicadores para la edición de objetivos.

- Certeza en la predicción de ventas: Con el objetivo de mejorar la calidad de los pronósticos de ventas, se medirá la certeza de los pronósticos que se realicen utilizando las distintas medidas de error explicadas en el capítulo 2, sección 2.3.2.3, estas son:1) Error Absoluto medio (MAE) y 2)Error cuadrático medio (MSE).

  Se realizará la medición de la exactitud actual de los métodos de predicción utilizados, para luego contrastarlos con los valores obtenidos mediante los procedimientos desarrollados en el presente trabajo.

- Tiempo en la ejecución de agrupación de tiendas: Actualmente no existe registro del tiempo involucrado en la realización de esta tarea, sin embargo este proceso no se ha realizado hace un par de años debido a las complicaciones asociadas a su realización.

  Se medirá el tiempo en base a la factibilidad de realización de este proceso a lo menos 2 veces al año.

- Tiempo en la ejecución de agrupación de productos: Actualmente no existe registro del tiempo involucrado en la realización de esta tarea, sin embargo este proceso se realiza dos veces al año.

  Se medirá el tiempo en base a la factibilidad de realización de este proceso a lo menos 4 veces al año.

- Recursos Humanos utilizados en actividades de catalogación: La utilización de recursos humanos se medirá en dinero involucrado por concepto de horas hombre.

  Actualmente la utilización de recursos humanos es estimada en $4.950.000 anualmente, monto que se desea disminuir con la implementación de este proceso.

3.2.5 Fuentes de información, accesibilidad y calidad de los datos

Existen dos fuentes primarias de información: a) Los sistemas de “Bussiness Intelligence” disponibles desde la Gerencia de Procesos Comerciales y b) El Sistema de Bases de Datos manejado desde la gerencia de Sistemas.

En una primera instancia la obtención de información se realiza directamente desde las herramientas de “Bussiness Intelligence” de las que dispone la Gerencia de Procesos Comerciales, sólo por la facilidad de obtención de esta información.

Sin embargo para el funcionamiento cabal del sistema, la consulta directa a las Bases de Datos transaccionales de ventas es la opción más rápida y de mayor confiabilidad.

- Descripción de las fuentes de información: En la actualidad existe un data warehouse funcionando para el almacenamiento de la información de ventas. En particular existen 3 cubos de información según el nivel de grano temporal al que se quiera acceder.

  Los datos es almacenada en un modelo entidad relación administrado mediante SQL server 2005.

  El modelamiento multidimensional es materializado usando un modelo de "*Snow Flake"* cuya *fact table* maneja alrededor de 100.000.000 de registros[42].

  El acceso al cubo de información puede ser mediante dos vías: El sistema de información BI y el acceso directo al repositorio de datos dependiente de la gerencia de sistemas.

  Sistema BI: Sistema de visualización de consultas OLAP al cubo de información mediante la utilización de Excel. La conexión al cubo se realiza directamente desde Excel y la visualización se realiza mediante una tabla dinámica.

  Cubo de información: Realizando consultas directamente al cubo de información de ventas, constituido como se señaló anteriormente.

- Accesibilidad a los datos: A continuación se caracteriza la accesibilidad a los datos desde el punto de vista de la velocidad en el acceso y la facilidad para la manipulación de estos

  Sistema BI: La velocidad de acceso a la información es muy lenta, esto debido a: 1) la cantidad de usuarios que tienen acceso a los cubos de información y 2) porque Excel no está diseñado para la manipulación grandes cantidades de datos[43].

  Cubo de información: Rápido acceso a los datos requeridos, alrededor de 2 horas para los datos de venta diaria de todo Easy.

- Caracterización de los datos: A continuación, se busca entregar una idea general de la calidad de los datos con los que se cuenta y dar una mirada global del negocio desde el punto de vista de los datos.

  Calidad de los datos: Los datos serán examinados desde dos aristas: 1) Completitud y coherencia en los datos recogidos desde el sistema operacional y 2) de acuerdo a las necesidades expresadas en el punto 3.1, en su capacidad para la realización de pronóstico de demanda y de asignación de productos por tiendas.

  Los datos recogidos corresponden a datos de ventas, en particular datos sobre cantidad vendida y monto vendido por producto, por tienda para cada día desde el 2006.

[42] Fuente: Entrevista a Luis Osorio, Analista programador
[43] Access es el software para estos requerimientos

Con esto se definen los siguientes conceptos para la medición de la calidad de los datos:

- Dato completo: Un registro es considerado completo si presenta cifras positivas tanto en cantidad vendida como en monto vendido (es decir se realiza una transacción que involucró salida de producto e ingreso de dinero), o si presenta valor cero tanto en cantidad vendida como en monto vendido (es decir no se realizó una transacción).
- Dato inconsistente: Dado que la información es sobre ventas, se considerará que un valor negativo en cantidad vendida no es consistente con lo que representa el registro[44].

En la tabla 3.5 se resume la calidad de los datos considerando los conceptos de completitud y consistencia anteriormente explicados.

Tabla 3.5: Completitud y coherencia de los datos recogidos

Tienda	Datos incompletos	Cifras Negativas	Total de datos	Porcentaje Defectuosos
E502	157	131	18.752	1,5%
E503	3.294	337	192.560	1,9%
E504	8.519	414	220.603	4,0%
E507	3.863	254	194.557	2,1%
E508	8.526	332	171.857	5,2%
E510	7.501	298	139.465	5,6%
E512	8.393	476	262.295	3,4%
E513	145	-	406	35,7%
E514	3.607	360	182.551	2,2%
E517	4.347	189	137.807	3,3%
E518	3.870	441	159.320	2,7%
E520	3.886	346	161.180	2,6%
E521	4.290	232	125.367	3,6%
E522	3.418	159	98.378	3,6%
E524	4.845	138	89.954	5,5%
E525	3.668	198	98.660	3,9%
E529	4.360	188	155.386	2,9%
E591	4.114	233	117.504	3,7%
E592	4.454	230	100.710	4,7%
E760	3.637	210	147.863	2,6%
E781	4.299	241	95.478	4,8%
Total general	93.193	5.407	2.870.653	3,4%

Fuente: Elaboración propia, basado en información de la gerencia de procesos comerciales

Se puede ver que, desde este punto de vista, los datos con los que se cuenta son de buena calidad ya que el porcentaje de datos que presentan anormalidades alcanza solo el 3,4% del total de datos disponibles, con un máximo de 5,6% para la tienda E510.

[44] La existencia de registros negativos corresponde a la devolución de productos, lo que no entrega información relevante para el problema a estudias.

El error más reiterativo es el correspondiente a datos incompletos representando un 3,4% del total de los datos mientras que los datos inconsitentes solo representan 0,19%.

Capacidad de pronóstico de los datos: Para que sea posible realizar pronósticos, es necesario que los datos a utilizar sean constituyentes de series de tiempos[45] y que además sean representativas de la demanda por productos.

De la observación de los datos entregados por la empresa se observa que cerca del 80% de los productos tienen una rotación menor que 1 unidad a la semana, mientras el 15% presenta una rotación de menos de 1 al día. Esto indica que será necesario especificar el nivel de agregación de los datos de modo que la proporción de valores vacios no sea mayor que la proporción de valores distintos de cero.

Tabla 3.6: Representación de productos

Ventas semanales	% del total de productos	% del total ventas
menos de 1	81,34%	59,3%
menos de 7	15,22%	31,5%
menos de 30	2,79%	6,9%
más de 30	0,65%	2,3%

Fuente: Elaboración propia, basado en información de la gerencia de procesos comerciales

Descripción del negocio: Los datos recogidos muestran la existencia de un marcada estacionalidad en la venta durante los meses de Marzo y Abril, como se muestra en la figura 3.3:

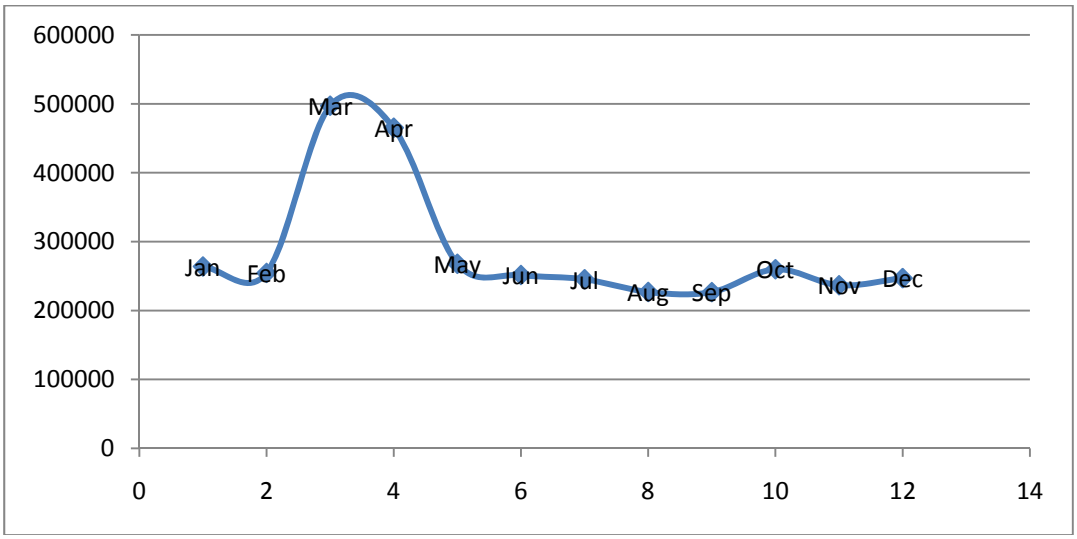


Figura 3.3: Estacionalidad en las ventas del corral de herramientas, ventas agregadas, Fuente: elaboración propia, basado en información de la gerencia de procesos comerciales

Por último se presentan las tiendas en función del monto vendido y el porcentaje que este monto representa del total vendido en la categoría en el país.

[45] Referirse al capítulo 2 punto 2.3.2

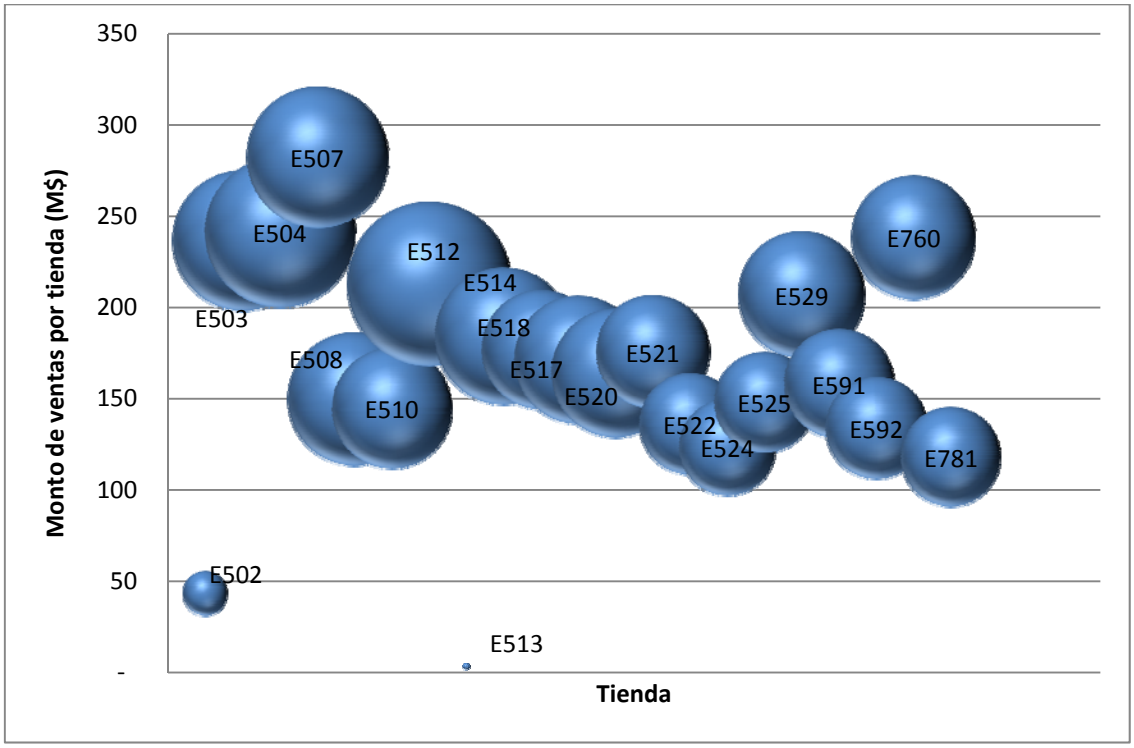


Figura 3.4: Monto de venta y porcentaje del total de artículos vendidos por tienda, Fuente: elaboración propia, basado en información de la gerencia de procesos comerciales

De donde se desprende que las tiendas más importantes para la sección son las tiendas E507, E504, E503 y E512, mientras las tiendas E502 y E513 presentan la menor incidencia sobre las ventas de la categoría.

3.2.6 Caracterización de los requerimientos de mecanización

Según el diagrama 2.12, los requerimientos de mecanización se dividirán en requerimientos de mecanización de datos, requerimientos de mecanización de procesos y requerimientos de mecanización de distribución.

A su vez cada requerimiento se describirá con 3 niveles de profundidad: nivel conceptual, lógico y físico de realización, según lo explicado en la sección 2.12.

a) Requerimientos de Mecanización de Datos: Se presenta la construcción de los modelos conceptual, lógico y físico de datos.

El modelamiento conceptual de datos se realiza mediante el diseño de un modelo entidad relación y el modelamiento lógico de datos se realizará mediante un modelo.

Modelo conceptual de datos: En la figura 3.5 se presenta el modelo entidad relación para el modelamiento conceptual de datos.

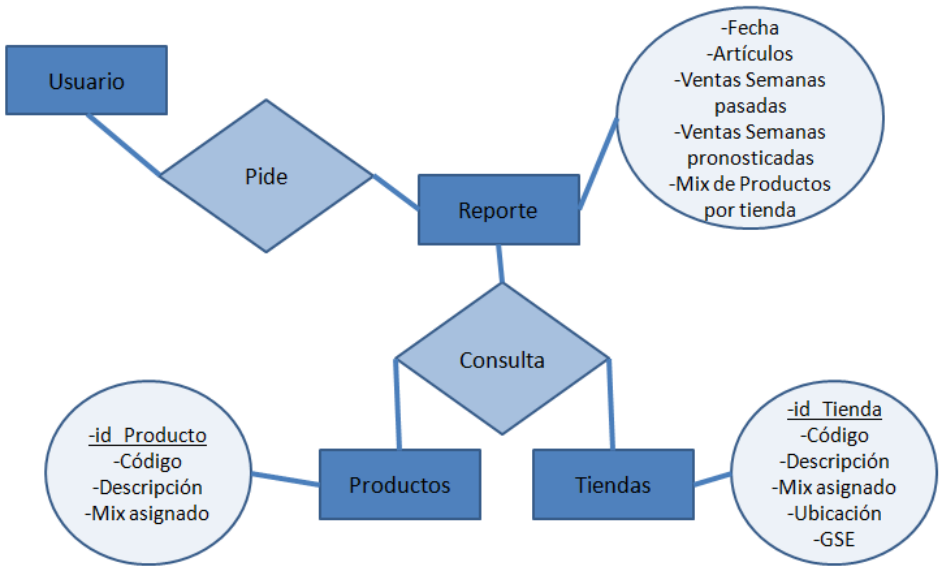


Figura 3.5: Modelo entidad relación del proceso comercial. Fuente: Elaboración propia

Modelo lógico de datos: Se propone un modelamiento multidimensional en base de datos relacionales.

Para esto se recomienda la construcción de un modelo de estrella con dos *fact table*, esto debido a que responden a distintas consultas. La *fact table* minería, responde a las necesidades internas del sistema en sus procesos de minería de datos y el grano de información es a nivel de temporada. Por otra parte, la fact table responde a las consultas que realizará el usuario y el grano de información es semanal.

Además se recomienda las siguientes tablas dimensiones: Tiempo, Producto y Tienda.

Por último, se proponen estructuras de *Snow Flake* para las tablas dimensionales Producto y Tienda, ya que tienen características que podrían modificarse sin la necesidad de que se modifiquen las relaciones existentes para el resto del Warehouse.

Modelo Físico de datos: Físicamente la Base de datos estará alojada en un servidor especialmente habilitado para este propósito, el que tendrá como cliente a dos computadores pertenecientes a los usuarios del deparartamento de procesos comerciales.

Requerimientos de Mecanización: Se requiere un sistema que recolecte la información de ventas desde la fuente de información transaccional para luego de procesada, cargarla a la base de datos según la caracterización del modelo lógico de datos especificado en la figura 3.4

b) Requerimientos de Mecanización de Procesos: El modelamiento de los procesos involucrados en el desarrollo del proyecto será realizado mediante la aplicación de la metodología de casos de uso y diagramas de flujos de datos (DFD), por la

capacidad que presenta este tipo de herramientas en modelar no sólo el proceso, sino también caracterizar el flujo de datos, en particular necesario para el desarrollo de este trabajo.

El modelo conceptual de procesos será diseñado mediante un diagrama de contexto. El nivel de diseño lógico será realizado mediante un diagrama de flujo de datos de nivel 0 y para cada uno de los casos de uso que se desprendan de este DFD se generará un DFD de nivel 1 para la determinación del diseño físico de datos.

Modelo Conceptual de Procesos: En la figura 3.6, se presenta el Diagrama de Contexto del sistema de Apoyo a la toma de decisiones que se diseñara.

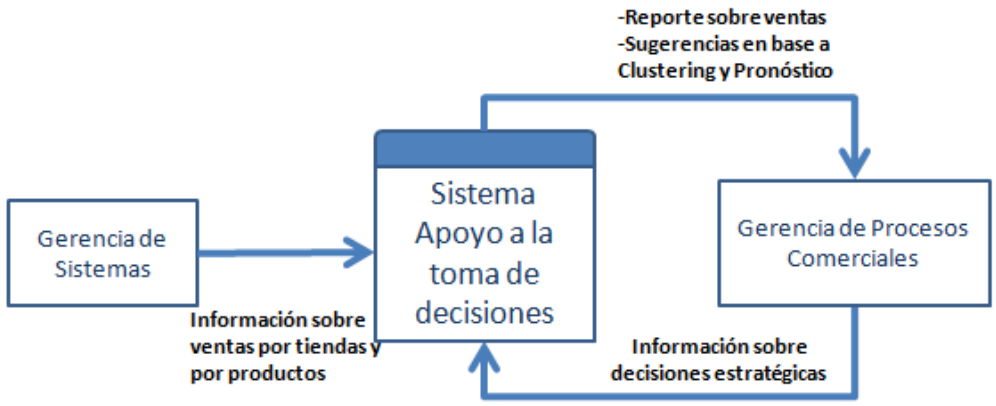


Figura 3.6: Diagrama de contexto, sistema de Apoyo a la toma de decisiones. Fuente: Elaboración propia

Modelo Lógico de procesos: En la figura 3.7, se presenta el Diagrama de Flujo de Datos de nivel de 0 del proceso a implementar como solución:

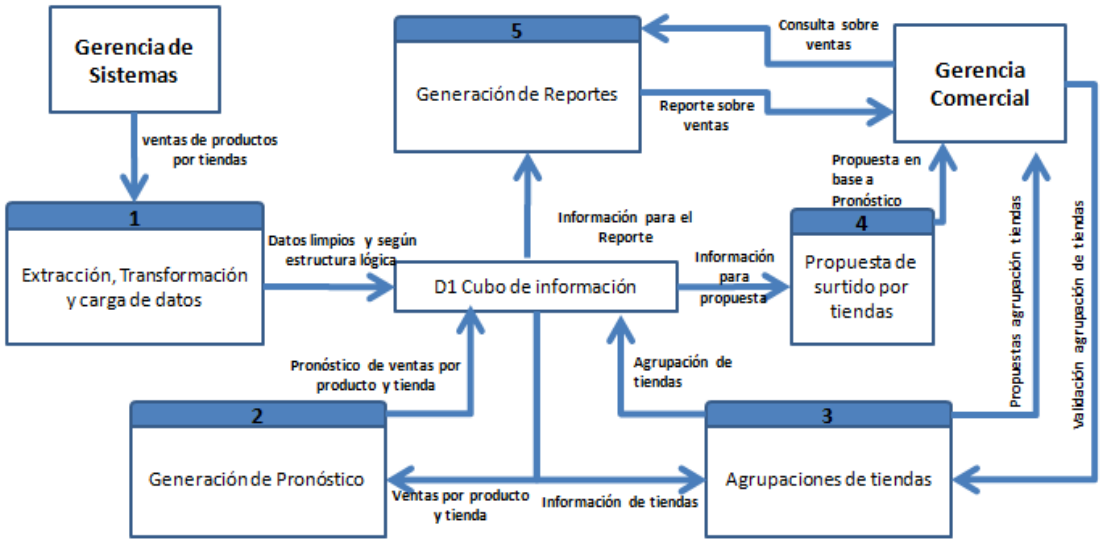


Figura 3.7: Diagrama de contexto, sistema de Apoyo a la toma de decisiones. Fuente: Elaboración propia

Se desprenden 5 casos de uso que comprenden el sistema de apoyo a la toma de decisiones: extracción, transformación y carga de datos (ETL), Generación de

pronósticos, Agrupación de tiendas, Propuesta de surtido por tiendas y Generación de Reportes.

Modelo físico de Procesos: A continuación, se especificarán los DFD de nivel 1 para cada caso de uso identificado en el modelo conceptual de procesos, especificando además el *trigge*r de cada uno de los casos de uso:

1.-ETL: Se ejecuta una vez por semana de manera autónoma. Recibe como input la información diaria de ventas de producto por tienda para luego cargar esta información al cubo de información.

En el diagrama 3.8, se presenta el DFD de nivel 1 para este caso de uso.

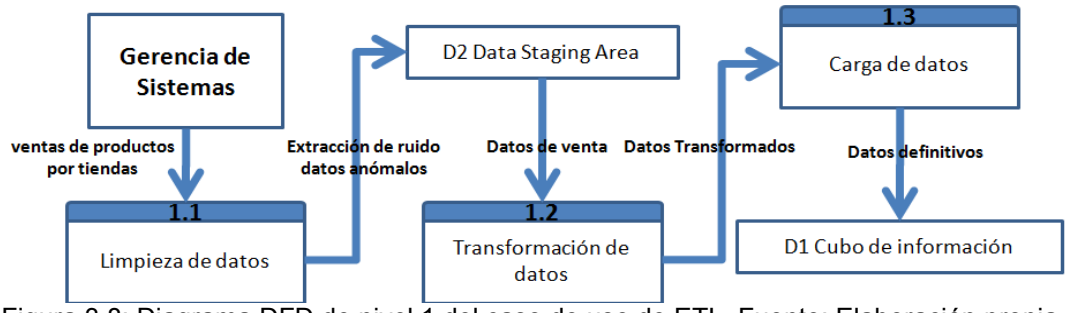


Figura 3.8: Diagrama DFD de nivel 1 del caso de uso de ETL. Fuente: Elaboración propia

2.-Generación de pronóstico: Se ejecuta 1 vez por semana justo después de terminado el proceso de ETL. Recibe como input los datos de ventas por tienda existentes en el cubo de información y como output genera pronósticos de venta para las próximas 8 semanas.

En el diagrama 3.9, se presenta el DFD de nivel 1 para este caso de uso.

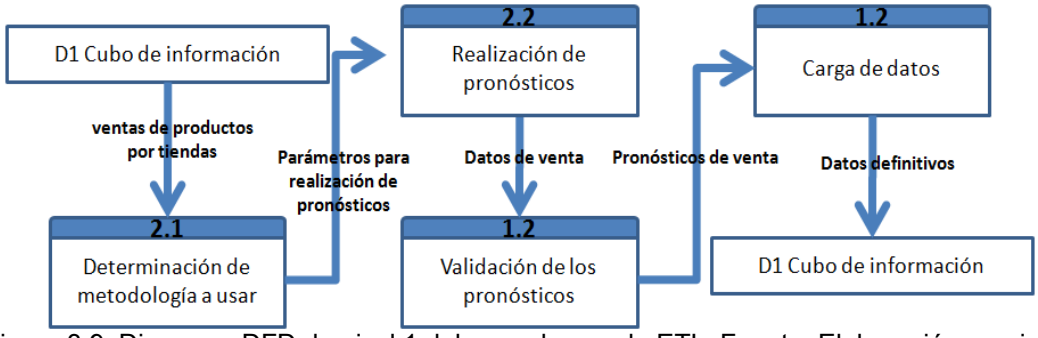


Figura 3.9: Diagrama DFD de nivel 1 del caso de uso de ETL. Fuente: Elaboración propia

3.-Agrupaciones de tiendas: Se ejecuta por requerimiento del usuario, el proceso contempla la recolección de los datos necesarios para la aplicación del minado de datos y la entrega de la información requerida.
En el diagrama 3.10, se presenta el DFD de nivel 1 para este caso de uso.

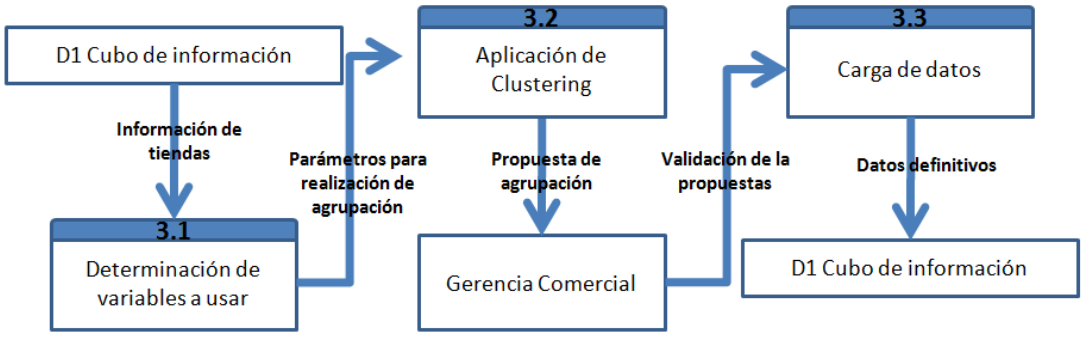


Figura 3.10: Diagrama DFD de nivel 1 del caso de uso de ETL. Fuente: Elaboración propia

4.-Propuesta de surtido por tiendas: Se ejecuta por requerimiento del usuario, el proceso contempla la recolección de los datos necesarios para la aplicación del minado de datos y la entrega de la información requerida.

En el diagrama 3.11, se presenta el DFD de nivel 1 para este caso de uso.

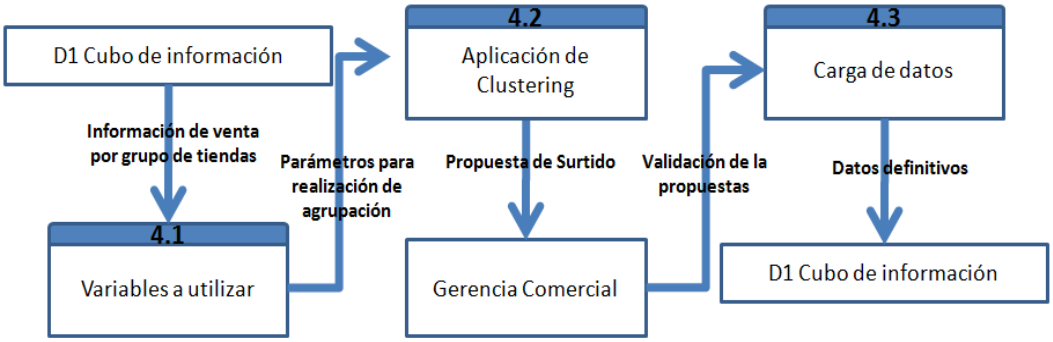


Figura 3.11: Diagrama DFD de nivel 1 del caso de uso de ETL. Fuente: Elaboración propia

5.-Generación de reportes: Se ejecuta por requerimiento del usuario, el proceso contempla la recolección de los datos necesarios para la aplicación del minado de datos y la entrega de la información requerida.

En el diagrama 3.12, se presenta el DFD de nivel 1 para este caso de uso.

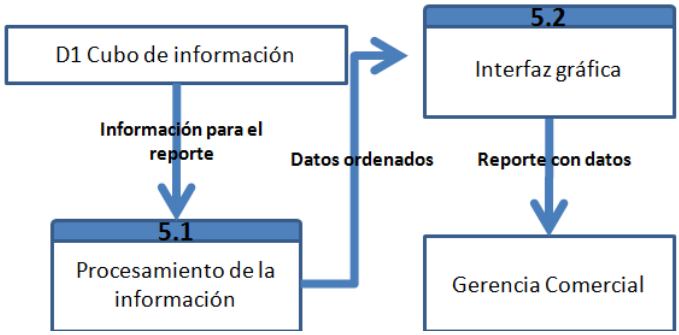


Figura 3.12: Diagrama DFD de nivel 1 del caso de uso de ETL. Fuente: Elaboración propia

Requerimientos de Mecanización: Se requiere un sistema que permita la realización de cada uno de estos casos de uso.

Por un lado los procesos automáticos, ETL y carga de los datos deberán desarrollarse internamente, por otro lado se requiere de una interfaz gráfica con el usuario para la realización de los procesos como agrupaciones de tiendas, propuestas de surtidos por tiendas y de generación de reportes.

c) **Requerimientos de Mecanización de distribución**: El desarrollo de los requerimientos de distribución resulta sencillo dada la centralización tanto de los datos como de los usuarios.

Modelo conceptual de distribución: A pesar de que los datos se generan en cada una de las tiendas, estos son finalmente almacenados centralmente en las oficinas de Easy S.A.

Por otro lado todos los clientes del sistema de apoyo a la toma de decisiones se encuentran físicamente en las oficinas centrales de Easy S.A.

Dado lo expuesto la distribución del sistema debe ser centralizado, pues disminuye costos de implementación y tiempos de ejecución tanto en la recopilación de la información como en consulta al sistema.

Modelo lógico de distribución: El sistema deberá ser una red entre los computadores clientes y un servidor central que recopile la información, realice los procesos especificados y permita la visualización de los reportes.

Modelo físico de distribución: Se deberá generar una red entre los clientes y el servidor del sistema que permita el acceso a la información y la consulta de reportes a los usuarios.

Requerimientos de Mecanización: Se requiere de una red, la que deberá conectarse con la red principal de Cencosud, para la obtención de la información relevante al sistema de apoyo a la toma de decisiones.

4. DISEÑO Y CONSTRUCCIÓN DE LA SOLUCIÓN

El presente capítulo presenta el diseño de la solución propuesta en el capítulo 3, además de describir técnicamente la implementación realizada.

Como solución se propone la creación de un sistema mecanizado para: 1) La recolección de datos relevante sobre ventas, su procesamiento y final almacenamiento en un *data warehouse* y 2) La aplicación de herramientas de *data mining* para la generación de reportes atingentes a las problemáticas descritas en el capítulo 3[46].

La figura 4.1 esquematiza la solución propuesta.

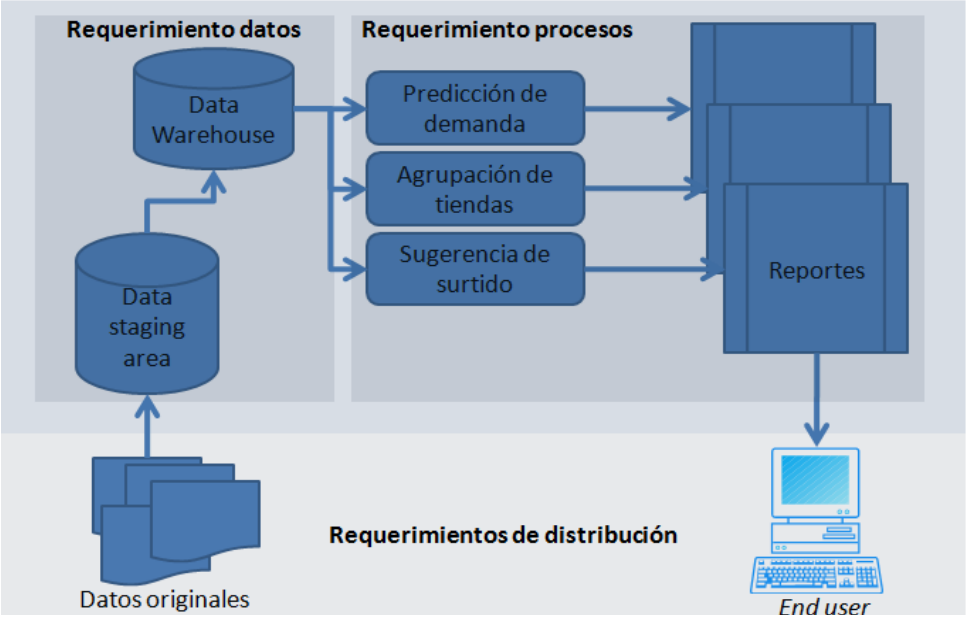


Figura 4.1: Esquema de la solución de la solución propuesta. Fuente: Elaboración propia

A continuación, se explicará la construcción de la solución presentada y su implementación.

4.1 Requerimientos de distribución

Se distingue, por un lado, el diseño de la solución y por otro la implementación física de ésta.

4.1.1 Diseño de la solución: Como se explicó en el punto 3.7.3, se propone una distribución centralizada de los procesos y las herramientas físicas en las oficinas corporativas de Easy S.A. ya que:

- Las actividades concernientes a la gerencia de procesos comerciales ocurren en las oficinas corporativas de Easy S.A.

[46] Capítulo 3 punto 3.2

- Los datos se encuentran almacenados en servidores ubicados en las oficinas corporativas de Easy S.A.
- Los usuarios del sistema se encuentran físicamente en las oficinas centrales de Easy S.A.

Se propone la generación de una red pequeña, de tres computadores, para la implementación de la solución: un computador que cumpla como servidor de los otros dos que serán las terminales que utilizarán los clientes del sistema.

La figura 4.2 presenta la arquitectura de la solución y su relación con la red principal de Easy S.A.

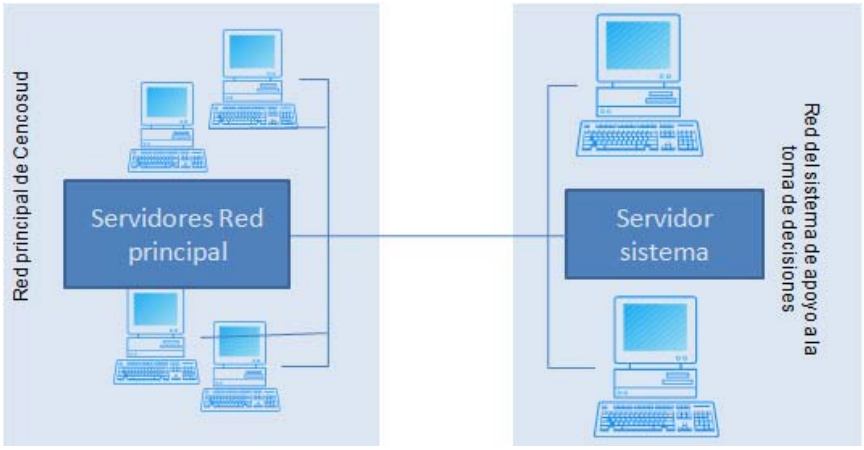


Figura 4.2: Diseño de la solución a los requerimientos de distribución. Fuente: Elaboración propia.

4.1.2 Implementación de la solución: A continuación se presentan los requerimientos para la construcción de la red a crear y de hardware y software para los equipos a utilizar.

Considerando que Cencosud cuenta con técnicos para la realización de la red se necesitarán como insumos en la construcción de la red[47]:

- s*witch* para la conexión de los equipos y la generación de la red.
- *router* para la conexión entre la pequeña red creada y la red principal de Cencosud, permitiendo acceso del servidor a las bases de datos para los procesos de actualización de datos.
- Costo de la adquisición de cables para la realización de las conexiones.

No existen requerimientos específicos para los equipos clientes del servidor excepto la instalación de Java y el servidor Tomcat para la realización de las consultas.

Para el servidor que alojará al *data warehouse y* realizará las operaciones de ETL y minado de información, se recomienda la adquisición de un computador las siguientes características enunciadas a continuación:

[47] Ver Anexo D “Cotización de equipos para la realización de la red”

Software: Las especificaciones consideran el soporte para el funcionamiento de los procesos diseñados.

- Motor de base de datos: MySQL
- Servidor OLAP: Mondrian
- Servidor Web: Apache y Tomcat
- Software de programación: Java, Python, R, C

Hardware: Las especificaciones presentadas consideran el funcionamiento del software principalmente

- Fuentes de poder: Redundantes, para evitar cortes de servicio por cortes de energía eléctrica.
- Memoria RAM: 1GB
- Procesador: 2 GHz
- Memoria en disco duro: 100 GB

La estimación del espacio en disco responde a las siguientes consideraciones:
1.- La base de datos cargada para los 2 años de información que se tienen para la sección 13 ocupa 1 GB (incluidos el DSA y el *Data warehouse*), lo que implica 0,5 GB por sección por año
2.- Existen 24 categorías.
3.- La velocidad de generación de información es estable dado que es información diaría.
4.- El sistema operativo requiere de 2GB aproximadamente.
5.- Se considera un 15% de holgura sobre el requerimiento de memoria.

Con lo anterior se calcula un disco duro de 100 GB para la posible expansión del *Data warehouse* al resto de las secciones de productos [48] más el flujo de información de los próximos 5 años.

Estas características responden a requerimientos de continuidad en el funcionamiento, tanto para las funciones de servir a los clientes como para el correcto funcionamiento de los procesos de ETL y minado de datos.

4.2 Requerimientos de mecanización de datos

En ésta sección, se introducirá tanto el diseño de la solución como su implementación.

4.2.1 Diseño de la solución

La solución diseñada se esquematiza en la figura 4.3 y contempla dos bases de datos: 1) la base de datos del DSA y 2) la base de datos para la realización del *data warehouse*[49].

[48] Son 24 secciones en total

La figura 4.3 muestra ambas bases de datos, su interrelación y el flujo de información existente entre las distintas tablas.

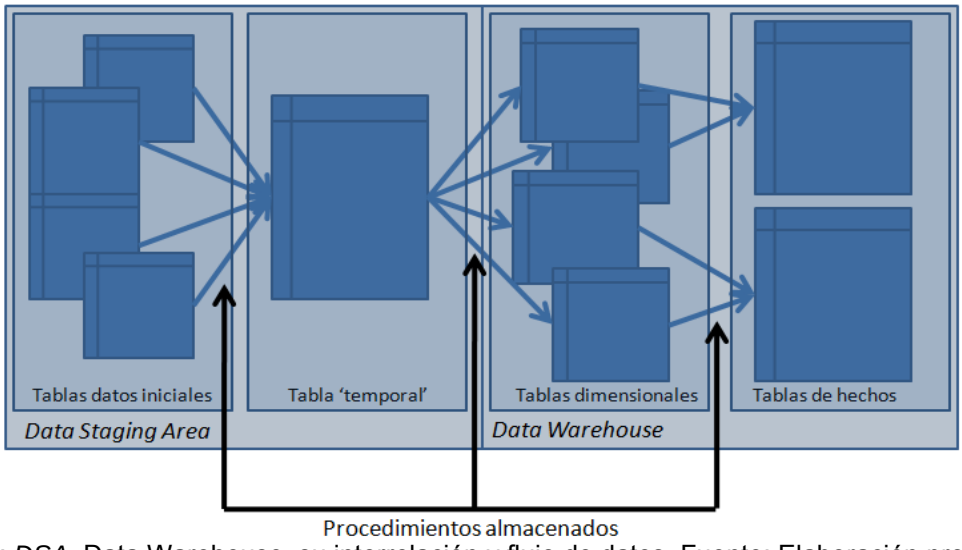


Figura 4.3: *DSA*, Data Warehouse, su interrelación y flujo de datos. Fuente: Elaboración propia

Como se observa en la figura 4.3, las bases de datos se encuentran compuestas por:

- Data Staging Area, compuesta por las tablas con datos originales[50] y la tabla temporal de precarga a las *fact table*.

  En la figura 4.4 se observa el diseño del DSA, donde se nota que cada tabla se encuentra aislada relacionalmente una de otra esto pues los datos que poseen cada una de estas tablas son completamente independiente unas de otras por lo que no existe la necesidad de generar un modelo relacional de los datos.

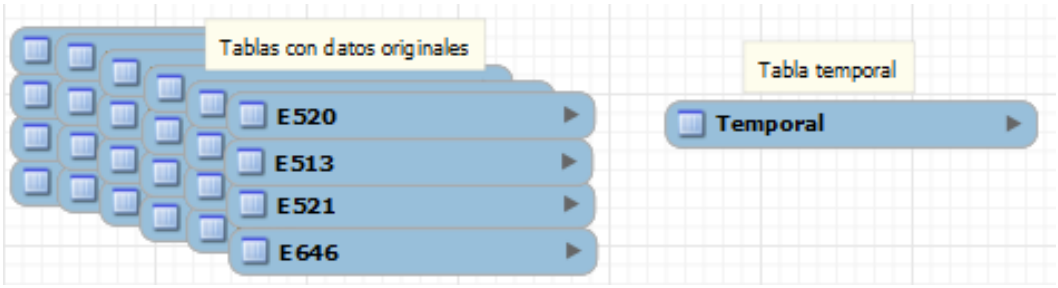


Figura 4.4: Diseño del DSA. Fuente: Elaboración propia

  La tabla temporal se justifica ya que se llenará más de una *fact table*, cada una con diferentes medidas.

[49] Referirse al capítulo 3.7.1 "Requerimientos de mecanización de datos"
[50] Previamente modificadas para tener el formato de tablas diseñado. Referirse al Anexo H Base de Datos

Además de la tabla temporal existen 23 tablas de datos por tienda, cada una con los datos originales de carga para cada tienda. Se ejemplificarán las tablas de datos originales por una genérica, aprovechando que todas tienes los mismos datos y solo varían en el campo tienda.

A continuación se describen las tablas y sus campos:

Tabla temporal: Repositorio de datos agregados y limpios, fuente de los datos para la carga del *data warehouse*.

- *Campo año*: Campo numérico entero (*integer*) para la identificación del año de registro de la transacción.

- *Campo semana*: Campo numérico entero entre 1 y 52[51] (*integer*) para la identificación de la semana del año en la que fue realizada la transacción.

- *Campo tienda*: Campo no númerico (*varchar(5)*) con el código de tienda en donde se realizó la transacción.

- *Campo Id_producto*: Campo numérico (*integer*) con el código de identificación del producto en la transacción.

- *Campo producto*: Campo no numérico (*varchar(255)*) con la descripción del producto en la transacción.

- *Campo cantidad_vendida*: Campo numérico (*integer*) con la especificación del número de unidades involucradas en la transacción para las semana y tienda correspondiente al registro.

- *Campo monto_vendido*: Campo numérico (FRACCIÓN) con la especificación del dinero involucrado en la transacción para las semana y tienda correspondiente al registro.

Tablas de datos originales: Repositorio de datos desagregados por tienda. Sobre estos datos es que sa aplican los programas para el ETL.

- *Campo año*: Campo numérico entero (integer) para la identificación del año de registro de la transacción.

- *Campo mes:* Campo no numérico (*varchar(3)*) para la identificación del mes de registro de la transacción.

- *Campo día:* Campo numérico entero (*integer*) para la identificación del día de registro de la transacción.

[51] Número de semanas de un año

- *Campo semana*: Campo numérico entero entre 1 y 52[52] (*integer*) para la identificación de la semana del año en la que fue realizada la transacción.

- *Campo tienda*: Campo no númerico (*varchar(5)*) con el código de tienda en donde se realizó la transacción.

- *Campo Id_Producto*: Campo numérico (*integer*) con el código de identificación del producto en la transacción.

- *Campo Producto*: Campo no numérico (*varchar(255)*) con la descripción del producto en la transacción.

- *Campo cantidad_vendida*: Campo numérico (*integer*) con la especificación del número de unidades involucradas en la transacción.

- *Campo monto_vendido*: Campo numérico (FRACCIÓN) con la especificación del dinero involucrado en la transacción.

- Data Warehouse, compuesta por las tablas dimensionales y 3 "fact tables", de las que dos responderán a consultas específicas de datos consolidados y la tercera se utilizará como repositorio de datos utilizados para los procesos de *clusterización* y pronóstico de demanda.

  En la figura 4.5 se muestra la estructura del Data Warehouse propuesto donde se distinguen las tablas dimensionales Tiendas, Producto y Fecha, y las *fact tables* Mining, Predicciones e Histórico.

  Para las tablas dimensionales Tiendas y Productos se propone un modelo "*Snow Flake*"[53], esto debido a que tanto productos como tiendas tienen características que pueden variar con el tiempo sin que esto signifiquen cambios en sus características intrínsecas.
  Para la tabla Producto los atributos categoría, rubro, subrubro y asignación pueden variar en función de necesidades comerciales sin que cambien sus atributos intrínsecos (características física o presentación) o históricos (ventas en dinero y cantidad).

  Para la tabla Tiendas el atributo agrupación puede variar sin que necesariamente cambien sus características físicas (utilización de espacio) ni las características socioeconómicas de las comunas donde se encuentren inmersas.

  Siguiendo la estructura del modelamiento estrella, las tablas que constituyen el *data warehouse* se encuentran relacionadas mediante llaves foráneas, por lo que antes de llenar las *fact tables* es necesario llenar las tablas dimensionales.

[52] Número de semanas de un año
[53] Ver Anexo E "Base de datos"

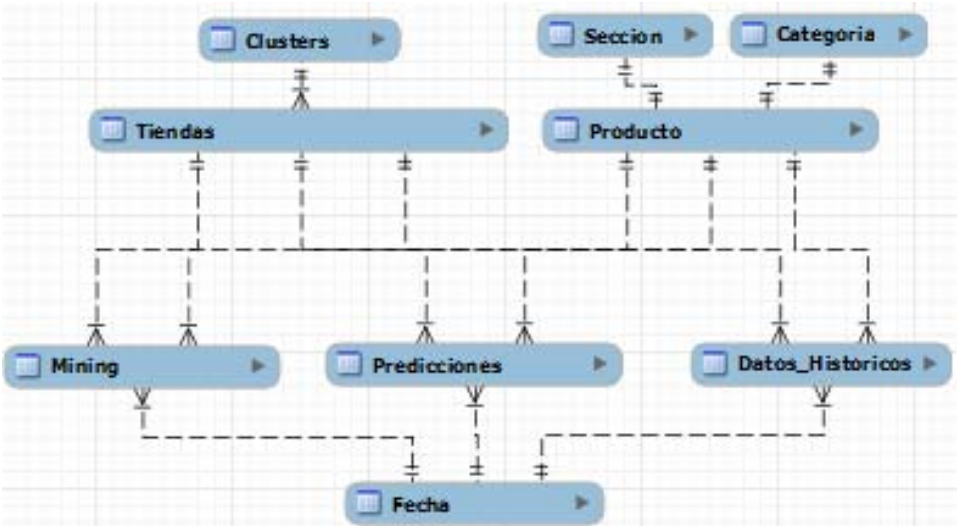


Figura 4.5: Diseño del Data Warehouse. Fuente: Elaboración propia

A continuación, se describen los campos constituyentes de cada tabla dimensional y de las *fact table*[54].

Tabla Fecha: Posee la información calendario de años, meses, día del mes, día de la semana y número de semana del año para los años desde el 2006 al 2013, sus campos son:

- *Campo Id_fecha:* Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo año:* Campo numérico entero (*integer*) para la identificación del año de registro de la transacción.
- *Campo mes:* Campo no numérico (*varchar(3)*) para la identificación del mes de registro de la transacción.
- *Campo día:* Campo numérico entero (*integer*) entre 1 y 31 para la identificación del día del mes en que se registró la transacción.
- *Campo día_semana:* Campo no numérico (*varchar(9)*) para la identificación del día de la seman en que registró la transacción.
- *Campo semana_año*: Campo numérico entero entre 1 y 52 (*integer*) para la identificación de la semana del año en la que fue realizada la transacción.

Existe un problema con la determinación de la semana del año. Dado que una semana tiene 7 días y un año 365 días, se produce un desfase entre el fin de la semana 52 y el fin del año.

Para corregir esto se modifico las semanas 51 y 52 del año 2012 y la semana 1 del año 2013.

[54] Para ver una descripción más detalladas de las tablas del *snow flake* de las tablas dimensionales referirse al Anexo E “Base de datos”.

A la semana 51 se le agregan los días lunes, martes y miércoles de la semana 52 mientras la semana 52 comienza el jueves y terina el sábado 30 del año 2012, finalmente la semana 1 del año 2013 se le agrega el domingo 31 del año 2012.

Estas modificaciones se realizan considerando los días de mayor actividad comercial de modo de no afectar las proyecciones que se realicen con estos datos.

Tabla Tienda: Contiene la información característica de la tienda, con sus datos se realizarán los procedimientos de *clusterización*. Sus campos se describen a continuación.

- *Campo código_tienda:* Campo no numérico (v*archar(4)*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo descripción_tienda*: Campo no numérico (*varchar(255)*) con el nombre alternativo de la tienda.
- *Campo comuna_ubicación*: Campo no numérico (*varchar(255)*) con el nombre de la comuna donde se encuentra ubicada la serie.
- *Campo GSE:* Son 5 campos numéricos enteros (*integer*), los que tienen el número de hogares pertenecientes al GSE 1, GSE 2, GSE 3, GSE4 y GSE 5 para la ubicación de la tienda[55].
- *Campo espacio utilizado por categoría*: 47 campos numéricos fraccionarios (*integer*) con la información de metros cuadrados ocupados por categoría para cada tienda[62].
- *Campo id_mix:* Campo numérico entero con el número el identificador del cluster al que pertenece la tienda[62].

Tabla Producto: Contiene la información característica de cada producto catalogado, con sus datos se realizarán los procedimientos de predicción de demanda y *clusterización*. Sus campos se describen a continuación.

- *Campo código_producto:* Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo descripción_producto*: Campo no numérico (*varchar(255)*) con el nombre alternativo de la tienda.
- *Campo sección id_seccion*: Campo numérico (*integer*) con el identificador de la sección a la que pertenece nombre de la comuna donde se encuentra ubicada la serie.
- *Campo categoria id_categoría*: Campo numérico (*integer*) con el identificador de la sección a la que pertenece nombre de la comuna donde se encuentra ubicada la serie.
- *Campo id_cluster*: Campos numérico entero (*integer*) con la información de asignación del producto a *cluster* de tiendas.

[55] Ver Anexo E “Base de datos”

Tabla Historico: *Fact table* con la información histórica de ventas para cada producto, por tienda y para cada fecha pasada. Sus campos se describen a continuación:

- *Campo id_historico:* Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo fecha id_fecha*: Campo numérico (*integer*) identificador de la fecha en la que se realizó la transacción, llave foránea de la tabla Fecha.
- *Campo codigo_tienda*: Campo no numérico (*varchar(5)*) identificador de la tienda en la que se realizó la transacción, llave foránea de la tabla Tiendas.
- *Campo producto id_producto*: Campo numérico (*integer*) identificador el producto para el que se realizó la transacción, llave foránea de la tabla Producto.
- *Campo cantidad_vendida*: Campo numérico (*integer*) con la especificación del número de unidades involucradas en la transacción registrada.
- *Campo monto_vendido*: Campo numérico (FRACCIÓN) con la especificación del dinero involucrado en la transacción registrada.

Tabla Prediccion: *Fact table* con la información de predicción de demanda realizada para los próximos 8 períodos para cada producto, por tienda y para cada fecha pasada. Sus campos se describen a continuación:

- *Campo id_prediccion:* Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo fecha id_fecha*: Campo numérico (*integer*) identificador de la fecha para la que se realizó la predicción, llave foránea de la tabla Fecha.
- *Campo codigo_tienda*: Campo no numérico (*varchar(5)*) identificador de la tienda para la que se realizó la predicción, llave foránea de la tabla Tiendas.
- *Campo producto id_producto*: Campo numérico (*integer*) identificador el producto para el que se realizó la predicción, llave foránea de la tabla Producto.
- *Campo pronostico_cantidad*: Campo numérico (*integer*) con la especificación del número de unidades involucradas en la transacción registrada.
- *Campo pronostico_monto*: Campo numérico (FRACCIÓN) con la especificación del dinero involucrado en la transacción registrada.

Tabla Mining: *Fact table* con la información sobre indicadores relevantes en la realización de predicciones para cada producto, por tienda y para cada fecha pasada. Sus campos se describen a continuación:

- *Campo id_mining*: Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.
- *Campo fecha id_fecha*: Campo numérico (*integer*) identificador de la fecha para la que se realizó la predicción, llave foránea de la tabla Fecha.

- *Campo codig_tienda*: Campo no numérico (*varchar(5)*) identificador de la tienda para la que se realizó la predicción, llave foránea de la tabla Tiendas.
- *Campo producto id_producto*: Campo numérico (*integer*) identificador el producto para el que se realizó la predicción, llave foránea de la tabla Producto.
- *Campo promedio*: Campo numérico (*integer*) con la especificación del número de unidades promedio vendidas en la semana correspondiente por producto y tienda.
- *Campo tendencia*: Campo numérico (FRACCIÓN) con la especificación de la tendencia de ventas observada para los datos históricos, por tienda y producto.

4.2.2 Propuesta de Implementación de la solución

Se distingue por un lado la implementación física de los dispositivos para el almacenamiento de datos y la arquitectura de estos y por otro los procesos que permitan el flujo de los datos entre las distintas componentes de la BD diseñada como solución.

a. Implementación física: Debido a que el sistema descrito en la figura 4.5 como solución es un sistema de almacenamiento de datos, se empleará el modelo de datos relacional para la construcción de esta base de datos.

Dentro de las ventajas que presenta el modelo relacional para el almacenamiento de datos, se puede mencionar:

- Sistema eficiente de indexación de registros.
- Independencia entre datos y tratamiento de los mismos.

Como motor de bases de datos se empleará MySQL, esto pues:

- Soporta bases de datos relacionales.
- Posee procedimientos almacenados, los que facilitan la administración y transformación de los datos.
- Dado que es un *freeware*, no aporta costos a la realización del prototipo.
- La experiencia del autor del trabajo, lo que ahorra tiempos en el aprendizaje de uso de otro motor de base de datos.

b. Implementación de procesos para la mecanización de procesos: En la figura 4.3 se aprecia en la que dentro de la solución se encuentra el flujo de datos ordenados entre las distintas tablas de la BD diseñada.

Se propone, en términos generales, la utilización de programas para el procesamiento de datos y de procedimientos almacenados para la realización de consultas y cargas de datos.

Esto para aprovechar las ventajas en indexación de las bases de datos y de procesamiento de datos de los archivos.

Los procesos necesarios para la mecanización de datos son: 1) Carga de datos desde archivos .csv hacia las tablas del DSA, 2) Transformación y carga de los datos desde las tablas de datos del DSA hacia la tabla temporal del DSA y 3) Carga final de los datos desde la tabla temporal del DSA hacia el *data warehouse.*

Carga de datos desde archivos .csv hacia tablas del DSA: Se propone la utilización de un programa simple, escrito en R, que cumpla con la tarea de ser un canal de paso entre los archivos escritos en csv. y las tablas de datos del DSA.

Esto facilita la mecanización del proceso ya que permite su autoejecución mediante un archivo .bat y un *trigger* temporal.

Transformación y carga de datos desde las tablas de datos del DSA hacia la tabla temporal del DSA: Se propone la utilización de un programa simple, escrito en R, que automatice el proceso de limpieza, explicado en el punto 4.3.1, y que realice la agregación semanal de datos por productos y por tiendas. Se propone la utilización de un programa en R y no la aplicación inmediata de un procedimiento almacenado, ya que el trabajo con archivos disminuirá los tiempos de ejecución.

Con este procedimiento se imita una carga por bash lo que hace el proceso más robusto y disminuye la posibilidad de fallo general del proceso, al ser una carga y procesamiento parcelado de los datos.

Otra ventaja de utilizar un programa escrito en R, es que se puede procesar el pronóstico de ventas inmediatamente en esta etapa, mediante un programa escrito en R descrito en el punto 4.3.2.

Carga final de los datos desde la tabla temporal del DSA hacia las "*fact tables*" del *data warehouse*: Se propone la utilización de un procedimiento almacenado, el que mediante un *trigger* temporal, realice la carga desde la tabla temporal del DSA hacia las tablas del *data warehouse*.

Un procedimiento almacenado se hace pertinente, ya que este proceso de carga comprende solo funciones de consulta de datos e inserciones, sin procesamiento de datos.

4.3 Solución de requerimientos de mecanización de procesos

En ésta sección se introducirá tanto el diseño de la solución como la propuesta de implementación de esta para cada uno de los procesos identificados en el punto 3.7.2, además de la implementación realizada.

4.3.1. Proceso de ETL

a) **Diseño de la solución,** este proceso fue diseñado en 5 etapas consecutivas, las que comprenden desde la recopilación de los datos hasta la carga final al data warehouse.

Estas etapas son:

- Entrega de formato para el manejo de los datos
- Carga de los datos al DSA
- Limpieza de los datos
- Transformación de los datos
- Carga de los datos al Data warehouse

Entrega de formato para el manejo de los datos: Los datos fueron obtenidos originalmente en formato Excel, como resultado de una consulta OLAP al cubo "ventas" del sistema informático de Easy[56]. El resultado de la consulta se presenta en la figura 4.2.

Estos datos fueron tratados en dos etapas sucesivas: 1) Eliminación de información no relevante y 2) Entrega de formato para la carga al DSA.

La consulta OLAP arroja mucha información irrelevante para el proceso de ETL: etiquetas con información sobre los datos (Rótulos de fila, semestre, trimestre), agregación jerárquica de los productos y resúmenes de información agregada mensual, trimestral, semestral y anual.

La figura 4.6. muestra los datos en la forma como son obtenidos.

1	Tip Locales	**Información sobre la tienda**		E503					
2									
3				Rótulos de columna	**Fecha de la información**				
4				2007					
5				Q1, 2007					
6				Jan, Q1					
7				2, Martes, Jan		8, Lunes, Jan		9, Martes, Jan	11, Jueves,
8	Rótulos de fila	**Información**		Q Act	Vta Act	Q Act	Vta Act	Q Act	Vta Act	Q Act
9	13 * Sección Ferreteria	**sobre SKU**		4	4	-1	-12	3	14	
10	1301 * Rubro Herramientas Manuales			4	4	-1	-12	3	14	
11	130101 * Subrubro Destornilladores							2	1	
12	130101001 * Dests/aisl.ppl/palet					**Información sobre ventas:**		2	1	
13	194475 * ACEITERA TUBOFLEX 150CC SAMOA					**Cantidad y monto**				
14	194476 * ACEITERA TUBOFLEX 300CC SAMOA									
15	194478 * LAPIZ CARPINTERO BICOLOR HEX.							2	1	
16	194479 * LAPIZ CARPINTERO CORRIENTE									
17	194690 * DESTORNILLADOR PLANO 3/16"" X 6"" STANLEY									
18	195235 * GRADA ESC.COMPLETA 1.20NEGRO									
19	195398 * TECLE PLUMA CAPACIDAD 2TONELADAS									
20	404858 * CORTADORA MANUAL 46CM SANSON QEP									
21	404865 * RPSTO P/CORTAD CLINKER C/CAR QEP									

Figura 4.6: Archivo CSV para la carga final al DSA. Fuente: Elaboración propia, basado en información de la gerencia de procesos comerciales

[56] Para más detalles del sistema referirse al punto capítulo 3 punto 3.6.

A continuación se necesita entregar el formato de columnas definidas para las tablas en el DSA, esto es la información que se muestra en la figura 4.6 debe ser transformada a la estructura que se observa en la figura 4.7 es decir: ID, código de tienda, código de producto, nombre del producto, año, mes, día de la semana, día del mes, cantidad vendida y dinero vendido.

```
688621;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Jueves;11;;
688622;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Viernes;12;;
688623;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Sabado;13;;
688624;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Domingo;14;;
688625;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Lunes;15;;
688626;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Martes;16;;
688627;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Miercoles;17;;
688628;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Jueves;18;;
688629;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Viernes;19;;
688630;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Sabado;20;;
688631;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Domingo;21;;
688632;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Lunes;22;;
688633;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Martes;23;;
688634;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Miercoles;24;;
688635;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Jueves;25;;
688636;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Viernes;26;;
688637;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Sabado;27;;
688638;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Domingo;28;;
688639;E502; 821192; NAVAJA ABSAROKA 2.3/4'' 12855 SHEFFIELD;2007; Jan;Lunes;29;;
```

Figura 4.7: Archivo CSV para la carga final al *DSA.* Fuente: Elaboración propia

La solución fue implementada, también en dos pasos. Primero se aplicó una Macros para limpiar los datos irrelevantes antes mencionados y luego un programa escrito en Python para la entrega del formato descrito en la figura 4.7 a los datos.

Como resultado final se obtuvo un archivo de texto plano, CSV, con 10 campos: id, tienda, código de artículo, descripción del artículo, año, mes, dia de la semana, dia, cantidad vendida y monto vendido.

Carga de los datos al DSA: Los datos son luego extraídos desde el archivo .csv para ser cargados al DSA a la tabla de la tienda correspondiente, previo al paso de limpieza.

La carga en las tablas del DSA se justifica como medio de respaldo de la información obtenida inicialmente.

En primera instancia se realizó este proceso manualmente, mediante el software MySQL, utilizando la función de importación de datos a través de un archivo CSV.

Debido al tamaño de los archivos a importar[57] fue necesario modificar el archivo php.ini. De esta forma se aumento el tamaño desde 2 MB a 100MB.

Si bien la carga se realizó manualmente en esta primera ocasión, la propuesta de automatización se encuentra especificada en el punto 4.2 mediante un programa escrito en R[58].

[57] Ver Tabla 4.1 “Archivos involucrados en el proceso de entrega de formato a los datos“
[58] Referirse al punto 4.2.2.b “Implementación de procesos para la mecanización de datos”

Cabe mencionar que además de permitir la mecanización del proceso se solucionan los problemas que presenta php de tiempo máximo de duración de una transacción y de tamaño máxima de carga en la actualización de datos, dado que la herramienta de carga de MySQL es una interfaz gráfica escrita en lenguaje Php.

Limpieza de datos: Se diseño un proceso mediante el cual los datos son limpiados de outliers, datos faltantes y quiebres de stock. A continuación, se describen los criterios utilizados para realizar estas tareas.

Detección de outlayers: Se consideró que cualquier registro por sobre dos o bajo dos desviaciones estándar con respecto al promedio de la serie de valores normalizados (18), constituye un error en la serie por lo que es eliminado y reemplazado por el promedio de la serie.

Con respecto a los datos faltantes se revisan dos condiciones: 1) si existe registro de cantidad vendida, entonces debe existir registro de monto de transacción y 2) si existe registro de monto de transacción, debe existir registro de cantidad vendida.

Para la corrección de los datos faltantes, se calcula un precio unitario promedio, correspondiente al cociente entre la suma del monto vendido de todos los registros completos[59] y la suma de todas las unidades vendidas de los registros completos[10].

Luego el dato faltante es calculado de la siguiente manera:

- Si el registro faltante es el de monto de transacción, este se reemplaza por el resultado de la multiplicación del registro de unidades vendidas por el precio unitario vendido.
- Si el registro faltante es el de unidades vendidas, este se reemplaza por el cociente entre el registro de monto de transacción y el precio unitario promedio.

Se decide considerar un quiebre de stock cada vez exista un registro de unidades vendidas igual a cero, entre dos registros mayores de cero antes y después del registro igual a cero.

El quiebre de stock será reemplazado por el promedio de unidades vendidas del día anterior y el día consecutivo al quiebre de stock y el valor del monto de transacción será igual al número de unidades vendidas calculado multiplicado por el precio unitario promedio.

Para la implementación de esta limpieza se propone la utilización de un programa escrito en R, explicado en el punto 4.2.2.b que implemente las operaciones de limpieza descritas en este punto.

Transformación de datos: Los datos de venta en dinero y cantidad fueron agregados a nivel semanal. Esto debido al grano requerido por el warehouse[60].

[59] E considera un registro completo cuando se cuenta con el valor de unidades vendidas y el de monto de transacción simultáneamente.

Para la realización de esta agregación se propone un programa escrito en R, el que realice consultas por datos de venta y cantidad vendida para un producto y una tienda especifica.

De esta forma se asegura la mayor velocidad en la manipulación de la información y se aprovecha la capacidad de indexación de la BD.

Carga de datos al *Data warehouse*[61]: Se propone la utilización de procedimientos almacenados ya que solo se comprenden consultas y cargas a la base de datos, sin manipulación de datos.

4.3.2. Proceso de generación de pronóstico de demanda

Se distingue, por un lado, el diseño del proceso de predicción de demanda y por otro lado la implementación de los programas y la carga de estos pronósticos al *Data warehouse*.

a. **Diseño del proceso:** Para la generación de pronósticos se implementa un proceso que consta de dos etapas:1) Consulta de los datos relevantes y 2) Realización de la predicción y carga de los resultados al *Data warehouse.*

Como fue explicado en el punto 4.2.2.b, este proceso se realizará en la etapa de carga de datos hacia la tabla temporal en el DSA, lo que presenta varias ventajas:

- No se necesita de otro paso para la realización de la predicción, por lo que queda completamente automatizado y aislado del usuario.
- Esta etapa se realiza por producto/tienda por lo que no se necesita de una consulta especial para la realización de la predicción. Lo que disminuye el tiempo total de predicción más carga de datos.

En la figura 4.8 se esquematiza la realización de este proceso en el sistema diseñado.

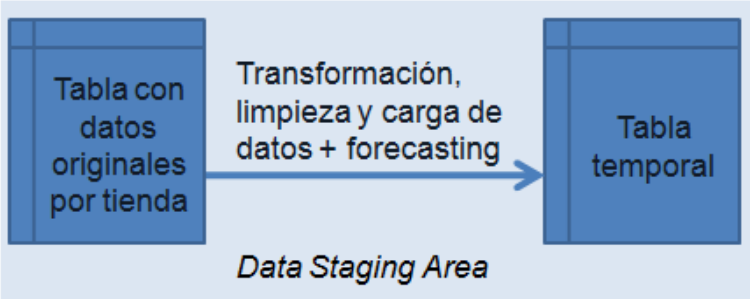


Figura 4.8. Proceso de generación de predicción de demanda. Fuente: Elaboración propia

b. **Implementación de la solución:** Para la realización de este proceso se utilizó el software R, en particular las bibliotecas *nnet* y *tseries* para la generación de pronósticos mediante redes neuronales artificiales, ARIMA y medias móviles.

---

[60] Ver Capítulo 3.7.1 “Requerimientos de mecanización de datos“
[61] Para mayor detalle referirse al punto 4.2.2.b

Se construye una subrutina en el programa de ETL, el que recibe como *input* una serie de datos a pronosticar y aplica el siguiente algoritmo para generar predicciones.

- Estimación de tendencia: La serie es removida de su tendencia, la cual es almacenada como información para la *fact table* Mining.
- Pronóstico mediante redes neuronales artificiales: Se prueba una serie de redes neuronales, dependiendo del número de rezagos correlacionados significativamente con el último dato de la serie[62].
  Se guarda la predicción que presente mejor comportamiento según parámetros a especificar en los experimentos.
- Pronóstico mediante modelo ARIMA: Se prueba una serie de modelos ARIMA, elegidos automáticamente[71], previa eliminación de tendencia y aplicación de una transformación logarítmica para la estabilización de la varianza.
  Se guarda el que presente mejor comportamiento según parámetros a especificar en los experimentos.
- Pronóstico mediante modelo medias móviles: Se prueba una serie de predicciones con medias móviles, previa eliminación de tendencia y aplicación de una transformación logarítmica de los datos para la estabilización de la serie.
- Se comparan los mejores resultados de pronóstico realizado por cada uno de los métodos antes descritos y se selecciona el que presente mejor ajuste según los indicadores de errores[63].
- Se entregan las predicciones como un vector de valores para las próximas 8 semanas.

Para ajustar una red neuronal, se utiliza la biblioteca *nnet*, en particular el comando *nnet(),* el que ajusta una red neuronal artificial de una capa oculta a un set de datos entregados, luego el método *predict()* realiza una predicción de valores para la serie entregada con los parámetros entregados por el método *nnet()* a un intervalo de tiempo entregado también como parámetro

Para ajustar un modelo ARIMA, se utiliza la biblioteca tseries, en particular el comando *arima()*, el que ajusta un proceso arima con p rezagos para el proceso AR y q rezagos para le proceso MA .

La predicción mediante medias móviles se realiza utilizando las funcionalidades básicas del software R.

La carga de datos final se realizará mediante una consulta directa a la base de datos, desde el programa de ETL.

Los resultados de la actuación de los distintos métodos frente a los datos pueden verse en el capítulo 5 “Experimentos y resultados”.

[62] Ver Anexo F “Programación”
[63] El método a utilizar se especificará mediante los experimentos en el capítulo 5

4.3.3. Proceso de agrupación de tiendas

Se distingue por un lado el diseño del proceso de agrupación de tiendas y por otro lado la implementación de los programas y la carga de estos pronósticos al *Data warehouse.*

a. **Diseño del proceso:** Este proceso se realizará a "pedido", es decir, a través de una consulta directa del usuario.

   La solución final se obtendrá mediante la interacción del experto en el negocio, con la ayuda del sistema el que entregará soluciones posibles para la agrupación de las tiendas según las variables seleccionadas, descritas en la la figura 4.9.

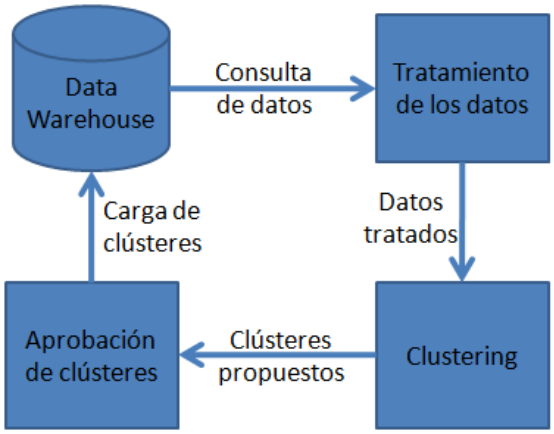


Figura 4.9: Proceso de generación de agrupaciones. Fuente: Elaboración propia

Este proceso consta de cuatro etapas:1) Consulta de los datos relevantes 2) Tratamiento de los datos, 3) Realización de las agrupaciones y 4) Aprobación de la propuesta y carga de los resultados al Data Warehouse.

El algoritmo para la ejecución del proceso se describe a continuación:

- Consulta de datos al Data Warehouse (primera etapa del proceso)
- Tratamiento de los datos (segunda etapa del proceso).
- Obtención del número de clusters mediante la aplicación de SOFM: Se presentan los resultados obtenidos mediante SOFM de manera de identificar los posibles clusters en el plano (tercera etapa del proceso).
- Agrupaciones mediante k-means: Se realizan un número de agrupaciones con k means, fijando el número de clusters obtenido en el paso anterior.
- Se rescatan las mejores soluciones obtenidas en base al indicador (distancia intragrupos)/(distancia intergrupos)[64] (tercera etapa del proceso).
- Estos resultados son luego presentados al experto en el negocio, junto con los criterios de validez de clusters[65] (tercera etapa del proceso).

[64] Ver Capítulo 2 punto 2.3.3.3
[65] Ver Anexo F "Programación"

- El experto en el negocio, elige la agrupación que encuentre pertinente y se genera un archivo con la clusterización (cuarta etapa del proceso).
- Se carga la agrupación escogida al Data Ware house (cuarta etapa del proceso).

b. Implementación de la solución: Para la realización de este proceso se utilizó el *software* R, en particular la biblioteca *RODBC* para la consulta de datos al Warehouse y las bibliotecas *kohonen*, *SOM* y *Cluster* para la implementación de los algoritmos de clustering k-means[66] y SOFM[67] . La carga de datos final se realizó mediante una consulta directa a la base de datos, realizada con una rutina en R.

Para la realización de las agrupaciones se realizó un programa escrito en R que consta de 4 etapas:

Consulta de los datos relevantes: La selección de los datos para la realización de agrupaciones se realiza mediante una interfaz gráfica, en la que se puede escoger de entre las siguientes variables para la realización de agrupaciones:

- Características físicas de las tiendas: Metros cuadrados utilizados en sala por cada categoría, para cada tienda.
- Características socioeconómicas: Ingreso promedio y GSE de las comunas en donde se encuentra cada tienda.
- Características comerciales: Ventas para cada tienda desagregadas por categorías.

Estas consultas arrojarán un archivo .csv, el que servirá para alimentar los programas de realización de clusterización.

Tratamiento de los datos, normalización y reducción: En esta primera etapa los datos descriptivos de las observaciones son normalizados, tanto por variables como por vector de características, y luego reducidos en su dimensionalidad.

En primera instancia se realiza una normalización por variables para luego normalizar cada observación por separado.

La normalización por variables se realiza para evitar que aquellas variables con mayores rangos de valores resten importancia a aquellas que presentan menor rango (16). A continuación se realiza una normalización observación, es decir, se normaliza el vector de características de cada observación, de esta forma se busca mantener las relaciones existentes entre atributos para cada observación.

El siguiente paso es la disminución de dimensiones. Esto es realizado mediante un análisis de componentes principales (ACP) para eliminar variables redundantes y distorsiones generadas por mediciones que entreguen poca diferencia.

[66] Ver Capítulo 2 sección 2.3.3.1
[67] Ver Capítulo 2 sección 2.3.3.2

Como consenso para el ACP se eligió utilizar las componentes principales que expliquen el 80% de la varianza de entre las observaciones o las primeras dos componentes principales en el caso de solo una observación explique más del 80% de la varianza.

Como salida de esta consulta asistida, se obtiene un conjunto de nuevas variables de características para cada observación (tienda).

Cálculo de clusters: En última instancia, la elección de clusters depende del conocimiento del experto de negocios. Es por esto que más que buscar una clusterización única de las tiendas, se pretende entregar un set de buenas opciones de agrupación de tiendas, de modo que el papel del experto en la elección sea asistida, no reemplazada.

Los puntos críticos en la realización de agrupaciones son la determinación del número óptimo de clusters y la realización de las agrupaciones.

Se propone el uso de dos métodos de clusterización: *Self Organizing Feature Maps* y *kmeans*.

Para la determinación del número óptimo de clusters se presentan los resultados del SOFM frente a los variables características de cada observación.

La representación gráfica que entrega el SOFM resulta fácilmente interpretable por el experto en negocio, quien puede determinar con mayor facilidad el número de clusters óptimo según sus restricciones particulares.

Desde el prisma comercial y logístico surgen restricciones que determinan el número de clusters con sentido desde el punto de vista del negocio. Estas limitaciones responden tanto a razones de manejo operacional como a razones de estrategia comercial.

Luego de determinado el número óptimo de clusters (*k*) se realiza el cálculo de *clusters* por medio de *kmeans.*

Se realizan agrupaciones con el número de clusters determinados en el punto (*k*) anterior más dos (*k+2*) y menos dos (*k-2*) con el objeto dejar una holgura a la estimación realizada.

Esto con el objeto de comparar los resultados obtenidos y determinar la clusterización que entregue mejores resultados, a partir de un juicio experto.

Los clusters son comparados entre sí mediante el siguiente indicador:

$$\frac{Varianza\ intragrupo}{Varianza\ intergrupo} \qquad (4.1)$$

Un *cluster* será mejor que otro en la medida que minimice el indicador[68] presentado en la ecuación 4.1.

Aprobación de clústeres: Luego de realizado el proceso de *clusterización* para cada número de potenciales *clusters*[69], se presentan los siguientes resultados al experto:

- La mejor solución obtenida para cada número posible clusters.
- El indicador 4.1 para cada solución encontrada.
- Correlación entre las variables descriptivas por cluster por cada solución.

El resultado de la agrupación es luego almacenado como un archivo separado por comas (.CSV) es que es luego importado al *data warehouse* una vez validados.

4.3.4. Proceso de generación de reportes

Se distingue, por un lado, el diseño del proceso de generación de reportes y por otro lado la implementación de las interfaces que permitan la visualización de los reportes.

a. **Diseño del proceso:** Este proceso se realizará a "pedido", es decir, a través de una consulta directa del usuario desde una interfaz gráfica, cuyo diseño se encuentra en la figura 4.10.

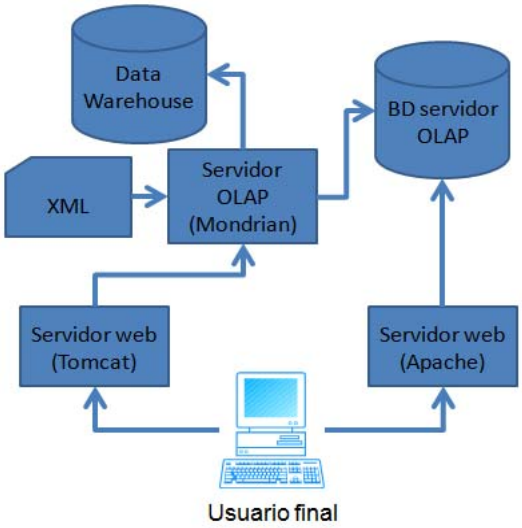


Figura 4.10 arquitectura de la generación de reportes. Fuente: Elaboración propia

A continuación se describen las partes del diseño de la generación de reportes:

Base de datos del Data Warehouse: Contiene la información necesaria para la generación de los reportes. Se une con el servidor OLAP mediante una conexión directa con la base de datos.

[68] Ver capítulo 2 sección 2.3.3.3
[69] Entre k-2 y k+2, con k el número de clusters identificado por el experto

Base de datos del servidor OLAP: Contiene la configuración del servidor. Tipos de reportes, consultas, puertos de conexión, etc.

Archivo .xml: Archivo con la estructura del modelo estrella. Mediante esta información el servidor OLAP es capaz de interpretar las consultas realizadas al Data Warehouse.

Servidor Web para reportes: Su principal función es la generación de la interfaz grafica para la visualización de la consulta OLAP.

Servidor web para la realización de las consultas: Su función principal es la interpretación y ejecución de comandos para la modificación de las bases de datos.

Servidor OLAP: Es la pieza central en la realización de las consultas. Su función es la de recibir la consulta del usuario y realizarla al Data Warehouse, en el contexto de la estructura del modelo estrella, para luego entregar el resultado al servidor web de reportes.

Interfaz gráfica: Tiene la función de reunir toda la información recolectada por medio de los servidores web y organizarla como una vista coherente con los requerimientos del usuario.

Se definió con el cliente la realización de dos reportes[70]: 1) Reporte de pronóstico de demanda, 2) Reporte de agrupación de

Mediante esta interfaz gráfica es posible la especificación de estos parámetros dinámicos en la consulta mdx.

b. Implementación de la solución: A continuación se describe la implementación de la solución en cuanto a software y las principales características y opciones de los distintos reportes a realizar.

Se escogió el software Jasper Server, definido como un "servidor de reportes"[71], como la herramienta principal para el manejo de la información de reportes y de generación de reportes.

Dentro de sus ventajas para este proyecto en particular, se cuenta:

- Reúne en un solo paquete tanto al servidor OLAP como al servidor web para reportes.
- Utiliza Mondrian como servidor OLAP, el que a su vez utiliza MySQL como motor de base de datos, el mismo que el que se utilizando para el almacenamiento de la información.

[70] Ver Anexo H "Entrevistas": " Reunión con Tomás Zavala del 28 de Mayo 2009 "
[71] www.jaspersoft.com

- Al ser un *freeware* no agrega costos a la realización del prototipo.
- La experiencia del realizador de la presente memoria, lo que reduce tiempos de implementación.

Como servidor web para la realización de consultas se escogió Apache, interprete de php. Nuevamente, privilegiando la experiencia del diseñador y para de esta forma disminuir tiempos de implementación.

Esto a pesar de que jasper server utiliza Tomcat, intérprete de jsp, para la realización de reportes.

Si bien los reportes son un número determinado, no variable y con campos fijos, existen variables que pueden querer ser determinadas de manera dinámica por el usuario.

Para esto se propone la creación de una interfaz gráfica para la realización de las consultas de reporte.

Los reportes son almacenados como un *string*, escrito en lenguaje MDX, en una tabla de la BD de Jasper Server. Mediante esta interfaz gráfica escrita en HTML, se realiza la determinación de las variables a incluir en los campos ya determinados para cada uno de los reportes preestablecidos.

Mediante código en php, es posible crear la consulta mdx dinámicamente, según los requerimientos del usuario e insertarla en la base de datos del servidor OLAP, de esta manera se logra que la consulta OLAP sea realizada dinámicamente.

A continuación se especifican las características de las consultas definidas y los campos dinámicos que pueden escogerse en cada uno de ellos.

5. EXPERIMENTOS Y RESULTADOS

En el presente capítulo se muestran los resultados obtenidos para cada una de las etapas de implementación del sistema de apoyo a la toma de decisiones descrito en el capítulo 4.

Se busca la presentación tanto de la implementación del sistema de apoyo a la toma de decisiones, como los resultados obtenidos a partir de la aplicación de los algoritmos de *data mining*.

En la primera parte de este capítulo, punto 5.1, se presentarán los resultados de los experimentos destinados a estudiar la factibilidad de la implementación del sistema de apoyo a la toma de decisiones, los procesos propuestos para esto, tiempos de ejecución y los principales problemas encontrados en el desarrollo.

Finalmente, punto 5.2, se presentarán los resultados obtenidos de la aplicación de los algoritmos de *data mining* a los datos almacenados en el sistema de apoyo a la toma de decisiones.

Como comentario preliminar cabe destacar que los resultados dan cuenta de la factibilidad en la implementación de lo propuesto en el capítulo 4, es decir un sistema que asista al experto en negocios, proporcionando datos comerciales actualizados y permitiendo extraer información de estos.

No se incluyen en este capítulo los experimentos para la propuesta de requerimientos de distribución, pues para esto se requiere de los equipos y la intervención de la red de Cencosud, ambos, fuera del alcance de la presente memoria.

Tampoco se incluyen experimentos para la realización de asignación de pronósticos por tiendas dado que: 1) No se logró consensuar una función objetivo para la generación de un problema de optimización a resolver y 2) no se obtuvieron datos de características físicas para los productos.

Todas las pruebas y experimentos fueron realizados en un equipo Packard Bell con procesador Intel® Pentium® Dual CPU T2390 de 1,87 Ghz, 2 MB de memoria RAM y Windows Vista® Home Basic edition como sistema operativo.

En la realización de todas las pruebas se utilizaron las mismas series de datos pertenecientes a 15 productos de 1 tienda, las que se describen en la tabla 5.1.

Tabla 5.21: Caracterización de los productos utilizados en la prueba de los métodos de predicción

Nombre del producto	Código	Agregación de los datos	Cantidad de datos	Años
CUCHILLO WHITE RIVER 12924 SHEFFIELD	821198	Semanal	121	2
JGO ACC 208PZ HET008 H MACHINERY	826428	Semanal	121	2
JUEGO 4 MULTIHERRAMIENTA HET005 SHEFFIEL	823851	Semanal	121	2
KIT FRESAS 30PCS HET006 H MACHINERY	824684	Semanal	121	2
KIT HTAS MANUAL 147PZ HET004 H MACHINERY	824683	Semanal	121	2
KIT HTAS MANUAL 205PZ HET011 H MACHINERY	824685	Semanal	121	2
LINTERNA LLAVERO HET009 GREATLITE	823852	Semanal	121	2
MALETA 47PZ BROCAS HET007 H MACHINERY	826427	Semanal	121	2
MALETA HERRAMIENTAS 82PZ HLS003 H MACH	826430	Semanal	121	2
MULTIHERRAMIENTA 8 IN 1 HET012 SHEFFIELD	823854	Semanal	121	2
NAVAJA ABSAROKA 2.3/4" 12855 SHEFFIELD	821192	Semanal	121	2
NAVAJA LANDER 3.1/4" 12856 SHEFFIELD	821193	Semanal	121	2
NAVAJA PITKIN 2.1/2" 12862H SHEFFIELD	821194	Semanal	121	2
SET HERITAGE 2 HERRTAS. 46007H SHEFFIELD	821197	Semanal	120	2
SET MALETA HTAS 51PZA HET003 H MACHINERY	826426	Semanal	121	2

Fuente: Elaboración propia

Estos productos fueron elegidos al azar, y representan un set de prueba para testear los métodos propuestos para la realización de pronósticos de demanda.

5.1 Resultados para la implementación de procesos

En esta sección se busca probar los mecanismos propuestos en el capítulo 4 para la implementación del sistema de apoyo a la toma de decisiones, tiempos involucrados en ejecución y dificultades en la implementación.

5.1.1 Proceso de mecanización de datos

Se realizaron 4 experimentos con el fin describir posibles fuentes de problemas en la implementación de la solución y de estimar tiempos de implementación.

a. **Entrega de formato para la carga hacia las tablas del DSA:** En esta etapa se transforman los datos desde el formato en Excel al formato diseñado para la tabla del DSA.

   En la tabla 5.2 se describen los archivos de entrada y los archivos de salida de cada uno de los 2 procesos realizados.

Tabla 5.2: Archivos involucrados en el proceso de entrega de formato a los datos

Descripción de los archivos	Formato de los archivos	Tamaño promedio de los archivos	Número de archivos
Datos originales de ventas obtenidos de la consulta OLAP	Excel (.xlsx)	4,45 MB	24
Datos solo para SKU, sin datos agregados	Texto plano (.CSV)	1,47 MB	23
Datos con formato para la carga en el *data staging areas*	Texto plano (.CSV)	44,16 MB	23

Fuente: Elaboración propia

Este proceso se realizó mediante un programa escrito en python con lo cual el tiempo de procesamiento para cada uno de los archivos fue de 5 minutos aproximadamente. Los resultados se presentan en la tabla 5.3

Tabla 5.3: Tiempos involucrados en la entrega de formato para la carga hacia el DSA

Descripción del proceso	Tiempo de ejecución (seg/archivo)
Eliminación de información no necesaria	20
Entrega de formato para la carga	30

Fuente: Elaboración propia

b. **Carga de datos desde los archivos .csv hacia las tablas del DSA:** Para el desarrollo de este experimento se utilizó la interfaz gráfica de MySQL para cargar los archivos en formato .csv a la tabla correspondiente.

La tabla 5.4 muestra los resultados obtenidos en el proceso de carga.

Tabla 5.4: Tiempos involucrados en la carga hacia el DSA

Nombre del archivo	Tamaño (MB)	Tiempo de Carga (seg)
E503.csv	59,49	290
E504.csv	59,12	305
E507.csv	56,83	289
E508.csv	60,69	317
E510.csv	48,64	222
E512.csv	56,04	260
E513.csv	0,23	2
E514.csv	51,4	255
E517.csv	50,49	232
E518.csv	48,29	237
E520.csv	47,9	230
Promedio	49,01	240

Fuente: Elaboración propia

En promedio los archivos pesan 49 MB y el tiempo de carga fue de 4 minutos aproximadamente.

Los problemas suscitados en esta etapa fueron:

- Tamaño máximo de carga mediante MySQL: Por defecto, el tamaño máximo de archivos .csv que se pueden cargar mediante MySQL es de 2 MB. Se aumento este número a 100MB modificando la configuración en el archivo php.ini.
- Tiempo de duración de una operación en php: Por defecto, después de 10 minutos php cierra la operación que se esté ejecutando. Para remediar esto se modificó la configuración del archivo php.ini y se aumento este tiempo hasta 5 horas.

c. **Transformación y carga de los datos a la tabla temporal:** Para este proceso se escribió un programa en R, el que implementa la limpieza de datos y las transformaciones descritas en el capítulo 4.3.1.

Se tomo una muestra de 15 productos para 1 tienda, formando un total de 15 series a tratar.

El tiempo en el tratamiento de las 15 series fue menor que unio segundo, no fue posible realizar un medición más precisa debido a la velocidad del proceso.

Dado lo anterior se estima el tiempo de procesamiento de las 15 series en 0,5 segundos.

d. Carga final de los datos desde la tabla temporal al Data Warehouse: Para este proceso se utiliza un procedimiento almacenado, el que, activado por un *trigger* temporal carga los datos a las *fact table*.

El tiempo utilizado en la realización de esta operación es homologable al de cualquier proceso de carga de datos entre tablas.

El tiempo en el tratamiento de las 15 series fue menor que un segundo, no fue posible realizar un medición más precisa debido a la velocidad del proceso.

Dado lo anterior se estima el tiempo de procesamiento de las 15 series en 0,5 segundos.

5.1.2 Proceso de mecanización de procesos

Se realizaron experimentos con el fin describir posibles fuentes de problemas en la implementación de la solución y de estimar tiempos de implementación.

a. **Proceso de ETL:** Este proceso incluye las tareas de Entrega de formato para el manejo de los datos, carga de los datos al DSA, limpieza de los datos, transformación de los datos y carga de los datos al *data warehouse.*

Estos puntos fueron testeados en el punto 5.1.1. y los tiempos de ejecución individuales para los procesos pueden ser observados en las tablas 5.2, 5.3, 5.4 y los puntos 5.1.1.b y 5.1.1.c.

A la luz de los resultados, no se hace necesario la utilización de software para asistir en las tareas de ETL. Esto se debe a que solo existe una única fuente de información y la carga a la base de datos se encuentra ligada con los procesos predicción de demanda.

b. Proceso de generación de Reportes: Para la implementación de este proceso se realizó un experimento de visualización de información histórica. La consulta se realizó sobre la *fact table* histórico.

En la figura 5.1 se muestra la consulta realizada, el campo ventas identifica cantidad de artículos vendidos y el campo monto expresa la transacción en dinero realizada (miles de pesos).

easy

Dimensions								Measures	
Tiempo	Año	Mes	Semana	Tiendas	Tienda	Productos	Producto	Ventas	Monto
All Tiempo.Fechas				All Tiendas.Tiendas		All Productos.Productos		413	2.920
All Tiempo.Fechas	2008			All Tiendas.Tiendas		All Productos.Productos		361	2.429
	2008	Apr		All Tiendas.Tiendas		All Productos.Productos		22	144
		Apr	14	All Tiendas.Tiendas		All Productos.Productos		9	69
						All Productos.Productos	194478	4	1
							194479	1	0
							194690	0	0
							195398	0	0
							404858	4	68
							452749	0	0
							478686	0	0
							585778	0	0
				All Tiendas.Tiendas	E513	All Productos.Productos		9	69
			15	All Tiendas.Tiendas		All Productos.Productos		6	13
			16	All Tiendas.Tiendas		All Productos.Productos		4	28
			17	All Tiendas.Tiendas		All Productos.Productos		3	34
		Aug		All Tiendas.Tiendas		All Productos.Productos		45	119
		Dic		All Tiendas.Tiendas		All Productos.Productos		14	185
		Feb		All Tiendas.Tiendas		All Productos.Productos		44	137

Figura 5.1 Consulta OLAP a la *fact table* histórico. Fuente: Elaboración propia

c. Proceso de generación de pronóstico de demanda: Para este proceso se diseñó un programa escrito en R, el que implementa el algoritmo descrito en el punto 4.3.4.

Se corrió el programa para 15 productos en 1 tienda, formando un total de 15 series a pronosticar.
En la tabla 5.5 se presentan los resultados obtenidos para el proceso de generación de pronóstico de demanda utilizando redes neuronales artificiales, medias móviles y la metodología ARIMA.

Tabla 5.5: Tiempos involucrados en la generación de predicción de demanda

Método de predicción	Número de series	Segundos en procesamiento	Tiempo por serie	Número total de series	horas
Redes neuronales artificiales	15	0,50	0,03	3000	1,67
Medias Móviles	15	0,50	0,03	3000	1,67
ARIMA	15	2,00	0,13	3000	6,67

Fuente: Elaboración propia

d. Proceso de agrupación de tiendas

Para este proceso se diseñó un programa escrito en R, el que implementa el algoritmo descrito en el punto 4.3.3. para la generación de agrupaciones mediante el algoritmo de SOFM y de *kmeans*.

Se realizaron tres grupos de experimentos, cada uno con las siguientes variables la *clusterización* para cada una de las 24 observaciones:

1. Variables socioeconómicas: Vectores con 5 campos
2. Variables de utilización de espacio en tiendas: Vectores con 48 campos
3. Variables socioeconómicas y de utilización de espacio en tiendas: Vectores con 53 campos.

En la tabla 5.6 se muestran los resultados obtenidos, para el proceso de agrupación mediante *kmeans* utilizando el software R y 10, 20, 30 y 40 semillas[72].

Tabla 5.6: Tiempos involucrados en la generación de agrupaciones utilizando metodología *Kmeans*

Agrupación por	Tiempo de ejecución (seg)			
	10 semillas	20 semillas	30 semillas	40 semillas
socioeconómicas	15	20	40	55
espacio en tiendas	12	20	30	52
socioeconómicas y uso de espacio en tiendas	18	29	44	57
Promedio	15	23	38	54

Fuente: Elaboración propia

5.2 Resultados de las operaciones de minado de datos

En esta sección se busca presentar los resultados arrojados por las herramientas de *data mining*, así como la calibración de parámetros de estos.

Se obtienen resultados positivos tanto para la agrupación de tiendas como para el pronóstico de demanda.

Como principales resultados de los experimentos se cuentan: 1) La generación de agrupaciones por tiendas y la corroboración de que estos grupos no presentan relación con la agrupación actualmente utilizada en la empresa y 2) se encuentran indicadas las herramientas de medias móviles y de redes neuronales artificiales para la predicción de demanda.

5.2.1 Proceso de agrupación de tiendas

A continuación, se presentan los resultados obtenidos para las distintas agrupaciones realizadas.

El proceso de selección de *clusters* se realizó según la descripción del Capítulo 4[73]:

- Determinación del número posible de *clusters* mediante inspección visual de mapa bidimensional de SOFM.
- Corroboración mediante *kmeans*
- Validación de *clusters* según criterios.

[72] El número de semillas indica el número de experimentos realizados antes de escoger una solución. 1 semilla implica 1 experimento con una solución inicial escogida al azar, diferente de las otras semillas.
[73] Ver Capítulo 4 punto 4.4.3

Agrupaciones realizadas según características de utilización de espacio en tiendas

1. SOFM: Se entrenó 4 grillas de 5x5. En todas las grillas se obtuvieron distribuciones similares de las observaciones sobre el plano bidimensional.

   En la figura 5.2 se muestra la grilla bidimensional y la distribución de las observaciones sobre ella.

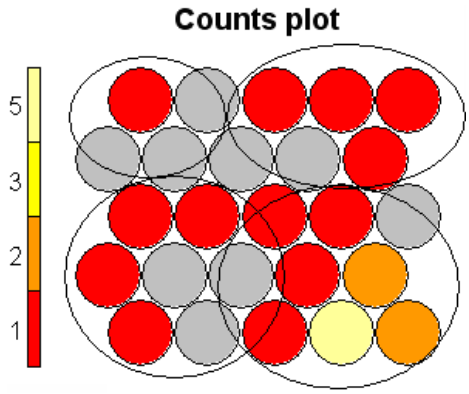


Figura 5.2 Distribución de las tiendas sobre la malla bidimensional de 5x5 en base a las características de distribución espacial de cada tienda. Fuente: Elaboración propia

La figura 5.2 muestra 4 posibles *clusters*, identificándose una observación aislada del resto y la sección del cuadrante derecho bajo con un gran número de observaciones.

En la tabla 5.7 se muestran las agrupaciones observadas en la figura 5.1.

Tabla 5.7: Clusterización obtenida para la agrupación por espacio utilizado en tiendas, SOFM

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Kennedy	La Reina	Copiapó	Antofagasta
Valparaíso	La Dehesa	La Serena	
Curicó	Maipú	Temuco	
Talca	Viña del Mar	La Florida	
Linares			
Chillan			
Los Ángeles			
Osorno			
Puerto Montt			
Quillota			
Quilicura			
Rancagua			
El Llano			
El Llano			
El Belloto			
Quilin			

Fuente: Elaboración propia

Probablemente el *cluster* 1 pueda ser subdividido, dado el gran número de observaciones que presenta.

Es por esto que se realizarán agrupaciones mediante el método de kmeans con 3, 4, 5 y 6 *clusters*.

2. K-means: Se realizaron 5 corridas del programa de clusterización[74] con 4, 3, 5 y 6 como número de clusters, cada agrupación se realizó con 20 semillas.

   Los mejores resultados se obtuvieron con 6 *clusters* y se presentan a continuación en la tabla 5.8.

Tabla 5.8: Clusterización obtenida para la agrupación por espacio utilizado en tiendas.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Kennedy	La Reina	El Llano	La Serena	Copiapó	Antofagasta
Valparaíso	La Dehesa	El Belloto	La Florida	Temuco	
Curicó	Maipú	Viña del Mar			
Talca	Quilín	Rancagua			
Linares	FISA				
Chilla					
Los Ángeles					
Osorno					
Puerto Montt					
Quillota					
Quilicura					

Fuente: Elaboración propia

Se observa que al disminuir el número de *clusters*, Antofagasta sigue quedando aislada del resto, por lo que se analizará su caso aisladamente.

3. Validación: Las agrupaciones observadas en el SOFM muestran coherencia con las obtenidas mediante el algoritmo de *kmeans*, con La Serena, La Florida, Copiapó, Temuco aisladas del resto de las tiendas y con Antofagasta completamente aislada.

   El *cluster* 1 de Kennedy, Vaparaíso, Curicó, Talca, Linares, Chillán, Los Ángeles, Osorno, Puerto Montt; Quillota y Quilicura se encuentran agrupados en ambos algoritmos al igual que La Reina, La Dehesa y Maipú y El Llano, El Belloto y Rancagua.

   Finalmente, en conjunto con el *Category Manager*, se obtiene la siguiente interpretación por *clusters*[75].

   Cluster 1: Agrupación de tiendas con mayor tamaño de salas.
   Cluster 2: Agrupación de tiendas con menor tamaño en salas.
   Cluster 3: Tiendas con un sesgo hacia las secciones destinadas al mejoramiento del hogar puertas adentro. Bazar hogar, Puertas, Ventanas y Molduras, Muebles y Textil hogar son las categorías que más espacio demandas.
   Cluster 4: Tienda que muestran un sesgo hacia las secciones de jardinería, exteriores y productos de temporada.
   Cluster 5: No presentan un sesgo particular, mostrando una división equitativa del espacio en sala.

[74] Ver Anexo D: "Programación"
[75] Ver Anexo C "Datos de la empresa" para ver la utilización de espacios por cluster y categoría

Cluster 6: Compuesto solamente por Antofagasta, donde se observa una tendencia clara hacia el espacio para la venta de insumos para obra gruesa.

Como conclusión se observa que la distribución de espacio en sala es bastante similar para todas las tiendas, lo que genera el tamaño total de la tienda sea una variable preponderante en la generación de *clusters*, creándose 3 grandes *clusters* y 3 *clusters* de observaciones aisladas, aunque cabe mencionar que las diferencias siguen siendo sutiles.

Agrupaciones realizadas según características socioeconómicas de las comunas en las que se encuentra cada tienda

1. SOFM: Se entrenó 4 grillas de 5x5, obteniéndose 4 muestras de distribución distintas. En todas las muestras se obtuvieron distribuciones similares de las observaciones sobre el plano bidimensional.

   En la figura 5.3 se muestra la grilla bidimensional y la distribución de las observaciones sobre ella.

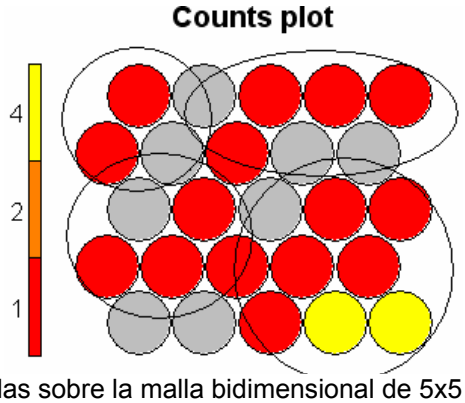


Figura 5.3 Distribución de las tiendas sobre la malla bidimensional de 5x5 en base a las características socioeconómicas de la comuna en la que se encuentran. Fuente: Elaboración propia

En la tabla 5.9 se muestran las agrupaciones observadas en la figura 5.2.

Tabla 5.9: Clusterización obtenida para la agrupación según características socioeconómicas de las tiendas, SOFM

Cluster 1	Cluster 2	Cluster 3	Cluster 4
La Florida	La Serena	Kennedy	Quilin
Copiapo	Antofagasta	La Dehesa	Quillita
	Temuco	La Reina	Quilicura
	Maipu	El Llano	El Belloto
			FISA
			Osorno
			Linares
			Talca
			Viña del Mar
			Valparaíso
			Rancagua
			Los Ángeles

Fuente: Elaboración propia

2. K-means: Se realizaron 5 corridas del programa de clusterización con 4, 3, 5 y 6 como numero de clusters, cada agrupación se realizó con 20 semillas.

   Los mejores resultados se obtuvieron con 6 clusters y se presentan a continuación en la tabla 5.10.

Tabla 5.10: Clusterización de tiendas según características socioeconómicas de las comunas en las que se encuentran

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Quilin	Copiapó	La Reina	Kennedy	Temuco	Antofagasta
El Llano	La Serena	La Dehesa			
FISA	La Florida				
Valparaíso	Maipú				
El Belloto					
Viña del Mar					
Rancagua					
Curicó					
Talca					
Linares					
Chillán					
Los Ángeles					
Osorno					
Puerto Montt					
Quillota					
Quilicura					

Fuente: Elaboración propia

3. Validación: Se observa que los resultados arrojados tanto por el SOFM como por el *kmeans* entregan soluciones similares, lo que valida en primera instancia los resultados.

   Por otra parte, en conjunto con el *Category Manager*, se obtiene la siguiente interpretación por *clusters*.

   Cluster 1: Es el *cluster* que agrupa a las comunas más pobres. Presenta el menor ingreso promedio por hogares de los 6 *clusters*. Además de contener a casi la totalidad de los hogares de los GSE 2, 3 y 4 del total de los hogares en la muestra.

   La distribución al interior del *cluster* muestra que cerca del 90% de los hogares pertenecientes a este *cluster* se encuentra en los GSE 3 y 4 y presenta el mayor porcentaje de hogares del GSE 1 que cualquier otro *cluster*.

Tabla 5.11: Distribución de los hogares pertenecientes al cluster 1

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	57	1.689	24.503	29.865	4.049
% del total de hogares	63,94%	75,63%	79,36%	38,84%	28,95%
% de hogares en el cluster	0,09%	2,81%	40,73%	49,64%	6,73%

Fuente: Elaboración propia

Cluster 2: Cluster de clase media. Si bien presenta un alto porcentaje de hogares en los GSE 4 y 5, muestra un alto porcentaje de los hogares de los GSE 1, 2 y 4.

Tabla 5.12: Distribución de los hogares pertenecientes al cluster 2

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	47	770	12.222	73.414	13.524
% del total de hogares	13,10%	8,62%	9,90%	23,87%	24,18%
% de hogares en el cluster	0,05%	0,77%	12,22%	73,43%	13,53%

Fuente: Elaboración propia

Cluster 3: Cluster que agrupa a las comunas con mayores ingresos después del cluster 4. Presenta el segundo mayor ingreso promedio por hogar además de presentar una gran cantidad de hogares en el GSE 4 y 5.

Tabla 5.13: Distribución de los hogares pertenecientes al cluster 3

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	1	118	1.373	9.543	12.902
% del total de hogares	0,07%	0,66%	0,56%	1,55%	11,53%
% de hogares en el cluster	0,00%	0,49%	5,73%	39,87%	53,90%

Fuente: Elaboración propia

Cluster 4: El más alto ingreso promedio. En escencia es igual al cluster 3, sin embargo posee un ingreso promedio cerca del 50% mayor que La Reina y La Dehesa, lo que termina por diferenciarlos en *clusters* distintos.

El 98% de los hogares que lo componen pertenecen al GSE 4 o 5.

Tabla 5.14: Distribución de los hogares pertenecientes al cluster 4

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	1	124	784	18.717	64.004
% del total de hogares	0,07%	0,35%	0,16%	1,52%	28,61%
% de hogares en el cluster	0,00%	0,15%	0,94%	22,38%	76,53%

Fuente: Elaboración propia

Cluster 5: Presenta la particularidad de que, si bien posee un ingreso promedo alto por hogar (el tercero más alto), tiene también la mayor concentración de hogares pertenecientes al GSE1. Esta desigualdad en los ingresos, lo hace quedar aislado del resto de los clusters

Tabla 5.15: Distribución de los hogares pertenecientes al cluster 5

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	294	4.273	36.404	359.834	3.236
% del total de hogares	20,59%	11,96%	7,37%	29,25%	1,45%
% de hogares en el cluster	0,07%	1,06%	9,01%	89,06%	0,07%

Fuente: Elaboración propia

Cluster 6: Presenta la particularidad de tener la población distribuida muy cercana a las medias para cada GSE. Además presenta un porcentaje casi idéntico de hogares por GSE considerando la muestra entera.

Lo anterior se corrobora observando que tiene un ingreso promedio medio en comparación con el resto de los *clusters*.

Tabla 5.16: Distribución de los hogares pertenecientes al cluster 6

	Número de hogares (2008)				
	GSE 1	GSE 2	GSE 3	GSE 4	GSE 5
Promedio intra	32	996	13.123	60.993	11.830
% del total de hogares	2,24%	2,79%	2,66%	4,96%	5,29%
% de hogares en el cluster	0,04%	1,15%	15,09%	70,13%	0,03%

Fuente: Elaboración propia

Agrupaciones realizadas según características socioeconómicas de las comunas en las que se encuentra cada tienda y de utilización de espacio en tiendas

1. SOFM: Se entrenó 4 grillas de 5x5, obteniéndose 4 muestras de distribución distintas. En todas las muestras se obtuvieron distribuciones similares de las observaciones sobre el plano bidimensional.

   En la figura 5.4 se muestra la grilla bidimensional y la distribución de las observaciones sobre ella, la tabla 5.17 lista las agrupaciones observadas.

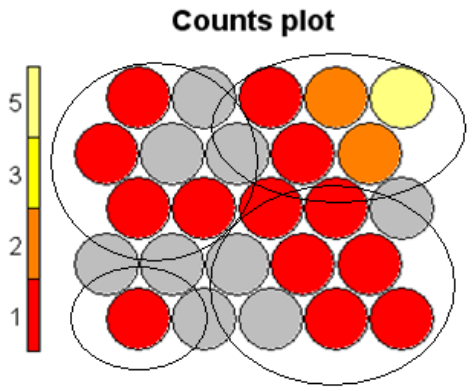


Figura 5.4 Distribución de las tiendas sobre la malla bidimensional de 5x5 en base a las características socioeconómicas de la comuna en la que se encuentran. Fuente: Elaboración propia

Tabla 5.17: Clusterización obtenida para la agrupación según características socioeconómicas y de utilización de espacio en las tiendas, SOFM

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Antofagasta	Copiapó	La Dehesa	Kennedy
	La Serena	La Reina	Linares
	Temuco	Quilin	Puerto Montt
	La Serena	FISA	Quillota
		El Belloto	Quilicura
			Valparaíso
			Chillan
			Curicó
			Talca
			El Llano
			El Belloto
			Los Ángeles
			Talca

Fuente: Elaboración propia

2. K-means: Se realizaron 5 corridas del programa de clusterización con 4, 3, 5 y 6 como numero de clusters, cada agrupación se realizó con 20 semillas.

   Los mejores resultados se obtuvieron con 4 clusters y se presentan a continuación en la tabla 5.18.

Tabla 5.18: Clusterización obtenida para la agrupación por espacio utilizado en tiendas y características socioeconómicas de las comunas en las que se encuentran las tiendas

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Kennedy	La Reina	Copiapó	Antofagasta
El Llano	La Dehesa	La Serena	
Valparaíso	Maipu	Temuco	
El Belloto	Quilin	La Florida	
Rancagua	FISA		
Curicó	Viña del Mar		
Talca			
Linares			
Chillán			
Los Ángeles			
Osorno			
Puerto Montt			
Quillota			
Quilicura			

Fuente: Elaboración propia

3. Validación: Se corrobora que las agrupaciones entregadas por el método de SOFM es muy coherente con la solución entregada por el método de *kmeans* programado.

   Por otra parte, en conjunto con el Category Manager, se obtiene la siguiente interpretación por clusters.

   Cluster 1: Presentan un nivel socioeconómico bajo, siendo esta su principal variable de agrupación.

   No se observan diferencias claras en tamaño de sala.

   Lo anterior es coherente con la observación de que las diferencias en tamaño de sala y distribución del espacio no generan grandes diferencias, por lo que la variable socioeconómica es preponderante en la agrupación.

   Cluster 2: Se observan tiendas con un nivel socioeconómico por sobre la media y con gran tamaño de sala.

   La excepción se observa en Viña del Mar y FISA, las que explican su inclusión en este cluster dado su gran tamaño de sala.

   Cluster 3: Tiendas con nivel igual nivel socioeconómico bajo y pequeñas en tamaño de salas y semejante distribución de espacial.

Muestra coherencia con las descripciones realizadas en los experimentos anteriores donde Copiapó, La Serena y Temuco siempre se encontraron juntas. La inclusión de La Florida en la agrupación responde a la similitud en el nivel socioeconómico de la tienda y el *cluster*.

Cluster 4: Contiene a Antofagasta y es consecuente con la descripción que se ha realizado en los experimentos anteriores, donde se encuentra siempre aislada del resto.

5.2.2 Proceso de generación de pronósticos de ventas

En la siguiente sección se presentan los resultados obtenidos para la realización de pronóstico de ventas. Se distinguen 2 etapas en la presentación de este proceso. Por un lado la presentación de los experimentos y por otra los resultados obtenidos.

5.2.2.1 Presentación de los resultados de la metodología de predicción

A continuación se presentan los resultados obtenidos para la realización de predicción de demanda.

Se realizarán pruebas con 15[76] productos para 1 tienda, formando un total de 15 series a tratar.

La efectividad de los métodos fueron medidos utilizando MAE y MSE[77]. En la tabla 5.19 y 5.20 se muestran los resultados obtenidos para cada uno de los experimentos de predicción realizados: Medias móviles, Redes Neuronales Adrtificiales y ARIMA.

Tabla 5.19: Resultados de la predicción realizada según métodos de predicción, MAE

	2 semanas	4 semanas	8 semanas	12 semanas	Pred RNA	ARIMA
821198	0,00625	0,00625	0,00625	0,00625	*0,00078125*	0,10625
826428	*0*	*0*	*0*	*0*	*0*	0,125
823851	0,04375	0,04375	*0,04375*	0,04375	0,125	0,1
824684	*0*	*0*	*0*	*0*	*0*	0,05625
824683	0,10625	0,08125	0,10625	0,0625	*0,01015625*	0,725
824685	*0*	*0*	*0*	*0*	*0*	0,225
823852	0,125	0,13125	0,13125	0,11875	*0,01484375*	0,1625
826427	*0,025*	*0,025*	*0,025*	*0,025*	*0,025*	0,025
826430	0,0875	0,09375	0,06875	0,04375	*0,01328125*	0,23125
823854	0,06875	0,0875	0,06875	0,075	*0,009375*	1,125
821192	*0*	*0*	*0*	*0*	*0*	0
821193	0,05	0,025	0,025	0,025	*0,003125*	0,06875
821194	*0*	*0*	*0*	*0*	*0*	0
821197	*0*	*0*	*0*	*0*	*0*	0
826426	0,1375	0,19375	0,20625	0,2625	*0,015625*	257,675

Fuente: Elaboración propia

[76] Ver tabla 5.1 para la descripción de las series a utilizar
[77] Ver Capítulo 2 punto 3.2.4

Tabla 5.20: Resultados de la predicción realizada según métodos de predicción, en términos del indicador MSE

	2 semanas	4 semanas	8 semanas	12 semanas	Pred RNA	ARIMA
821198	0,00625	0,00625	0,00625	0,00625	*0,00078125*	0,10625
826428	*0*	*0*	*0*	*0*	*0*	0,125
823851	0,04375	0,04375	0,04375	0,04375	*0,00546875*	0,1375
824684	*0*	*0*	*0*	*0*	*0*	0,05625
824683	0,15625	0,13125	0,15625	0,0625	*0*	7,725
824685	*0*	*0*	*0*	*0*	*0*	0,425
823852	0,3125	0,34375	0,34375	0,33125	*0*	0,2375
826427	0,025	0,025	0,025	0,025	*0*	0,025
826430	0,1375	0,14375	0,06875	0,04375	*0*	0,70625
823854	0,06875	0,0875	0,06875	0,075	*0*	24,5125
821192	*0*	*0*	*0*	*0*	*0*	0
821193	0,05	0,025	0,025	0,025	*0*	0,06875
821194	*0*	*0*	*0*	*0*	*0*	0
821197	*0*	*0*	*0*	*0*	*0*	0
826426	0,275	0,43125	0,38125	0,6875	*0*	536206,725

Fuente: Elaboración propia

5.2.2.2 Descripción de los experimentos

Se busca ejemplificar la metodología de realización de predicciones y comparar los resultados obtenidos para cada uno de los métodos utilizados.

Se ejemplificará con una serie de muestra los pasos realizados y se especificarán los resultados obtenidos para cada etapa[78]. La serie de datos elegida como muestra para la descripción de los experimentos será la 821198, CUCHILLO WHITE RIVER 12924 SHEFFIELD.

La figura 5.5 muestra los valores originales para la serie 821198.

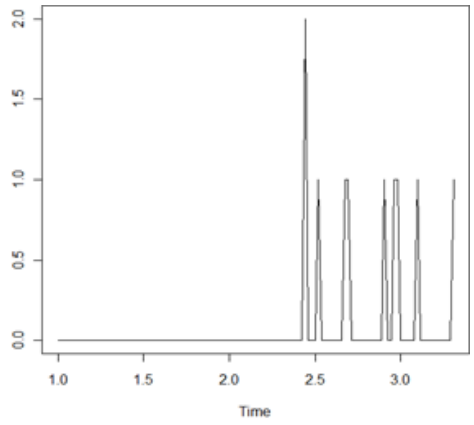


Figura 5.5 valores semanales de unidades vendidas para el artículo 821198. Fuente: Elaboración propia

A continuación se presentan los métodos utilizados en los experimentos para la metodología de redes neuronales artificiales, medias móviles y ARIMA.

[78] Para ver los resultados para todas las series de prueba, referirse al Anexo D, “Resultados”

Redes Neuronales Artificiales

Dado que las redes neuronales artificiales internalizan efectos no lineales, estas las series no requieren de mayor tratamiento que la normalización de sus valores, pues cualquier otro efecto será recogido en el proceso de entrenamiento.

Normalización de valores: Los datos normalizados, son presentados en la figura 5.6.

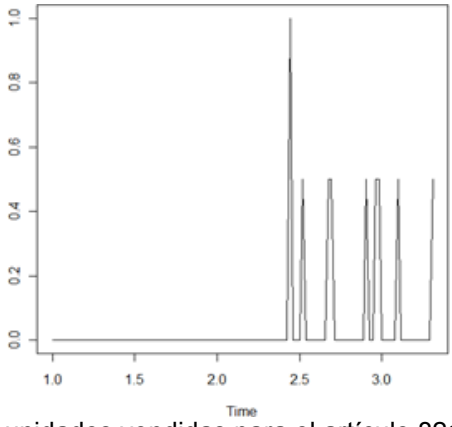


Figura 5.6 valores semanales de unidades vendidas para el artículo 821198. Fuente: Elaboración propia

Se utilizarán redes neuronales artificiales de 1 capa de entrada con 1 neurona de entrada, 1 capa intermedia con 20 neuronas y 1 capa de salida con 1 neurona de salida (14) (20).

Entrenamiento: Para la elección automática de los rezagos utilizar en el proceso de entrenamiento se utilizará el correlograma de la serie. Luego se escogerán aquellos rezagos que muestren correlaciones significativas con el valor presente de la serie.

En la figura 5.7 se muestra el correlograma[79] de la serie para el artículo 821198. Se observa que los rezagos significativos son los rezagos 4 y 12.

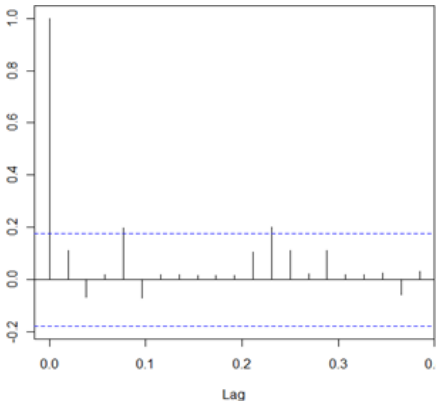


Figura 5.7 valores semanales de unidades vendidas para el artículo 821198. Fuente: Elaboración propia

[79] Ver Anexo B “Estadística”

Realización de predicciones: A continuación se realizan 2 redes neuronales artificiales (tantas como rezagos se utilicen), con las que se realizan predicciones para las próximas 8 semanas.

En la tabla 5.21 se muestran los resultados para las predicciones realizadas con las redes neuronales artificiales entrenadas con el cuarto rezago (RNA 4 rezago) y con el décimo segundo rezago (RNA 12 rezago).

Tabla 5.21: Resultados de la predicción realizada para el producto 821198 utilizando RNA

Semana a predecir	Predicción con RNA 4 rezago	Predicción con RNA 12 rezago	Valor a predecir
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	1

Fuente: Elaboración propia

ARIMA

La principal complicación se encuentra en hacer estacionaria a la serie. Para lo cual se propone la aplicación de las siguientes soluciones:1) Transformación logarítmica para la estabilización de la varianza, 2) estimación y eliminación de la tendencia mediante una regresión lineal y 3) diferenciación de la serie para detectar y eliminar estacionalidad.

Eliminación de tendencia: Se calculó la tendencia para cada una de las series mediante la realización de una regresión lineal. Este valor es almacenado y luego guardado antes de eliminarlo de la serie.

En la figura 5.8 se muestra el resultado de las operaciones de transformación logarítmica y eliminación de la tendencia.

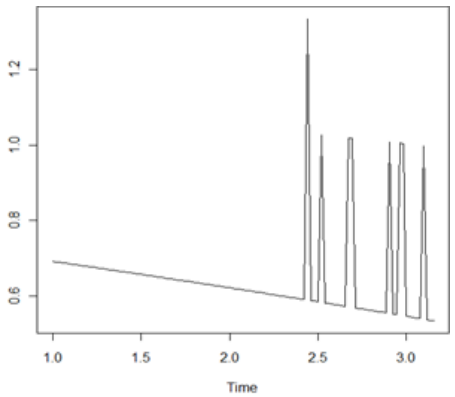


Figura 5.8 Serie de valores semanales de unidades vendidas para el artículo 821198 sin tendencia.
Fuente: Elaboración propia

Transformación logarítmica de los datos: Se realiza para la estabilización de la varianza de la serie.

Se utiliza la transformación logarítmica, basado en los resultados obtenidos para las series probadas, datos presentados en la tabla 5.7.

Cabe mencionar que la transformación logarítmica fue la siguiente:

$$Valor\ transformado = \log(Valor\ no\ transformado + 1) \qquad (5.1)$$

Esto dado a que la existencia de valores cero en la serie original, hacen explotar a menos infinito los valores de la serie transformada si no se le suma 1, con lo que no se cumple el objetivo de disminuir la varianza sino que se aumenta.

Ajuste del modelo ARIMA: Para esto se realiza una serie de pasos que se detallan a continuación:

- Estacionalidad de la serie: Se comprueba estacionalidad de la serie mediante el test de Dickey-Fuller Aumentado. Si es que no se logra, la serie es diferenciada tantas veces como sea necesario hasta lograr que sea estadísticamente estacionaria.

  El número de diferencias aplicadas determina el valor d para el modelo ARIMA a ajustar.

  A continuación se muestra los resultados obtenidos para la serie utilizada como muestra.

  Resultados del test de Dickey Fuller Aumentado
  Estadístico = -4,1393, razago = 1, p-value < 0.01
  La serie es estacionaria considerando 1 rezago y sin diferenciar la serie.

  No fue necesario diferenciar la serie por lo que el valor d se establece como cero para esta serie.

- Identificación del parámetro p del proceso AR(p) subyacente en la serie: Para esto se utilizo el comando *ar()* el que ajusta el numero de rezagos r del proceso AR(p) para una serie dada según el criterio de máxima verosimilitud.

  Para el caso de la serie utilizada como muestra se fija el parámetro p en 5.

- Identificación del parámetro q del proceso MA(q) subyacente en la serie: Para esto se recurre al correlograma parcial de la serie y se busca la máxima caída entre rezagos sucesivos, lo que marca el potencial parámetro q del proceso MA(q).

  En la figura 5.9 se observa el correlograma obtenido.

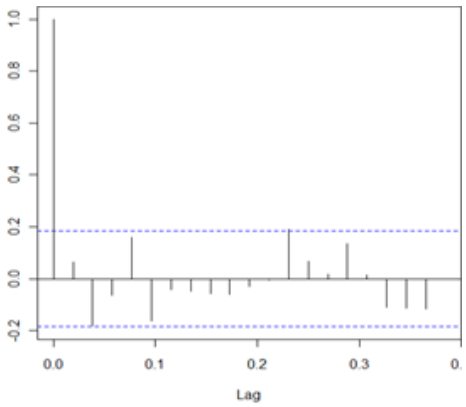


Figura 5.9 Correlograma parcial para la serie de valores semanales de unidades vendidas para el artículo 821198. Fuente: Elaboración propia

- El autocorrelograma simple muestra que el mayor decaimiento se obtiene luego del cuarto rezago por lo que se fija el valor de q en 5.

- Selección del mejor modelo: Se procede a escoger de entre un número de posibles modelos basados en los resultados obtenidos para los indicadores p, q y d. El criterio de selección será el criterio de selección de Akaike[80].

  Para el caso presentado como ejemplo se tiene:

  Los valores obtenidos en los pasos anteriores son:

  p=3, q=3 y d=0

  Se probarán los siguientes modelos: AR(5), AR(4), AR(3), MA(5), MA(4), MA(3), MA(5,0,5). Se eligirá el mejor en función del estadístico de Akaike.

  En la tabla 5.22 se muestran los resultados obtenidos para la serie 821198 .

Tabla 5.22: Comparación de modelos ARIMA en función del estadístico de AKAIKE

Modelo	indicador de Akaike
AR(5)	-152,4915
AR(4)	-148,6455
AR(6)	-150,7536
MA(5)	-154,9275
MA(4)	-153,8082
MA(6)	-154,8499
ARMA(5,0,5)	-156,6437

Fuente: Elaboración propia

Se escoge el modelo ARIMA(5,0,5) ya que presenta un estadístico de Akaike menor.

[80] Ver Anexo B “Estadística”

Realización de predicciones: En la tabla 5.23 se muestra el resultado de la predicción realizada para las próximas 8 semanas con un modelo ARIMA(5,0,5).

Tabla 5.23: Predicción realizada para la serie 821198

Predicción	A predecir
1	0
1	0
1	0
1	0
1	0
0	0
1	0
1	1

Fuente: Elaboración propia

Medias móviles

La metodología comprende la reducción de varianza mediante la aplicación de la función logaritmo de la misma manera que para el método ARIMA, y la aplicación de promedios móviles para 2, 4, 8 y 12 semanas.

En la tabla 5.24 se muestran los resultados de las predicciones realizadas mediante el método de medias móviles.

Tabla 5.24: Predicción realizada para la serie 821198

Semana a predecir	Valor a predecir	Predicción con información de las 2 últimas semanas	Predicción con información de las 4 últimas semanas	Predicción con información de las 8 últimas semanas	Predicción con información de las 12 últimas semanas
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	1	0	0	0	0

Fuente: Elaboración propia

6. Discusión y conclusiones

El presente trabajo tuvo como objetivo estudiar la factibilidad en la implementación de un sistema de apoyo a la toma de decisiones para la obtención de información relevante al departamento de procesos comerciales de Easy S.A.

La aplicación de un *data warehouse* para el almacenamiento de los datos operacionales resulta indispensable en el posterior uso de estos en algoritmos de *data mining*. Destacan el almacenamiento estructurado de datos y la estandarización de los procesos de limpieza.

Cabe señalar la importancia del experto en el negocio, sobre todo en el proceso de generación de agrupaciones de tiendas, para la obtención de resultados coherentes. Dado esto se buscó incorporarlo dentro de la elección de los parámetros utilizados por los algoritmos de *data mining*.

A continuación se presenta una breve discusión y presentación de los principales resultados obtenidos en el cumplimiento de los resultados, así como recomendaciones que surgen a partir de estos y trabajos a seguir para continuar mejorando la solución propuesta.

6.1 Conclusiones

El mayor valor del trabajo realizado se encuentra en la implementación de un sistema de administración de productos, con enfoque exclusivamente en los procesos de administración de categorías y que incorpora el conocimiento del experto de negocios en la aplicación del minado de información.

Esta herramienta permite la administración de las categorías por separado, generando agrupaciones de tiendas por categorías y la corroboración del cumplimiento de objetivos comerciales al poder revisar semana a semana tanto la venta realizada como las proyecciones de venta y la estructura de ocupación de espacio de cada tienda.

Se alcanzan la mayor parte de los objetivos planteados, en términos de lo expresado en el capítulo 1 estos son:

Se construyó la arquitectura conceptual de todos los aspectos relevantes en la construcción de un *data warehouse*, así como la construcción y medición de un prototipo funcional de este.

Se corrobora la factibilidad de la implementación de un sistema de fácil uso para la recolección, limpieza y minado de la información. Entregándose una arquitectura a nivel de procesos, datos y distribución de este sistema.

Se logra además entregar un sistema fácilmente escalable a otras secciones y categorías de productos.

Debido a la automatización en los procesos, se logra disminuir las horas hombres utilizadas tanto en la generación de agrupaciones de tiendas como en predicción de demanda.

Se propone una metodología para la generación de agrupaciones de tiendas, obteniendo agrupaciones coherentes e incluyendo en este proceso al *category manager* y su experiencia.

Se propone una metodología para la realización de predicción de demanda por productos en tiendas, con un error cercano a cero, según los parámetros establecidos en conjunto con el cliente.

6.2 Discusión

Mediante los experimentos realizados se comprueba la efectividad tanto de los procesos como de los *softwares* propuestos para implementar la solución propuesta.

R entrega la posibilidad de utilizar algoritmos y rutinas estadísticas y de *machine learning* con un buen nivel de programación y en constante actualización, así como la posibilidad de crear programas propios lo que lo hace muy flexible.

MySQL como motor de base de datos entrega buenos resultados para la cantidad de datos presentes en el problema. Además, los procedimientos almacenados, ofrecen la posibilidad de automatizar la administración de la base de datos.

No se hace necesario la utilización de un software para la integración de datos, como Kettle, ya que existe una única fuente de información y la carga de datos se encuentra ligada al proceso de ETL y generación de pronósticos.

En el proceso de pronóstico de demanda, tanto las redes neuronales artificiales como las medias móviles muestran resultados similares, muy cerca del error cero tanto para el indicador MSE como MAE, mientras que el método ARIMA muestra los peores resultados, incluso algunos completamente inaceptables (por ejemplo la predicción para la serie 826426).

Estos malos resultados del método ARIMA, puede deberse a distintos factores dentro de los cuales se cuentan: eliminación de la tendencia cuando esta no existía, mala interpretación de de los autocorrelogramas simple y parcial en la determinación de los parámetros P y Q, poca o nula significancia del modelo calculado, etc. Pero aunque estos problemas fuesen solucionados, el tiempo de ejecución del algoritmo resulta excesivo para la aplicación de éste a casi 3.000 series de tiempo.

Para el proceso de *clustering*, se comprueba la utilidad del método de clustering propuesto dados los buenos resultados del las agrupaciones resultantes. Esto es, se comprueba tanto la utilidad del análisis de componentes principales para la disminución de la dimensionalidad del problema, la efectividad de las SOFM en la visualización del problema y la corroboración de las observaciones a través del método de *kmeans*. Todo

lo anterior ejecutado en tiempos que hacen factible su aplicación dinámica a varias categorías distintas y con múltiples variables.

El análisis de componentes principales demostró ser eficiente para lograr resultados de clusterización válidos con una fracción de las variables, permitiendo obtener agrupaciones que aportan conocimiento.

Del proceso de agrupaciones se obtiene que la actual agrupación de tiendas no responde a ninguna agrupación, ni por variables de utilización de espacio, ni variables socioeconómicas lo que habre la posibilidad de revisar las agrupaciones según estrategias comerciales.

El proceso de asignación de surtido para los grupos de tiendas no fue posible de realizar por las siguientes razones: 1) Nunca se pudo consensuar una función objetivo, 2) falta de datos respecto del tamaño de los productos, esencial para obtener la rentabilidad por metro cuadrado.

Pese a esto la herramienta es flexible como para la implementación de una solución a este problema en el marco de la estructura de datos y procesos diseñado. Basta crear una rutina en R, la que puede ser ejecutada de manera similar al *clustering*.

En cuanto a los métodos de pronóstico de demanda probados se recomienda la utilización de medias móviles y redes neuronales artificiales, descartando la utilización del método ARIMA. Con los primeros dos métodos se obtienen los mejores resultados y en tiempos que permiten su ejecución masiva en series de tiempo.

El método ARIMA presenta una serie de complicaciones en su implementación que lo hacen poco atractivo para ser utilizado en la determinación de grandes cantidades de series de tiempos y con demandas tan intermitentes como las observadas en el presente trabajo. En primer lugar, el tiempo de ejecución hace imposible su aplicación en tiempos razonables, luego se observa que la eliminación de la tendencia, genera más ruido que estacionalización de la serie. Un tercer punto a destacar es que desde el punto de vista teórico la intermitencia de la demanda hace necesario modelos más complejos para la explicación los procesos subyacentes en los datos, lo que no es recomendable dados los costos de aplicar modelos complejos a grandes cantidades de series. Esto no es el *core bussiness* de Easy S.A. y no aporta mucho rédito económico, sobre todo si las series presentan poca rotación y por lo tanto pocas ventas.

Finalmente, el trabajo realizado presenta múltiple opciones de aplicación tanto en áreas de retail como en otras áreas del conocimiento, ciencias sociales, marketing, medicina, etc. Básicamente el modelo diseñado es aplicable a cualquier problema donde se necesite administrar grandes cúmulos de datos para encontrar información a partir de estos.

Un último punto a destacar es el bajo costo económico en su implementación, ya que el proyecto ha sido desarrollado íntegramente con herramientas *open source*.

6.3 Recomendaciones y trabajos futuros

Para el proceso de pronóstico de demanda se recomienda la utilización de medias móviles y redes neuronales artificiales, descartando la utilización del método automatizado ARIMA presentado en este trabajo. Mientras que para el proceso de generación agrupaciones se recomienda continuar la metodología presentada en la presente memoria.

Se presenta como interesante la aplicación de pronóstico de demanda para un nivel de agrupación distinto de los datos, por ejemplo a nivel de tiendas, de categorías o de productos para todas las tiendas.

Para la implementación final del proceso y del *data warehouse* se recomienda continuar con las herramientas computacionales propuestas en la construcción del prototipo.

Como trabajo futuro se recomienda continuar con la experimentación sobre los parámetros de las redes neuronales artificiales. Variables a probar son: número de neuronas de entrada a la red, número de neuronas en la capa oculta de la red, función de transferencia.

Con respecto a los métodos de agrupación se recomienda continuar las experimentaciones con las SOFM, específicamente con la una estructura rectangular en lugar de la grilla hexagonal utilizada en el presente trabajo.

Se presenta como interesante la aplicación de agrupaciones según variables de ventas en tiendas, por ejemplo a nivel de ventas de categorías por tiendas o a nivel de productos por tiendas.

7. Bibliografía

1. **Cencosud.** *Memoria anual 2008.* Santiago, Chile : s.n., Marzo de 2009.

2. *Forecasting aggregate retail sales: a comparison of artificial neuronal networks and traditional methods.* **Alon Ilan, Qi Min, Sadowski Robert.** 2001, Journal of retailing and consumer services, págs. 147-156.

3. **Aluja, Tomàs.** La minería de datos, entre la estadística y la inteligencia artificial. Barcelona : Universitat Politècnica de Catalunya, 2001. Vol. 25, 479-498.

4. *The KDD process for extracting useful knowledge from volumes of data.* **Fayyad Usama, Piatetsky-Shapiro Gregory,Smyth Padhraic.** 11, November de 1996, Comunication of the ACM, Vol. 39, págs. 27-34.

5. **Rebolledo, Victor.** *Plataforma para la extracción y almacenamiento del conocimiento extraido de los web data.* Santiago : s.n., 2008.

6. *The Data Warehouse toolkit: The complete guide to dimensional mideling.* **Kimball R., Moss M.** s.l. : Wiley, 2002.

7. **Villavicencio, Fredi Palominos.** Unidad III Modelamiento multidimencional, metodología Warehousing. Santiago : s.n.

8. **W. Inmon, M. Moss.** *Building the datawarehouse.* s.l. : Wiley, 2002.

9. **R. Kimball, J. Caserta.** *The Data Warehouse ETL Toolkit: Practical tecniques for extracting, cleaning, conforming and delivering data .* s.l. : Wiley, 2004.

10. **Juan, Velásquez.** Apunts del curso IN830 "Data Warehousing". Santiago : Universidad de Chile, 2007. Vol. Semestre Otoño.

11. *Sistemas de información orientados a la toma de decisiones: El enfoque multidimensional.* **Antonio Muñoz San Roque, Álvaro Sanchez Miralles, Isabel Dapena Bosquet.** Madrid : Asociación de ingenieros del ICAI, Junio de 2005, Anales de mecánica y electricidad, Vol. LXXXII, págs. 18-23.

12. **Juan D. Valásquez, Vasili Palade.** *Adaptative Web Site, A Knowledge Extraction From Web Data Approach.* s.l. : IOS Press, 2008.

13. **Mena, J.** *Data mining your website.* s.l. : Digital Press, 1999.

14. *Data mining: Statistics and more?* **Hand, D.J.** 2, s.l. : The american statistician, 1998, Vol. 52, págs. 112-118.

15. **Rao, C.R.** *Statistics and Truth.* New Delhi : CSRI, 1989.

16. **Carolina, Mejia Cols Aimara.** *Predicción del índice bursátil Caracas utilizando un híbrido entre metodología ARIMA y redes neuronales artificiales.* Mérida : Universidad de los Andes, 2006.

17. **Glenn, Fung.** *A comprehensible overview of basic clustering algorithms.* 2001.

18. **Lückeheide, Sandra.** *Segmentación de contribuyentes que declaran IVA utilizando técnicas de data mining.* 2007.

19. **Law, Au.** *A neural network model to forecast Japanese demand for travel to Hong Kong.* 1999.

20. *Predicción de series temporales con redes neuronales: Una aplicación a la inflación Colombiana.* **Santana, Juan Camilo.** 1, s.l. : Revista Colombiana de estadística, 2006, Vol. 29, págs. 79-92.

21. **Chicago, The University of Illinois at.** *Predicción del índice bursatil de Caracas utilizando un híbrido entre la metodología ARIMA y redes neuronales artificiales.* Mérida : s.n., 2006.

22. *Automatic seasonal auto regressive moving average models and unit root test detection.* **Siana Halim, Indriati N Bisono.** 4, Marzo de 2008, International Journal of Management Science and Engineering Manage, Vol. 3, págs. 266-274.

23. **Lucas, David Casado de.** *Protocolo para la identificación de modelos ARIMA en series temporales (según los pasos de Box-Jenkins).* Madrid : Departamento de Estadística Universidad Carlos III de Madrid, 2006.

24. **Patricio, Seguel.** *Identificación de oportunidades de negocio a partir de análisis de compra en una cadena de home improvement.* Santiago : s.n., 2006.

25. *Clustering of the self orginizing map.* **Vesanto Juha, Alhoniemi Esa.** IEEE transactions on neuronal networks.

26. **Siddheswar Ray, Rose H. Turi.** *Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation.* Victoria : School of Computer Science and Software Engineering Monash University, 2005.

27. **Rengifo, Hector Fabio Cadavid.** *Estrategias para la detección y corrección automática de errores en fuentes de datos.* 2007.

29. **Salvador, Figueras.** 5campus.org. *ucursos.ing.uchile.cl.* [En línea] 2001.

30. **Lon-Mu Liu, Siddhartha Bhattacharyya, Stanley L. Sclove, Rong Chen.** *DATA MINING ON TIME SERIES:AN ILLUSTRATION USING FAST-FOOD RESTAURANT FRANCHISE DATA.* Chicago : Department of Information and Decision Sciences, The University of Illinois at Chicago, 2001.

Anexo A: Presentación del retail en Chile

"*Retail*" es la denominación que se les entrega a la industria de venta a minoristas, supermercados, tiendas de conveniencia, tiendas de ropa, etc.

En el contexto latinoamericano de grupos con operaciones en retail, las empresas Chilenas juegan un rol destacado, situándose dentro de los más grandes retailers de la zona.

La tabla A.1, muestra a las Chilenas, Falabella, Cencosud y D&S dentro de los 10 más grandes retailer de la zona.

Tabla A.1: Ranking de los 10 retailers más grandes de Latino América

Top 10 en Latinoamérica	Ventas MMUS$	País
Walmex	33,711	México
Falabella	12,194	Chile
Cencosud	8,039	Chile
Liverpool	7,474	México
Natura	5,814	Brasil
Soriana	5,679	México
Lojas Americanas	5,587	Brasil
Grupo Elektra	4,856	México
Pão Açurar	4,306	Brasil
D&S	3,485	Chile

Fuente: elaboración propia con datos S&P, 2006

Reseña histórica del retail en Chile

El retail en Chile nace con pequeños empresarios, quienes montan empresas familiares con giro en la compra y venta de productos.

A finales de los 80 y principios de los 90 se vive la revolución del *retail* en Chile, amparado en las condiciones de estabilidad porlítica y económica esta industria crece notablemente.

En una primera etapa las empresas de la industria muestran una agresiva estrategia de crecimiento horizontal para luego evolucionar hacia una integración vertical, donde sustituyen proveedores y finalmente integrar servicios complementarios al cliente (el mercado financiero principalmente)

Importancia del retail como sector productivo

Desde los años 90 la industria ha vivido sostenidamente una gran expansión, donde las tasas de crecimiento del sector superan a las de la economía.

En la figura A.2 se muestra la evolución de las ventas, con una fuerte tendencia al alza para los años 2003 hasta el 2007.

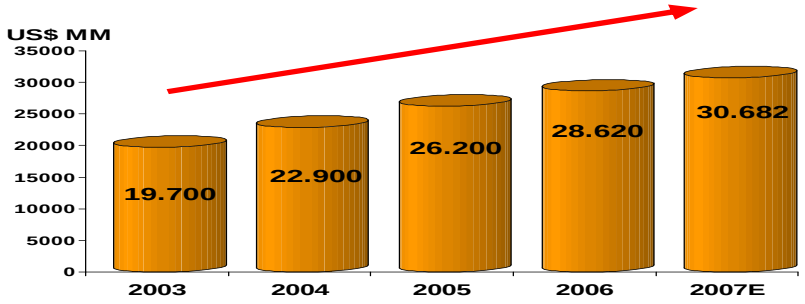


Figura A.2: Evolución de las ventas del sector retail en Chile. Fuente: Mercado del retail en Chile, Global Property Solutions 2007

De esta forma el *retail* se presenta como uno de los sectores productivos más importantes para el país, llegando a representar una importante fracción del PIB Chileno. De hecho, en el año 2007, el sector obtuvo ventas por US$3,1 millones, los que significo un 21,7% del PIB para ese año[81].

El *retail* chileno en Latinoamérica

El *retail* Chileno ha influenciado significativamente la realidad de la industria a nivel Sudamericano, como resultado de su política fuertemente expansionista que la lleva a tomar posiciones en el resto de los países de la región.

En la tabla A.2 se muestran los principales *retailers* y su presencia en el mercado latinoamericano.

Tabla A.2: Presencia de los retailers en Sudamérica

Retailer	Perú	México	Argentina	Colombia	Brasil
D&S	Presencia		Presencia	Presencia	
Ripley	Presencia				
Falabella	Presencia		Presencia	Presencia	
Cencosud	Presencia		Presencia	Presencia	Presencia

Fuente: Elaboración Propia

Estructura de negocios en la industria del retail

El retail puede clasificarse de acuerdo a los distintos canales de venta que presenta: 1) supermercados, 2) grandes tiendas, 3) ferreterías u "*home improvement*",

[81] Fuente AC Nielse, www.clnielsen.cl

4) farmacias y perfumerías, 5) comercio tradicional, 6) consumo local y 7) otros formatos.

La figura A.2 se puede ver que los supermercados, las tiendas por departamento, y las tiendas de home *improvement* reúnen más del 65% de las ventas totales de la industria.

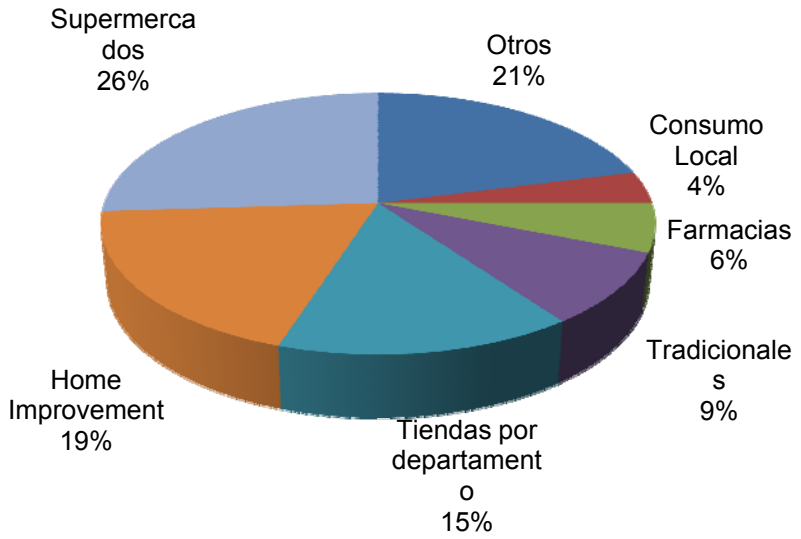


Figura A.2: Distribución de ventas según formato (año 2006). Fuente: AC Nielsen

La categoría "Otros" abarca tiendas especializadas en una línea de productos, como vestuario y tecnología, y está constituido en su mayoría de pequeñas empresas.

La categoría "Consumo local" incluye restaurantes y fuentes de soda, mientras que "Tradicionales" comprende negocios pequeños, incluyendo kioscos y botillerías. El sector de las "Farmacias" corresponde al más concentrado en el país.

Como se aprecia el formato de mejoramiento del hogar se presenta como uno relevante dentro de la industria con cerca del 19% de las ventas al año 2006.

Categoría de ferreterías u *Home improvement*

Este sector, si bien es de gran crecimiento, se encuentra sujeto a las variaciones de sectores sensibles a cambios en los ciclos económicos, como el de la construcción.

Se distinguen como los principales actores de este mercado a Cencosud con su marca Easy y Falabella mediante Sodimac, además de las cadenas especializadas en materiales de construcción, MTS y Construmart y las ferreterías independientes que son, en general, muy fragmentadas

La figura A.3 muestra la participación de mercado de los actores antes descritos para el año 2006, con un nivel de ventas avaluado en US$ 5800 Millones para la industria.

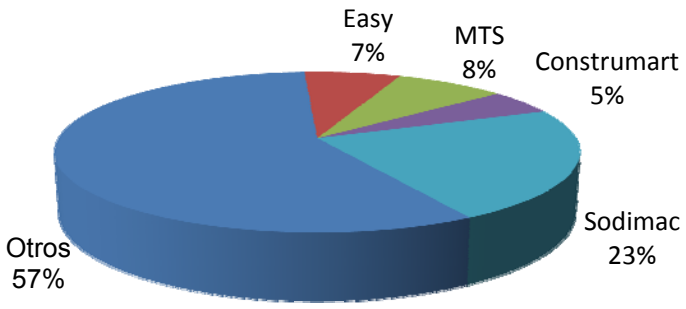


Figura A.3: Participación de mercado, Home Improvement.. Fuente: Fitch Ratings

Actualidad económica del retail en Chile

La actualidad del retail en Chile es de una disminución en la tendencia alsista que mostraban sus ventas hasta el año 2007. Esto explicado en gran medida por la situación de crisis económica mundial. Lo que ha obligado a la industria a moderar sus inversiones.

No obstante esto, dentro del ranking de las 250 mayores empresas del *retail* realizado por *Delloite Global Powers Retailing 2008*, se encuentran posicionadas 2 empresas chilena, Cencosud en la posición número 119 facturando 5.864 millones de dólares el año 2007 y Falabella en la posición número 159 facturando por su parte 4.302 millones de dólares, siendo número 2 y 5 en Latinoamérica, respectivamente.

Por último y no menos relevante ambos aparecen en el ranking top 50 de crecimiento entre 2001 y 2006, con un crecimiento de 28,6% para Cencosud que le vale el puesto número 13 y un 24,3% el puesto 12 para Falabella.

Anexo B: Estadística

B.1 Clasificación de los métodos estadísticos para el pronóstico de series de tiempo

Estos se dividen principalmente en dos grandes grupos: Cuantitativos y Cualitativos.

Los métodos Cualitativos se caracterizan por la utilización del criterio, experiencia, buen juicio e intuición en la elaboración de pronósticos. Su uso es apropiado cuando el tiempo para la elaboración del pronóstico es poco, cuando los datos son de naturaleza poco confiable o el acceso a ellos es difícil o cuando el pronóstico es pensado a largo plazo[82]. Sin embargo es un método de poca utilidad práctica para la empresa toda vez que el conocimiento sigue concentrado en los expertos.

Ejemplos de métodos cualitativos son: Método Delphi, Analogías históricas, Consensos de panel, etc.

Los métodos Cuantitativos responden a la aplicación de herramientas matemáticas a los datos históricos relacionados con las variables que se desean pronosticar. Estos se pueden a su vez subagrupar en métodos Univariantes y métodos Multivariantes.

Los métodos univariantes trabajan con los datos históricos de la variable a predecir, mientras los métodos multivariantes trabajan identificando otras variables relacionadas a la variable a pronosticar. Una vez identificadas estas variables se construye un modelo matemático que cuantifica el impacto de cada variable determinante en el valor de la predicción.

Como ejemplo de métodos uinivariantes podemos mencionar: Extrapolación de curvas de tendencia, Suavisamiento Exponencial, Método de Holt – Winters, Método de Box – Jenkins (ARIMA).

Como ejemplo de métodos multivariantes podemos mencionar: Regresiones múltiples, Modelos econométricos, Modelo de funciones de Transferencia.

B.2 Modelo ARIMA utilizados en la metodología de Box Jenkins para la predicción de series de tiempo

La metodología de Box.Jenkins propone la utilización de 4 modelos matemáticos de predicción: Modelo Auto Regresivo (AR), Modelo de promedios móviles (MA), Modelo Auto regresivo de promedios móviles (ARMA) y Modelos Auto regresivos integrado (ARIMA).

a) Modelo Auto Regresivo (AR): El valor de la observación dependerá de las p observaciones anteriores.

[82] Esto debido a que los métodos matemáticos o de Machine Learning pierden certeza a medida que el tiempo de pronóstico aumenta.

$$y_t = \phi_1 * y_{t-1} + \phi_2 * y_{t-2} + \dots + \phi_p * y_{t-p} + \varepsilon_t \quad (2.7)$$

Con

- y_t el valor del pronóstico
- *p* el número de observaciones pasadas (rezagos utilizados)
- ϕ_i un conjunto de regresores de las observaciones pasadas
- ε_t el error en la predicción para la observación *t*

Equivalentemente utilizando el operador de Retardo B se llega a la forma abreviada de un proceso AR:

$$(1 - \phi_1 * B - \phi_2 * B^2 - \square - \phi_p * B^p) * y_t = \varepsilon_t \quad (2.8)$$

O equivalentemente:

$$\Phi_p(B) * y_t = \varepsilon_t \quad (2.9)$$

b) Modelo de promedios Móviles (MA): En este modelo el pronóstico de la variable se hace en función de la combinación lineal de los q errores de predicción previos.

$$y_t = \varepsilon_t - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - \dots - \theta_q * \varepsilon_{t-q} \quad (2.10)$$

Con

- y_t el valor del pronóstico
- q el numero de errores pasados
- Θ_i un conjunto de regresores
- ε_t el error en la predicción para la observación t

Utilizando el operador de retardo B:

$$y_t = \varepsilon_t * (1 - \theta_1 * B - \theta_2 * B^2 - \square - \theta_q * B^q) \quad (2.11)$$

O equivalentemente:

$$y_t = \varepsilon_t * \Theta_q(B) \quad (2.12)$$

c) Modelo Autorregresivo de Promedios móviles (ARMA): En este caso el modelo propuesto es una combinación lineal de los modelos AR y MA.

$$y_t = \phi_1 * y_{t-1} + \phi_2 * y_{t-2} + \dots + \phi_p * y_{t-p} + \varepsilon_t - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - \square - \theta_q * \varepsilon_{t-q} \quad (2.12)$$

Con

- p el número de observaciones pasadas (rezagos utilizados)
- Φ_i un conjunto de regresores de las observaciones pasadas
- y_t el valor del pronóstico
- q el numero de errores pasados
- Θ_i un conjunto de regresores
- ε_t el error en la predicción para la observación t

O mediante el uso del operador rezago B:

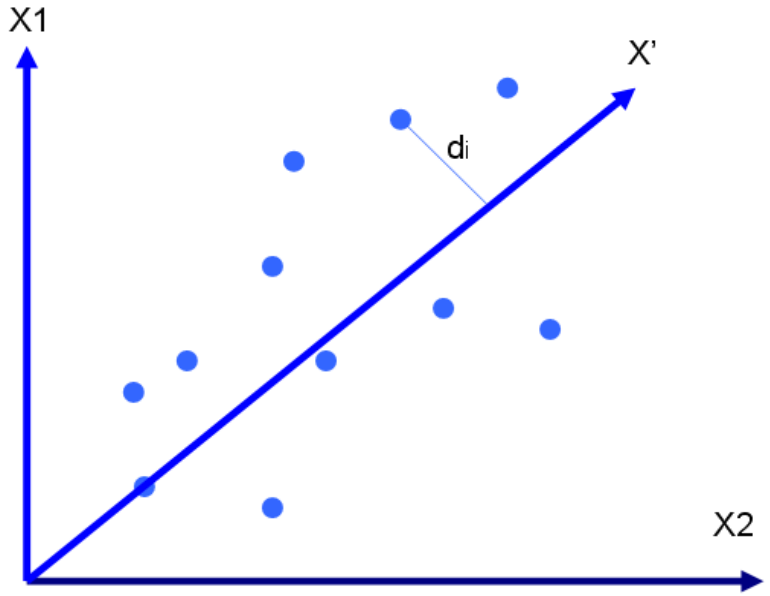
$$\Phi_p(B) * y_t = \varepsilon_t * \Theta_q(B) \qquad (2.12)$$

B.3 Análisis de Componentes Principales

El análisis de componentes principales es una técnica estadística que busca reducir la dimensionalidad de un conjunto de observaciones.

La idea general de este método es la de crear nuevas variables descriptivas, correlacionadas con las variables originales, rotando el sistema de coordenadas de las variables originales de forma de lograr un nuevo sistema de coordenadas, ortogonal al original, que capture la varianza de las observaciones originales.

Figura B.1: Cambio de coordenadas en el método ACP



Fuente: Apuntes estadística, Nancy Lacourly, 2004, Universidad de Chile

El método matemático subyacente busca replicar lo que se muestra en la figura B.1, crear un nuevo sistema de referencias que describan las variables minimizando la varianza que presentan en el sistema de coordenadas inicial.

Este nuevo sistema se busca minimizando la distancia di, la distancia de cada punto a la nueva dimensión.

La metodología ACP comprende los siguientes pasos:

- Análisis de la matriz de covarianzas y creación de las nuevas coordenadas: La importancia de este paso radica en extraer las correlaciones existentes entre las variables originales y, a partir de estas correlaciones, generar el nuevo sistema de coordenadas.

Sea *X*, la matriz de las *j* variables observadas para las *m* observaciones y *S* la matriz de correlaciones de *X*.

Dado que *S* es simétrica, es diagonalizable de la forma $S = T*\lambda*T'$ con $T=\{t_1,t_2,\ldots,t_j\}$, $\lambda =(\lambda_1, \lambda_2,\ldots,\lambda_j\}$ los autovalores de la matriz de correlaciones y $T*T'=T'*T$=Identidad (I).

Se definen las nuevas variables, *Y*, como la siguiente transformación lineal:
$Y_j= X*t_j$ para todo *j* entre (1, *j*) y para todo i entre (1, *m*).

Lo que se obtiene con esta transformación en una rotación del sistema referencial original, de forma que la corvarianza entre Y_1 y $X_1\ldots.X_p$ es simplemente λ_1*t_1'.

$$Cov(Y,X) = \frac{Y'}{X} \underset{Y=X*T}{\Longrightarrow} Cov(Y,X) = \frac{T'*X'*X}{n}$$

$$Cov(Y,X) = T'*\left(T*\lambda*\mathrm{T}'\right) \Rightarrow \mathrm{Cov(Y,X)} = \lambda*\mathrm{T} \quad (B.9)$$

- <u>Selección de los factores</u>: El objetivo de este método es el de disminuir la dimensionalidad del problema, manteniendo un nivel de información que sea relevante para concluir sobre el fenómeno observado.

  Como se observa en la ecuación B.9, el nivel de relación entre las nuevas dimensiones (*Y*) y las originales (*X*) está representado por el autovalor de la matriz *T*, es decir los valores de λ de cada nueva dimension Y_j.

  Este valor λ indica la varianza de las variables originales (*X)* explicada por la nueva coordenada Y_j .

  Luego, el porcentaje total explicado por las *n* primeras componentes principales será:

$$Varianza\ explicada = \frac{\sum_{i=1}^{n}\lambda_i}{traza(\lambda)} \qquad (B.10)$$

  La elección de los factores se realiza en función de la varianza que se desea explicar, de forma que si se quiere representar W% de la varianza del problema original, el número de componentes principales, n, será:

$$n\ tal\ que\ W \geq \frac{\sum_{i=1}^{n}\lambda_i}{traza(\lambda)} \qquad (B.11)$$

- <u>Análisis de la matriz factorial e interpretación de los factores</u>: Una vez escogidos los factores a utilizar, se construye la matriz factorial la que corresponde a la matriz de correlaciones entre cada componente principal y cada variable del problema original.

Esta herramienta permite identificar las relacines existentes entre cada variable observada (X) y las variables subyacentes encontradas (Y), de esta forma se puede interpretar cada nueva componente principal como un conjunto de variables originales.

- Calculo de las puntuaciones factoriales: las puntuaciones factoriales son las transformaciones de los observaciones a las nuevas componentes principales.

  Es decir la coordenada de la observación n para la componente principal i es la combinación lineal de las j variables originales con las covarianzas entre las componentes principales y las variable originales, según la siguiente formulación:

$$Y_{in} = Cov(Y_i, X_1) * X_{1n} + \ \dots\dots + Cov\left(Y_i, X_j\right) * X_{jn} \qquad (B.11)$$

  De esta forma se ha reducido el número de variables observadas a un numero menor de variables subyacentes en el problema, que son capaces de explicar la cantidad de varianza reducida.

  Además se han transformado las observaciones originales a este nuevo sistema de coordenadas, sobre las cuales se pueden realizar los análisis de clusters o de clasificación que se requieran.

Anexo C Datos de la empresa

El siguiente anexo contiene la información característica de la empresa tanto para sus tiendas como para sus productos.

La tabla C.1 muestra enumera las tiendas junto con su código de identificación en el sistema SAP y el tamaño en metros cuadrados de sala.

Tabla C.1 Tiendas de Easy S.A. en Chile

Código Tienda	Nombre Tienda	Tamaño m2
E534	Antofagasta	8964
E525	Chillán	9001
E760	Curicó	8545
E592	El Belloto	10240
E781	El Llano	6314
E513	FISA	12726
E502	Kennedy	3184
E514	La Dehesa	9062
E510	La Florida	11398
E521	La Reina	11003
E512	La Serena	9607
E524	Linares	6980
E529	Los Ángeles	9357
E503	Maipú	12260
E585	Osorno	9207
E507	Puerto Montt	7631
E518	Quilin	11046
E646	Quillota	7901
E504	Rancagua	9992
E591	Talca	9394
E517	Temuco	10331
E508	Viña del mar	14043
E520	Valparaiso	6122

Fuente: elaboración propia, datos de la gerencia de procesos comerciales

Anexo D Cotización de equipos y software para la realización de la red

El presente capítulo presenta las cotizaciones para los equipos y software para la implementación de la solución propuesta en el capítulo 4 sección 4.1.a.

La cotización se realizó de acuerdo a los requerimientos indicados en el capítulo 4 sección 4.2.b

En la tabla D.1 se presenta la cotización para el equipo servidor y en la tabla D.2 se presenta la cotización para los requerimientos de software.

Tabla D.1 Cotización para el equipo servidor del sistema

Modelo	Precio	Características
Server HP Proliant ML115 AMD Opteron	$ 319.138	Procesador: Dual-Core AMD Opteron 2,2 GHz
		Memoria RAM: 1 GB, máximo 8GB
		Disco Duro: SATA 160 Gb, máximo 4
		No posee redundancia fuente de poder
ProLiant ML110 G5	478713	Procesador: Intel Pentium Dual Core Xenon 2,33 GHz
		Memoria RAM: 1 GB, máximo 8GB
		posee redundancia de fuente
Dell PowerEdge R200	746670	Procesador: Dua-Core E2180, 2 GHz
		Memoria RAM: 1 GB
		Discos Duros: 2 SATA 160 GB
		posee redundancia de fuente

Fuente: Elaboración propia con datos de www.pcfactory.cl y www.sym.cl

Tabla D.2: Cotización para software estipulados como requisito

Software	Software	Precio	Requerimientos y licencias
Requerimiento	Software	Precio	-----
Motor de Base de datos	MySQL	Gratis	-----
Servidor OLAP	Mondrian	Gratis	-----
Servidor Web	Apache	Gratis	-----
	Tomcat	Gratis	-----
Programación	Java	Gratis	-----
	Python	Gratis	-----
	C	Gratis	-----
	R	Gratis	-----
Sistema operativo	Wundows XP	$ 106.372	Procesador: 0,3 GHz
			Memoria RAM: 128 MB
			Disco Duro: 1,5 GB
			Licencias: 1
	Windows vista business	$ 107.447	Procesador: 1 GHz
			Memoria RAM: 1GB
			Disco Duro: 1,5 GB
			Licencias: 1
	Windows 2003 small business server	$ 341.915	Procesador: 0,5 GHz
			Memoria RAM: 256 MB
			Disco Duro: 4 GB
			Licencias: 1 Server + 5 Clientes

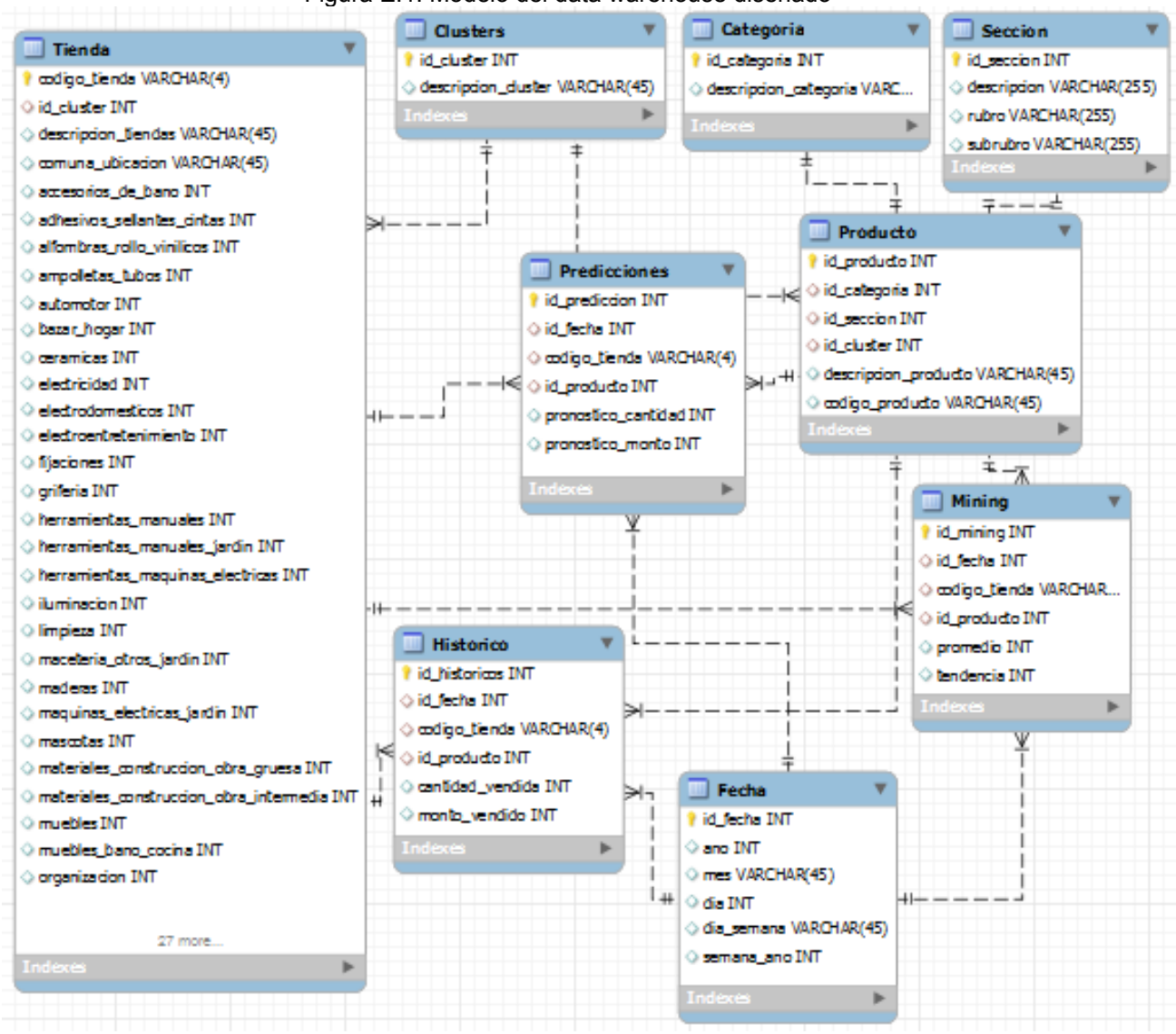
Fuente: Elaboración propia con datos de www.pcfactory.cl

Anexo E: Base de datos

En el siguiente anexo se especificará en detalle la base de datos construida.

La figura E.1 muestra la base de datos del data warehouse con todos sus campos.

Figura E.1: Modelo del data warehouse diseñado



Fuente: Elaboración propia

A continuación se describen las tablas de Cluster, Categoria y Secciones, no descritas en el capítulo 4 seccion 4.4.1 b.

Tabla Cluster: Contiene la información sobre agrupaciones de tiendas existentes. Sus campos se describen a continuación:

Campo id_cluster: Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.

Puede tomar un número entero entre 1 y N, con N el número de combinaciones posibles entre los *clusters*. De esta forma se permite que un producto este asociado a todas las tiendas o solo a un grupo de tiendas.

Por ejemplo, si se estima que existen 3 clusters, existinrán entonces 7 *id_cluster*, identificando las agrupaciones de los clusters 1, 2 y 3 (todas las tiendas), 3, 1, 2, 3 (cada cluster por separado) y los grupos de clusters 1 y 2, 1 y 3, 2 y 3 (subgrupos de clusters).

Campo descripción_cluster: Campo no numérico (*varchar(255)*) con la descripción del *cluster*.

Tabla Categoría: Contiene la información sobre categorías existentes. Sus campos se describen a continuación:

Campo id_categoria: Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.

Campo descripción_categoria: Campo no numérico (*varchar(255)*) con la descripción de la categoría.

Tabla Seccion: Contiene la información sobre Secciones existentes. Un registro cosntituye toda una línea jerárquica, sección, rubro y subrubro. Sus campos se describen a continuación:

Campo id_sección: Campo numérico (*integer*) identificador de la tupla, constituye la llave primaria de la tabla y llave foránea para el resto de las tablas.

Campo descripción_sección: Campo no numérico (*varchar(255)*) con la descripción de la categoría.

Campo rubro: Campo no numérico (*varchar(255)*) con la descripción del rubro.

Campo sub-rubro: Campo no numérico (*varchar(255)*) con la descripción del sub-rubro.

Anexo F: Programación

En el presente anexo se busca especificar los elementos de programación relevantes en el proceso de construcción y de minado de información.

Para lo anterior se adjunta el código utilizado junto con la explicación de lo realizado en cada rutina programada.

E.1 Pronóstico mediante redes neuronales artificiales

El algoritmo contempla los siguientes pasos:

1. Normalización de valores
2. Generación del correlograma de la serie
3. Entrenamiento de la serie
4. Realización de predicciones
5. Re-escalamiento de los datos
6. Comparación de los resultados
7. Elección de la mejor predicción

E.2 Pronóstico mediante medias móviles

El algoritmo contempla los siguientes pasos:

1. Transformación logarítmica de los datos
2. Generación de predicciones
3. Re-escalamiento de los datos
4. Comparación de los resultados
5. Elección de la mejor predicción

```
#recibe como parámetro la serie a predecir
forecastingMA<-function(serie){
serie.original<-as.matrix(serie)
largo.original<-length(serie.original)

#transformacion logarítmica para estabilizar variaza de la serie
for (i in 1:largo.original) serie<-lg(serie.original+1)
a.predecir<-serie[(largo.original-7):largo.original]
serie<-serie[1:(largo.original-8)]

#idices de errores a utilizar para cuantificar el error
errores<-matrix[1:8]
errores[1]=0.2
errores[2]=0.2
errores[3]=0.15
errores[4]=0.15
errores[5]=0.1
errores[6]=0.1
errores[7]=0.05
errores[8]=0.05
```

```
#prediccion 2 semanas
prediccion2<-matrix[1:10]
for (i in 1:2) prediccion2[i]<-serie[(largo.original-2+i)]
for (i in 1:6) prediccion2[i+2]<-(prediccion2[i+1]+prediccion2[i])/2
pred.2<-prediccion2[3:10]
MSE.2<-0
for (i in 1:8) MSE.2<-MSE.2+errores[i]*(a.predecir[i]-pred.2[i])^2

#prediccion con 4 semanas
prediccion4<-matrix[1:12]
for (i in 1:4) prediccion4[i]<-serie[(largo.original-4+i)]
for (i in 1:8) prediccion4[i+4]<-(prediccion4[(i+3)]+prediccion4[i+2]+prediccion4[(i+1)]+prediccion4[(i)])/4
pred.4<-prediccion4[5:12]
MSE.4<-0
for (i in 1:8) MSE.4<-MSE.2+errores[i]*(a.predecir[i]-pred.4[i])^2

#prediccion con 8 semanas
prediccion8<-matrix[1:16]
for (i in 1:8) prediccion8[i]<-serie[(largo.original-8+i)]
for (i in 1:8) prediccion8[i+8]<-
(prediccion8[(i+7)]+prediccion8[i+6]+prediccion8[(i+5)]+prediccion8[(i+4)]+prediccion8[(i+3)]+prediccion8[(
i+2)]+prediccion8[(i+1)]+prediccion8[(i)])/8
pred.8<-prediccion4[9:16]
MSE.8<-0
for (i in 1:8) MSE.8<-MSE.8+errores[i]*(a.predecir[i]-pred.8[i])^2


#predicción con 12 semanas
prediccion12<-matrix[1:20]
for (i in 1:12) prediccion12[i]<-serie[(largo.original-12+i)]
for (i in 1:8) prediccion12[i+12]<-(prediccion12[(i+11)]+prediccion12[i+10]+prediccion12[(i+9)]
+prediccion12[(i+8)]+prediccion12[(i+7)]+prediccion12[(i+6)]+prediccion12[(i+5)]
+prediccion12[(i+4)]+prediccion12[(i+3)]+prediccion12[(i+2)]+prediccion12[(i)])/12
pred.12<-prediccion4[13:20]
MSE.8<-0
for (i in 1:8) MSE.12<-MSE.12+errores[i]*(a.predecir[i]-pred.12[i])^2

#elección del major prnóstico en function del MSE
error<-matrix[1:4]
error[1]<-MSE.2
error[2]<-MSE.4
error[8]<-MSE.8
error[12]<-MSE.12
if(MSE.2==min(error)) {pred<-pred.2, MSE<-MSE.2}
else if (MSE.4==min(error)) {pred<-pred.4, MSE<-MSE.4}
else if (MSE.8==min(error)) {pred<-pred.8, MSE<-MSE.8}
else pred<-{pred.12, MSE<-MSE.12}

#re-escalamiento de la predicccion elegida

for (i in 1:8) pred<-exp(pred)-1



#Lleno la lista de retorno
answer<-list(pred,MSE)
names(answer)[[1]]<-"prediccion"
```

```
names(answer)[[2]]<-"MSE"
return(answer)}
```

E.3 Pronóstico mediante ajuste de método ARIMA

El algoritmo contempla los siguientes pasos:

1. Eliminación de la tendencia en la serie
2. Transformación logarítmica de los datos
3. Diferenciación de la serie
4. Identificación del parámetro p del proceso AR subyacente
5. Identificación del parámetro q del proceso MA subyacente
6. Selección del mejor modelo AR
7. Selección del mejor modelo MA
8. Selección del mejor modelo ARMA
9. Comparación y elección entre modelos
10. Realización de predicciones
11. Re-escalamiento de datos

E.4 Agrupaciones mediante la utilización de *self organizing maps*

El algoritmo contempla los siguientes pasos:

1. Normalización de valores
2. Análisis de componentes principales
3. Generación de agrupaciones
4. Diagrama de mapa bidimensional con conteo de observaciones por centroide.
5. Entrega de las soluciones

```
#Parámetros: Archivo con datos y el tamaño de malla
clustering.SOFM.acp<-function(archivo ,malla)
{

#Carga de paquetes y funciones auxiliares
library(class)
library(kohonen)
source("ACP.R")
source("agregacion.R")
source("norma.R")

#Lectura y limpieza de datos
datos<-read.csv2(archivo,header=FALSE)
carga<-datos[,-1]
carga<-norma(carga)
tiendas<-as.matrix(datos[,1])
datos.n.s<-as.matrix(ACP(carga,0.8))
filas<-nrow(datos.n.s)
columnas<-ncol(datos.n.s)

clusterizacion<-list()
centros<-list()

#realizacion del entrenamiento de la malla
```

```
som<-som(datos.n.s,grid=somgrid(malla,malla,"hexagonal"))
clusterizacion[[1]]<-som$unit.classif
centros[[1]]<-som$count
diagrama<-plot(som, tip3="count")
gruposc<-as.matrix(clusterizacion[[1]])
grupos<-agregacion(tiendas,gruposc)

#Lleno la lista de retorno
answer<-list(grupos,grafico)
names(answer)[[1]]<-"grupos"
names(answer)[[2]]<-"grafico"
return(answer)}
```

E.5 Agrupaciones mediante la utilización del método de *kmeans*

El algoritmo contempla los siguientes pasos:

1. Normalización de valores
2. Análisis de componentes principales
3. Generación de agrupaciones
4. Búsqueda de la solución óptima
5. Búsqueda de la solución óptima según las restricciones
6. Entrega de las soluciones

```
#Parámetros: Archivo con datos, número mínimo y máximo de clusters y el número de soluciones iniciales para cada solución con k clusters
clustering.kmeans.acp<-function(archivo,min,max,corridas)
{

#Carga de paquetes y funciones auxiliares
source("ACP.R")
source("dist.inter.R")
source("agregacion.R")
source("dist.intra.R")
source("norma.R")

#Lectura y limpieza de datos
datos<-read.csv2(archivo,header=FALSE)
carga<-datos[,-1]
carga<-norma(carga)
tiendas<-as.matrix(datos[,1])
datos.n.s<-as.matrix(ACP(carga,0.8))
filas<-nrow(datos.n.s)
columnas<-ncol(datos.n.s)

#Inicialización de variables auxiliares
clusterizacion.op<-list()
clusterizacion.min<-list()
clusterizacion.max<-list()
clusterizacion.op.rest<-list()
wss.op<-1e06
wss.min<-1e06
wss.max<-1e06
wss.op.rest<-1e06
k.op<-0
```

```
k.min<-min
k.max<-max
k.op.rest<-0
#Conteo de observaciones distintas
for (i in 1:(filas-1)){
for (j in (i+1):filas){
if (sum(datos.n.s[i,]==datos.n.s[j,])==columnas)
{filas<-filas-1
i<-i+1}
else
{filas<-filas}
}}

#Existe un problema con el método si es que se tiene un numero de observaciones menores que centro,
es necesario identificar el número de observaciones distintas
if (filas==nrow(datos.n.s)){

#si son todas las observaciones distintas se usa este método
#guardo la corrida j
wss<-matrix(2:(filas-1))
clusterizacion<-list()
for (i in 1:length(wss)){
k.means<-kmeans(datos.n.s,centers=(i+1),nstart=corridas)
wss[i]<-
(dist.inter(datos.n.s,k.means$cluster))/(as.numeric(dist.intra(datos.n.s,k.means$cluster,k.means$centers))
)
clusterizacion[[i]]<-k.means$cluster
}

#busco el óptimo en base a la dist.intragrupo/dist.intergrupos
for (i in 1:length(wss)){
if (wss[i]<wss.op) {
k.op<-i+1
wss.op<-wss[i]
clusterizacion.op[[1]]<-clusterizacion[[i]]
}}

#Agrupacion óptima con mínimo como número de clusters
wss.min<-wss[(min-1)]
clusterizacion.min[[1]]<-clusterizacion[[(min-1)]]

#Agrupacion óptima con máximo como número de clusters
wss.max<-wss[(max-1)]
clusterizacion.max[[1]]<-clusterizacion[[(max-1)]]

#Agrupacion óptima conclusters entre máximo y mínimo como número de clusters
for (i in min:max){
if (wss[i-1]<wss.op.rest) {
k.op.rest<-i
wss.op.rest<-wss[i-1]
clusterizacion.op.rest[[1]]<-clusterizacion[[i-1]]
}}}

else{

#si son todas las observaciones iguales se usa este método
#guardo la corrida j
```

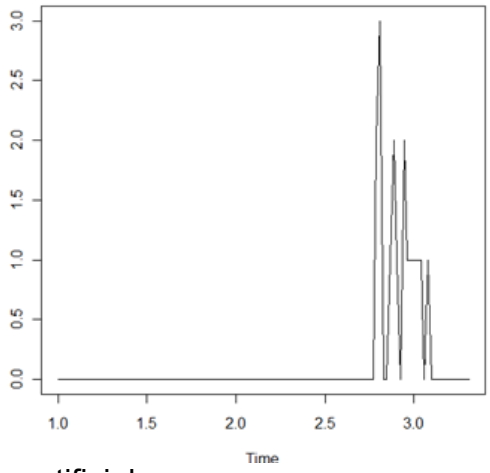
```
wss<-matrix(2:filas)
clusterizacion<-list()
for (i in 1:length(wss)){
k.means<-kmeans(datos.n.s,centers=(i+1),nstart=corridas)
wss[i]<-
(dist.inter(datos.n.s,k.means$cluster))/(as.numeric(dist.intra(datos.n.s,k.means$cluster,k.means$centers))
)
clusterizacion[[i]]<-k.means$cluster
}

#busco el óptimo en base a la dist.intragrupo/dist.intergrupos
for (i in 1:length(wss)){
if (wss[i]<wss.op) {
k.op<-i+1
wss.op<-wss[i]
clusterizacion.op[[1]]<-clusterizacion[[i]]
}}

#Agrupación óptima con mínimo como número de clusters
wss.min<-wss[(min-1)]
clusterizacion.min[[1]]<-clusterizacion[[(min-1)]]

#Agrupación óptima con máximo como número de clusters
wss.max<-wss[(max-1)]
clusterizacion.max[[1]]<-clusterizacion[[(max-1)]]

#Agrupación óptima con clusters entre máximo y mínimo como número de clusters
for (i in min:max){
if (wss[i-1]<wss.op.rest) {
k.op.rest<-i
wss.op.rest<-wss[i-1]
clusterizacion.op.rest[[1]]<-clusterizacion[[i-1]]
}}}

gruposminc<-as.matrix(clusterizacion.min[[1]])
gruposmaxc<-as.matrix(clusterizacion.max[[1]])
gruposc<-as.matrix(clusterizacion.op[[1]])
gruposrestc<-as.matrix(clusterizacion.op.rest[[1]])
gruposmin<-agregacion(tiendas,gruposminc)
gruposmax<-agregacion(tiendas,gruposmaxc)
grupos<-agregacion(tiendas,gruposc)
gruposrest<-agregacion(tiendas,gruposrestc)
wss.res<-list(wss,wss.op,wss.op.rest,wss.min,wss.max)

#Lleno la lista de retorno
answer<-list(gruposmin,gruposmax,grupos,gruposrest,wss.res,k.op,k.op.rest)
names(answer)[[1]]<-"grupos.min"
names(answer)[[2]]<-"grupos.max"
names(answer)[[3]]<-"grupos.op"
names(answer)[[4]]<-"grupos.op.rest"
names(answer)[[5]]<-"indicadores"
names(answer)[[6]]<-"clusters.op"
names(answer)[[7]]<-"clusters.op.rest"
return(answer)}
```

Anexo G Resultados

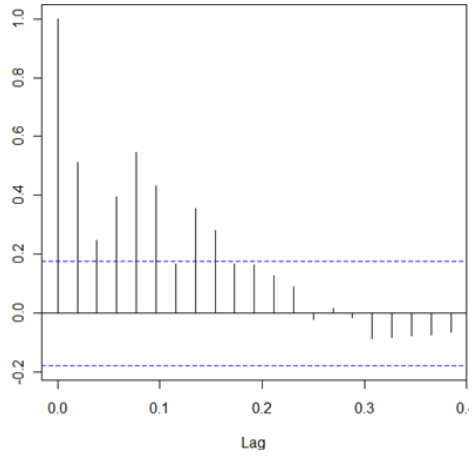
G.1 Pronóstico de demanda

En el siguiente anexo se busca mostrar los resultados obtenidos en el proceso de prueba en el muestreo de 15 series escogidas

Serie 826428
Serie Original



Método de redes neuronales artificiales
Autocorrelograma simple

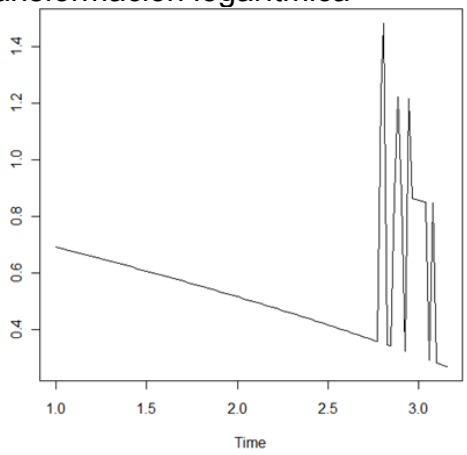


Predicciones

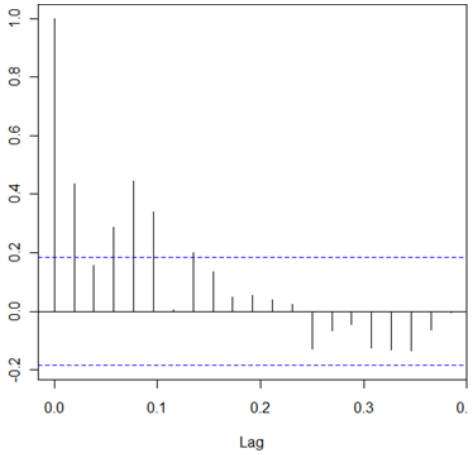
Semana	1	2	3	4	5	7	8	A predecir
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(6)	AR(5)	AR(4)	MA(12)	MA(13)	MA(11)	ARMA(6,0,12)
Akaike	-86,30	-79,68	-81,01	-121,84	***-120,11***	-108,85	-119,31

Predicciones

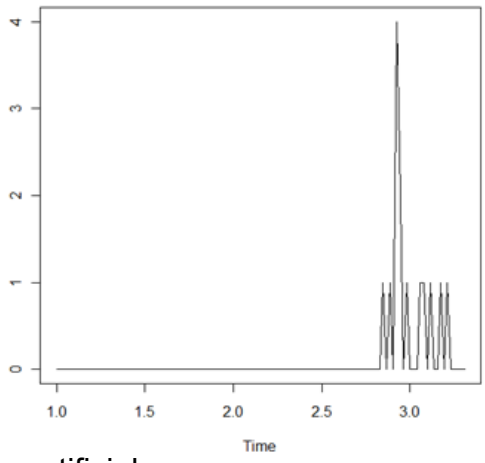
Semana	Predicción	A predecir
1	1	0
2	1	0
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0

Método de Medias móvile

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

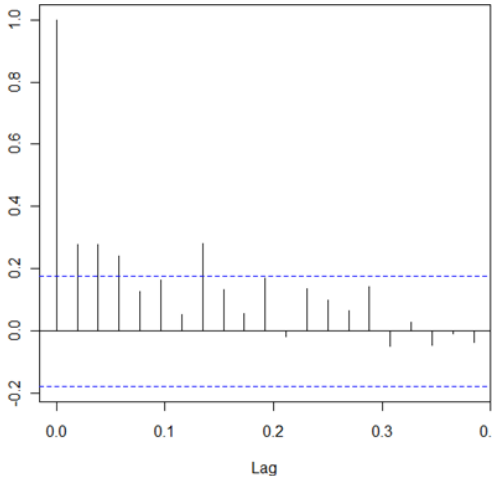
Serie 823851

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

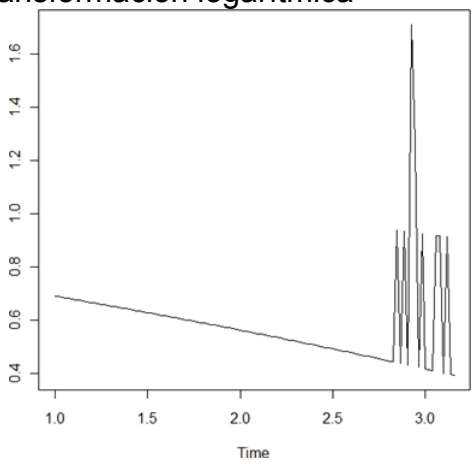


Predicciones

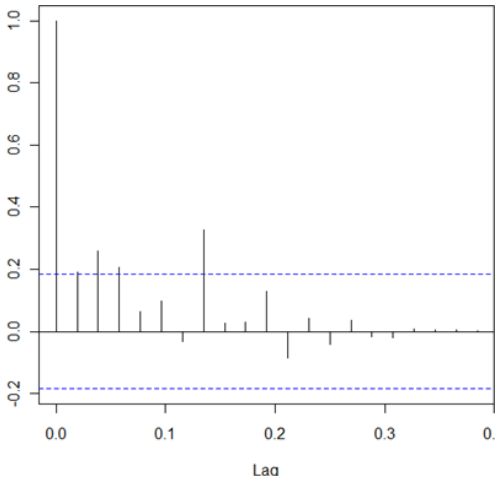
Semana	1	2	3	7	A predecir
1	0	0	0	0	1
2	0	0	0	0	0
3	0	0	0	0	1
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(7)	AR(6)	AR(8)	MA(5)	MA(6)	MA(7)	ARMA(7,0,7)
Akaike	-96,53	-81,01	***-95,08***	-84,41	-84,88	-94,96	-86,84

Predicciones

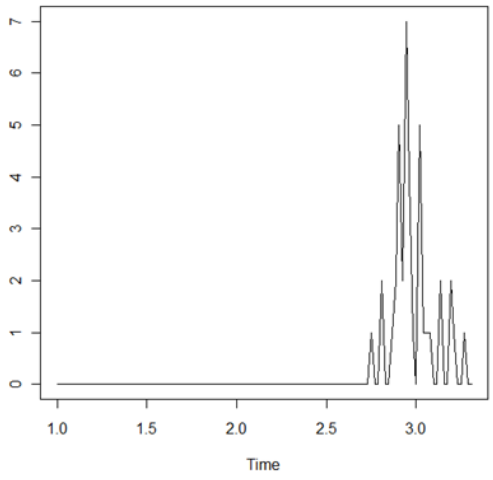
Semana	Predicción	A predecir
1	1	1
2	1	0
3	1	1
4	1	0
5	2	0
6	1	0
7	1	0
8	2	0

Método de Medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
1	0	0	0	0
0	0	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

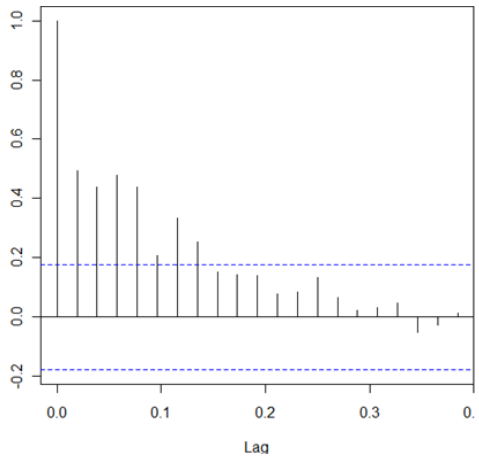
Serie 824683

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

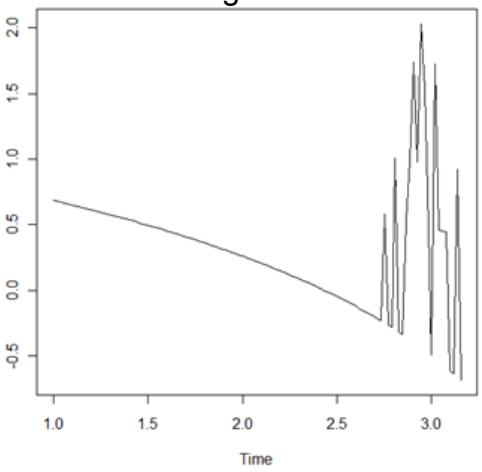


Predicciones

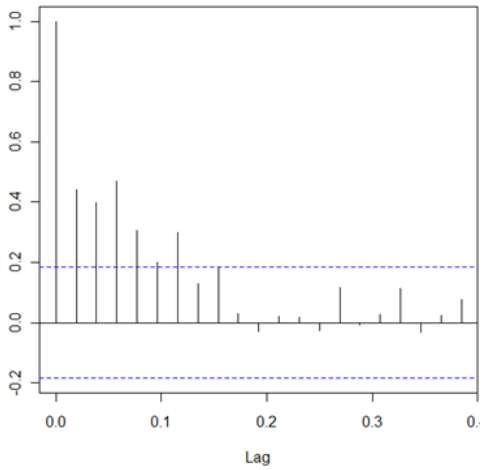
Semana	1	2	3	4	5	6	7	A predecir
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(3)	AR(2)	AR(4)	MA(8)	MA(7)	MA(9)	ARMA(3,0,9)
Akaike	97,98	109,08	99,95	***89,78***	90,31	88,51	91,71

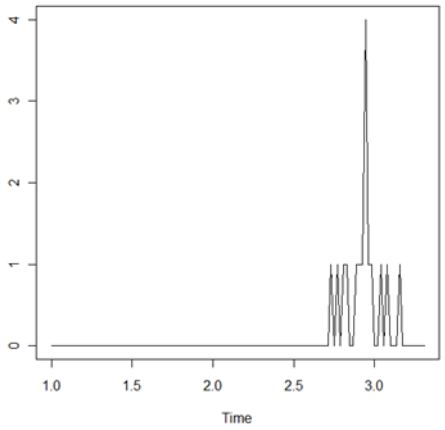
Predicciones

Semana	Predicción	A predecir
1	1	1
2	1	0
3	1	1
4	1	0
5	2	0
6	1	0
7	1	0
8	2	0

Método de medias móvile

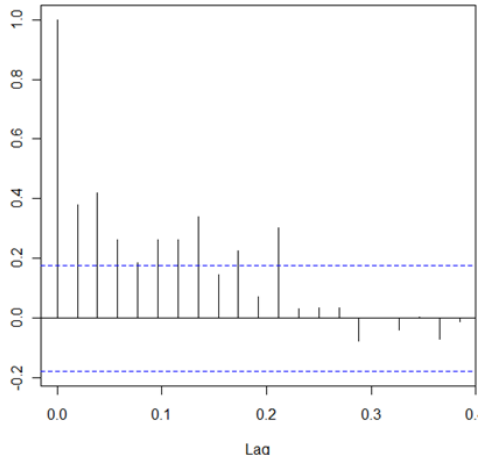
A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	1	0	1	1
2	0	0	0	1
1	1	1	1	1
0	1	1	1	1
0	0	1	0	1
1	0	1	0	1
0	1	0	0	1
0	0	0	0	0

Serie 824685
Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

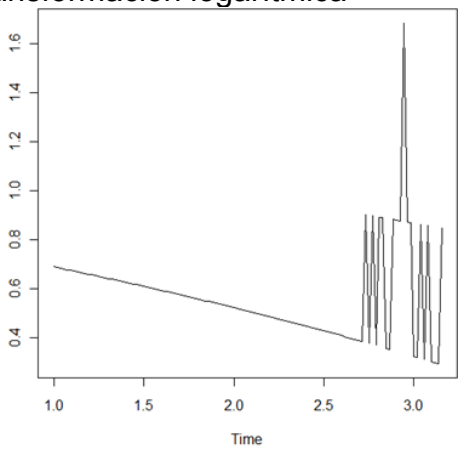


Predicciones

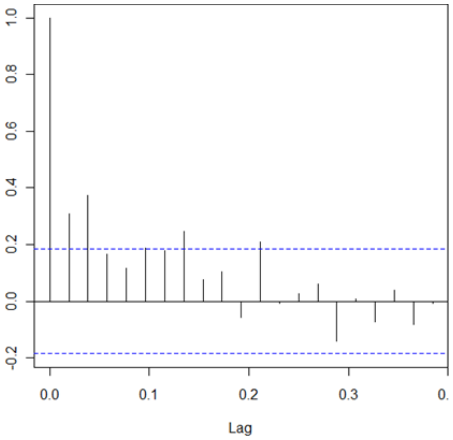
Semana	1	2	3	5	6	7	9	10	A predecir
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(2)	AR(1)	AR(3)	MA(9)	MA(8)	MA(10)	ARMA(2,0,8)
Akaike	-79,52	-70,35	-77,55	***-82,32***	-83,89	-81,43	-80,32

Predicciones

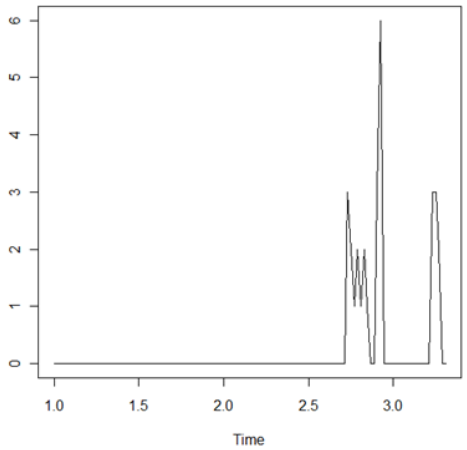
Semana	Predicción	A predecir
1	1	0
2	2	0
3	2	0
4	2	0
5	2	0
6	2	0
7	2	0
8	2	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

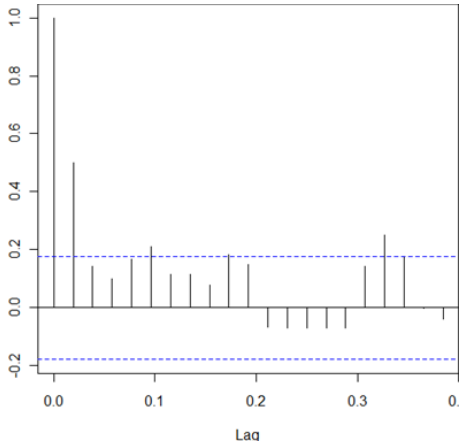
Serie 823852

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

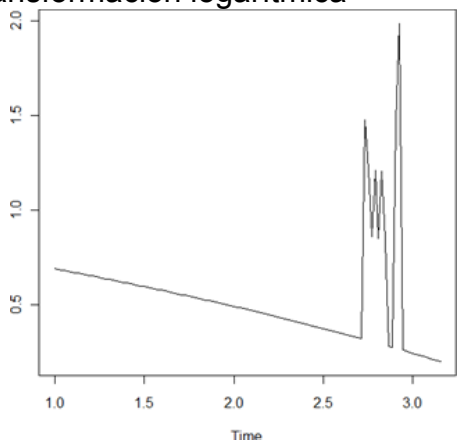


Predicciones

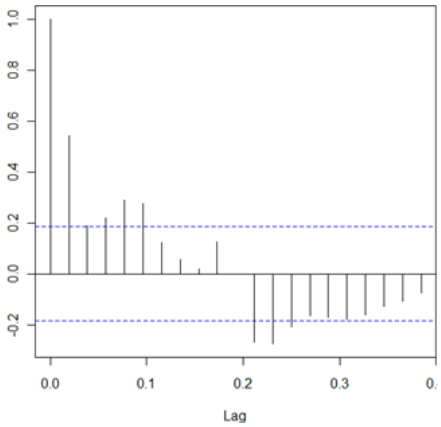
Semana	1	5	9	17	18	A predecir
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	3
5	0	0	0	0	0	3
6	0	0	0	0	0	2
7	0	0	0	0	0	0
8	0	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelo

Modelo	AR(11)	AR(10)	AR(12)	MA(9)	MA(8)	MA(10)	ARMA(11,0,10)
Akaike	-30,75	-30,46	-30,53	-20,57	-22,54	***-50,90***	-32,33

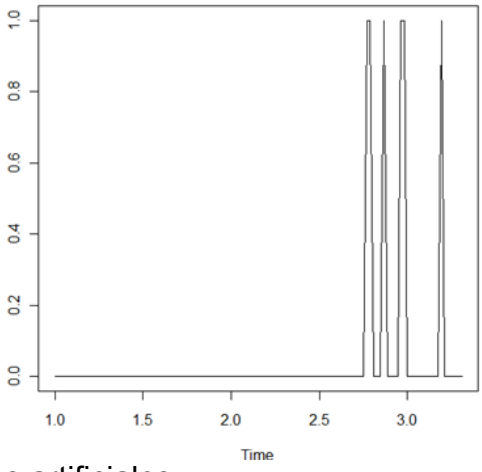
Predicciones

Semana	Predicción	A predecir
1	1	0
2	1	0
3	2	0
4	2	3
5	1	3
6	1	2
7	1	0
8	2	0

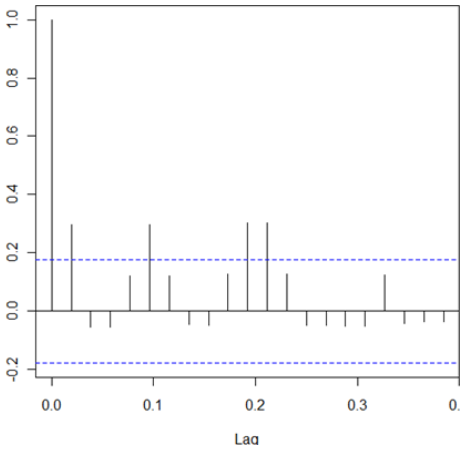
Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
3	0	0	0	0
3	0	0	1	0
2	0	1	3	0
0	1	2	2	0
0	1	2	1	0

Serie 826427
Serie Original



Método de redes neuronales artificiales
Autocorrelograma simple

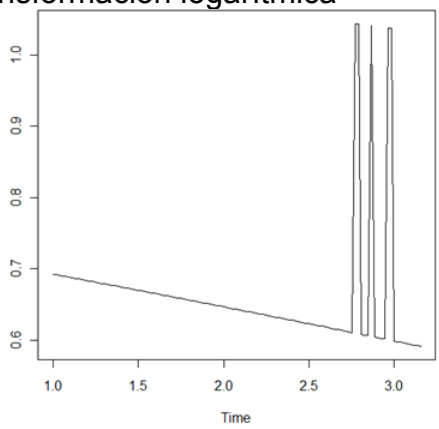


Predicciones

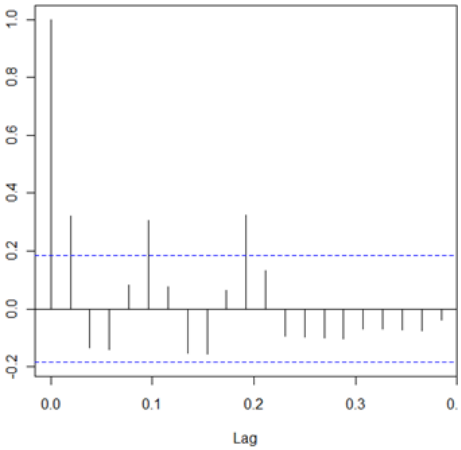
Semana	1	5	10	11	A predecir
1	0	0	0	0	0
2	0	0	0	0	1
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(5)	AR(4)	AR(6)	MA(1)	MA(2)	MA(3)	ARMA(5,0,1)
Akaike	-247,27	-242,85	-246,59	-244,83	-243,32	***-241,83***	-246,22

Predicciones

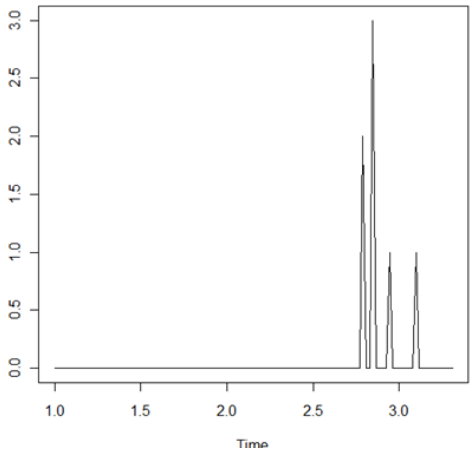
Semana	Predicción	A predecir
1	0	0
2	0	1
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

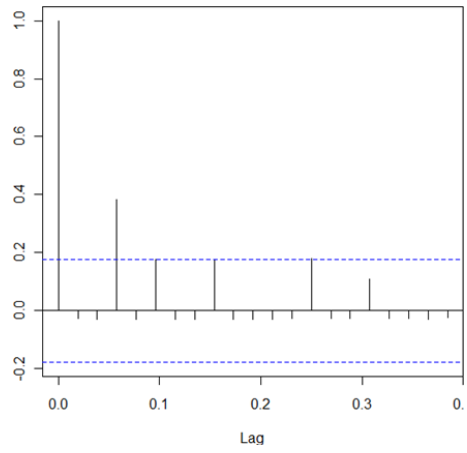
Serie 824684

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

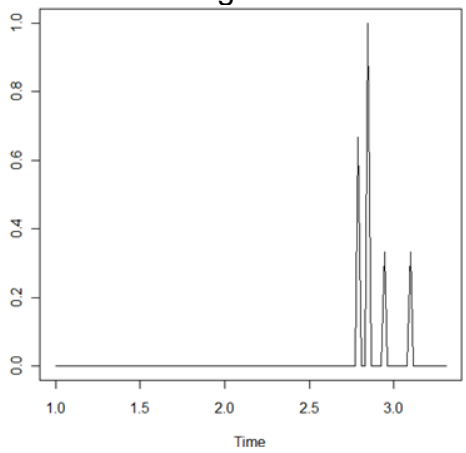


Predicciones

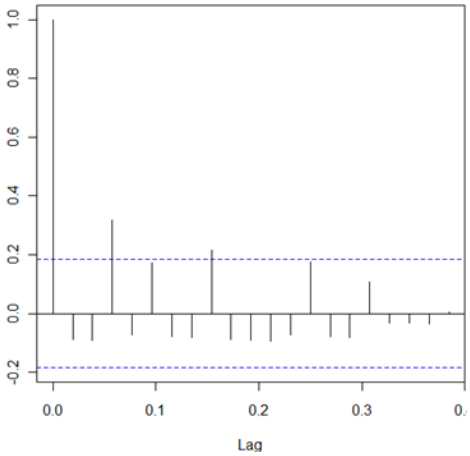
Semana	3	A predecir
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(6)	AR(5)	AR(7)	MA(3)	MA(4)	MA(2)	ARMA(6,0,3)
Akaike	-156,03	-152,57	-154,04	-156,59	-155,55	-139,37	***-156,87***

Predicciones

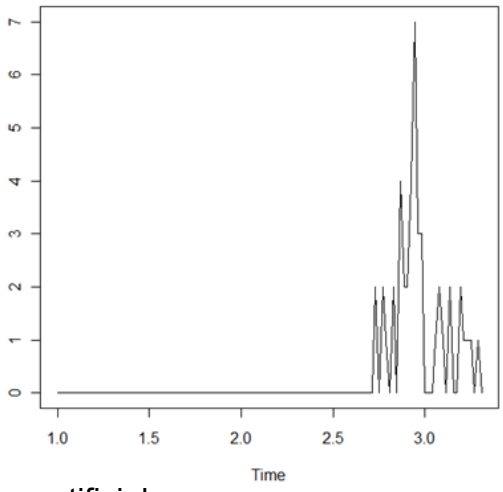
Semana	Predicción	A predecir
1	0	0
2	1	0
3	0	0
4	0	0
5	1	0
6	1	0
7	1	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

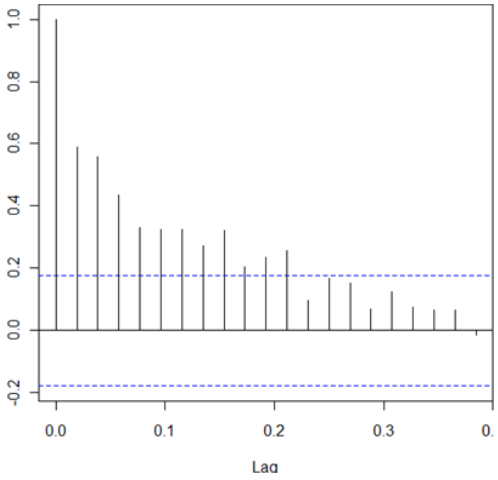
Serie 826430

Serie Original



Método de redes neuronales artificiales

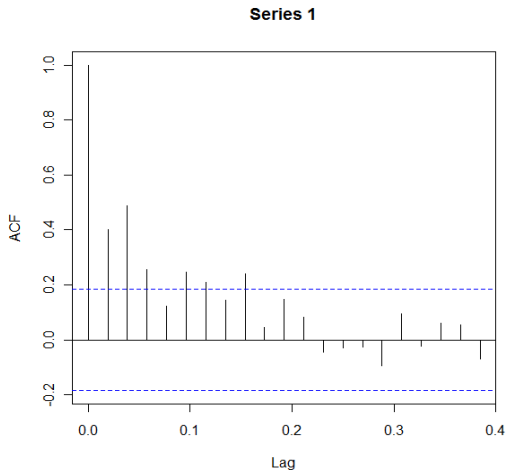
Autocorrelograma simple



Predicciones

Semana	1	2	3	4	5	6	7	8	9	10	11	A predecir
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	0	0	0	0

Método ARIMA
Serie sin tendencia y con transformación logarítmica
Autocorrelograma parcial



Modelos

Modelo	AR(7)	AR(6)	AR(8)	MA(11)	MA(10)	MA(12)	ARMA(8,0,12)
Akaike	102,79	***109,89***	101,75	93,17	94,82	78,88	90,89

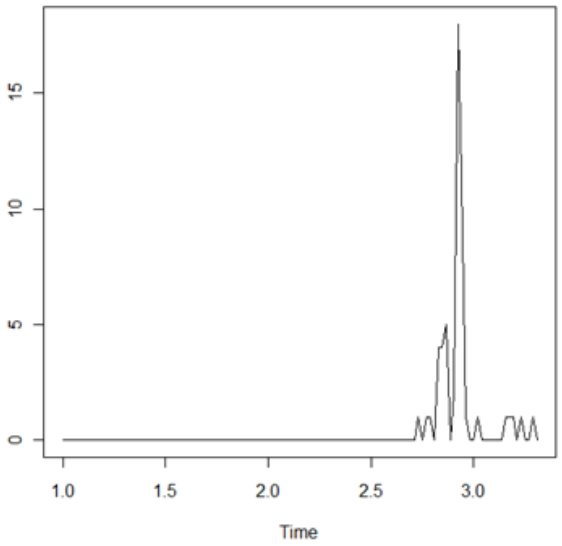
Predicciones

Semana	Predicción	A predecir
1	0	0
2	2	2
3	2	1
4	1	1
5	1	1
6	0	0
7	0	1
8	1	0

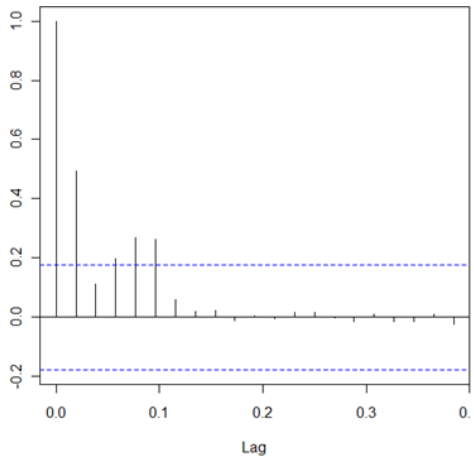
Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	1	1	1	1
2	1	0	0	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
0	1	1	1	1
1	1	1	0	1
0	1	1	0	1

Serie 823854
Serie Original



Método de redes neuronales artificiales
Autocorrelograma simple

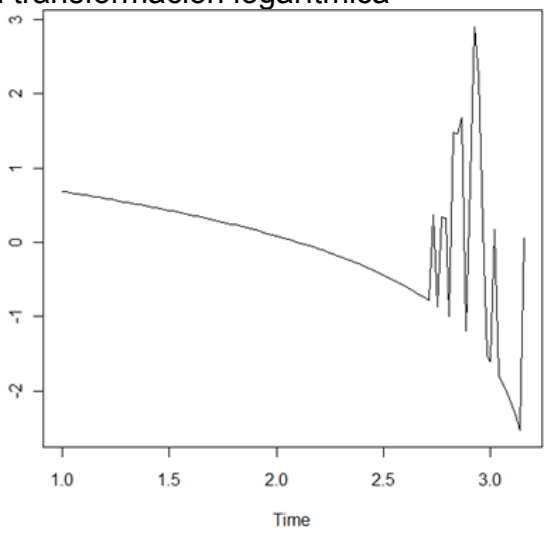


Predicciones

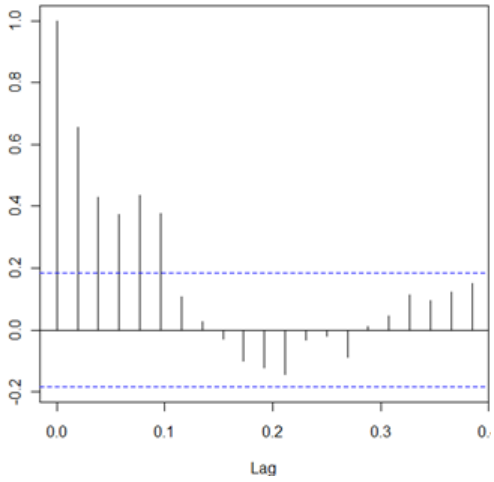
Semana	1	3	4	5	A predecir
1	0	0	0	0	1
2	0	0	0	0	1
3	0	0	0	0	0
4	0	0	0	0	1
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	1
8	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(10)	AR(9)	AR(11)	MA(5)	MA(4)	MA(6)	ARMA(9,0,5)
Akaike	187,04	186,98	187,57	184,13	***216,07***	184,76	177,40

Predicciones

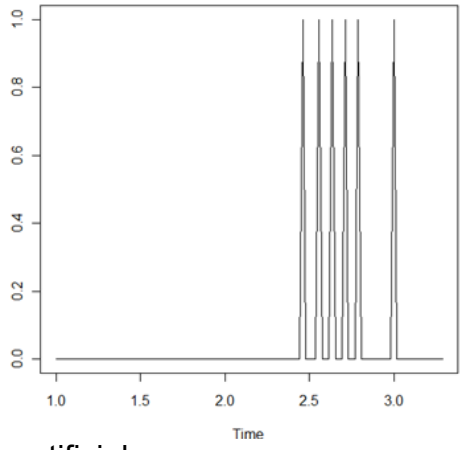
Semana	Predicción	A predecir
1	8	1
2	1	1
3	1	0
4	2	1
5	23	0
6	31	0
7	17	1
8	22	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
1	0	0	0	0
1	0	0	1	0
0	0	1	1	0
1	0	1	0	0
0	0	1	0	0
0	0	0	0	0
1	0	0	0	0
0	1	0	0	0

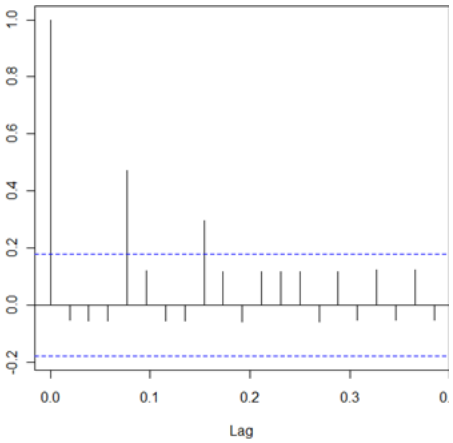
Serie 821192

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

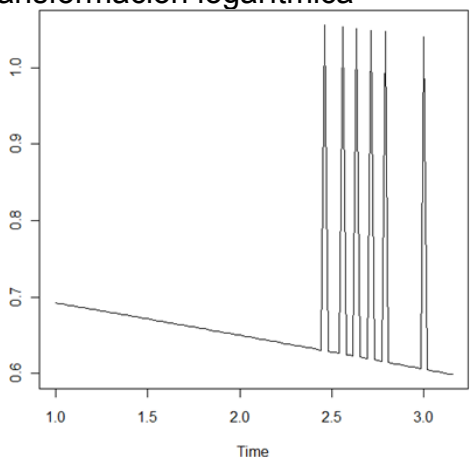


Predicciones

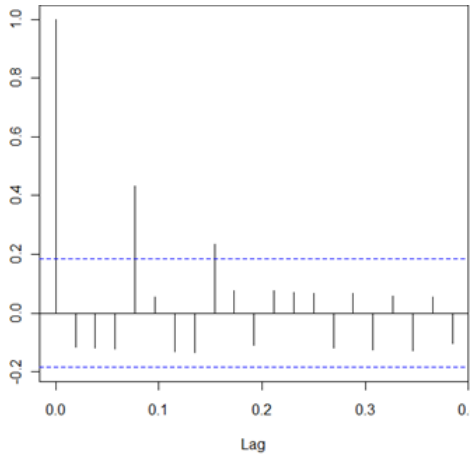
Semana	4	8	A predecir
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(5)	AR(6)	AR(4)	MA(3)	MA(2)	MA(4)	ARMA(4,0,3)
Akaike	-227,74	-226,96	-225,82	-206,79	***-208,76***	-222,76	-224,36

Predicciones

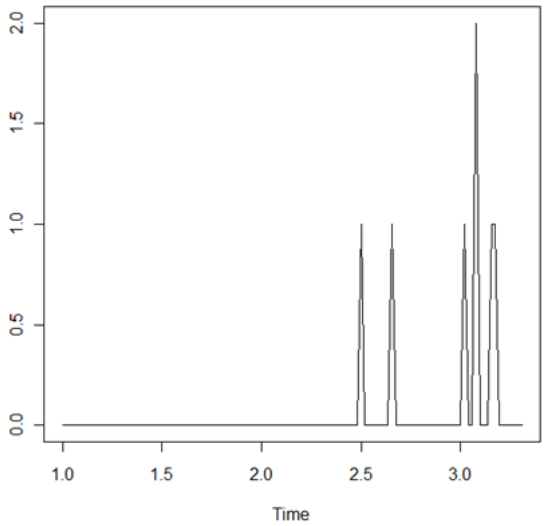
Semana	Predicción	A predecir
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

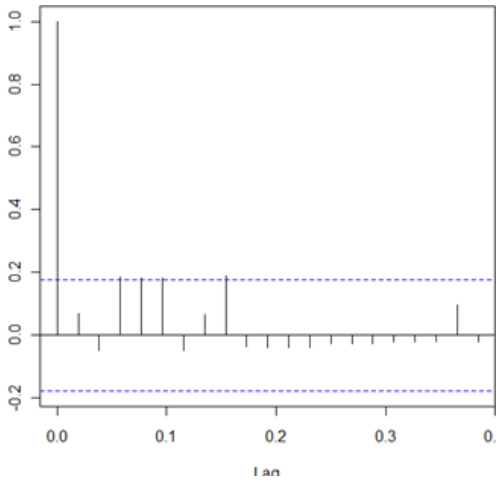
Serie 821193

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

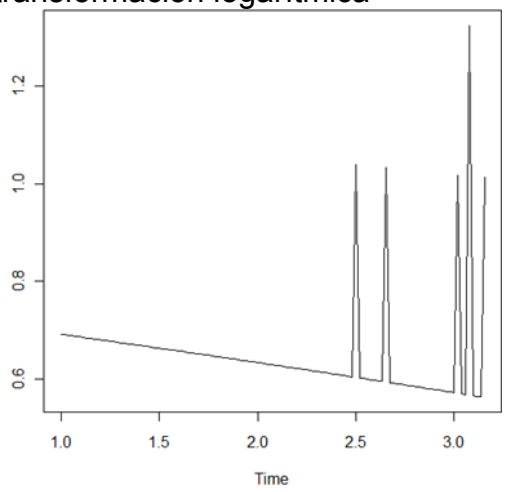


Predicciones

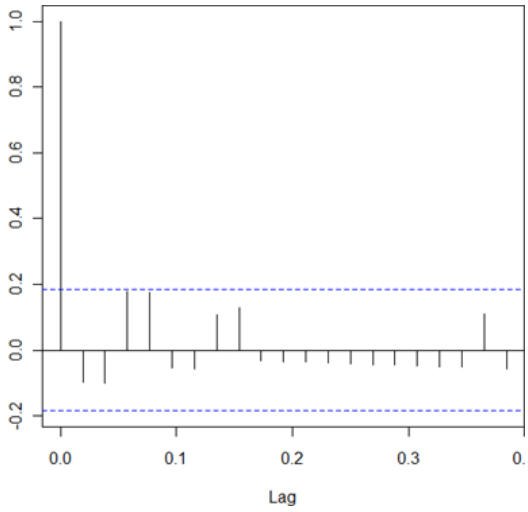
Semana	3	4	5	8	A predecir
1	0	0	0	0	1
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(5)	AR(6)	AR(4)	MA(3)	MA(2)	MA(4)	ARMA(4,0,3)
Akaike	***-227,74***	-226,96	-225,82	-206,79	-208,76	-222,76	-224,36

Predicciones

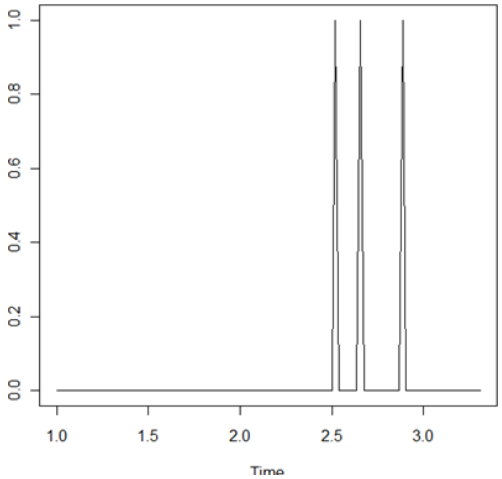
Semana	Predicción	A predecir
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
1	0	0	0	0
0	0	0	1	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

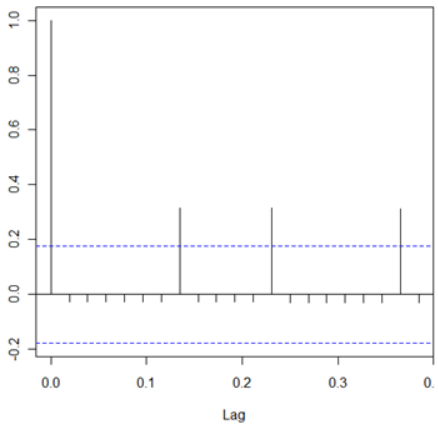
Serie 821194

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

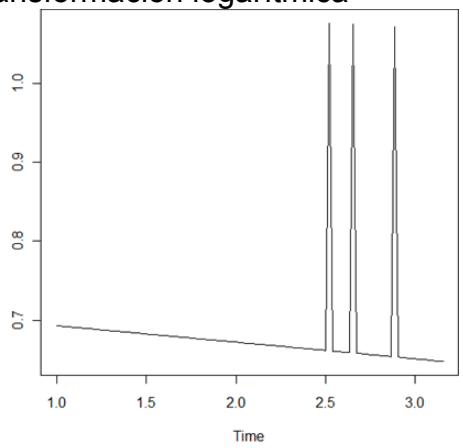


Predicciones

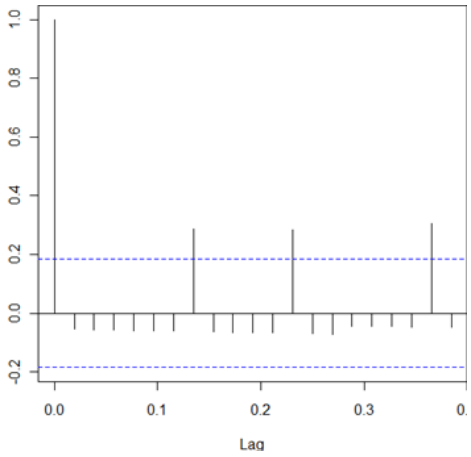
Semana	7	12	19	A predecir
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(14)	AR(13)	AR(15)	MA(7)	MA(6)	MA(5)	ARMA(14,0,7)
Akaike	***-288,75***	-286,82	-286,81	-290,96	-279,03	-289,52	-285,54

Predicciones

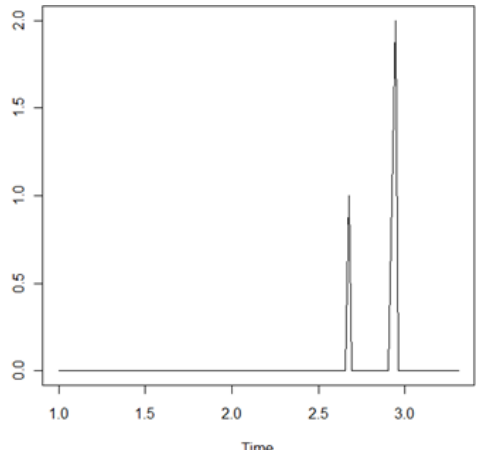
Semana	Predicción	A predecir
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

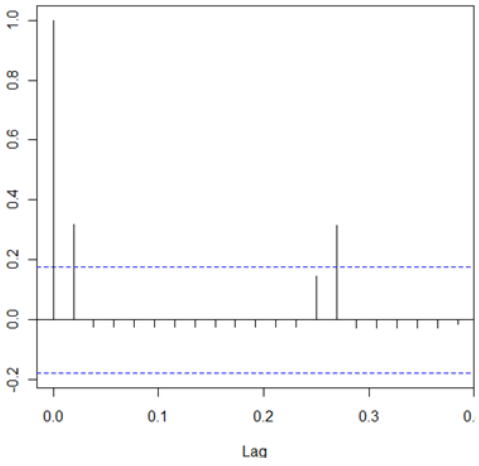
Serie 821197

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

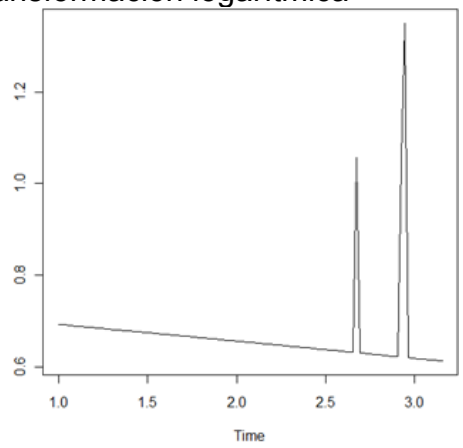


Predicciones

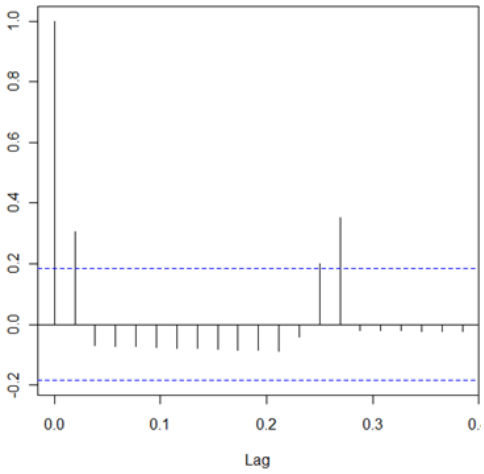
Semana	1	14	A predecir
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(2)	AR(1)	AR(3)	MA(1)	MA(2)	MA(3)	ARMA(2,0,1)
Akaike	***-239,00***	-237,43	-237,01	-240,62	-238,81	-237,07	-237,03

Predicciones

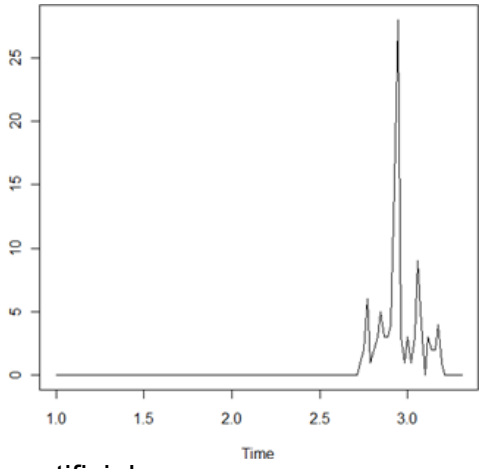
Semana	Predicción	A predecir
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

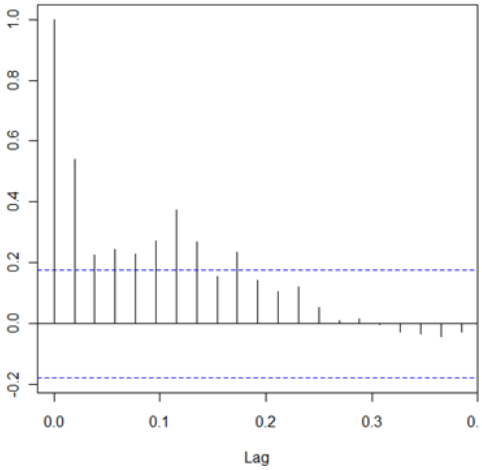
Serie 826426

Serie Original



Método de redes neuronales artificiales

Autocorrelograma simple

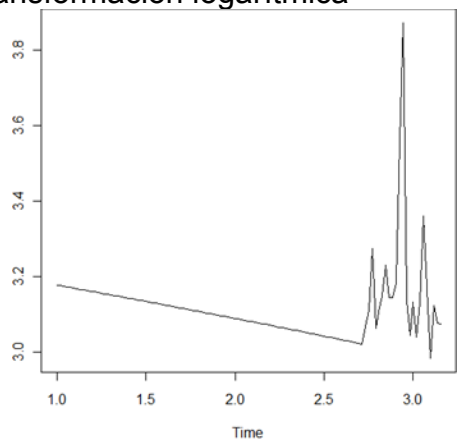


Predicciones

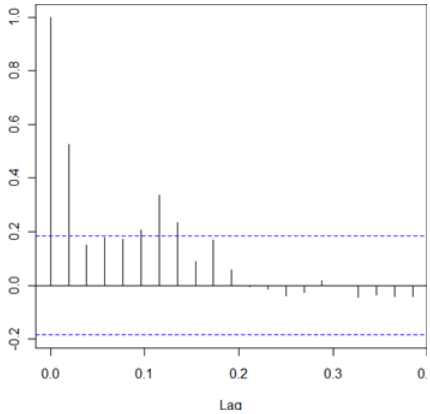
Semana	1	2	3	4	5	6	7	9	A predecir
1	0	0	0	0	0	0	0	0	4
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0

Método ARIMA

Serie sin tendencia y con transformación logarítmica



Autocorrelograma parcial



Modelos

Modelo	AR(4)	AR(3)	AR(5)	MA(9)	MA(8)	MA(10)	ARMA(5,0,9)
Akaike	-236,40	***-238,18***	-239,56	-257,00	-256,99	-256,87	-250,05

Predicciones

Semana	Predicción	A predecir
1	1823	4
2	1888	1
3	2091	0
4	2073	0
5	2213	0
6	2342	0
7	2324	0
8	2478	0

Método de medias móviles

A predecir	8 semanas	4 semanas	2 semanas	12 semanas
4	2	1	2	3
1	3	3	3	2
0	2	2	2	2
0	2	1	0	2
0	1	1	0	2
0	1	0	0	2
0	1	0	0	1
0	1	0	0	1

Anexo H: Entrevistas

En el siguiente anexo se presentarán las reuniones mantenidas durante los meses de trabajo en la empresa.

H.1 Reunión 05.03.09 Acta Reunión Nathaly Sanhueza

Presentes: Nathaly Sanhueza, Estudios Comerciales Cencosud Supermercados S.A.

Objetivo: Presentación del trabajo de Datamart a realizar, comentar sobre experiencias similares realizadas en Supermercados, conseguir contactos, líneas de acción futuras

Conclusiones

Principales variables utilizadas pora clusterizacion de tiendas: Metros Cuadrados de tienda, Ventas por local, Región, Comuna.
Principales variables utilizadas para clusterizacion por producto: Ventas dentro de la categoría, Unidades por presentación, precio, elasticidad precio.
No es claro si es mejor comenzar clusterizando las tiendas y luego asignar surtido tipo o, clusterizar artículos y luego generar clusters.
Se estableció el contacto para resolver futuras dudas.

Comentarios:

Si bien existen experiencias similares en Supermercados, el enfoque es netamente estadístico, por lo que no todo es homologable.
El área y tema específico del trabajo a realizar en Easy tampoco esta muy desarrollado en Supermercados en Cencosud, de hecho no existe una metodología clara para trabajar en esta área.
A pesar de los puntos antes mencionados, nathaly indicó que con los estudios realizados se han conseguido buenos resultados.

H.2 Reunión 18.03.09 Validación de Datos a utilizar

Presentes: Tomás Zavala, subgerente de precios y surtidos

Objetivo: Validar datos a utilizar e identificar grano de la información, estimar costos de RRHH involucrados en proceso de catalogación.

Desarrollo:

A.- Presentación de los reportes en los que se va a trabajar:

1.- Forecasting de productos: A nivel de producto por tienda.
2.-Determinación del mix por grupos de tiendas.

B.- Presentación de los datos a utilizar en cada reporte.

Forecasting:

Venta de productos por tienda
Correlación entre ventas cruzadas

Determinación del mix por grupos de tiendas:

Datos Ubicación: Zona geográfica, Región, Ciudad, Comuna.
Datos Demográficos: habitantes por Zona, Región, Ciudad, Comuna.
Metros cuadrados en Sala: Tienda, Categoría, Familia.
Elasticidad precio por producto.
Utilidad por metro cuadrado.

C.- Descripción de los datos obtenidos

Calidad (criterios utilizados para definir calidad)
Bajo nivel de Outliers (menos del 1%)
Se eliminan datos que salgan fuera de la media + 4*desviación estándar.
Manejo de singularidades (Quiebres de Stock, detección y manejo)
Mejor forma de identificarlos
Como tratarlos de mejor forma.

D.- Estimación de Costos de RRHH involucrados en proceso de catalogación.

Se estima que el costo en RRHH alcanza los $4.950.000 anualmente, considerando el trabajo de 12 personas durante 1 semana y media para la realización de la catalogación.

Tabla F.1: Estimación del costo anual en RRHH para procesos de catalogación

Horas hombre involucradas anualmente	Costo Hora hombre ($/hora)	Total anual involucrado
660	7.500	4.950.000

Fuente: Elaboración propia con datos de la gerencia de procesos comerciales

H.3 Reunión 28.05.09

Presentes: Tomás Zavala, subgerente de precios y surtidos

Objetivo: Definir agregación de datos, definir reportes a realizar, entrevista con gerencia de sistemas.

Desarrollo:

A.- Agregación de datos:

Datos a nivel de semana, pronóstico a 8 semanas
B.- Reportes a realizar

1.- Forecasting:
- A nivel de productos por tienda
- Plazo del pronóstico 8 semanas
- Se entregará el mejor resultado de acuerdo al método con mejor ajuste

2.-Clustering
- Validación de datos se realizará manualmente
- Variables de decisión será físicas por tienda y de ventas en las tiendas

3.- Determinación del mix por grupos de tiendas:
- Sugerencias a nivel de grupo de tiendas
- Sugerencia a nivel de SKU

H.4 Reunión 4.08.09

Presentes: Claudio Larrea, Category Manager.

Objetivo: Probar método de trabajo de clustering propuesto. Corroborar valides de resultados de agrupación.

Desarrollo:

A. Se presento mapa de SOFM y se calculo el número ideal de clusters.
B. Se realizo Kmeans para la determinación de clusters
C. Se corroboro la validez de los resultados obtenidos y se caracterizo cada cluster.

Resultados:

Se realizó una clusterizaión con las variables de espacio por categorías en tiendas.

Se obtienen cinco clusters de tiendas con identificación clara de características entre sus integrantes y la existencia de un cluster denominado “raro” pues no deberían estar juntas según el experto en negocio, pues tienen ventas distintas y se encuentran en comunas con características socioeconómicas distintas.

Cluster 1: Tiendas con ubicación rural.
Cluster 2: Tiendas con muy poca venta, ubicación en regiones.
Cluster 3: Tiendas con poco surtido, ubicación en regiones.
Cluster 4: Tiendas no ubicadas junto a Jumbo ni en Malls, con dificultad para el acceso.
Cluster 5: Tiendas con muchas ventas y gran surtido.
Cluster 6: No se identificaron características claras.

Conclusiones:

El método de trabajo fue entendido y aprobado por el usuario.