VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN: GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

MODELO DE PREDICCIÓN DE DATOS PARA DETERMINAR PERFILES DE COMPORTAMIENTOS DE CONSUMO DE LOS CLIENTES DE SUPERMERCADOS SANTAMARÍA UTILIZANDO MINERÍA DE DATOS MEDIANTE LA SEGMENTACIÓN DE SUCURSALES.

AUTORES: TOAPANTA CHANCUSI, DIEGO GONZALO
VIZUETE OÑATE, LEANDRO REYNALDO

DIRECTOR: ING. SOLÍS ACOSTA, EDGAR FERNANDO

SANGOLQUÍ

2019

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**MODELO DE PREDICCIÓN DE DATOS PARA DETERMINAR PERFILES DE COMPORTAMIENTOS DE CONSUMO DE LOS CLIENTES DE SUPERMERCADOS SANTAMARÍA UTILIZANDO MINERÍA DE DATOS MEDIANTE LA SEGMENTACIÓN DE SUCURSALES**" fue realizado por los señores **Toapanta Chancusi Diego Gonzalo y Vizuete Oñate Leandro Reynaldo** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido, por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 02 de diciembre del 2019

Firma:

.....

Ing. Solís Acosta, Edgar Fernando. Mgs

CC: 1803005071

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Nosotros **Toapanta Chancusi Diego Gonzalo** con cedula de ciudadanía 1717654873 y **Vizuete Oñate Leandro Reynaldo** con cédula de ciudadanía 1718413253, declaramos que el contenido, ideas y criterios del trabajo de titulación: **MODELO DE PREDICCIÓN DE DATOS PARA DETERMINAR PERFILES DE COMPORTAMIENTOS DE CONSUMO DE LOS CLIENTES DE SUPERMERCADOS SANTAMARÍA UTILIZANDO MINERÍA DE DATOS MEDIANTE LA SEGMENTACIÓN DE SUCURSALES,** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos y técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciado las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 02 de diciembre del 2019

Firma

.....

Ing. Diego Toapanta
C.C: 1717654873

Firma

.....

Ing. Leandro Vizuete
C.C: 1718413253

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Nosotros **Toapanta Chancusi Diego Gonzalo y Vizuete Oñate Leandro Reynaldo**, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **MODELO DE PREDICCIÓN DE DATOS PARA DETERMINAR PERFILES DE COMPORTAMIENTOS DE CONSUMO DE LOS CLIENTES DE SUPERMERCADOS SANTAMARÍA UTILIZANDO MINERÍA DE DATOS MEDIANTE LA SEGMENTACIÓN DE SUCURSALES,** en el repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad

Sangolquí, 14 de enero del 2020

Firma

.....

Ing. Diego Toapanta
C.C: 1717654873

Firma

.....

Ing. Leandro Vizuete
C.C: 1718413253

DEDICATORIAS

A Dios, por permitirme culminar esta etapa tan importante de mi especialización profesional, por ser mi fuente de inspiración y perseverancia para el cumplimiento de mis objetivos, por mostrarme que el fin de una etapa solo es el inicio de una nueva con más retos por tomar para enfrentarlos con el mejor de los ánimos.

A quienes han sabido brindarme su apoyo y respaldo constante, a mi padre y mis amigos que se han convertido en la familia que se elige, por su acompañamiento y aliento en cada instante de mi vida.

DIEGO GONZALO TOAPANTA CHANCUSI

A Dios, por ser una fuente de inspiración y superación constante, por concederme la paciencia y perseverancia para cumplir mis metas.

A quienes han sabido brindarme su apoyo y respaldo constante, a mis padres y mis todos mis seres queridos, por su acompañamiento y aliento en cada instante de este camino llamado vida.

LEANDRO REYNALDO VIZUETE OÑATE

AGRADECIMIENTOS

Agradezco a Dios por guiar mi camino durante todas mis etapas académicas dándome el ánimo para culminar todos y cada uno de los retos que he iniciado.

A mi padre Manuel Toapanta que es una fuente inagotable de ejemplo de perseverancia y quien ha dejado su legado en mi mente que no existen imposibles y que todo es posible cuando nos lo proponemos con trabajo honesto y constante.

A todas las personas que siempre han puesto su confianza, buenos deseos y que de manera alentadora han compartido sus conocimientos y por quienes soy ahora a nivel personal y profesional.

DIEGO GONZALO TOAPANTA CHANCUSI

Agradezco a Dios por guiar mi camino durante todas mis etapas académicas dándome el ánimo para culminar todos y cada uno de los retos que he iniciado.

A mis padres que son una fuente inagotable de ejemplo de trabajo, honestidad y perseverancia quienes han dejado su legado en mi mente donde la premisa del éxito es la constancia.

A todas las personas y tutores que siempre han puesto su confianza, buenos deseos y que de manera alentadora han compartido sus conocimientos.

LEANDRO REYNALDO VIZUETE OÑATE

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Las empresas dedicadas al sector de autoservicios[1] invierten grandes sumas de dinero en campañas de publicidad masiva. Esto busca alcanzar a la mayor cantidad de potenciales clientes y vender sus productos. Dichas estrategias comerciales se basan en el uso de correos masivos (emailing), llamadas telefónicas, mensajería instantánea, publicaciones en redes sociales, propaganda en radio y televisión. No obstante, se ha detectado que estos métodos de captación de clientes no tienen los resultados esperados. El presente proyecto tuvo como propósito mejorar la efectividad de sus estrategias comerciales. La empresa de retail ecuatoriana cuenta con una generosa cantidad de datos producto de la facturación de sus ventas. Motivo por el cual se usaron herramientas de técnicas de minería de datos y segmentación de información. De esa manera se encontraron patrones de comportamiento en base a la información histórica de consumo de los clientes. Así como también segmentar a los clientes según sus hábitos de compra. Para el respectivo tratamiento de datos, se siguieron los lineamientos de las metodologías CRISP-DM[2] y KDD-Process[3] dando como resultado la generación de una metodología propia enfocada en el desarrollo de la investigación. Una vez concluido este proyecto el personal de marketing en conjunto con la alta gerencia lograron evaluar las mejores estrategias comerciales según el segmento al que pertenecen sus clientes y sus comportamientos de compra.

---

1 Autoservicio: es un tipo de tienda donde el cliente puede elegir y recoger personalmente las mercancías que desea adquirir, a diferencia de las tiendas departamentales.

2 CRISP-DM: Metodología de estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos. Es el modelo analítico más usado

3 KDD-Process: Metodología para la implementación de proyectos de inteligencia de negocios que consiste de nueve pasos los cuales son iterativos la cual no depende de una herramienta de software especifica

ABSTRACT

The companies, which work in the self-service industry, invest large sums of money in mass advertising campaigns. This seeks to reach as many potential customers as possible and subsequently to sell their products. These commercial strategies are based on the use of mass mailings (emailing), telephone calls, instant messaging, social media publications, radio, and television advertisements. However, these customer acquisition methods do not bring the expected results. The purpose of this project is to improve the effectiveness of its commercial strategies. The Ecuadorian retail company has a generous amount of data due to the invoicing of its sales, which is why the use of data mining techniques and information segmentation tools takes the strategies to the next level. In this way, behavior patterns were found based on the customers' historical consumption information, as well as customers segments according to their buying habits. For the respective data processing, the guidelines of the CRISP-DM and KDD-Process methodologies were followed, resulting in an own methodology focused on the development of the research. Once this project was completed, the marketing staff together with the senior management managed to evaluate the best commercial strategies according to the segment to which their customers belong and their purchasing behavior.

CAPÍTULO I

INTRODUCCIÓN

1.1 Antecedentes

Para la empresa de retail ecuatoriana a través de la incorporación de nuevas herramientas tecnológicas desde el año 2016, ha podido incrementar el volumen de sus transacciones de ventas mensuales a aproximadamente veinticinco millones de registros entre todas las sucursales. Esta situación ha permitido que surja la necesidad de elaborar modelos de análisis de datos que permitan predecir los comportamientos de consumo de productos con mayor o menor demanda, con la finalidad de satisfacer las necesidades de los clientes en las diferentes sucursales, y elaborar de manera efectiva campañas de marketing que promuevan elevar el margen de ventas. La empresa ha buscado la manera de innovar con el objetivo de ser un referente de compra dentro del mercado de autoservicios, para esto ha incorporado estrategias de fidelización de clientes, ventas corporativas, mecanismos de pago que disminuyan los tiempos de compras de sus clientes, promociones complejas y seguimiento postventa a clientes; este último aspecto al realizar encuestas se ha detectado que varios de sus clientes tuvieron un mismo inconveniente, algunos productos que la cadena oferta no se encuentran en todas las sucursales a las que el cliente frecuenta.

1.2 Justificación e Importancia

La presente investigación surgió en base a la necesidad de que en la empresa de retail ecuatoriana no existen modelos de perfiles de comportamientos personalizados de clientes que permitan elaborar estrategias de consumo relacionados al estudio del:

- Cumplimiento del presupuesto asignado a cada sucursal para mejorar la gestión operativa.

- Tener una segmentación optima de clientes mediante el análisis RFM[4] optimizando los análisis de comportamiento de consumo de clientes evitando el uso de modelos basados en el manejo empírico
- Análisis descriptivo en base a los datos de compras de los clientes para mejorar la planificación de la demanda diaria con la información que se encuentra en el Data Warehouse (DWH).[5]
- Mejorar la gestión de campañas de marketing según el segmento de mercado en base a la interpretación de la información de consumos de clientes.
- Incorporar un plan piloto en base al análisis RFM enfocado al incrementar la rotación de productos fuera de temporada enfocado a segmentos de clientes.
- Mejorar la interpretación de los datos históricos mediante los recursos tecnológicos adecuados y el personal de capacitado.

Hoy en día el conocer el comportamiento del consumidor y anticiparse a la decisión de compra permitirá diseñar estrategias de marketing para un determinado producto, mejorar el concepto de category[6] por sucursal. Además, los consumidores son imprevisibles delante una novedad, promociones, remates y ofertas que pueden reaccionar con rechazo o desconfianza. El

4 RFM: es una técnica utilizada para grupos o segmentos de clientes existentes que se nutre del comportamiento histórico transaccional, el acrónimo proviene de Recency (reciente), Frequency (frecuencia), Monetary value (Valor monetario o importe)

5 DWH: es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

6 Category: es una administración de producto de los distribuidores tomando a las categorías como grupo de productos.

comportamiento del consumidor se define como la serie de actividades que desarrolla una persona cuando busca, compra, evalúa, dispone y usa un bien para satisfacer sus necesidades y deseos.

¿Qué tan importante es conocer al cliente? Para poder responder con éxito esta inquietud, se debe conocer a los consumidores para venderles lo que se ajusten a sus necesidades y no caer en el error de intentar vender lo que nosotros queremos por motivo que el cliente o consumidor elige los productos que mejor satisfagan sus necesidades o deseos, para eso hay que conocer su comportamiento de consumo para desarrollar estrategias con productos interesantes, ofertas, promociones, estrategias de marketing, etc.

Con la investigación desarrollada se determinó el comportamiento de consumo, el conocimiento de los mercados y sus necesidades por sucursal, anticiparse a los hechos de compra del cliente, ofertar productos más apetecibles con características específicas para el cliente, saber que publicidad hacer y donde emitirla para influir más en el comprador.

Adicional, el estudio del comportamiento del consumidor supone dar respuesta a una serie de interrogantes: ¿qué compran?, ¿Por qué?, ¿Cuánto?, ¿Cómo? ¿Dónde?

1.3 Planteamiento del problema

No se ha logrado determinar patrones de consumo en los segmentos de mercado de los clientes, esto provoca que el análisis de la información se delegue a empresas externas, permitiendo que el recurso intangible más valioso sufra el riesgo de caer en empresas dedicadas al mismo giro de negocio.

Entre los principales problemas que se han identificado son:

- No poseer un seguimiento de cumplimiento de presupuesto por sucursales.
- Tener una mala segmentación de ventas por sucursales.
- No realizan análisis predictivos en base a las compras de los clientes e interpretación de datos históricos.
- El no tener campañas personalizadas por cada segmento de mercado.
- Tener una baja rotación de productos.

Para desglosar la problemática se ha incorporado un análisis mediante el diagrama de espina de pescado conocido también como diagrama de causa-efecto, representado en la figura 1.

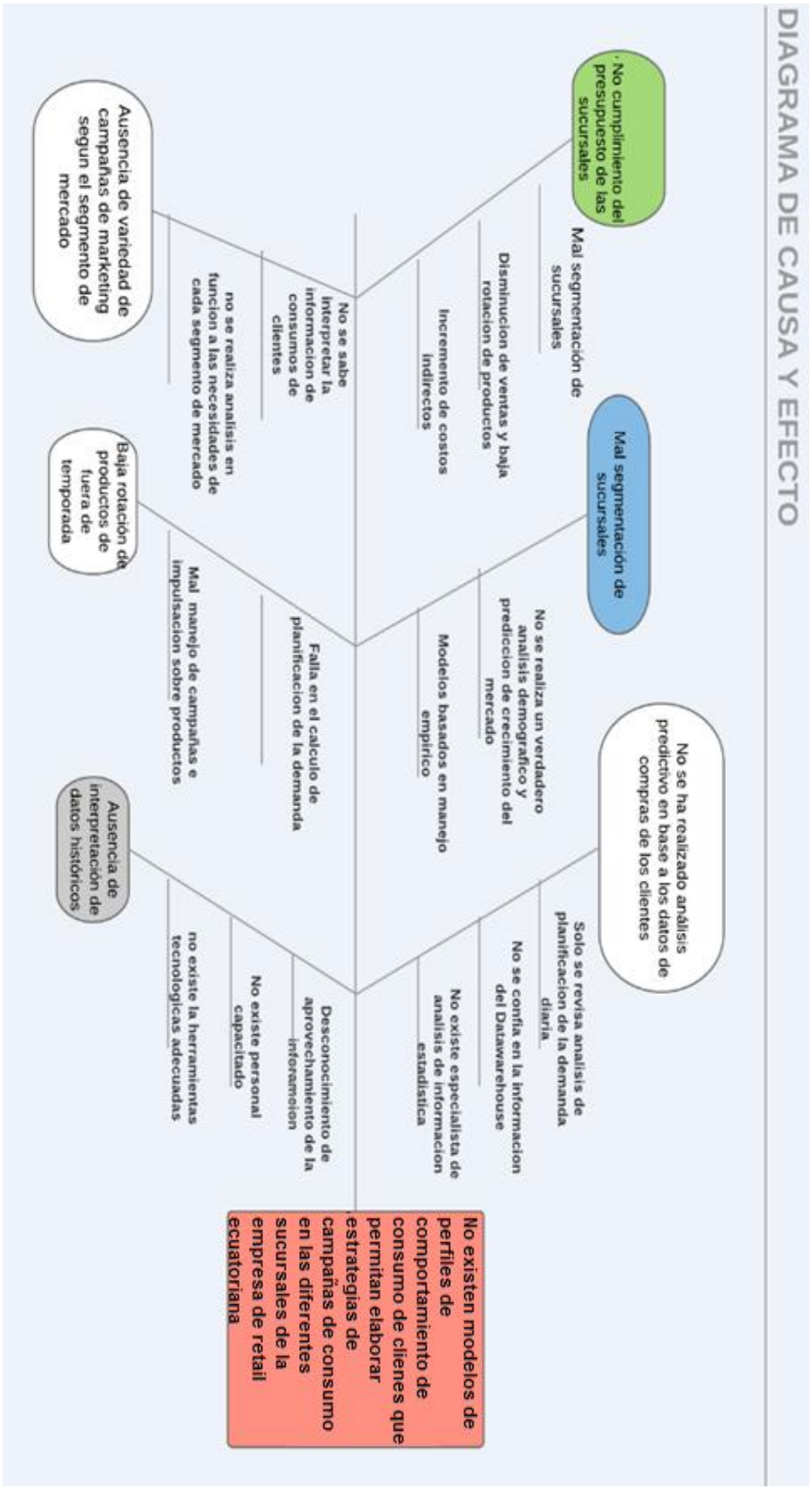

Figura 1 Diagrama Causa Efecto. Fuente: Ilustración Propia

En la actualidad la planificación de la demanda se basa en muestras de información histórica de la empresa y plantea propuestas sugeridas de reabastecimiento de inventarios. Sin embargo, las empresas ya no se manejan de esta manera, sino que, realizan análisis de su información para proyectar como van a ser sus ventas en el próximo mes de tal manera que puedan distribuir de mejor manera sus inventarios permitiéndoles reducir significativamente los costos de almacenamiento de los productos.

1.4 Objetivo general

Elaborar un modelo de comportamiento de consumo de clientes segmentados en un análisis RFM mediante la analítica de datos en base a las ventas históricas de la empresa de retail ecuatoriana que permitan elaborar estrategias comerciales eficientes y focalizadas.

1.5 Objetivos específicos

O.E.1 Fortalecer el área de inteligencia de negocios generando conocimiento mediante el uso de herramientas de analítica de datos, visualizadores de cuadros de mandos y reportes analíticos que ayuden a la toma de decisiones.

O.E.2 Recolectar, gestionar y analizar los datos de ventas históricas mediante herramientas de analítica de datos.

O.E.3 Diseñar el modelo de comportamiento de consumo de clientes por el tipo de segmento generados por el análisis de compras.

1.6 Formulación del problema

Para la investigación correspondiente se realizó las etapas de análisis, diseño e implementación de un modelo de comportamiento de consumo de clientes se requirió que se respondan las siguientes preguntas de investigación para cada objetivo específico:

O.E.1 – RQ1.1: ¿Existen herramientas que permitan el análisis de minería de datos para elaborar toma de decisiones?

O.E.1 – RQ1.2: ¿Existe un proceso análisis de datos modelos de comportamiento de consumo de clientes?

O.E.2 – RQ2.1: ¿Cómo se puede elaborar modelos de comportamiento de consumo tomando en cuenta toda la información de clusterización de ventas de las sucursales?

O.E.2 – RQ2.2: ¿Existe un análisis de la información de ventas basadas en el comportamiento de consumo de los clientes?

O.E.3 – RQ3.1: ¿Se puede diseñar un modelo de comportamiento de consumo en base a las ventas históricas de los clientes?

O.E.3 – RQ3.2: ¿La utilización de técnicas de minería de datos permitirá desarrollar modelos de comportamiento de consumo?

CAPÍTULO II
FUNDAMENTACIÓN TEÓRICA

2.1 Marco teórico

La implementación de un modelo de comportamiento de consumo de clientes basado en las ventas históricas fortaleció el forecast de ventas.[7] Se definió la variable dependiente, que se refiere a la interpretación del consumo de ventas de clientes teniendo un aprendizaje de dicha información y la variable independiente que buscó un análisis de comportamiento de ventas históricas. La generación de un modelo de comportamiento de consumo de clientes permitió generar nuevas promociones y descuentos de productos con menos volúmenes de ventas ya sea por ticket promedio o cantidad de productos segmentados por tipo de cliente, tomando como fuente de información un sector de puntos de venta entregados para el respectivo análisis. La fundamentación teórica permitió generar la congruencia con la hipótesis, para esto se realizó un análisis de la teoría usando las variables del problema, con la finalidad de investigar jerárquicamente cada categoría hasta llegar a la categoría que comprende y explica las variables dependiente e independiente del tema de estudio, para esto se propone la siguiente jerarquía de estudio indicados en la figura 2.

7 Forecast de ventas: El forecast consiste en la estimación y monitorización de las ventas futuras para un producto, utilizando diferentes herramientas como los datos históricos de venta.

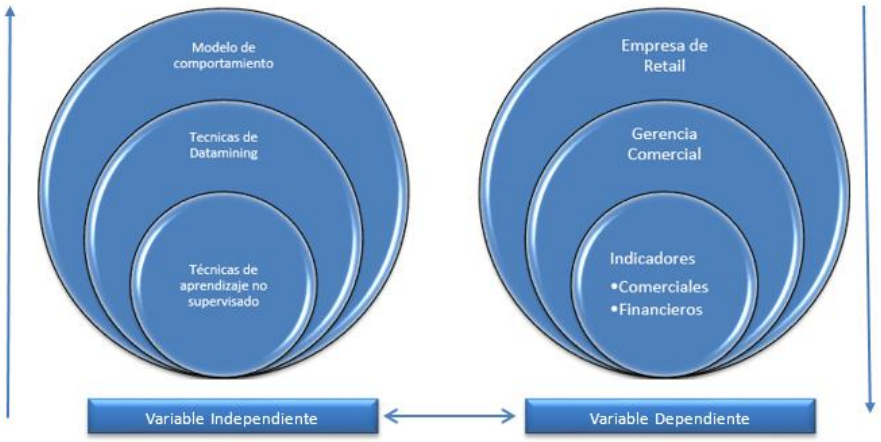

Figura 2 Categorización de Variables. Fuente: Ilustración Propia

2.2 Fundamentación de las variables independientes

Modelo de Comportamiento

Los modelos de comportamiento del consumidor buscan dar una respuesta estructurada previsible y planificada sobre el comportamiento de un conjunto de consumidores con características similares. Son útiles para analizar cómo ciertos grupos de consumidores toman decisiones y ayudan a realizar estrategias de marketing más útiles y más eficientes.

- **Análisis RFM**

  El análisis de RFM es una técnica que permite identificar a clientes actuales que tienen más posibilidades de responder a una nueva oferta. Esta técnica es muy común en el marketing directo. El análisis de RFM se basa en la siguiente teoría simple[8]:

  - Identificar a los clientes con más posibilidades de responder a una nueva oferta es la actualidad. Los clientes que han realizado adquisiciones recientemente tienen

8https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/rfm/rfm_intro.xml.html

más posibilidades de volver a adquirir nuevos productos que aquellos clientes que han adquirido productos en el pasado.

- Identificar la frecuencia en que los clientes han adquirido más productos tienen más posibilidades de responder que aquellos que han adquirido menos productos.
- Identificar la cantidad total invertida, a la que se le denomina valor monetario. Los clientes que han invertido más cantidad en todas las compras que en el pasado, tienen más posibilidades de responder que aquellos que han invertido menos.

Técnicas de Datamining

Es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación).

Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en el análisis adicional o, por ejemplo, en el aprendizaje automático y análisis predictivo.

Las técnicas de minería de datos se dividen en: técnicas de aprendizaje no supervisado y técnicas de aprendizaje supervisado.

Técnicas de aprendizaje no supervisado

En las técnicas de aprendizaje no supervisado no tienen una variable de salida para predecir, solo se tiene variables de entrada. En vez de ajustar el modelo a las variables de entrada para predecir la variable de salida, estas técnicas buscan descubrir patrones dentro de los volúmenes de información.

- **Análisis de componentes principales PCA[9]**

  El análisis de componentes principales (principal component analysis) o PCA es una de las técnicas de aprendizaje no supervisado, las cuales suelen aplicarse como parte del análisis exploratorio de los datos. A diferencia de los métodos de aprendizaje supervisado, donde contamos con un grupo de variables o características (X=X1,X2,...,Xp) medidas sobre un conjunto de observaciones n, con la intención de obtener predicciones sobre una variable respuesta Y asociada, en los no supervisados solo contamos con un número de variables de las cuales nos interesa conocer o de las que queremos extraer información, por ejemplo, sobre la existencia de subgrupos entre las variables u observaciones. Una de las aplicaciones de PCA es la reducción de dimensionalidad (variables), perdiendo la menor cantidad de información (varianza) posible: cuando contamos con un gran número de variables cuantitativas posiblemente correlacionadas (indicativo de existencia de información redundante), PCA permite reducirlas a un número menor de variables transformadas (componentes principales) que expliquen gran parte de la variabilidad en los datos. Cada dimensión o componente principal generada por PCA será una combinación lineal de las variables originales, y serán además independientes o no correlacionadas entre

9 PCA: el análisis de componentes principales, en español se utiliza las siglas ACP mientras tanto que en inglés se la conoce como PCA.

sí. Los componentes principales generados pueden utilizarse a su vez en métodos de aprendizaje supervisado, como regresión de componentes principales o partial least squares.[10] Para poder tener un mejor entendimiento el proceso PCA hace referencia y utilización de procesos estadísticos y matemáticos como:

- **Desviación estándar:** La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido. La desviación estándar se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.[11]
- **Varianza:** En el ámbito de la estadística, solemos emplear la varianza de variables aleatorias estando vinculada a la dispersión de la misma. Las medidas de dispersión en matemáticas se encargarán de expresar la variabilidad de una distribución a través de un número. La varianza establecerá la variabilidad de la variable aleatoria, siendo la varianza en estadística es la raíz cuadrada de la desviación estándar, siendo una media de las frecuencias con la media elevadas al cuadrado.[12]

10 https://rpubs.com/Cristina_Gil/PCA
11https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/what-is-the-standard-deviation/
12 https://es.plusmaths.com/la-varianza-en-estadistica.html

- **Covarianza:** La covarianza es un valor estadístico que nos indica la variación producida por dos variables aleatorias que varían de forma conjunta respecto a sus medias. Es decir, sabremos cómo se comporta una variable dependiendo de cómo lo haga la otra. En estadística se usan dos valores para representar cada una de las variables: X e Y [su covarianza queda representada como COV (X,Y)]. Dependiendo de lo que haga una de las dos variables, la otra variable se comportará de una manera u otra. Lo vemos a continuación:
  - Covarianza (X,Y)<0: sucede cuando X sube e Y baja. Existe una relación negativa.
  - Covarianza (X,Y)>0: sucede cuando X sube e Y sube. Existe una relación positiva.
  - Covarianza (X,Y)=0: sucede cuando X sube e Y baja. No existe relación entre X e Y.[13]

2.3 Fundamentación de las variables dependientes

Empresa de Retail

Para las empresas dedicas al negocio del retail, su estructura u organización pueden contener a todas las tiendas y locales comerciales que habitualmente se encuentran en cualquier ciudad o población con venta directa al público. Sin embargo, su uso se halla más bien ligado a las grandes cadenas de locales comerciales, el ejemplo más común del retail lo constituyen:

13 https://www.economiasimple.net/glosario/covarianza

- Los supermercados.
- Las tiendas departamentales.
- Tiendas para el mejoramiento del hogar.
- Farmacias.
- Tiendas de venta de indumentaria y moda
- Librerías
- También está muy relacionado con las cadenas de franquicias, centrales de compras y el comercio online dentro de este grupo se le ha denominado e-retail.[14]

Gerencia comercial

Es el área encargada de preparar los planes, pronósticos y presupuestos de ventas, calculando para el efecto tanto las cifras históricas y metas corporativas cuanto la demanda puntual del mercado en sus diferentes canales, considerando para el efecto no solo el crecimiento monetario, sino garantizando porcentajes de penetración y participación del mercado.

- Planificar las acciones de las diferentes áreas a su cargo, tomando en cuenta los recursos necesarios y disponibles para llevar a cabo dichos planes y presupuestos.
- Conocimiento muy amplio y detallado de nuestros productos, con todas sus características y aplicaciones.
- Proponer, desarrollar e implementar metas y objetivos con las áreas de su responsabilidad a través de la ejecución de programas y planes de acción dirigidos a alcanzar los objetivos

14 https://www.peru-retail.com/que-es-retail/

propuestos, así como la definición de estándares de desempeño para todos los miembros del equipo comercial.

- Determinar el tamaño y la estructura de la fuerza de ventas, así como su perfil de competencia y su sistema de remuneración e incentivos.
- Revisar la descripción de funciones de cada uno de los equipos integrados en su área de responsabilidad.
- Participar activamente de los procesos de reclutamiento, selección y capacitación de los vendedores y determinar conjuntamente con Talento Humano los planes de capacitación de toda la fuerza de ventas, jefes de productos y asistentes comerciales
- Vista a partir de los objetivos corporativos, diseñar, planificar, implementar y controlar la puesta en marcha de la estrategia comercial, creando y definiendo para el efecto la política comercial nacional y velando porque el cumplimiento de esta se desarrolle identificando oportunidades de negocio que creen valor en la relación con los diferentes canales y sus respectivos clientes, y teniendo como enfoque principal, el cumplimiento del presupuesto anual de ventas y rentabilidad.

Indicadores Financieros

Las empresas de autoservicios realizan un gran esfuerzo económico para consolidar sus indicadores financieros de tal manera que les permita tener información en tiempo real para la toma de decisiones. Los indicadores financieros utilizados son el retorno sobre la inversión, utilidad neta sobre la inversión, pasivo financiero, endeudamiento, gastos, ventas, flujo de caja, etc. De estos indicadores para la investigación se usarán solo los indicadores pertenecientes a la

parte comercial, de los cuales algunos sirven para determinar la efectividad de su fuerza de ventas (frecuencia de compra, recencia y total de consumo).

Indicadores Comerciales

Los indicadores comerciales permiten medir la rentabilidad de la empresa ya que nos provee de datos de ventas totales, ventas por región, ventas por agencia, margen total, margen comercial, plazos y gastos. Estos indicadores permiten realizar comparaciones entre los ingresos reales y los ingresos proyectados para determinar la situación actual de la empresa y tomar acciones correctivas a tiempo para la consecución de los presupuestos de ventas mensuales y anuales de la empresa. El nivel de granularidad de los indicadores de ventas de la empresa se lo puede reducir al punto que permite determinar la efectividad de una campaña específica de ventas, lo cual es útil ya que esto permite determinar si una determinada estrategia de ventas tuvo o no la efectividad esperada.

Up selling

El up selling una técnica de marketing y de ventas. Esta consiste en ofrecer a un cliente un producto similar al que tiene la intención de comprar, normalmente se basa en ofrecerle al cliente un producto más caro a lo que quiere comprar, obteniendo con ello una ganancia mucho más grande.

Cross Selling

El cross selling también llamado venta cruzada, consiste en ofrecer a un cliente potencial un nuevo producto o servicio que sea complementario al producto que quiere comprar o que anteriormente ya ha comprado, subiendo de manera exponencial por cliente las ganancias de

ventas. Esto no quiere decir que sea vender por vender más productos. Lo que significa es conseguir vender de forma correcta, es decir, conocer bien las necesidades de nuestros clientes y satisfacerlos correctamente.

2.4 Antecedentes del estado del arte

Para esta sección se planteó el uso de la revisión inicial de literatura[15]. La cual se basa en cinco aspectos:

- Motivación de la revisión
- Grupo de control
- Elaboración de la cadena de búsqueda en bases digitales como IEEE, Science Direct.
- Determinación de los filtros adicionales para la reducción del espectro de búsqueda en bases de datos digitales indexadas.
- Revisión de estudios primarios.

2.5 Definición de la estrategia de búsqueda

Revisión inicial: se realizó la búsqueda de palabras claves, contenidos, resúmenes entre los cuales permitan constatar que estudios se han realizado y sobre todo que estén relacionados al tema planteado.

15 Revisión inicial de literatura: es una actividad que se lleva a cabo como parte del desarrollo de una propuesta de tesis de investigación o disertación, utiliza un escrito más dinámico que analiza y discute informes generalmente científicos publicados en un área del conocimiento

Validación cruzada de estudios: se planteó incluir criterios de inclusión y exclusión para obtener como resultado de información acorde, veraz y oportuna en un grupo de control relacionado al tema que se desea investigar.

Integración de grupo de control: el grupo de control se conformó por los estudios que cumplieron con los criterios de inclusión y exclusión. Para determinar el grupo de control se utilizó una matriz de correlación entre los títulos de los estudios, introducciones, conclusiones y palabras claves como se muestra en la Tabla 1. Los estudios seleccionados para el grupo de control son los siguientes:

Tabla 1

Estudios por grupo de control.

GRUPO DE CONTROL	*TITULO*	*PALABRAS CLAVE*
GC1	Segmenting customers by transaction data with concept hierarchy	Customer segmentation; Concept hierarchy; Hierarchical clustering
GC2	Segmentation of telecom customers based on customer value by decision tree model	Customer value; Customer lifecycle; Loyalty; Credit; Decision tree model
GC3	Literature Review Application of data mining techniques in customer relationship management	Data mining; Customer relationship management; Literature review
GC4	Customer data mining for lifestyle segmentation	Retailing; Clustering; Segmentation; Lifestyle
GC5	Sales forecasting using extreme learning machine with applications in fashion retailing	Fashion sales forecasting; Extreme learning machine; Artificial neural network

GRUPO DE CONTROL	*TITULO*	*PALABRAS CLAVE*
		Backpropagation neural networks; Decision support system

Formulación de cadena de búsqueda: Para la construcción de la cadena de búsqueda se usaron las palabras que más se repiten en cada contexto definido a partir de los estudios del grupo de control, para el presente estudio se definió los siguientes contextos: Minería de datos, clusterización, segmentación, análisis predictivo y modelos de comportamiento como se muestra en la tabla 2.

Tabla 2.

Cuadro de repetición de palabras clave

CONTEXTO	*PALABRA CLAVE*	*GC1*	*GC2*	*GC3*	*GC4*	*GC5*	*FRECUENCIA DE REPETICIÓN*
Minería de datos	Datamining	X		X			**2**
Clusterización	Clustering		X		X	X	**3**
Análisis Predictivo	predictive analytics	X		X	X	X	**4**
Modelo de comportamiento	behavior model	X	X	X		X	**4**
Segmentación	Segmentation	X	X		X		**3**

Por consiguiente, de acuerdo a la frecuencia de repetición de las palabras clave que se encuentra en la tabla 2; se definió la cadena de búsqueda en el repositorio digital de la IEEE como se puede observar en la tabla 3.

Tabla 3.

Cuadro de cadenas de búsqueda

PALABRAS CLAVES	*CADENA DE BUSQUEDA*	*RESULTADOS*
Predictive analytics Behavior model	(predictive analytics AND behavior model)	142 artículos técnicos

PALABRAS CLAVES	*CADENA DE BUSQUEDA*	*RESULTADOS*
Datamining **Clustering** **Segmentation** **Predictive analytics**	(((((datamining) AND clustering) OR segmentation) AND predictive analytics))	17 artículos técnicos
Datamining **Clustering** **Segmentation** **Predictive analytics** **Behavior model**	(((((datamining) AND clustering) OR segmentation) AND predictive analytics) AND behavior model)	3 artículos técnicos

Revisión de estudios primarios: en esta sección es donde de los artículos técnicos finales que se obtuvo mediante la cadena de búsqueda, se describe brevemente que es lo que se investigó con respecto al tema de tesis planteado.

Data-driven segmentation of consumers' purchase behavior in the retail industry; George Carmichael; Yu-Wang Chen; Cheng Luo; 2018 4th International Conference on Information Management (ICIM)

En el mundo moderno, los enfoques de mercadotecnia tradicionales se están abandonando gradualmente a favor de análisis de negocios basados en datos debido a la mejora de la eficiencia y la relevancia del consumidor. La segmentación de clientes está liderando el camino en el análisis basado en marketing, ya que permite agrupar a los consumidores según sus comportamientos de compra. Esta investigación presenta la aplicación de un modelo de segmentación de clientes en un mercado minorista estadounidense dinámico y competitivo, con el objetivo de producir estrategias de marketing perspicaces. Se monitorearon 994 consumidores con 168,621 transacciones durante un período de cuatro años utilizando los datos del escáner de la tienda IRI. Cada consumidor se segmentó utilizando el análisis de clúster en cuatro segmentos, a saber: (1) "Cazadores de ofertas", (2) "Exploradores oportunistas", (3) "Operadores aversos a la promoción" y (4) "explotadores oportunistas" identificados en el Mercado de sal-merienda de EE. UU.

Leveraging Cloud-Based Predictive Analytics to Strengthen Audience Engagement; **U. Shakeel ; M. Limcaco; SMPTE Motion Imaging Journal**

Para hacer crecer su negocio y aumentar su audiencia, los distribuidores de contenido deben comprender los hábitos de visualización y los intereses de los consumidores de contenido. Por lo general, esto requiere resolver problemas de cómputo difíciles, como procesar rápidamente grandes cantidades de datos en bruto de sitios web, redes sociales, dispositivos, catálogos y fuentes de canales secundarios. Afortunadamente, los distribuidores de contenido actuales pueden aprovechar la escalabilidad, la rentabilidad y el modelo de pago por uso de la nube para abordar estos desafíos. En este documento, mostramos a los distribuidores de contenido cómo usar las tecnologías de la nube para crear soluciones analíticas predictivas. Examinamos los patrones arquitectónicos para optimizar la entrega de medios, y discutimos cómo evaluar la experiencia general del consumidor en base a fuentes de datos representativos. Finalmente, presentamos implementaciones concretas de servicios de aprendizaje automático basados en la nube y mostramos cómo utilizar los servicios para perfilar la demanda de la audiencia, para orientar las recomendaciones de contenido y para priorizar la entrega de los medios relacionados.

Modeling and Predicting the Active Video-Viewing Time in a Large-Scale E-Learning System; Tao Xie ; Qinghua Zheng ; Weizhan Zhang ; Huamin Qu; IEEE Access

Muchos estudios sobre la extracción de grandes datos de aprendizaje se centran en los patrones de acceso de los usuarios y las conductas de video visión, mientras que se presta menos atención al tiempo activo de visualización de videos. Este documento identifica esta unidad de análisis completamente diferente, modela la medida en que los factores influyen en ella y predice aún más

cuando un usuario abandona permanentemente un curso. El objetivo es proporcionar nuevos conocimientos y tutoriales sobre el análisis de datos y la construcción de sub-espacios de características para los analistas del aprendizaje, los investigadores de la inteligencia artificial en las comunidades de educación y minería de datos. Con este fin, recopilamos datos de visualización de video de un sistema de aprendizaje electrónico a gran escala y usamos la función de riesgo proporcional de Cox para modelar el tiempo de salida. Los modelos incluyen principalmente las interacciones entre variables, el supuesto de no linealidad y la segmentación por edad. Finalmente, utilizamos los cocientes de riesgos recogidos de las co-variables del modelo como las características de aprendizaje y predecimos qué usuarios tienden a abandonar prematura y permanentemente un curso utilizando algoritmos eficientes de aprendizaje automático. Los resultados muestran que, primero, el modelado se puede utilizar como una tecnología de extracción y selección de características eficiente para problemas de clasificación y que, en segundo lugar, la predicción puede identificar efectivamente el tiempo de salida de los usuarios utilizando solo unas pocas variables. Nuestro método es eficiente y útil para analizar cursos en línea abiertos masivos. A los artículos antes mencionados se han añadido dos más, ya que hace referencia a trabajos realizados al análisis de componente PCA y sobre la metodología CRISP-DM.

Data Mining model in the discovery of trends and patterns of intruder attacks on the data network as a public-sector innovation. In 2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 55-62). IEEE; Mayra Macas; Lidia Lagla; Walter Fuertes; Graciela Guerrero; Theofilos Toulkeridis

La innovación en el sector público se refiere al desarrollo de mejoras importantes en la administración pública y sus servicios correspondientes. Uno de esos servicios públicos es la

seguridad social, cuyo proceso central ha sido la seguridad de la información de sus servicios ofrecidos. El objetivo del presente estudio ha sido el análisis de las tendencias y el descubrimiento de patrones de comportamiento en los ataques a la red de datos de una institución del sector público. Para cumplir dicho objetivo, se implementó un modelo sobre algoritmos y técnicas de minería de datos, basado en el proceso estándar de la industria cruzada para la metodología de minería de datos. El modelo utiliza un sistema de detección y prevención de intrusos de red libre y de código abierto (IDS / IPS) para la captura de los registros de los ataques a la red de datos de la organización. Esto ha sido seguido por una evaluación cuantitativa de varios algoritmos de detección de intrusos que conducen a la selección de J48 y REPTree como algoritmos de minería de datos con un nivel de insolencia en casos debidamente clasificados por el error absoluto más bajo. Los datos se procesaron y sirvieron como entrada para la construcción de reglas. Las reglas resultantes del árbol de decisión se han basado en el principio de calcular la ganancia de información a través de la entropía y minimizar el error que surge de la varianza. Estas reglas fueron el producto de aplicar el aprendizaje automático en los registros analizados y posteriormente se tradujeron y reprogramaron al IDS / IPS para evaluar la eficiencia del modelo. Los resultados demuestran una mejora significativa de alrededor del 67% en la detección de ataques en relación con los IDS tradicionales. En consecuencia, extrapolamos una gran diferencia en el comportamiento y las tendencias con el uso de un sistema tradicional en comparación con el generado por Data Mining.

Formalistic Modelling Based on Pattern Recognition Applied to the Knowledge and Human Talent Sector in Ecuador; Adrían Guayasmín; Walter Fuertes; Mauricio Campaña; Theofilos Toulkeridis

El propósito de este estudio ha sido analizar el conjunto de datos sobre la educación recibida por los graduados de secundaria en Ecuador. El resultado puede permitir el reconocimiento de patrones de comportamiento y la relación correspondiente entre su empleo y las actividades económicas asociadas. Inicialmente, realizamos una evaluación cualitativa de las metodologías y las herramientas de minería de datos disponibles gratuitamente. Posteriormente, el análisis de datos se aplicó en el repositorio central del sector Conocimiento y Talento Humano de la entidad rectora ecuatoriana. Se han realizado varios algoritmos de clasificación, así como técnicas de asociación utilizando el proceso de minería de datos CRISP-DM con R-Studio. Para la agrupación, utilizamos el algoritmo K-means. Para la asociación, el algoritmo A-piori se ha utilizado para encontrar posibles ocurrencias en el conjunto de datos. Luego, se han aplicado la plataforma de integración de datos Pentaho y PostgreSQL, para implementar el estudio. Los resultados demuestran la generación dinámica de la composición social, con algunos problemas detectados, como el desempleo, el subempleo, la ocupación informal e incluso un excedente de profesionales en ciertas áreas, dando una conclusión de la relación con su actividad económica. Finalmente, la información obtenida puede permitir el desarrollo de políticas públicas que podrían mejorar la gestión del conocimiento, así como la matriz productiva del país.

2.6 Marco Conceptual

Para el análisis del presente proyecto se tomaron en cuenta herramientas informáticas y matemáticas que se adapten de mejor manera a la investigación, por ello se menciona por cada herramienta utilizada una breve descripción.

Lenguaje R

R es el software de referencia en el mundo de la Estadística, la herramienta más potente y eficiente del mercado. Se puede desarrollar análisis en un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Inicialmente, R fue desarrollado por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland, en 1993, pero actualmente R es responsabilidad del R Development Core Team. Entre las principales fortalezas para el procesamiento de datos se pueden mencionar:

- R es gratuito con aproximadamente de 6800 paquetes o librerías disponibles.
- Es un programa estadístico que tiene el acompañamiento científico y la actualización de documentación para entender el comportamiento y análisis de cada paquete.
- Puede realizar el procesamiento y análisis de grandes volúmenes de datos con ayuda de herramientas como Hadoop y Spark.

Oracle – PL/SQL

Las siglas PL/SQL significa Procedural Language/Structured Query Language. Es un lenguaje de programación incluido dentro de la base de datos del motor de base de datos Oracle. Esta

herramienta surge ante la necesidad de ampliar las posibilidades del SQL, ya que es un lenguaje de consulta y no un lenguaje de programación. La ventaja respecto a otros lenguajes, para trabajar con una base de datos, es poder trabajar directamente en el servidor de base de datos. Es un lenguaje de programación concebido por y para Oracle, con lo que se puede crear:

- Procedimientos[16]
- Funciones.[17]
- Disparador.[18]
- Paquetes y bloques. [19]

Algunas pinceladas de las características del lenguaje PL/SQL serían:

- Utilización de constantes, variables y cursores.
- Estructuras de control: bucles, ordenes condicionadas
- Funciones predefinidas: aritméticas, lógicas, relacionales.
- Tratamiento de excepciones
- Posibilidad de utilización de comentarios

16 Procedimientos (procedure): Un procedimiento almacenado permite agrupar en forma exclusiva parte de algo específico que se desee realizar o, mejor dicho, el SQL apropiado para dicha acción.
17 Funciones (functions): tiene las mismas características que un procedimiento almacenado. La diferencia estriba que una función devuelve un valor al retornar.
18 Disparador (triggers): objeto que se asocia con tablas y se almacena en la base de datos, los eventos que hacen que se ejecute un trigger son las operaciones de inserción (INSERT), borrado (DELETE) o actualización (UPDATE), ya que modifican los datos de una tabla
19 Paquetes y Bloques (package): tienen el objetivo de agrupar procedimientos y funciones de forma lógica

CAPÍTULO III
MEMORIA TÉCNICA METODOLÓGICA

3.1 Metodología de Investigación

La presente investigación tomó como referencia la metodología Design Science[20] basada en los resultados. Sin embargo, se combinará y adoptará lo mejor de metodologías como KDD-Process[21], CRISP-DM[22] para la analítica de datos diseñada para pequeñas y medianas empresas. Esta, está comprendida por seis fases y que algunas de ellas son bidireccionales lo que permitirá la comunicación entre las distintas fases sin la necesidad que se cumplan secuencialmente, las cuales se visualizan en la figura 3.

20 Desing Science: Es una metodología de investigación de la tecnología de la información basada en los resultados, que ofrece directrices específicas para la evaluación y la iteración dentro de los proyectos de investigación.
21 KDD-Process: Metodología para la implementación de proyectos de inteligencia de negocios que consiste de nueve pasos los cuales son iterativos la cual no depende de una herramienta de software especifica
22 CRISP-DM Metodología para la implementación de proyectos de inteligencia de negocios que consiste de nueve pasos los cuales son iterativos la cual no depende de una herramienta de software especifica

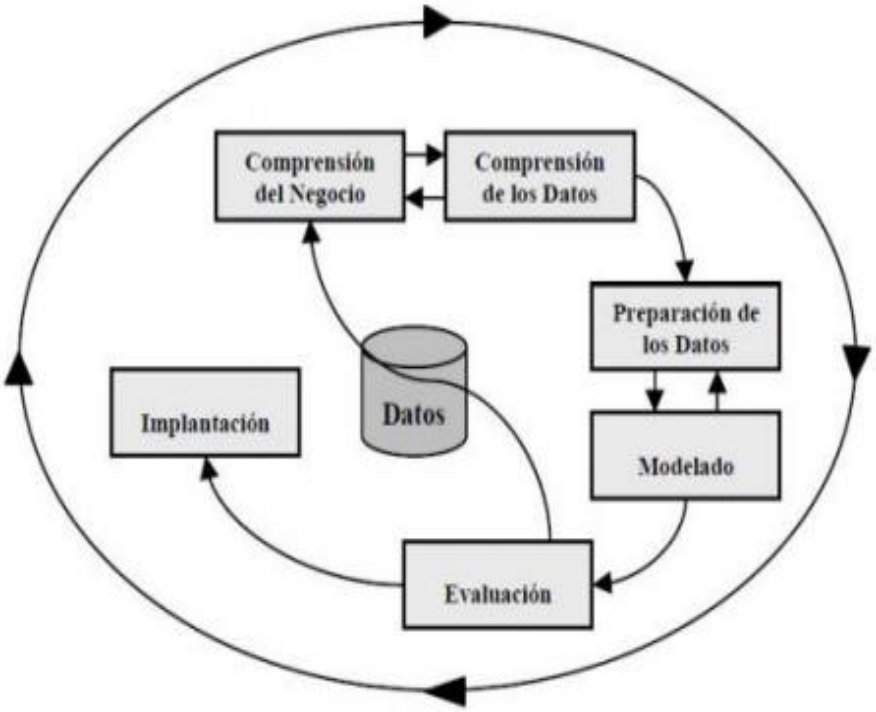

Figura 3. Metodología para minería de datos CRISP-DM, Fuente: Rüdiger Wirth, Jochen Hipp. Fases CRISP-DM.

Las fases que comprende la metodología son:

- **Comprensión del negocio:** corresponde a los objetivos, metas y requerimientos comprendidos por los dueños del proceso o negocio.
    - Establecimiento de los objetivos del negocio; se debe especificar el contexto inicial, objetivos y criterios de éxito.
    - Evaluación de la situación; comprende el inventario de recursos, requerimientos, supuestos, terminologías propias del negocio.
    - Establecimiento de los objetivos de la minería de datos.
    - Generación del plan del proyecto; se especifican el plan, herramientas, equipo y técnicas.

- **Comprensión de los datos:** esta fase es importante por motivo que se deben conocer los datos del negocio, por tal razón se deben tener en cuenta los conceptos.
    - Recopilación inicial de datos.
    - Descripción de los datos.
    - Exploración de los datos.
    - Verificación de calidad de datos.
- **Preparación de los datos:** es la obtención de la vista minable o los dataset.
    - Selección de los datos.
    - Limpieza de datos.
    - Construcción de datos.
    - Integración de datos.
    - Formateo de datos.
- **Modelado:** es la aplicación de técnicas de minería de datos a los dataset.
    - Selección de la técnica de modelado.
    - Diseño de la evaluación.
    - Construcción del modelo.
    - Evaluación del modelo.
- **Evaluación:** esta fase permite que los modelos de la fase anteriores han sido útiles a las necesidades del negocio.
    - Evaluación de resultados.
    - Revisar el proceso.
    - Establecimiento de los siguientes pasos o acciones.

- **Despliegue:** permite el explotar de la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización.
    - Planificación de despliegue.
    - Planificación de la monitorización y del mantenimiento.
    - Generación de informe final.
    - Revisión del proyecto

3.2 Ejecución del proceso de investigación

La investigación implantará su propia metodología con el fin de resolver la problemática ya descrita (ver sección 1.3), las fases que se abarcaran en esta investigación se pueden identificar en la figura 4.

Figura 4. Metodología Propia para Desarrollo Investigación. Fuente: Ilustración propia.

- **Justificación y Viabilidad:** el objetivo de la investigación en ciencia del diseño es desarrollar soluciones basadas en la tecnología para problemas empresariales importantes y relevantes donde para establecer esta fase se incorporará una revisión inicial de literatura

- **Recolección y Análisis de Datos:** en esta etapa se analiza las fuentes de datos con las que cuenta la organización y que puedan ser útiles para la tarea a realizar, donde se debe establecer que como fuente primordial para esta investigación se tomara de los datos de las ventas de los clientes obtenidos del DWH de la organización.

Figura 5. Pre-procesamiento de datos. Fuente: Ilustración propia

- **El Diseño del Modelo:** la investigación de la ciencia del diseño se produjo en un modelo de comportamiento tomado como base la información histórica de las ventas de los clientes. En esta etapa se diseñó cual iba a ser el modelo de base de datos, que técnicas de datamining se adaptaron mejor a los objetivos planteados y que tipo de visualización de resultados se necesitaron.
- **Implementación del Modelo:** para cada etapa de la implementación se ejecutaron varias técnicas como programación SQL, ejecución de técnicas de datamining y ejecución de técnicas de agrupación.
- **Análisis de Resultados:** la evaluación de resultados va a depender del modelo a implementar para métodos de clasificación. Una vez obtenidos los resultados estos deben

ser analizados conjuntamente con expertos de la organización quienes van a determinar si los mismos son los óptimos y sobre todo lo que se está buscando.

- **Comunicación de la Investigación:** una vez obtenidos y validados los resultados estos deben ser compartidos con las áreas gerenciales de la organización a fin de que esta se convierta en una herramienta que permita mejorar la toma de decisiones.

CAPÍTULO IV
ANALISIS, DISEÑO E IMPLEMENTACIÓN

4.1 Justificación y Viabilidad

Para este caso de estudio se realizó el análisis de las ventas de manera masiva, el mismo que sirvió al negocio tener una manera más clara de cómo sus clientes tienen diferentes tendencias y comportamientos de consumo en todas las sucursales; lo que permitirá mediante los resultados entregados mejorar las estrategias comerciales de descuento, mejoramiento de planes de fidelización, optimizar que productos se pueden otorgar y ofrecer a ciertos segmentos de mercado.

Se determinó que es muy factible y viable la investigación por motivo que se tienen las herramientas necesarias, la información disponible de las ventas por cliente de un año, la realización de análisis estadísticos para datos masivos, análisis de información RFM, análisis de componentes principales y la generación de resultados segmentados por clientes de acuerdo a su calificación RFM.

4.2 Recolección y Análisis de datos

Para realizar el modelo de comportamiento de consumo, se tomaron varios aspectos de la información como: clientes que realizaron al menos una compra en la empresa, árbol de categorización o jerarquización de productos, nombres y atributos de productos, nombres de sucursales y los datos bases que son las ventas históricas de alrededor de un año desde junio 2018 a junio 2019.

Para fines de optimización de consulta de datos, se procedió a crear nuevas estructuras o tablas de base de datos para poder agrupar y consolidar la información de manera:

- Resumen de total de unidades vendidas agrupadas por sucursales, departamento, clase y subclase.
- Resumen de ventas agrupadas por departamento, clase y subclase.
- Agregación de ventas y cantidad por ítem vendidos en un día.
- Completitud de datos de clientes que poseían inconsistencias a nivel de fecha de nacimiento, cedula y nombres del mismo.

4.3 Implementación del Modelo

Se realizó la diagramación y creación del modelo de base de datos de cómo deberían ir relacionadas las tablas para tener una mejor comprensión de la necesidad planteada. A continuación, se describen las tablas creadas que se van a utilizar durante el desarrollo del proyecto descritas en la tabla 4 y adiciona al listado de tablas que se utilizaron en la investigación se muestra el diagrama de base de datos representados entre las figuras 6 y 10.

Tabla 4.

Cuadro de tablas creadas

TABLA	*DESCRIPCION*
VEN_CLIENTES	Contiene información de los clientes para la facturación.
SA_VENTAS	Contiene las transacciones de compras de clientes.
ITEM_PROVEEDOR	Contiene información de los ítems por proveedor.
ITEM_MASTER	Contiene el catálogo de los productos a nivel de presentación.
SUBCLASS	Contiene el maestro detalle de los ítems por subclase.
DIVISION	Contiene el maestro detalle de los ítems por división.
DEPT	Contiene el maestro detalle de los ítems por departamento.

TABLA	*DESCRIPCION*
CLASS	Contiene el maestro detalle de los ítems por clase.
STORE	Contiene el maestro detalle de las sucursales
BI_RESUM_VENTA_DEPT	Contiene la información del total de ventas por departamento en cada sucursal.
BI_RESUM_VTA_STORE_DEPT_CLASS_SUBCLASS	Contiene la información del total de ventas por departamento, clase, subclase por sucursal.
RFM_RETAIL	Contiene la información del análisis RFM por cliente.
INTERVALO_MONETARY_SCORE	Contiene la información de los rangos de ponderación de valores monetarios usados en el cálculo de análisis RFM.
INTERVALO_RECENCIA_SCORE	Contiene la información de los rangos de ponderación de valores para la el análisis de la última compra de clientes usados en el cálculo de análisis RFM
INTERVALO_FRECUENCIA_SCORE	Contiene la información de los rangos de ponderación de valores para la frecuencia de compra usados en el cálculo de análisis RFM
BI_RESUM_VENTA_DEPT_CLASS_SUBCLASS	Contiene la información del total de ventas por departamento, clase, subclase por cliente.
AGG_VTS_STORE_DAY	Contiene la información del total de ventas, numero de compras en las sucursales por día.
AGG_VTS_CLI_DAY_F	Contiene la información del total de ventas, numero de compras de clientes por día.
VEN_CLIENTES_UPD	Contiene la información de los clientes para facturación, actualizando el segmento al que pertenece.
RFM_CLI_SEG	Contiene la información de la relación del cliente con el segmento al que pertenece posterior al análisis RFM.

TABLA	*DESCRIPCION*
RFM_SEGMENTOS	Contiene la información del maestro de segmentos.
BI_VENTA_DEPT_CLI_SEG	Contiene la información del total de compras de cada cliente por departamento.

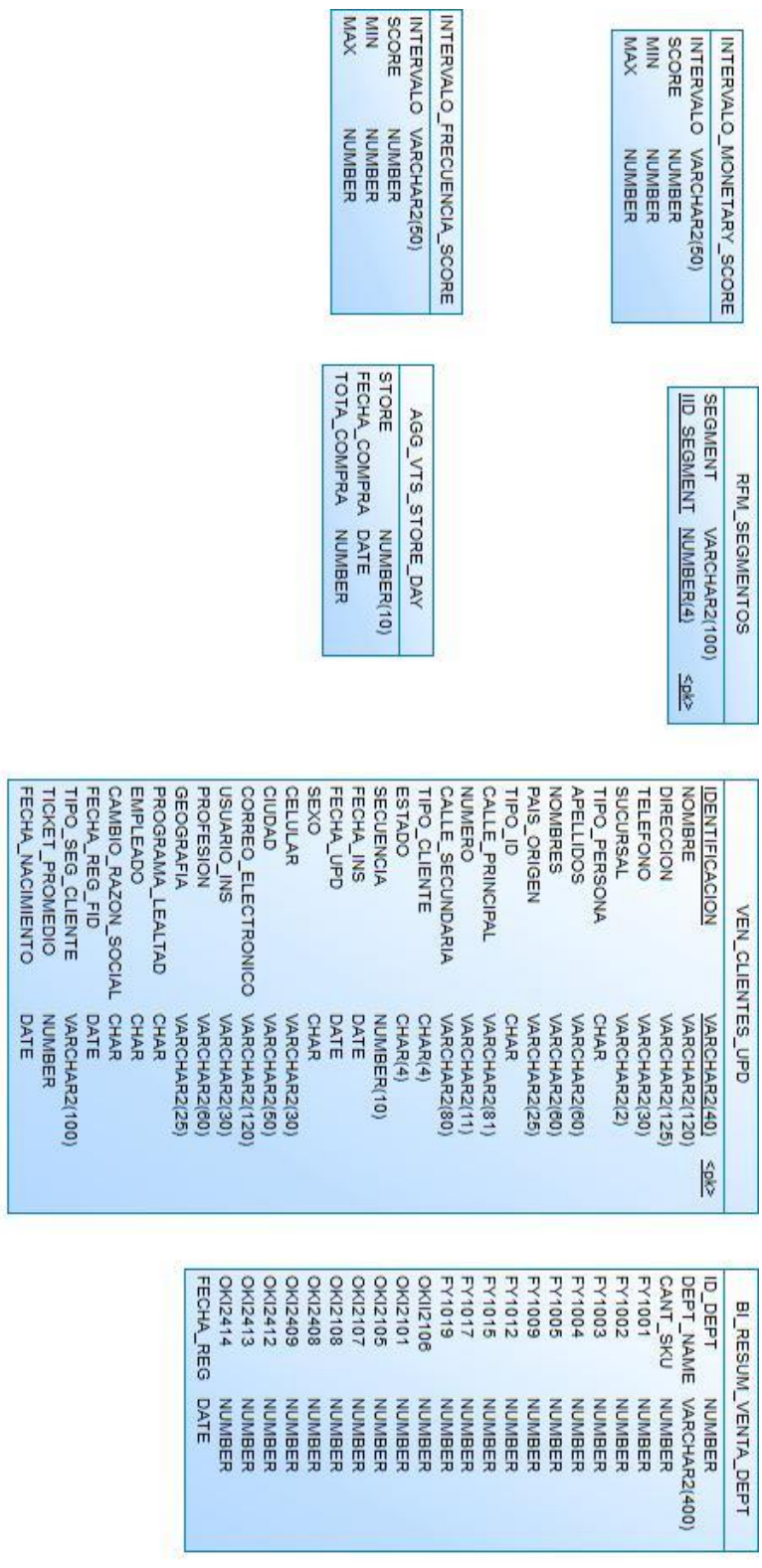

Figura 6. Modelo de Base de datos – parte 1. Fuente: Ilustración Propia

Figura 7. Modelo de Base de datos – parte 2. Fuente: Ilustración Propia

Figura 8. Modelo de Base de datos – parte 3. Fuente: Ilustración Propia

BI_VENTA_DEPT_CLI_SEG		
CUSTOMER_ID	VARCHAR2(40)	<pk>
SEGMENT_ID	NUMBER	
N_A_12	NUMBER	
PLANIFICACION_FAMILIA	NUMBER	
TECNOLOGIA	NUMBER	
MANICURE_Y_PEDICURE	NUMBER	
CONGELADOS	NUMBER	
PROGRAMA_PREMIOS_INSTANTANEOS	NUMBER	
N_A_21	NUMBER	
N_A_28	NUMBER	
CUIDADO_DEL_BEBE	NUMBER	
APARATO_RESPIRATORIO	NUMBER	
N_A_34	NUMBER	
N_A_30	NUMBER	
MASCOTAS	NUMBER	
NO_USABLES	NUMBER	
CLIENTES___OBSEQUIOS_MED	NUMBER	
SERVICIOS_OPERACIONALES	NUMBER	
N_A_25	NUMBER	
ALIVIO_DEL_DOLOR_OTC	NUMBER	
PROTECCION_SOLAR_Y_REPELENTES	NUMBER	
N_A_29	NUMBER	
PERFUMERIA	NUMBER	
SHAMPOO_ACONDICIONADOR_Y_TRATAMIENTO	NUMBER	
N_A_26	NUMBER	
FOTO_SHOP	NUMBER	
N_A_9	NUMBER	
N_A_6	NUMBER	
N_A_20	NUMBER	
MEDIAS	NUMBER	
N_A_23	NUMBER	
N_A_7	NUMBER	
INCONTINENCIA	NUMBER	
CUIDADO_ORAL	NUMBER	
COMPUTACION___INTERNET	NUMBER	
ANTIPARASITARIOS	NUMBER	
N_A_14	NUMBER	
SOLUCIONES_HOSPITALARIAS	NUMBER	
ROPA_INFANTIL	NUMBER	
DEPILACION	NUMBER	
N_A_19	NUMBER	
N_A_5	NUMBER	
ASISTENCIAS	NUMBER	
N_A_31	NUMBER	
N_A_11	NUMBER	
AUTOLIQUIDABLES	NUMBER	
PANALES	NUMBER	
CUIDADO_DE_LA_PIEL	NUMBER	
N_A_3	NUMBER	
CONFITERIA	NUMBER	
N_A_18	NUMBER	
JABONES_PRODS__BANO	NUMBER	
ALBUMES_Y_CROMOS	NUMBER	
DESODORANTES_PRODS__PIES	NUMBER	
N_A_4	NUMBER	
PAPEL_HIGIENICO_Y_FACIALES	NUMBER	
N_A_8	NUMBER	
SUMINISTROS_CRM	NUMBER	
PRODUCTOS_NATURALES	NUMBER	
N_A_32	NUMBER	
HORMONAS	NUMBER	
N_A_13	NUMBER	
ARTICULOS_DE_TEMPORADA	NUMBER	
TINTES_Y_COLORACION	NUMBER	
N_A_15	NUMBER	
N_A_16	NUMBER	
PRIMEROS_AUXILIOS	NUMBER	
N_A_33	NUMBER	
INSUMOS_MEDICOS	NUMBER	
APARATO_CARDIOVASCULAR	NUMBER	
RECETARIO	NUMBER	
N_A_10	NUMBER	
ABASTOS	NUMBER	
N_A_27	NUMBER	
N_A_1	NUMBER	
N_A_2	NUMBER	
DIGESTIVOS_OTC	NUMBER	
GRIPE_OTC	NUMBER	
N_A_35	NUMBER	
MERCHANDISING_SUMINISTROS	NUMBER	
CUIDADO_FAMILIAR_OTC	NUMBER	
BEBIDAS_NO_ALCOHOLICAS	NUMBER	
PAPELERIA_Y_UTILES_DE_OFICINA	NUMBER	
N_A_24	NUMBER	
LENCERIA	NUMBER	
APARATO_DIGEST_Y_METABOL	NUMBER	
SNACKS	NUMBER	
MATERIAL_PUBLICITARIO_IMPRESO	NUMBER	
FECHA_REG	DATE	

BI_RESUM_VENTA_DEPT_CLASS_SUBCLASS	
ID_DEPT	NUMBER
DEPT_NAME	VARCHAR2(400)
ID_CLASS	NUMBER
CLASS_NAME	VARCHAR2(400)
ID_SUBCLASS	NUMBER
SUBCLASS_NAME	VARCHAR2(400)
CANT_SKU	NUMBER
FY1001	NUMBER
FY1002	NUMBER
FY1003	NUMBER
FY1004	NUMBER
FY1005	NUMBER
FY1009	NUMBER
FY1012	NUMBER
FY1015	NUMBER
FY1017	NUMBER
FY1019	NUMBER
FECHA_REG	DATE

Figura 9. Modelo de Base de datos – parte 4. Fuente: Ilustración Propia

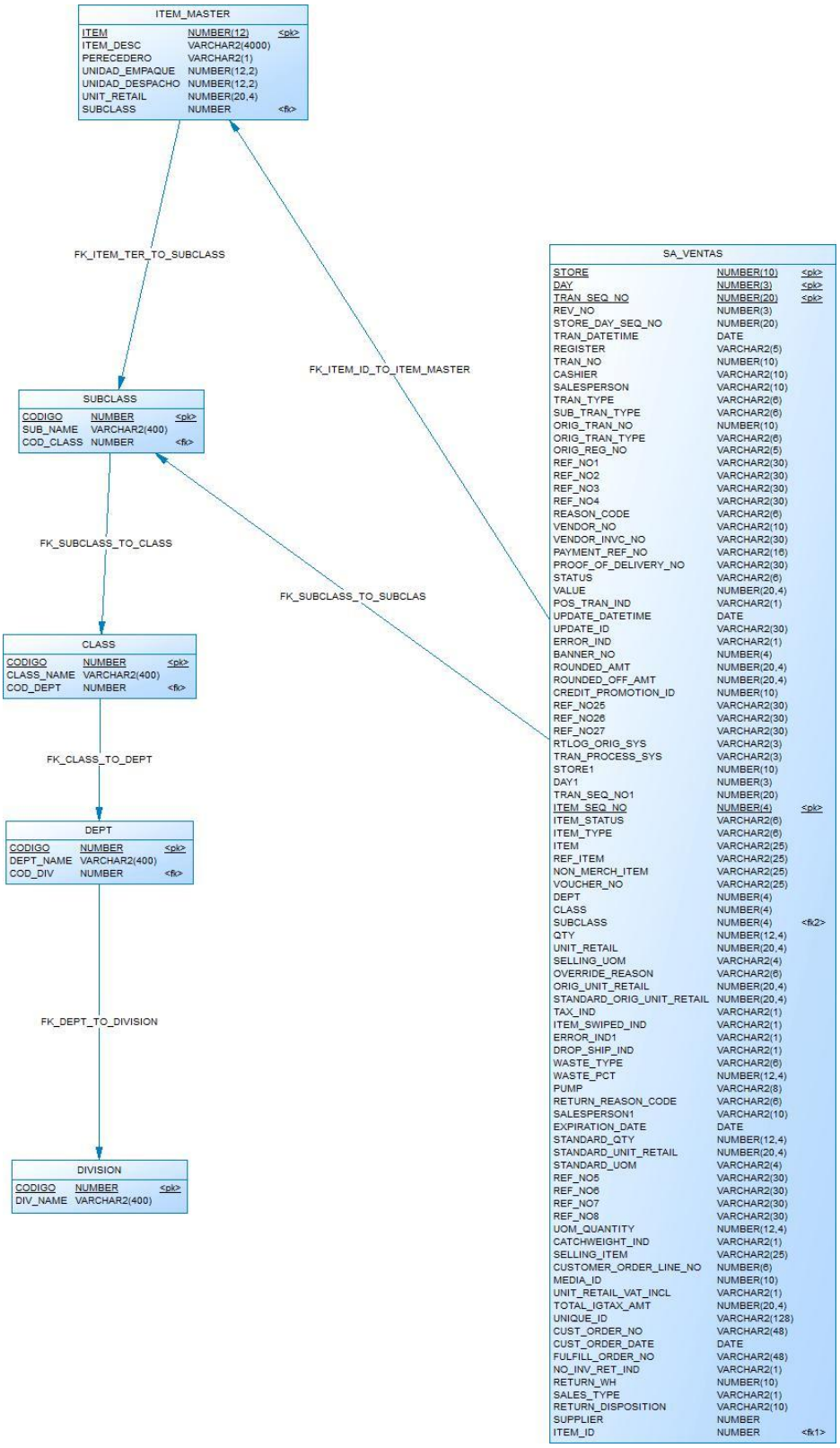

Figura 10. Modelo de Base de datos – parte 5. Fuente: Ilustración Propia

- **Creación Base de Datos**

Se utilizó el motor de base de datos Oracle Express 18c, el cual contenía una copia de los datos y objetos del esquema DW de la base de datos del Data Warehouse de la empresa de retail ecuatoriana cuyo motor de base es Oracle Enterprise 12c. Para la creación de los distintos objetos de base de datos utilizados para los procesos ETL y repositorio de datos se utilizó lenguaje PL/SQL propio de Oracle. Con la herramienta Developer PL/SQL como IDE de administración e implementación, dichas representaciones son mostradas en las figuras 11 y 12 respectivamente.

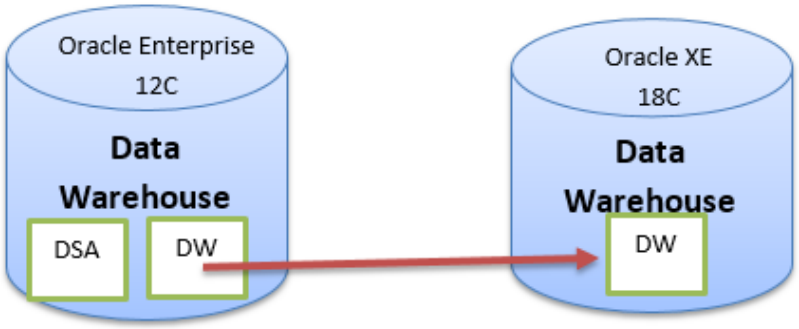

Figura 11. Esquema de Bases de Datos.Fuente: Ilustración Propia

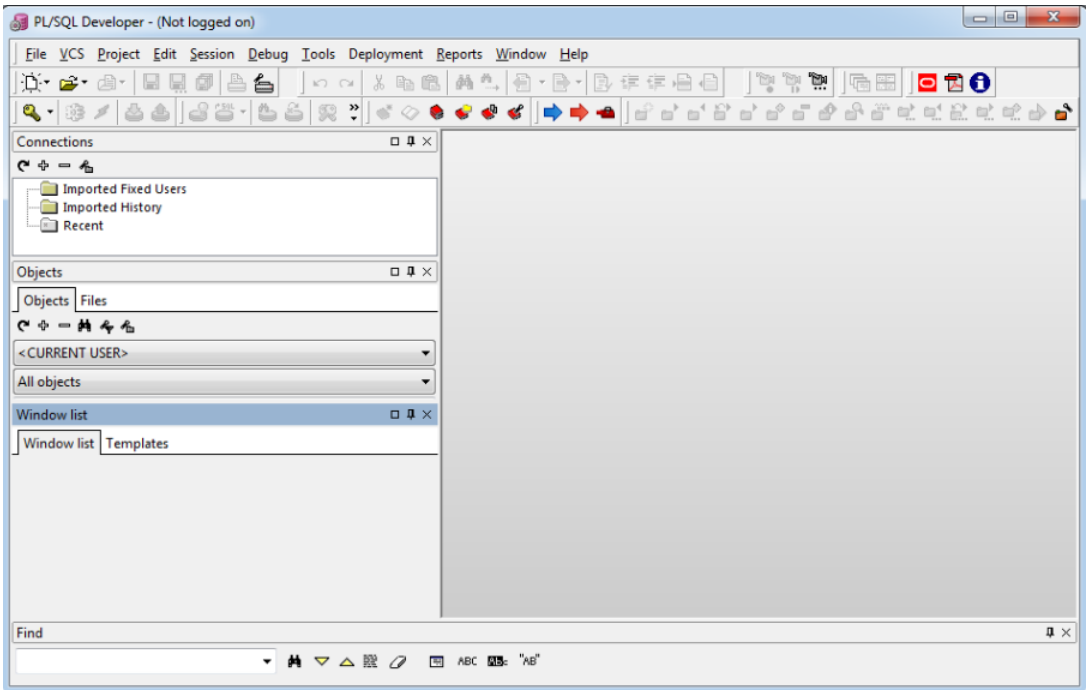

Figura 12. Lenguaje de Programación incrustado en Oracle PL/SQL. Fuente: PL/SQL

- **Creación de dimensiones, factos y agregaciones**

STORE: tabla que se usó para el catálogo de sucursales de la empresa retail; la misma que contiene la información de a nivel de código de sucursal, nombre de sucursal, tamaño de la sucursal, día de apertura de la sucursal, números telefónicos, etc.; dichos campos son ilustrados en la figura 13.

STORE	STORE_NAME10	STORE_NAME3	STORE_NAME_SECONDARY	STORE_CLASS	STORE_MGR_NAME	STORE_OPEN_DATE	FAX_NUMBER
26	FF1026	121	GRANDE	A	1026	1/12/1996	1790710319027
1	FY1001	106	GRANDE	A	1001	10/5/1985	1790710319002
2	FY1002	107	GRANDE	A	1002	1/11/2001	1790710319003
3	FY1003	108	GRANDE	A	1003	10/5/1985	1790710319004
4	FY1004	109	MEDIANA	A	1004	1/10/1985	1790710319005
5	FY1005	110	PEQUEÑA	A	1005	1/10/1985	1790710319006
9	FY1009	111	MEDIANA	A	1009	1/5/2004	1790710319010
12	FY1012	112	GRANDE	A	1012	4/12/2011	1790710319015
15	FY1015	113	GRANDE	A	1015	1/5/2003	1790710319018
17	FY1017	114	GRANDE	A	1017	1/8/1985	1790710319020
19	FY1019	115	GRANDE	A	1019	29/7/2010	1790710319022
20	FY1020	116	MEDIANA	A	1020	11/3/2011	1790710319012
21	FY1021	117	GRANDE	A	1021	1/12/2002	1790710319017
23	FY1023	118	MEDIANA	A	1023	1/10/1985	1790710319024
24	FY1024	119	MEDIANA	A	1024	1/12/1996	1790710319025
25	FY1025	120	GRANDE	A	1025	1/12/2001	1790710319026
28	FY1028	122	MEDIANA	A	1028	1/12/1996	1790710319029
29	FY1029	123	MEDIANA	A	1029	1/12/1995	1790710319030
31	FY1031	124	MEDIANA	A	1031	1/10/1997	1790710319032
32	FY1032	125	MEDIANA	A	1032	1/4/1998	1790710319033
33	FY1033	126	MEDIANA	A	1033	1/4/1998	1790710319034
34	FY1034	127	MEDIANA	A	1034	24/11/1999	1790710319035
271	FY1036	128	PEQUEÑA	A	1036	1/8/2002	1790710319038
148	FY1037	129	MEDIANA	A	1037	24/11/1999	1790710319037
531	FY1038	130	GRANDE	A	1038	1/3/2007	1790710319039
791	FY1039	131	MEDIANA	A	1039	1/4/2005	1790710319041
631	FY1040	132	GRANDE	A	1040	14/8/1985	1790710319021

Figura 13. Datos de la tabla STORE. Fuente: PL/SQL

VEN_CLIENTES: tabla que se usó para el catálogo de clientes la misma que almacena la información como la identificación, nombres y apellidos, dirección, teléfonos, fecha de nacimiento, etc.; los datos se muestran en la figura 14.

IDENTIFICACION	NOMBRE	DIRECCION	TELEFONO
1002158804001	CEDENO CRISTINA	HNOS. MIDEROS 272 IBARRA	099701191
1002178505001	ANDRADE MIRIAN JACQUELINE	CONJUNTO PUEBLO BLANCO	2074081
1002185674001	VILLEGAS CABRERA ANABEL DEL ROCIO	ULPIANO BETANCOURT S N Y FRANCISCO GUARDER	2870248
1002200069001	CABEZAS CRIOLLO EDITH NOEMI	IBARRA LA BOLA VERDE	022493076
1002256780001	VEGA MALES MARTHA CECILIA	OVIEDO 953 Y CHICA NARVAEZ	062605725
1002260063001	GUZMAN PROANO LENIN ALEJO	ARUPOS 154 Y TULIPANES	2476500
1002261350001	CHICAIZA VIRACOCHA PEDRO RENE	11 DE NOVIEMBRE MZ 29 LT 4	022492863
1002263802001	QUILCA VARGAS EDMUNDO JAVIER	GALTE Y AV. RODRIGO DE CHAVEZ	
1002296117001	PABON DE JESUS OSCAR XAVIER	MOSQUERA NARVAEZ	022228025
1002296810001	TITUANA QUIGUANGO LUIS EFRAIN	NOGALES 1458 Y FARSALIAS	098694972
1002305348001	CANAMAR MOLINA CESAR MANUEL	AV. ELOY ALFARO Y LAS FRUTILLAS	022228025
1002305959001	MALDONADO PADILLA LUIS TARQUINO	OTAVALO	022228025
1002321055001	CALPA JUAN	MARIANA DE JESUS	2527079
1002325742001	GUATEMAL FLORES LUIS PABLO	GUABOS LT.1-14 Y M. M. LIZARZA	022460660
0100333681001	RIOS MARCA JOSE OLEGARIO	RIOFRIO 1412 Y NICARAGUA	2525145
1002326591001	CACUANGO FERNANDO	CACHA 1340 Y GEOVANNY CALLES	2821001
1002332599001	POZO PATRICIO	PIFO - CHAUPI MOLINO	2381668
1002353025001	GUERRA HELEN	VANCOUVER 440 Y ALEMANIA	2563148
1002355574001	RUIZ VEGA EDGAR HIPOLITO	URB. PUERTA DEL SOL E12-113	2421000
1002360533001	CHALACAN RAMOS MARCO VINICIO	SANGOLQUI AV 10 DE DICIEMBRE Y COTAGCOCHA	2872821
1002374187001	AMBULANCIAS VIDA SALUD	QUESERAS DEL MEDIO E11 221 Y AV. GRAN COLOME	2524185
1002397550001	MANOSALVAS CORNEJO GIOVANNY MANUEL	VICENTE LOPEZ Y MACHALA	2291386
1002449633001	YEPEZ POZO HOLGER WASHINGTON	ISLA BALTRA 321 Y AV GENERAL RUMINAHUI	022863401
1002453767001	LA CANDELA	AV. DE LA PRENSA 157-189	2295366
1002460077001	PADILLA PADILLA WILMER JOSELITO	AV. 10 DE AGOSTO Y AV. AMAZONAS	023454021
1002499976001	CHIRIBOGA FLORES FELISA XIMENA	VICENTE ANDAGUIRRE OE5-59	022258223
1002513578001	S I M S A	AV. 6 DE DICIEMBRE N45-367 Y PIO VALDIVIESO	022430672

Figura 14. Datos de la tabla VEN_CLIENTES. Fuente: PL/SQL

ITEM_MASTER: tabla que se utilizó para el catálogo de ítems en la misma se encontró todas las características como el código del ítem, descripción o nombre del ítem, unidad de embalaje, precio de venta, etc.; los datos son ilustrados en la figura 15.

ITEM	ITEM_DESC	PERECEDERO	UNIDAD_EMPAQUE	UNIDAD_DESPACHO	UNIT_RETAIL
199155	CHUNCHIS SCH-058-450 CHUNCHI S/M	N	1,00	1,00	
206231	OXCARBAZEPINA (MK) TABS. REC. 300MG C/20 SUEL	Y			0,3500
210527	ACCES MAQUILLA ECOTOOLS 1202 BROCHA PARA M	N			8,1600
205265	JABON BARRA GLICERINA MISTY FLOR DE PLUMERIA	N			1,8800
210599	SOMBRAS JOLIE SEXTETO FRENCH MARTINI # 11 BL	N	1,00	1,00	
210323	CINTILLO SCUNCI 59280A BANDA ELASTICA 2PZS	N			2,0300
202427	PRODUCTOS UNIBALL SIGNO AROMA X 60 UNDS SUI	N	60,00	60,00	
207135	INSECTICIDAS RAID GOLD CUCARACHAS SPF/360CM	N			6,7500
205565	JUGO NATURAL DHOY NARANJA ZANAHORIA 335 ML	Y			1,2500
202521	BODY COCOON 1482 BODY STRAP TERM TANGA XL	N	12,00	1,00	
205330	REVISTAS MANUALIDADES CINTAS	N	1,00	1,00	
209004	MONOS QUILTEX MONO PETIT 016 TS	Y	12,00	1,00	
211105	GIVENCHY ANGE OU DEMON EDT 50 ML VAPO R.372	N	1,00	1,00	
202656	HELIOCARE GEL ULTRA PROTEC 90 UVB 50ML	N			30,8000
205709	BIBERONES CARLITOS PP 4-5 ONZAS BI-003-4	N	12,00	1,00	
205718	MONOS QUILTEX MONO ACOLCHADO 008 T3	N	12,00	1,00	
209089	PAPEL HIG HUMEDO POMPI FRSH POMPI FRESH ADI	N	12,00	1,00	
209100	ZAPATONES DESECHABLES C/100 SUELTAS	N			0,2200
209415	TOTTO GORRAS AC60IND185 LYNN STDO	N	1,00	1,00	
205594	CARAMELOS WONKA RUNTS	Y	24,00	1,00	
209490	AFEITADO ESPUMAS BIC REFRESH 150 ML	N	12,00	1,00	
212505	S-TRIAFEM CREMA 100GR	N	12,00	1,00	
205985	ARADOS HCT CAJA X 30 50MG 12.5MG SUELTAS	Y			0,6400
205986	ARADOS HCT CAJA X 30 100MG 25MG SUELTAS	Y			0,9800
209614	CERVEZA BRAHMA 300 ML	Y	24,00	1,00	
211567	PREMIOS CLIENTES ALBUM EASY ALBUM	N	1,00	1,00	
211642	LAPICES DE COLOR BIC PINTURAS BICOLOR CB X 24	N			6,2300
206000	TINTES COVER YOUR GRAY MEDIUM BROWN STICK	N			6,6600
203272	UNICLAR SPRAY NASAL INF 18 GR S/ALCOHOL	Y			22,9800

Figura 15. Datos de la tabla ITEM_MASTER. Fuente: PL/SQL

ITEM_PROVEEDOR: tabla que se utilizó para el catálogo de relación entre los ítems y los distintos proveedores en la misma se encontró todas las características como el proveedor padre, proveedor hijo, descripción o nombre del proveedor, código del ítem que está asignado a la sucursal, etc.; los datos son ilustrados en la figura 16.

PROVEEDOR_PADRE	PROVEEDOR_HIJO	DESC_PROV_HIJO	ITEM	LEAD_TIME_PROVEEDOR	ITEM_DESC	SUB_NAME
79,00	408651,00	ALVAREZ BARBA S.A. - Sin División	90910	3	PERFUMES PROBADORES HOMBRES DIOR HOMME	PERFUMES PROBADORES HOM
208,00	408409,00	KIMBERLY-CLARK EC S.A - Sin División	90912	4	PANAL HUGGIES NATURAL CARE NATURAL CARE XG	PANAL HUGGIES NATURAL CARE
208,00	408409,00	KIMBERLY-CLARK EC S.A - Sin División	90913	4	PANAL HUGGIES NATURAL CARE NATURAL CARE GR	PANAL HUGGIES NATURAL CARE
208,00	408409,00	KIMBERLY-CLARK EC S.A - Sin División	90914	4	PANITOS HUMEDOS HUGGIES R NACIDO C	PANITOS HUMEDOS HUGGIES.
226,00	10363411,00	LAS FRAGANCIAS C - SIN DIVISIÓN	90930	3	LANCOME MIRACLE EDP 30ML VAPO 802940	LANCOME MIRACLE
12598,00	407543,00	TARSIS S.A. - Sin División	90937	6	POP COLORS LABIAL ONLY YOU BRILLO LABIAL CREN	POP COLORS LABIAL ONLY YOU
76,00	408163,00	ABBOTT LABS DEL EC - ABBOTT NUTRICIONAL	90953	5	PEDIASURE SUPLEMENTO POLVO CHOCOLATE T/40(	PEDIASURE SUPLEMENTO
262,00	407264,00	PFIZER CIA LTDA.	90954	6	UNASYN SUSP.250 MG/ 5 ML F/120ML	UNASYN
262,00	407264,00	PFIZER CIA LTDA.	90955	6	ZITROMAX SUSP INF 1200MG /5ML F/30ML	ZITROMAX
238,00	408622,00	MERCK C.A. - BIOPHARMA	90971	5	ARTREN CAPS. SR 150MG C/10 SUELTAS	ARTREN
324,00	408681,00	NOVARTIS EC S.A. - SANDOZ	90974	3	ITRACARE CAPS. 100MG C/15 SUELTAS	ITRACARE
185,00	407801,00	BIDELSA S.A. - Sin División	96056	7	SET PARA BEBE PIGEON SET CUIDADO DE LA SALUD	SET PARA BEBE PIGEON.
203,00	1000269,00	JOHNSON & JOHNSON DEL EC - Sin División	96073	3	JABON BARRA NEUTROGENA DEEP CLEAN X 80 GR.	JABON BARRA NEUTROGENA GE
9596,00	408503,00	LETERAGO DEL EC S.A. - ROEMMERS	96075	5	UDOX COMP. 500 MG C/3 SUELTAS	UDOX
9596,00	408503,00	LETERAGO DEL EC S.A. - ROEMMERS	96076	5	UDOX SUSP. 200 MG F/30 ML	UDOX
218,00	408363,00	LABS INDUNIDAS - INDUNIDAS	96081	5	BIO T AMP. BEBIBLES 10 ML C/5 SUELTAS	BIO T
217,00	408464,00	LABOTAROTIOS H.G.	96082	4	CLOTRAZIL CREMA VAGINAL 2% T/20 GR.	CLOTRAZIL
217,00	408464,00	LABOTAROTIOS H.G.	96083	4	HEPALIDIN FORTE CAPS. C/20 SUELTAS	HEPALIDIN
217,00	408464,00	LABOTAROTIOS H.G.	96084	4	UROCAP CAPS. 500 MG C/50 SUELTAS	UROCAP
9242,00	407918,00	CONAIR JHERRY - Sin División	96195	3	ELASTICOS SCUNCI CONAIR 16775-A ELAST X 18 COL	ELASTICOS
16605,00	408039,00	DUTRAEC S.A. - Sin División	96207	7	BOCADITO DUTRAEC ALFAJOR INDIVIDUAL 15GR	BOCADITOS DUTRAEC1
16605,00	408039,00	DUTRAEC S.A. - Sin División	96208	7	BOCADITO DUTRAEC MANIZADO INDIVIDUAL 24GR	BOCADITOS DUTRAEC1
16605,00	408039,00	DUTRAEC S.A. - Sin División	96209	7	BOCADITO DUTRAEC GUAYABA INDIVIDUAL 30GR	BOCADITOS DUTRAEC1
16605,00	408039,00	DUTRAEC S.A. - Sin División	96210	7	BOCADITO DUTRAEC DULCE DE LECHE INDIVIDUAL 3	BOCADITOS DUTRAEC1

Figura 16. Datos de la tabla ITEM_PROVEEDOR. Fuente: PL/SQL

RFM_SEGMENTOS: tabla que se utilizó para almacenar la información de los tipos de segmentos como son el código y el nombre, estos datos sirvieron para hacer el descarte de los clientes que se encuentren en cada uno de los distintos segmentos y los mismos que servirán para el análisis RFM, estos datos son ilustrados en la figura 17.

SEGMENT	IID_SEGMENT
En riesgo	8
Potencial Leal	2
Nuevos Clientes	5
Others	4
Prometedor	1
Clientes Leales	9
No puedo perderlos	6
Perdido	3
A punto de dormir	10
Necesita Atencion	7

Figura 17. Datos de la tabla RFM_SEGMENTOS. Fuente: PL/SQL

DIVISION: tabla que se utilizó para almacenar el catálogo de la categorización de la división de los ítems y que contiene los códigos y nombres respectivamente, los datos de la división son ilustrados en la figura 18.

CODIGO	DIV_NAME
1	ALIMENTOS Y BEBIDAS
2	FOTO SHOP
3	CONSUMO
4	SUMINISTROS
5	TARJETAS ILIMITADAS
6	PRODUCTOS PROGRAMA LEALTAD
7	COMIDA RAPIDA
8	RECETARIO
9	BAZAR
10	MANTENIMIENTO
11	MEDICINAS
12	PROMOCIONES AUTOLIQUIDABLES

Figura 18. Datos de la tabla DIVISION. Fuente: PL/SQL

DEPT: tabla que se utilizó para almacenar el catálogo de la categorización de los departamentos de los ítems y que contiene los códigos de la división al cual pertenece, el código y nombre respectivamente del departamento, los datos de la división son ilustrados en la figura 19.

CODIGO	DEPT_NAME	COD_DIV
1	APARATO LOCOMOTOR	11
2	GRIPE OTC	11
3	DIGESTIVOS OTC	11
4	CUIDADO DEL BEBE	3
5	DESODORANTES PRODS. PIES	3
6	INCONTINENCIA	3
7	ACCESORIOS DE BEBE	3
8	TINTES Y COLORACION	3
9	CUIDADO ORAL	3
10	LIBROS	9
11	SNACKS	1
12	MERCHANDISING SUMINISTROS	4
13	TARJETAS DESCUENTO	5
14	FOTO SHOP	2
15	MASCOTAS	3
16	ASISTENCIAS	9
17	COMPLEMENTOS Y SUPLEMENTOS	11
18	ROPA INFANTIL	9
19	PROTECCION SOLAR Y REPELENTES	3
20	PILAS - BATERIAS - LINTERNAS	3

Figura 19. Datos de la tabla DEPT. Fuente: PL/SQL

CLASS: tabla que se utilizó para almacenar el catálogo de la categorización de las clases de los ítems y que contiene los códigos del departamento al cual pertenece, el código y nombre respectivamente de la clase, los datos son ilustrados en la figura 20.

CODIGO	CLASS_NAME	COD_DEPT
1	COLIRIO VASOCONSTRICTOR OTC	84
2	ANTIGOTOSOS	1
3	CORTICOSTEROIDES TOPICOS	29
4	ANTIEPILEPTICOS	85
5	PSICOLEPTICOS	85
6	ESTOMATOLOGICOS	50
7	PRODUCT ANTIDIABETICOS	50
8	FARMAC DESORD GASTROINT GENERICOS	50
9	ANALGESICOS Y ANTIPIRETICOS	42
10	ANTIRREUMATICOS TOPICOS	1

Figura 20. Datos de la tabla CLASS. Fuente: PL/SQL

SUBCLASS: tabla que se utilizó para almacenar el catálogo de la categorización de las subclases de los ítems y que contiene los códigos de la clase al cual pertenece, el código y nombre respectivamente de la subclase, los datos de la tabla son ilustrados en la figura 21.

CODIGO	SUB_NAME	COD_CLASS
1	FLAGYL-ANTIINFEC	305
2	6-COPIN	471
3	ALZATEN	90
4	AMPECU	539
5	ANAUTIN	471
6	DAVESOL	395
7	RINSOL	396
8	TEGRETOL	4
9	NAPAFEN	95
10	AMPLIURINA	310
11	CISTIL	310
12	ZINNAT	539
13	IMIGRAN	95
14	FOSFOCINA	539
15	FEMSTAT	305
16	FLORATIL	98
17	RENITEC	543
18	LOPID	476
19	LOCERYL	309
20	VASOACTIN	475

Figura 21. Datos de la tabla SUBCLASS. Fuente: PL/SQL

INTERVALO_RECENCIA_SCORE: tabla donde se almacenó la información correspondiente a los intervalos y pesos para la característica de recencia; los rangos o intervalos son valores por defecto recomendados por la librería RFM de la herramienta R como se muestra en la figura 22.

INTERVALO	SCORE	MIN	MAX
<=25	5	0	25
>25 & <=56	4	26	56
>56 & <= 106	3	57	106
>106 & <=184	2	107	184
>184	1	185	9999999999

Figura 22. Datos de la tabla INTERVALO_RECENCIA_SCORE. Fuente: PL/SQL

INTERVALO_FRECUENCIA_SCORE: tabla donde se almacenó la información correspondiente a los intervalos o pesos para la característica de frecuencia; los rangos o intervalos son valores por defecto recomendados por la librería RFM de la herramienta R como se muestra en la figura 23.

INTERVALO	SCORE	MIN	MAX
<=1	1	0	1
.>1 & <=1	2	2	2
>2	3	3	3
>3 & <= 6	4	4	6
>6	5	7	9999999

Figura 23. Datos de la tabla INTERVALO_FRECUENCIA_SCORE. Fuente: PL/SQL

INTERVALO_MONETARY_SCORE: tabla donde se almacenó la información correspondiente a los intervalos o pesos para la característica de valores monetarios; los rangos o intervalos son valores por defecto recomendados por la librería RFM de la herramienta R como s muestra en la figura 24.

INTERVALO	SCORE	MIN	MAX
<= 5,3	1	0	5,3
> 5,3 & <= 15,37	2	5,31	15,37
> 15,37 & <= 36,44	3	15,38	36,44
> 36,44 & <= 92,13	4	36,45	92,13
> 92,13	5	92,14	99999999999999

Figura 24. Datos de la tabla INTERVALO_MONETARY_SCORE. Fuente: PL/SQL

SA_VENTAS: tabla transaccional donde se utilizó la información correspondiente a las ventas realizadas en las sucursales, fecha de venta, ítem comprado, cantidad de ítems comprados, tipo de transacción e identificación del cliente; como se muestran en la figura 25.

TRAN_SEQ_NO	REV_NO	STORE_DAY_SEQ_NO	SUPPLIER	ITEM_ID	TRAN_DATETIME	REGISTER	TRAN_NO	CASHIER	SALESPERSON
2188416897	1	21903325	408315	100054017	1/1/2019 18:55:00	11	1211659	1002179685	1002179685
2188416897	1	21903325	10352946	100164832	1/1/2019 18:55:00	11	1211659	1002179685	1002179685
2188416898	1	21903325	407867	100087527	1/1/2019 16:48:00	11	1211660	1002179685	1002179685
2188416898	1	21903325	408315	209101	1/1/2019 16:48:00	11	1211660	1002179685	1002179685
2188416899	1	21903325	10582818	168967	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	408511	254998	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	10322435	167006	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	408144	7290	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	407604	100117230	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	407944	166699	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	10582818	168967	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416899	1	21903325	10352946	100164832	1/1/2019 20:28:00	13	1211661	1758649493	1758649493
2188416900	1	21903325	407287	300262	1/1/2019 16:57:00	9	1211662	1722705256	1722705256
2188416900	1	21903325	408665	17028	1/1/2019 16:57:00	9	1211662	1722705256	1722705256
2188416900	1	21903325	408665	17029	1/1/2019 16:57:00	9	1211662	1722705256	1722705256
2188416900	1	21903325	10352946	100164832	1/1/2019 16:57:00	9	1211662	1722705256	1722705256
2188416901	1	21903325	407287	300261	1/1/2019 16:42:00	7	1211663	1720798220	1720798220

Figura 25. Datos de la tabla SA_VENTAS. Fuente: PL/SQL

BI_VENTA_DEPT_CLI_SEG: tabla donde se almacenó la información correspondiente a la matriz de clientes sobre las categorías y las diferentes categorías existentes en las ventas asemejando a una matriz donde se encuentre agrupada la información, los datos se ilustran en la figura 26.

CUSTOMER_ID	SEGMENT_ID	CONGELADOS	PLANIFICACION_FAMILIA	TECNOLOGIA	MANICURE_Y_PEDICURE	PROGRAMA_PREMIOS_INSTANTANEOS
0502159742	3	0	0	0	0	0
0502160203	5	0	0	0	0	0
0502161102	9	0	0	0	0	0
0502161110	10	0	0	0	0	0
0502164213	8	0	0	0	0	0
0502175623	9	0	0	0	0	0
0502182496	7	0	0	0	0	0
0502197825	5	0	4,23	0	0	0
0502204969	10	0	0	0	0	0
0501511489	4	0	0	0	0	0
0501514707	6	0	0	0	0	0
0501522585	3	0	0	0	0	0
0501531248	2	0	0	0	0	0
0501544522	1	0	0	0	0	0
0501554075	3	0	0	0	0	0
0501565378	4	0	0	0	0	0
0501575377	3	0	0	0	0	0
0501580609	9	0	0	0	0	0
0501586267	10	0	0	0	0	0
0501598700	3	0	0	0	0	0

Figura 26. Datos de tabla BI_VENTA_DEPT_CLI_SEG. Fuente: PL/SQL

BI_RESUM_VTA_STORE_DEPT_CLASS_SUBCLASS: tabla donde se almacenó la información al total de las unidades vendidas y el total de las ventas agrupadas por el catálogo de ítems por sucursal, los datos son ilustrados en la tabla 27.

STORE	ID_DEPT	DEPT_NAME	ID_CLASS	CLASS_NAME	ID_SUBCLASS	SUBCLASS_NAME	CANT_VENDIDA	TOTAL_VENDIDA
1	3	DIGESTIVOS OTC	401	ANTIACID ANTIFLAT OTC	992	FLATLIVE	304	3,840(
1	3	DIGESTIVOS OTC	401	ANTIACID ANTIFLAT OTC	1186	ANTIAX	7387	313,140(
1	3	DIGESTIVOS OTC	401	ANTIACID ANTIFLAT OTC	1236	MILPAX	417	3047,010(
1	3	DIGESTIVOS OTC	401	ANTIACID ANTIFLAT OTC	1569	DIGESTA	10536	284,040(
1	3	DIGESTIVOS OTC	401	ANTIACID ANTIFLAT OTC	2015	PEPTOCID	2499	394,670(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	5456	COTONETES FYBE	94	193,500(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	5569	TALCO FISHER PRI	6	21,840(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	5842	VASELINA JOHNSO	117	393,240(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6033	CREMA HIDRATANT	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6161	CREMA HIDRATANT	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6311	COLONIA BEBE ARI	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6368	TALCO TUINIES	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6687	REPELENTE ANGEI	10	29,120(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6782	COTONETES COPI	15	46,200(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	6983	TALCO JOHNSON.	274	1134,930(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	7040	COLONIA JOHNSOI	52	371,280(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	7083	CREMA HIDRATANT	26	100,540(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	7297	TALCO PARA MI BE	30	48,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	7431	COLONIA ANGELIN	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	7742	ACEITE BEBE.	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	8353	REPELENTE JOHN	12	62,720(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	8406	CREMA HIDRATANT	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	8656	COTONETES CERC	0	0,000(
1	4	CUIDADO DEL BEBE	407	PRODUCTOS DE BEBE	8960	ACEITE SANASANA.	0	0,000(
1	4	CUIDADO DEL BEBE	408	ESTUCHES OBSEQUIOS BEBE	3679	ESTUCHES NIVEA E	0	0,000(
1	4	CUIDADO DEL BEBE	408	ESTUCHES OBSEQUIOS BEBE	5101	KIT ASEO PEQUENI	0	0,000(
1	4	CUIDADO DEL BEBE	408	ESTUCHES OBSEQUIOS BEBE	6263	ESTUCHES CREMA	0	0,000(
1	4	CUIDADO DEL BEBE	408	ESTUCHES OBSEQUIOS BEBE	8912	KIT ESCOLAR.	0	0,000(

Figura 27. Datos de tabla BI_RESUM_VTA_STORE_DEPT_CLASS_SUBCLAS. Fuente: PL/SQL

BI_RESUM_VENTA_DEPT_CLASS_SUBCLASS: tabla donde se almacenó la información del catálogo de productos contra la cantidad de productos vendidos en cada sucursal, los datos son ilustrados en la figura 28.

ID_DEPT	DEPT_NAME	ID_CLASS	CLASS_NAME	ID_SUBCLASS	SUBCLASS_NAME	CANT_SKU	FY1001	FY1002	FY1003	FY1004	FY1005	FY1009
78	CUIDADO DE LA PIEL	572	DESMAQUILLANTES GB	5206	DESMAQUILLANTES POND´S GB	2	0	0	0	0	0	0
78	CUIDADO DE LA PIEL	572	DESMAQUILLANTES GB	6759	DESMAQUILLANTES POMYS GB	1	0	0	0	0	0	0
78	CUIDADO DE LA PIEL	572	DESMAQUILLANTES GB	7721	DESMAQUILLANTES OPUS GB	12	442	137	348	173	34	76
78	CUIDADO DE LA PIEL	572	DESMAQUILLANTES GB	9135	DESMAQUILLANTES CALA GB.	4	22	0	18	22	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	804	CHOCOLATE HAPPY STOP.	3	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	1457	ARCOR.	6	52	10	7	18	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	1546	LAMBETZ.	2	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	1943	CHOC LA UNIVERSAL.	3	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	2089	CONFITECA NAVIDAD.	1	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	2505	CANDISNEY.	3	517	177	362	155	72	198
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	2873	NESTLE CLASSIC.	2	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	2959	CHOCOGOLOZIN.	2	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	3087	RINLETS.	2	7	3	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	3696	KINDER.	1	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	4334	BOMBONES DE LA VIUDA.	2	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	4990	RITTER.	4	220	81	137	100	0	79
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	5323	SHOCK.	2	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	5435	BOMBON MARIETTA.	3	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	5980	GALAK.	6	217	16	74	0	11	49
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	6341	JUBILEU.	9	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	6703	CHOCOLATE MINKA.	3	37	6	6	13	0	6
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	7104	GUYLIAN.	18	573	160	291	49	0	41
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	7594	S/AZUCAR AHORA SI BACCI.	1	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	7737	SOPRAFINO.	3	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	7862	ENTREDULCES.	23	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	8020	KATHY	1	0	0	0	0	0	0
79	CONFITERIA	47	CHOCOLATES GRANDES Y OBSEQUIO	8329	ILUSION.	1	0	0	0	0	0	0

Figura 28. Datos de tabla BI_RESUM_VENTA_DEPT_CLASS_SUBCLAS. Fuente: PL/SQL

AGG_VTS_CLI_DAY: tabla donde se almacenó la información agregada por día sobre el total de ítems comprados, total del valor monetario comprado por cliente; los datos son ilustrados en la figura 29.

CUSTUMER_ID	ORDER_DATE	REVENUE	CANTIDAD
1703324630	1/1/2019	63,54	6
1719819789	1/1/2019	48,08	9
1710644269	1/1/2019	117,36	6
1716586811	1/1/2019	32,14	3
1712895117	1/1/2019	78,9	24
0400429569	1/1/2019	7,36	16
1701076828	1/1/2019	241,12	10
0701118499	1/1/2019	71,55	5
1710649953	1/1/2019	33,65	24
1719040212	1/1/2019	70,68	70
1702939594	1/1/2019	31,41	3
0600515258	1/1/2019	8,44	11
1704455623	1/1/2019	248,19	36
1715469035	1/1/2019	50,08	5
1801821867	1/1/2019	2,35	1
1712287901	1/1/2019	117,85	56
1707826143	1/1/2019	34,65	3
1704322930	1/1/2019	2623,06	36

Figura 29. Datos de tabla AGG_VTS_CLI_DAY. Fuente: PL/SQL

RFM_RETAIL: tabla donde se almacenó la información de recencia, frecuencia y valor monetario por cliente; adicional existen los datos de la fecha de última compra así como la ponderación por cada categoría y la ponderación final; los datos son ilustrados en la figura 30.

IDENTIFICACION	RECENCIA	FRECUENCIA	MONETARIO	FECHA_ULTIMA_COMPRA	SCORE_R	SCORE_F	SCORE_M	RFM_SCORE
1700528613	10	4	41,7400	20/6/2019	5	4	4	544
1700533266	12	12	57,5600	18/6/2019	5	5	4	554
1700533803	10	22	81,1000	20/6/2019	5	5	4	554
1700534728	95	1	11,1000	27/3/2019	3	1	2	312
1700545740	32	11	138,0900	29/5/2019	4	5	5	455
1700547738	62	4	44,8100	29/4/2019	3	4	4	344
1700549718	18	11	282,7200	12/6/2019	5	5	5	555
1700551698	220	1	0,1700	22/11/2018	1	1	1	111
1700552639	197	2	23,8000	15/12/2018	1	2	3	123
1700562588	25	1	4,9700	5/6/2019	5	1	1	511
1700562968	118	1	29,6900	4/3/2019	2	1	3	213
1700563305	225	3	27,4300	17/11/2018	1	3	3	133
1700563644	127	2	8,1200	23/2/2019	2	2	2	222
1700564618	146	1	5,5000	4/2/2019	2	1	2	212
1700575770	30	9	31,9100	31/5/2019	4	5	3	453
1700575986	297	1	7,1200	6/9/2018	1	1	2	112
1700579285	217	1	0,7000	25/11/2018	1	1	1	111
1700582586	139	1	1,2600	11/2/2019	2	1	1	211
1700586215	25	15	394,6500	5/6/2019	5	5	5	555
1700587775	134	1	6,5300	16/2/2019	2	1	2	212
1700590472	42	3	14,5300	19/5/2019	4	3	2	432
1700592346	102	2	0,3500	20/3/2019	3	2	1	321
1700602749	194	1	4,9800	18/12/2018	1	1	1	111
1700608050	14	14	87,9500	16/6/2019	5	5	4	554
1700609017	236	1	2,5200	6/11/2018	1	1	1	111
1700615451	27	17	201,5800	3/6/2019	4	5	5	455
1700619321	58	7	97,9300	3/5/2019	3	5	5	355
1700620360	80	6	121,6200	11/4/2019	3	4	5	345
1700624776	125	1	7,0000	25/2/2019	2	1	2	212

Figura 30. Datos de tabla RFM_RETAIL. Fuente: PL/SQL

- **Procesos ETL**

Para los procesos de carga de información o conocidos como ETL, se realizó y desarrolló programación SQL que fueron creados a nivel de base de datos para que su procesamiento sea más óptimo y el consumo de recursos sea menor, teniendo en cuenta el concepto de ETL que es la extracción, transformación y carga de información utilizada para cada tabla creada, que se detalla a continuación:

Se pobló de datos a la tabla **BI_RESUM_VTA_STORE_DEPT_CLASS_SUBCLAS** tomando como cursores bases las fuentes de las tablas dept, class, subclass y store; de este último

se filtró para el análisis solo las sucursales donde existe un número mayor de ventas con el fin de evitar que en el comportamiento del cliente no exista una dispersión de datos. A continuación, se procede con el inicio de los bucles de iteración tomando como referencia primero las sucursales, posteriormente los departamentos, luego la clase de productos y al final las subclases obteniendo como resultado los datos del total de ventas y unidades por cliente agrupados por departamento, clase, subclase por sucursal. Este proceso se ilustra a continuación en la figura 31.

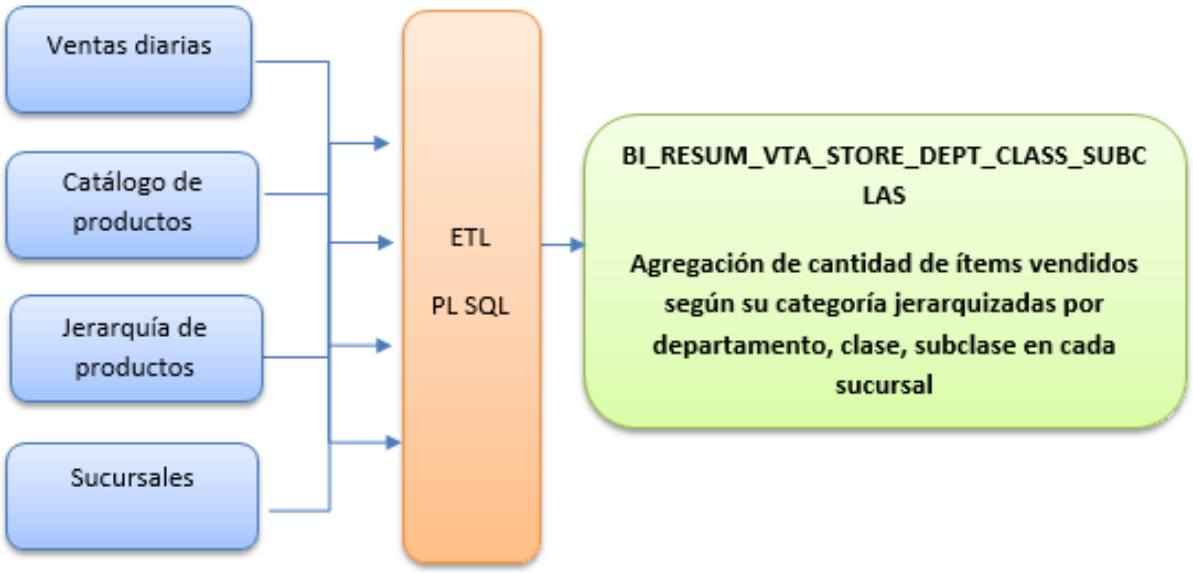

Figura 31. Ilustración de ETL para llenado de tabla BI_RESUM_VTA_STORE_DEPT_CLASS_SUBCLAS. Fuente: Ilustración Propia

Se pobló de datos a la tabla **BI_RESUM_VENTA_DEPT_CLASS_SUBCLASS** tomando como cursores bases las fuentes las tablas dept, class y subclass. A continuación, se procedió con el inicio de los bucles de iteración tomando como referencia primero los departamentos, posteriormente la clase de productos y al final las subclases obteniendo como resultado el total de la cantidad de ítems que está asociado al proveedor tomando en cuenta la jerarquización de

productos comparado con la cantidad de productos de ese proveedor vendidos en cada sucursal. Este proceso se ilustra a continuación en la figura 32.

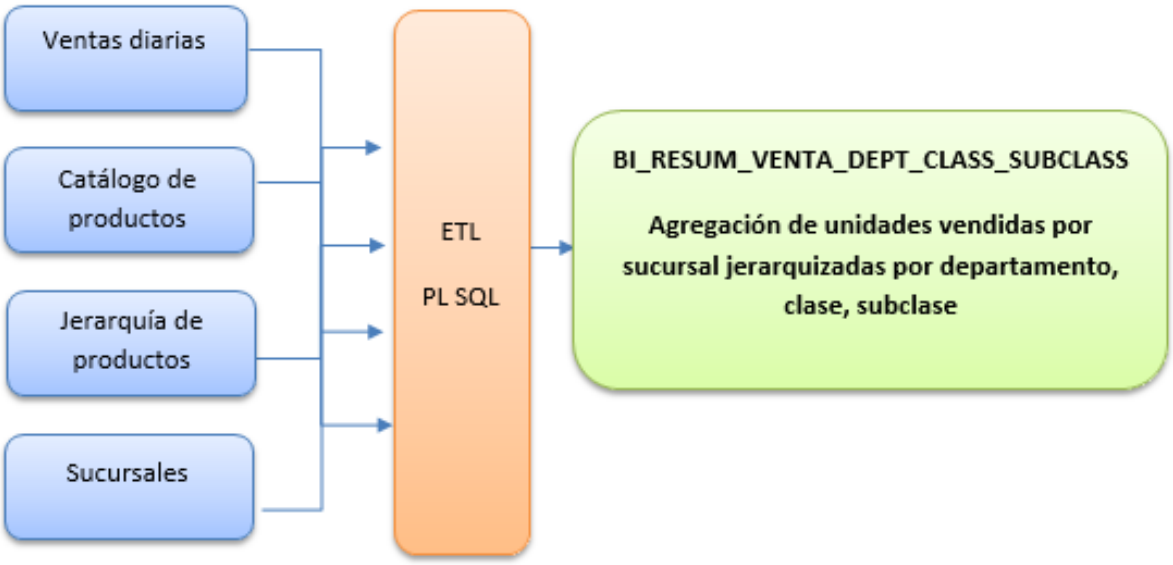

Figura 32. Ilustración de ETL para llenado de tabla BI_RESUM_VENTA_DEPT_CLASS_SUBCLASS. Fuente: Ilustración Propia

Para poblar la tabla **AGG_VTS_CLI_DAY** se realizó la obtención de los datos del total de compra y el total de ítems de las ventas agrupados por día y cliente, esta información permitirá tener una visión global de las ventas diarias de los clientes. Este proceso se resume en la figura 33.

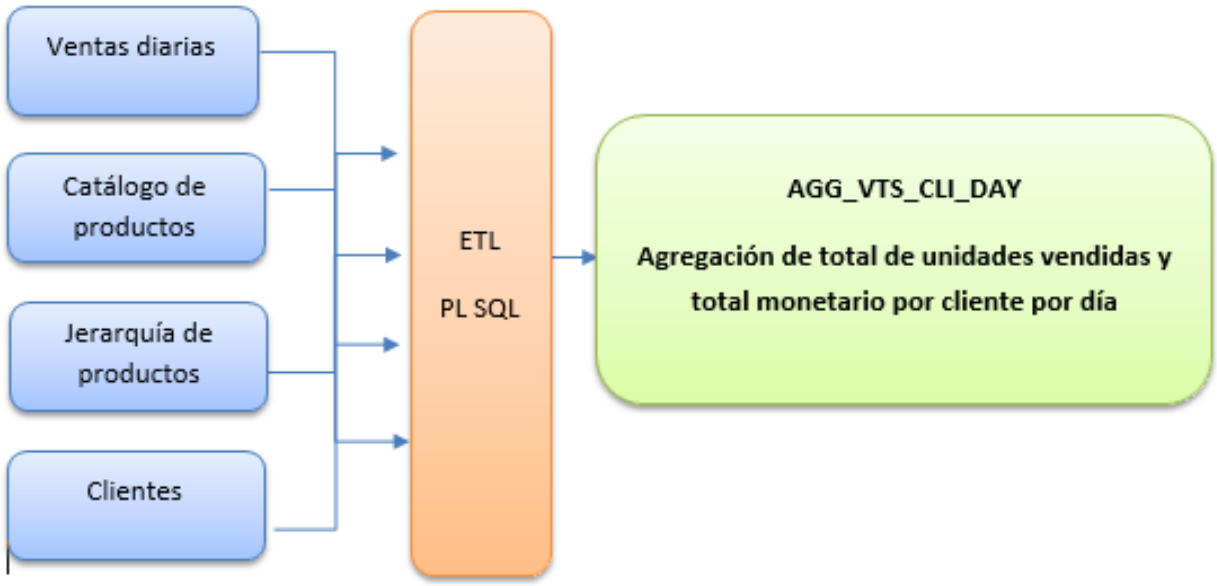

Figura 33. Ilustración de ETL para llenado de tabla AGG_VTS_CLI_DAY. Fuente: Ilustración Propia

Para poblar de datos a la tabla **RFM_RETAIL** mediante sentencia SQL, se realizó la obtención de clientes únicos donde su número de identificación no sea nula; a continuación se necesita realizar la actualización de datos para que cada persona se le asigne la puntuación o ponderación correspondiente a la recencia, frecuencia y valor monetario utilizando una sentencia de repetición automática desde el primer hasta el último registro por cliente. Para la carga de la información correspondiente a las ponderaciones en la tabla **RFM_RETAIL** se realizó tres actualizaciones de los clientes:

- Para la recencia se tomó como fecha de análisis el último día de corte de la información 30/06/2019 que se encuentra en la tabla **AGG_VTS_CLI_DAY**, con dicha fecha se realizó la resta de fechas entre la última fecha de corte y la última fecha de compra por cada cliente obteniendo como resultado el número de días.

- Para la frecuencia se realizó el conteo de todas las compras por cliente, con dicho dato se realizó la actualización del total de compras.
- Se finalizó la carga la información del valor monetario con la sumatoria del total de compra realizado por cada cliente.

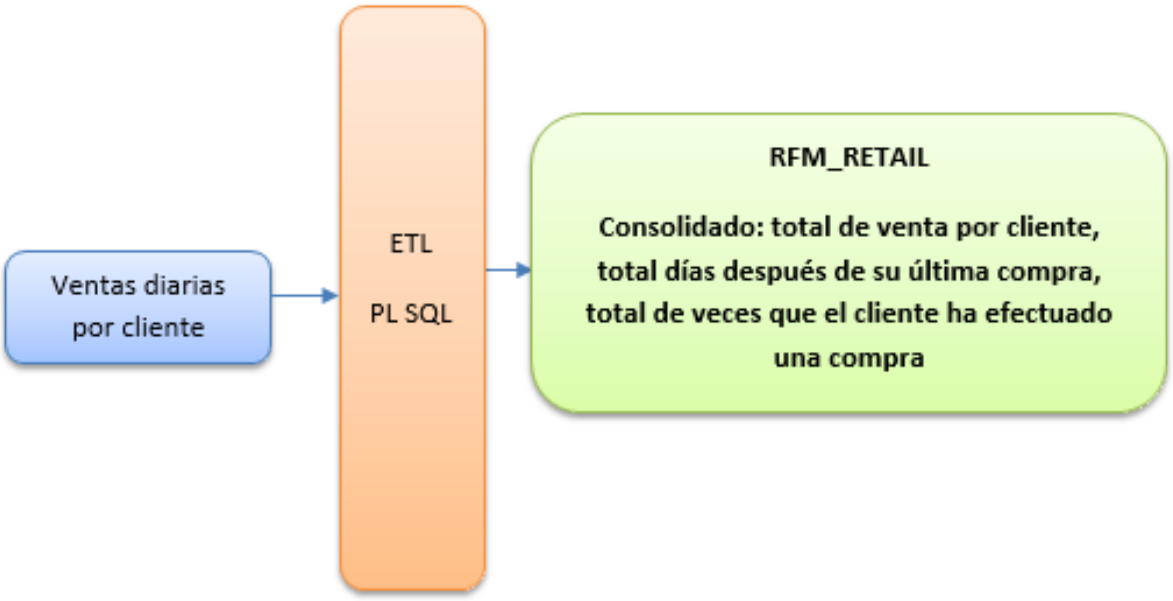

Figura 34. Ilustración de ETL para llenado de tabla RFM_RETAIL. Fuente: Ilustración Propia

Para poblar la tabla **BI_VENTA_DEPT_CLI_SEG** se realizaron dos cursores bases tomando como referencia los departamentos de la categorización y todos los clientes que han realizado al menos una compra en la cadena de supermercados. Para cada cliente se realiza la sumatoria del total de compra realizado en las sucursales 1, 2, 3, 4, 5, 9, 12, 15, 17 y 19 así como que el ítem llevado en la compra conste en el catálogo del maestro detalle del ítem.

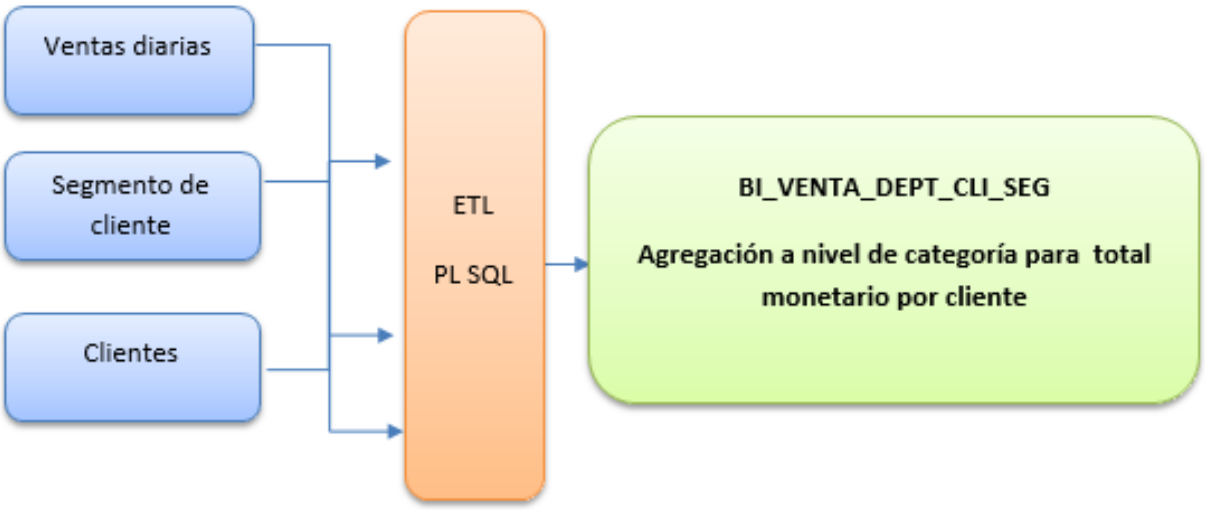

Figura 35. Ilustración de ETL para llenado de tabla BI_VENTA_DEPT_CLI_SEG. Fuente: Ilustración Propia

CAPÍTULO V
ANÁLISIS Y RESULTADOS

5.1 Análisis de Resultados

Para la investigación de datos masivos de las ventas, se realizó el análisis RFM con dos tipos de librerías en la herramienta estadística R, con el fin de comparar los resultados obtenidos y saber cuál fue la mejor opción y que se adapte a las necesidades de la empresa de retail ecuatoriana.

- **Análisis RFM en R con librería RJDBC**

Una vez que se procesó la información de recencia, frecuencia y valor monetario, se procedió a realizar el análisis mediante R con la librería RJDBC, ejecutando las sentenciasen en el orden correspondiente:

Se cargaron las librerías de R:

- Library (RJDBC)
- Library (rJava)

Se cargó el driver que se encuentra en extensión jar ejecutando la sentencia en R:

- driver <- JDBC(driverClass = "oracle.jdbc.OracleDriver","C:\\Users\\Diego\\Documents\\jdbc\\ojdbc8.jar")

Se realizó la conexión a la base de datos del DWH, ejecutando la sentencia en R, este paso depende tanto de la instalación de las librerías como del driver

- conexion <- dbConnect(driver,"jdbc:oracle:thin:@localhost:1521:xe","dw","dw1234")

Una vez finalizada la conexión de la base de datos, se debe cargó desde la tabla RFM_RETAIL en una variable de tipo consulta, esto permitió guardar toda la información de la tabla sin tener la necesidad de estar consultando n-veces a la base de datos.

- consulta <-dbReadTable(conexion,"RFM_RETAIL")

Una vez realizada la conexión de base de datos, se generaron diferentes quantiles, los mismos que están especificados por los valores: 0%, 20%, 40%, 60%, 80% y 100%; para ello se ejecutaron las siguientes sentencias.

- Recencia: quantile(consulta$RECENCIA,probs = c(0, 0.20, 0.40, 0.60, 0.80, 1), na.rm = TRUE)
- Frecuencia: quantile(consulta$FRECUENCIA,probs = c(0, 0.20, 0.40, 0.60, 0.80, 1), na.rm = TRUE)
- Valor Monetario: quantile(consulta$MONETARIO,probs = c(0, 0.20, 0.40, 0.60, 0.80, 1), na.rm = TRUE)

Para finalizar la generación del análisis RFM, se imprimieron los datos dando como resultado la tabla 5.

Tabla 5.

Cuadro de Análisis RFM

PROBABILIDAD	*RECENCIA*	*FRECUENCIA*	*MONETARIO*
0%	0,00	0,00	0,01
20%	25,00	1,00	5,30
40%	56,00	1,00	15,37
60%	106,00	3,00	36,44
80%	184,00	6,00	92,13
100%	394,00	297,00	4340238,79

- **Análisis RFM en R con librería RFM**

Para el segundo análisis con la librería RFM como primer pasó se realizó la exportación hacia un archivo Excel los datos de las transacciones de ventas por día y cliente como se indica en la figura 36.

	A	B	C
1	customer_id	order_date	revenue
2	1700528613	20/6/2019	41,74
3	1700533266	18/6/2019	57,56
4	1700533803	20/6/2019	81,1
5	1700534728	27/3/2019	11,1
6	1700545740	29/5/2019	138,09
7	1700547738	29/4/2019	44,81
8	1700549718	12/6/2019	282,72
9	1700551698	22/11/2018	0,17

Figura 36. Transacciones de ventas de clientes por día. Fuente: Ilustración Propia

A continuación, se debe importó la librería RFM mediante la sentencia:

- install.packages("rfm")

Una vez finalizada la instalación de la librería RFM, se procedió a cargar el archivo de las transacciones de ventas de clientes por día en R a la variable **rfmCSV** mediante la ejecución de la sentencia:

- rfmCSV <- read_delim("C:/Users/Diego/Desktop/rfmCSV.csv",";", escape_double = FALSE, col_types = cols(customer_id = col_character()),locale = locale(decimal_mark = ","),trim_ws = TRUE)

La librería RFM necesitó de la declaración de una fecha de corte, la misma que sirvió para establecer el número de días de la última fecha de compra por cliente, para este caso en mención

se utilizó la fecha de corte 2019-06-30 asignando a la variable **analysis_date** que es cuando se comenzó a realizar el análisis RFM, se ejecutó la siguiente sentencia:

- analysis_date <- lubridate::as_date('2019-06-30', tz = 'UTC')

Una vez que se estableció la fecha de corte, se procedió a realizar el análisis RFM de los datos transaccionales, el resultado se vió reflejado en una tabla propia declarando en la variable **rfm_result**, para ello se ejecutó la sentencia:

- rfm_result <- rfm::rfm_table_order(rfmCSV,customer_id, order_date, revenue, analysis_date)

A continuación, se establecieron segmentos de clientes los mismos que fueron creados de acuerdo a la experiencia de los asesores comerciales de la empresa de retail, los cuales fueron guardados en la variable **segment_names,** para ello se ejecutó la sentencia:

- segment_names <- c("Champions", "Loyal Customers", "Potential Loyalist","New Customers", "Promising", "Need Attention", "About To Sleep","At Risk", "Can't Lose Them", "Lost")

Para establecer las ponderaciones por cada tipo del análisis RFM, se tomaron en cuenta valores mínimos y máximos establecidos conjuntamente con el área comercial y que son fueron guardados en las variables **recency_lower, recency_upper, frequency_lower, frequency_upper, monetary_lower** y **monetary_upper:**

- recency_lower <- c(4, 2, 3, 4, 3, 2, 2, 1, 1, 1)
- recency_upper <- c(5, 5, 5, 5, 4, 3, 3, 2, 1, 2)

- frequency_lower <- c(4, 3, 1, 1, 1, 2, 1, 2, 4, 1)
- frequency_upper <- c(5, 5, 3, 1, 1, 3, 2, 5, 5, 2)
- monetary_lower <- c(4, 3, 1, 1, 1, 2, 1, 2, 4, 1)
- monetary_upper <- c(5, 5, 3, 1, 1, 3, 2, 5, 5, 2)

Una vez que las variables calculadas que se necesitaban para el proceso de segmentación de paquete RFM, se ejecutó la sentencia:

- rfm::rfm_segment(rfm_result$rfm,segment_names, recency_lower, recency_upper, frequency_lower, frequency_upper, monetary_lower, monetary_upper)

Se generó el gráfico de la media frecuencia para saber cuál fue la segmentación más óptima del análisis FRM ejecutando la sentencia:

- rfm::rfm_plot_median_frequency(rfm::rfm_segment(rfm_result,segment_names, recency_lower, recency_upper, frequency_lower, frequency_upper, monetary_lower, monetary_upper))

La sentencia ejecutada anteriormente dio como resultado que la segmentación de clientes leales es la que posee una mayor frecuencia de compra frente a las demás como se indica en la figura 37.

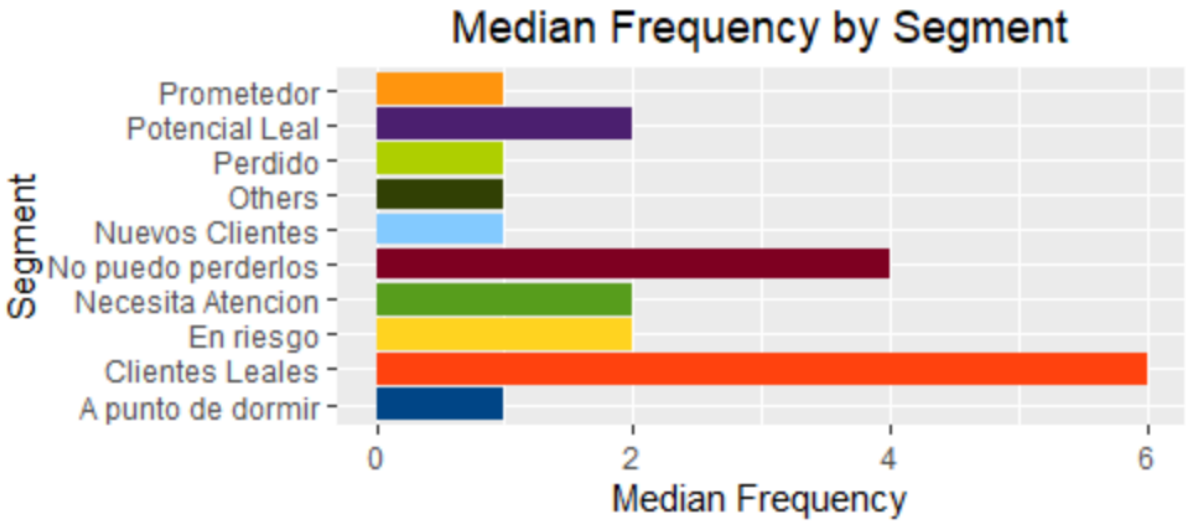

Figura 37. Frecuencia Media por Segmentación de clientes. Fuente: Herramienta R

A continuación se procedió a visualizar por cada segmento de clientes los valores calculados en el proceso RFM, para ello se ejecutó la siguiente sintaxis:

- RFM_RESULT_CLI_SEG<-rfm::rfm_segment(rfm_result,segment_names, recency_lower, recency_upper, frequency_lower, frequency_upper, monetary_lower, monetary_upper)

El resultado de la ejecución de la función RFM_RESULT_CLI_SEG del proceso RFM se puede visualizar en el gráfico 38.

customer_id	segment	rfm_score	transaction_count	recency_days	amount	recency_score	frequency_score	monetary_score
0000043000001	Clientes Leales	555	59	9	1495570.82	5	5	5
0000045000001	No puedo perderlos	145	5	206	10882.36	1	4	5
00001494	Perdido	111	1	265	5.75	1	1	1
0000279	Clientes Leales	434	2	37	501.09	4	3	4
0001861467	Perdido	111	1	255	7.12	1	1	1
00033689	Perdido	112	1	237	51.60	1	1	2
0003376	Perdido	111	1	257	3.08	1	1	1
000707158	Potencial Leal	532	3	10	20.37	5	3	2
0007721	Others	115	1	191	964.56	1	1	5
000979948	En riesgo	244	4	171	358.83	2	4	4
0010253	Others	214	1	175	535.50	2	1	4

Figura 38. Segmentación de clientes. Fuente: Herramienta R

A continuación, se importaron los datos obtenidos en la herramienta R hacia la base de datos, dichos datos se guardan en la tabla RFM_CLI_SEG, la misma que consta con la información por cada cliente, segmento al que pertenece, ponderación o calificación del análisis RFM, número de transacciones, número de días de la última compra, total de la venta y así como las ponderaciones de frecuencia, recencia y valor monetario. Los datos obtenidos son visualizados en la figura 39.

CUSTOMER_ID	SEGMENT	RFM_SCORE	TRANSACTION_COUNT	RECENCY_DAYS	AMOUNT	RECENCY_SCORE	FREQUENCY_SCORE	MONETARY_SCORE
0502159742	Perdido	112	1	199	17,74	1	1	2
0502160203	Nuevos Clientes	511	1	9	5,36	5	1	1
0502161102	Clientes Leales	335	2	78	786,22	3	3	5
0502161110	A punto de dormir	312	1	91	41,76	3	1	2
0502161888	A punto de dormir	311	1	59	15,46	3	1	1
0502163298	En riesgo	232	2	123	66,39	2	3	2
0502163348	Others	442	5	37	44,80	4	4	2
0502163447	Potencial Leal	413	1	47	113,60	4	1	3
0502163868	Clientes Leales	344	4	103	311,18	3	4	4
0502164031	Prometedor	411	1	40	2,97	4	1	1
0502164213	En riesgo	233	2	143	72,27	2	3	3
0502164619	Perdido	211	1	141	4,83	2	1	1
0502165053	A punto de dormir	311	1	103	3,36	3	1	1
0502165665	Others	114	1	252	249,39	1	1	4
0502165863	Perdido	211	1	184	10,48	2	1	1
0502165889	Perdido	211	1	149	5,50	2	1	1
0502165905	Others	115	1	297	1003,73	1	1	5
0502165913	En riesgo	232	2	140	68,47	2	3	2
0502166184	En riesgo	134	2	210	461,11	1	3	4

Figura 39. Datos importados a la tabla RF_CLI_SEG. Fuente: PL/SQL

Para manera informativa se ejecutó la sentencia SQL de agrupación de datos para saber el total de clientes por segmento como se muestra en la figura 40, constatando la coherencia con el gráfico de frecuencia media (ver figura 37).

SEGMENT	TOTAL_CLIENTES
Clientes Leales	101107
Perdido	55766
Others	40833
Potencial Leal	37121
En riesgo	36344
A punto de dormir	16779
Necesita Atencion	10873
Prometedor	7346
Nuevos Clientes	4830
No puedo perderlos	862

Figura 40. Datos agrupados del total de clientes por segmento. Fuente: PL/SQL

- **Análisis de Componentes Principales (PCA)**

En la variable datos se cargó la información de las ventas de todos los clientes solo del segmento de **clientes leales** que es igual al identificador nueve (9), realizando un match de datos entre la información obtenida por el análisis RFM (RFM_CLI_SEG) y la tabla de agregación de ventas de clientes por categoría (BI_VENTA_DEPT_CLI_SEG); para ello se tomó como filtro el segmento de tipo de clientes.

CUSTOMER_ID	SEGMENT_ID	N_A_12	PLANIFICACION_FAMILIA	TECNOLOGIA	MANICURE_Y_PEDICURE	CONGELADOS	PROGRAMA_PREMIOS_INSTANTANEOS	N_A_21	N_A_28	CUIDADO_DEL_BEBE
0502159742	3	0	0	0	0	0	0	0	0	0
0502160203	5	0	0	0	0	0	0	0	0	0
0502161102	9	0	0	0	0	0	0	0	0	0
0502161110	10	0	0	0	0	0	0	0	0	0
0502164213	8	0	0	0	0	0	0	0	0	0
0502175623	9	0	0	0	0	0	0	0	0	0
0502182496	7	0	0	0	0	0	0	0	0	0
0502197825	5	0	4,23	0	0	0	0	0	0	0
0502204969	10	0	0	0	0	0	0	0	0	0
0501511489	4	0	0	0	0	0	0	0	0	0
0501514707	6	0	0	0	0	0	0	0	0	0
0501522585	3	0	0	0	0	0	0	0	0	0
0501531248	2	0	0	0	0	0	0	0	0	0
0501544522	1	0	0	0	0	0	0	0	0	0
0501554075	3	0	0	0	0	0	0	0	0	0
0501565378	4	0	0	0	0	0	0	0	0	0
0501575377	3	0	0	0	0	0	0	0	0	0
0501580609	9	0	0	0	0	0	0	0	0	0
0501586267	10	0	0	0	0	0	0	0	0	0
0501598700	3	0	0	0	0	0	0	0	0	0
0501601660	3	0	0	0	0	0	0	0	0	0
0501602536	9	0	4.42	0	0	0	0	0	0	0

Figura 41. Datos agrupados de las ventas de clientes leales por todas sus categorías. Fuente: PL/SQL

Se guardó la información de la tabla BI_VENTA_DEPT_CLI_SEG en la variable datos, mediante la ejecución de la sentencia:

- attach(datos)

A continuación, se debe realizó la revisión de la correlación de datos, para ello se ejecutó la sentencia:

- cor(datos)

Luego de que se realizó la correlación, se estandarizaron los datos para posteriormente regularizar en base a su media, esto se lo realizó con la ejecución de las sentencias:

- scaledatos=scale(datos)
- acp=prcomp(scaledatos)

Ahora se debe visualizar los datos de la variable de ACP ya que contiene todos los componentes principales dando como resultado 42 agrupaciones, para ello se ejecutó las siguientes sentencias:

- acp
- summary(acp)

Los datos que se visualizan de todos los componentes principales, se pueden ver en la tabla 6.

Tabla 6.

Cuadro de Componentes Principales

TIPO	***PC1***	***PC2***	***PC3***	***PC4***	***PC5***	***PC6***	***PC7***
Standard deviation	61.039	101.484	0.99202	0.71942	0.64687	0.56239	0.49924
Proportion of Variance	0.8871	0.02452	0.02343	0.01232	0.00996	0.00753	0.00593
Cumulative Proportion	0.8871	0.91161	0.93504	0.94737	0.95733	0.96486	0.97079
TIPO	***OPC8***	***PC9***	***PC10***	***PC11***	***PC12***	***PC13***	***PC14***
Standard deviation	0.44963	0.4201	0.37510	0.30982	0.29183	0.23585	0.22638
Proportion of Variance	0.00481	0.0042	0.00335	0.00229	0.00203	0.00132	0.00122
Cumulative Proportion	0.97561	0.9798	0.98316	0.98544	0.98747	0.98880	0.99002
TIPO	***PC15***	***PC16***	***PC17***	***PC18***	***PC19***	***PC20***	***PC21***
Standard deviation	0.22098	0.21784	0.21228	0.20655	0.19059	0.16668	0.14324
Proportion of Variance	0.00116	0.00113	0.00107	0.00102	0.00086	0.00066	0.00049
Cumulative Proportion	0.99118	0.99231	0.99338	0.99440	0.99526	0.99592	0.99641
TIPO	***PC22***	***PC23***	***PC24***	***PC25***	***PC26***	***PC27***	***PC28***
Standard deviation	0.13781	0.12719	0.12105	0.11673	0.10471	0.10045	0.09947
Proportion of Variance	0.00045	0.00039	0.00035	0.00032	0.00026	0.00024	0.00024
Cumulative Proportion	0.99687	0.99725	0.99760	0.99792	0.99819	0.99843	0.99866
TIPO	***PC29***	***PC30***	***PC31***	***PC32***	***PC33***	***PC34***	***PC35***

Standard deviation	0.09615	0.09023	0.08410	0.07992	0.07685	0.06883	0.06489
Proportion of Variance	0.00022	0.00019	0.00017	0.00015	0.00014	0.00011	0.00010
Cumulative Proportion	0.99888	0.99907	0.99924	0.99940	0.99954	0.99965	0.99975
TIPO	***PC36***	***PC37***	***PC38***	***PC39***	***PC40***	***PC41***	***PC42***
Standard deviation	0.05840	0.05480	0.04821	0.03284	0.02025	0.01453	0.01001
Proportion of Variance	0.00008	0.00007	0.00006	0.00003	0.00001	0.00001	0.00000
Cumulative Proportion	0.99983	0.99990	0.99996	0.99998	0.99999	100.000	100.000

Una vez que se visualizó todos los componentes principales, solo se toman para el análisis los componentes CP1, CP2 y CP3 por motivo que sus datos se encuentran más correlacionados entre sí. Sin embargo, se procede a calcular la desviación estándar de los componentes principales uno, dos y tres guardando la información en las variables **desv_stand, desv_stand2** y **desv_stand3.**

- desv_stand=acp[[1]]
- desv_stand2=acp[[2]]
- desv_stand3=acp[[3]]

Con el valor de la desviación estándar se procede a calcular la varianza de los tres componentes y guardar el valor en las variables **varianza, varianza2** y **varianza3.**

- varianza=desv_stand^2
- varianza2=desv_stand^2
- varianza3=desv_stand^2

Ahora se procedió a calcular los tres componentes principales y guardarlos en las variables **CP1**, **CP2** y **CP3**

- CP1=acp[[2]][,1]
- CP2=acp[[2]][,2]
- CP3=acp[[2]][,3]

Una vez calculados los componentes principales, se procedió con el cálculo de la composición de los tres componentes principales, que indica como las variables están relacionadas entre sí, este cálculo es guardado en la variable **Comp_prin.**

- Comp_prin = cbind(CP1,CP2,CP3)

Para poder realizar el análisis de componentes principales, se han tomado en cuenta los componentes CP1, CP2 y CP3 ya que sus categorías son las que más se asemejan entre sí, para poder visualizar la matriz final de componente se ejecutó el comando:

- Comp_prin

La visualización de los datos se puede verificar en la tabla 7, entre los cuales son la categoría de ítems, el valor del componente principal uno (CP1), dos (CP2) y tres (CP3).

Tabla 7.

Cuadro de Categorías de Productos de Componentes Principales

CATEGORIA DE ITEMS	*CP1*	*CP2*	*CP3*
planificacion_familia	1.635479e-01	0.0107786603	0.0083714483
Tecnología	1.613500e-01	0.0103635045	0.0080123645
manicure_y_pedicure	1.628715e-01	0.0084870821	0.0065422809
Congelados	1.634951e-01	0.0098153032	0.0076329962
cuidado_del_bebe	1.623027e-01	0.0034282766	0.0025400686
aparato_respiratorio	1.582257e-01	-0.0077404105	-0.0057219166
Mascotas	1.361006e-01	-0.0072096488	-0.0058287922
alivio_del_dolor_otc	1.628117e-01	0.0019670083	0.0017989425
proteccion_solar_y_repelentes	1.620273e-01	0.0059559709	0.0045240943

CATEGORIA DE ITEMS	*CP1*	*CP2*	*CP3*
Perfumería	1.255776e-01	-0.0050320041	-0.0039265024
shampoo_acondicionador_y_trata	1.625743e-01	0.0033936703	0.0028544644
Medias	1.630843e-01	0.0101137916	0.0078098919
Incontinencia	1.141209e-01	-0.1262608118	-0.1015666172
cuidado_oral	1.632369e-01	0.0035243805	0.0029484400
Antiparasitarios	1.598157e-01	0.0090264490	0.0069347428
soluciones_hospitalarias	1.597842e-01	0.0089868514	0.0069695297
ropa_infantil	1.487988e-01	0.0045707214	0.0034308702
Depilación	1.621469e-01	0.0085176041	0.0065498960
Autoliquidables	8.767293e-03	-0.7955317918	-0.5789941642
Panales	1.524285e-01	0.0005189145	0.0002621147
cuidado_de_la_piel	1.628424e-01	0.0025937737	0.0020428805
Confitería	1.635946e-01	0.0097536828	0.0075762771
jabones_prods__bano	1.629109e-01	0.0048590849	0.0037167563
albumes_y_cromos	1.423633e-01	0.0125236058	0.0098589644
desodorantes_prods__pies	1.632800e-01	0.0080922951	0.0062115849
papel_higienico_y_faciales	1.607839e-01	-0.0046344247	-0.0037031668
productos_naturales	1.599458e-01	-0.0108692004	-0.0085276113
Hormonas	1.579588e-01	0.0045864531	0.0034353165
articulos_de_temporada	1.625519e-01	0.0054116888	0.0042087229
tintes_y_coloracion	1.596983e-01	0.0035306605	0.0026416332
primeros_auxilios	1.634354e-01	0.0053350475	0.0041099757
insumos_medicos	1.632384e-01	0.0097828503	0.0076250449
aparato_cardiovascular	1.464234e-01	-0.0397545741	-0.0300069952
Recetario	-4.579290e-06	-0.5895728745	0.8076760569
Abastos	1.634649e-01	0.0066284339	0.0051285432
digestivos_otc	1.634118e-01	0.0056727475	0.0043991491
gripe_otc	1.630409e-01	0.0040361475	0.0030645116
cuidado_familiar_otc	1.631683e-01	0.0027678116	0.0020889462
bebidas_no_alcoholicas	1.635917e-01	0.0101551076	0.0078970639
Lencería	1.564955e-01	0.0039228987	0.0030074561
aparato_digest_y_metabol	1.597408e-01	-0.0071855228	-0.0054277906
Snacks	1.635786e-01	0.0101857800	0.0079161643

5.2 Comunicación de la Investigación

Se efectuó la comunicación de la investigación identificando las coordenadas de los componentes principales así como con sus categorías a las que están correlacionadas. Para ello se crearon las variables **individuos**, **individuos2** y **individuos3** que permitió realizar el análisis entre el componente principal uno (1) y dos (2), componente principal uno (1) y tres (3) y componente principal uno (2) y dos (3) respectivamente; ejecutando las sentencias:

- individuos= acp$x[,1:3]
- individuos= acp$x[,1:2]
- individuos= acp$x[,2:1]

Se cargó la librería **ade4** mediante la sentencia:

- library(ade4)

Se realizó el gráfico del componente principal uno (1) con respecto al componente principal dos (2) que se visualiza en la figura 42; se ejecutaron las sentencias:

- x11()
- s.corcircle(Comp_prin[,-3], sub="CP1 y CP2", possub="topright")

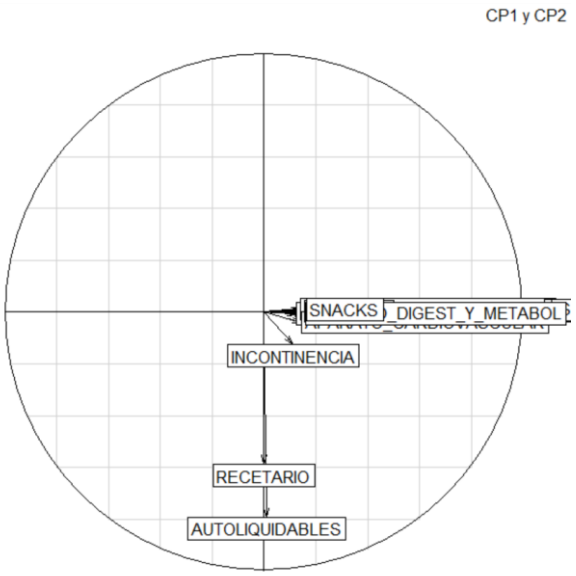

Figura 42. Datos Segmentados por Categorías de Productos de los componentes CP1 Y CP2. Fuente: Herramienta R

A continuación, se generó el gráfico de los datos por coordenadas de clientes de los componentes principales CP1 y CP2 mediante los siguientes comandos:

- x11()
- s.label(individuos[,-3], label=row.names(datos), sub="Coordenadas de los individuos", possub="topright")

Los datos visualizados en el grafico 43 corresponden a como los clientes están distribuidos en los componentes principales CP1 y CP2.

Figura 43. Datos de coordenadas de clientes de los componentes principales CP1 y CP2. Fuente: Herramienta R

Una vez que se finalizó los resultados el primer componente con respecto al segundo, se realizó el gráfico del componente principal uno (1) con respecto al componente principal tres (3) que se visualiza en la figura 44; para ello se ejecutaron las sentencias:

- x11()
- s.corcircle(Comp_prin[,-2], sub="CP1 y CP3", possub="topright")

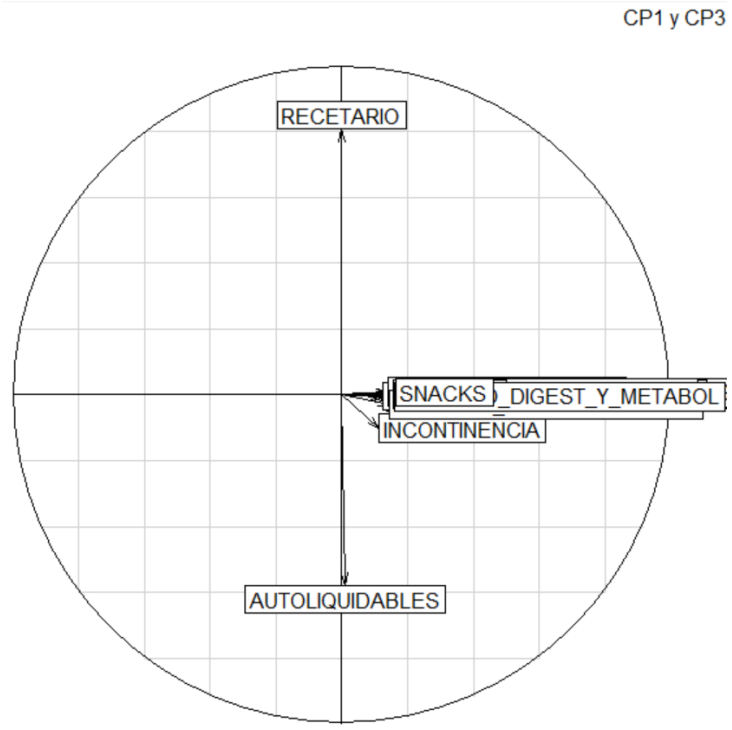

Figura 44. Datos Segmentados por Categorías de Productos de los componentes CP1 Y CP3. Fuente: Herramienta R

A continuación, se generó el gráfico de los datos por coordenadas de clientes de los componentes CP1 y CP3 mediante los siguientes comandos:

- x11()
- s.label(individuos[,-2], label=row.names(datos), sub="Coordenadas de los individuos", possub="topright")

Los datos visualizados en el gráfico 45 corresponden a como los clientes están distribuidos en los componentes principales CP1 y CP3.

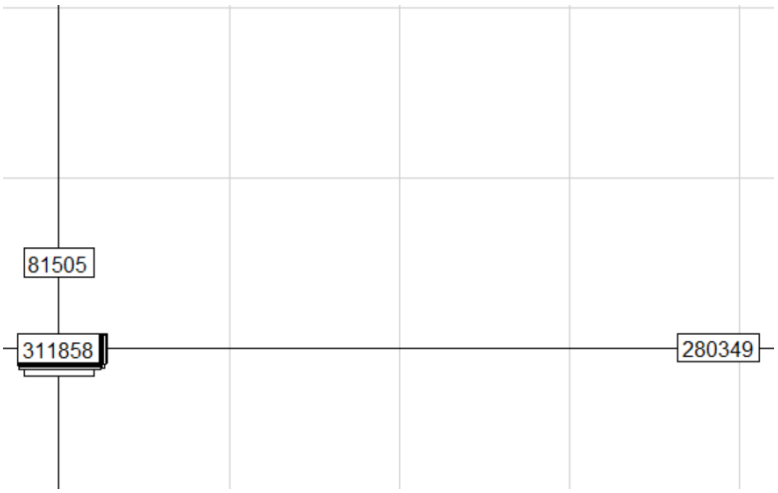

Figura 45. Datos de coordenadas de clientes de los componentes principales CP1 y CP3. Fuente: Herramienta R

Una vez que se finalizó los resultados el primer componente con respecto al segundo, se realizó el gráfico del componente principal uno (1) con respecto al componente principal tres (3) que se visualiza en la figura 46; para ello se ejecutaron las sentencias:

- x11()
- s.corcircle(Comp_prin[,-2], sub="CP1 y CP3", possub="topright")

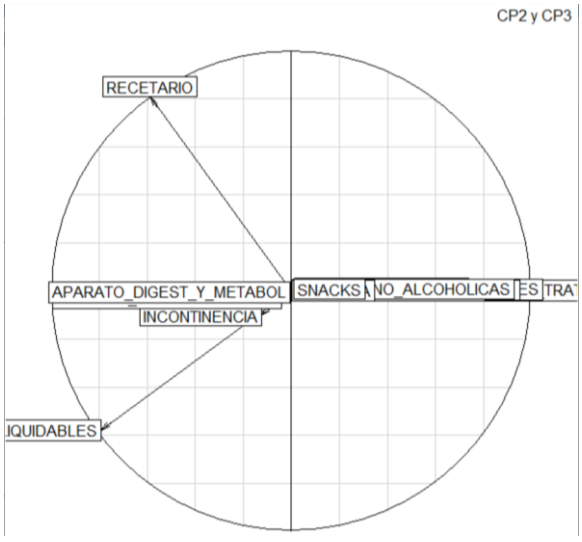

Figura 46. Datos Segmentados por Categorías de Productos de los componentes CP2 Y CP3. Fuente: Herramienta R

A continuación, se generó el gráfico de los datos por coordenadas de clientes de los componentes CP2 y CP3 mediante los siguientes comandos:

- x11()
- s.label(individuos[,-1], label=row.names(datos), sub="Coordenadas de los individuos", possub="topright")

Los datos visualizados en el grafico 47 corresponden a como los clientes están distribuidos en los componentes principales CP2 y CP3.

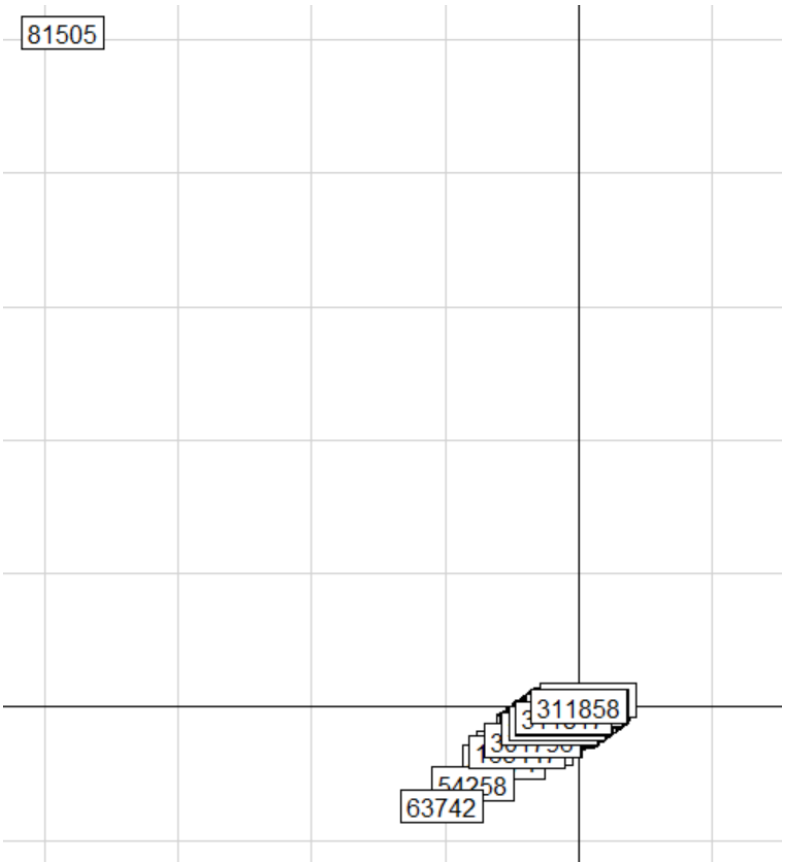

Figura 47. Datos de coordenadas de clientes de los componentes principales CP2 y CP3. Fuente: Herramienta R

CAPÍTULO VI
CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

- Al realizar la segmentación por tipos de clientes según preferencias de consumo permitirá a la organización elaborar nuevos perfiles de clientes adicionales sobre los que ya se tienen implementados como por ejemplo clientes institucionales.
- Al conocer al cliente de una manera más personalizada, el presente estudio permitirá la realizar mejoras a sistemas de software como el CRM complementando al giro del negocio.
- Al disponer la organización datos referenciales ayudará a que la toma de decisiones sea más eficiente creando un ambiente de confianza para el uso de los datos por parte de la alta gerencia.
- El uso de técnicas de segmentación como es el análisis RFM enfocado a los datos, permitirá al negocio tomar decisiones estratégicas en cuanto a la dirección que deben estar enfocadas ya sean estas de manera cross selling (cruce de categorías) o up selling (estrategias de fidelización de clientes).
- El análisis de componentes principales permitió identificar de manera aterrizada las correlaciones entre datos como también es una herramienta para la visualización de datos, mediante la reducción de la dimensionalidad de categorías del portafolio de la empresa.
- El análisis de comportamiento de consumo de clientes se enfocó en el segmento de clientes leales, puesto bajo la premisa de la empresa son los clientes que actualmente mantienen el giro de negocio.
- Se evidenció que las categorías que se acercan más hacia el círculo de correlación, tienen una mayor probabilidad de adquirir productos de esa categoría, ya que su valor monetario,

frecuencia de compra y recencia tienen un comportamiento favorable hacia la empresa de retail.

- Con el análisis de componentes principales, se podrá evidenciar que el manejo empírico de estrategias comerciales que viene efectuando la empresa de retail ya no está acorde a la época de transformación digital y manejo de información. Este análisis permitirá tener otro enfoque de como los datos de la empresa pueden ayudar a la toma de decisiones de la gerencia comercial sean más asertivas.

6.2 Recomendaciones

- Tener en cuenta que el estudio investigativo es una guía para la organización en cuanto al manejo de datos masivos e históricos.
- Generar nuevas alternativas de promociones según el tipo de cliente, que permita la creación de productos personalizados de acuerdo a sus preferencias.
- Generar mejores propuestas de agrupamiento de productos de exhibición para complementar los recorridos en tienda.
- Mejorar la eficiencia en las campañas de fidelización según el tipo de cliente, ya que se dispone de las preferencias de consumo.
- Repotenciar el área de investigación informática y estadística enfocada hacia el análisis de datos masivos de la organización.
- Proponer estudios que tengan un manejo más investigativo que empírico, que permitirá generar conocimiento dentro de la organización.

BIBLIOGRAFÍA

- Segmenting customers by transaction data with concept hierarchy (2012). Fang-Ming Hsu, Li-Pang Lu, Chun-Min Lin, www.elsevier.com/locate/eswa
- Segmentation of telecom customers based on customer value by decision tree model (2012). Shui Hua Han, Shui Xiu Lu, Stephen C.H. Leung, www.elsevier.com/locate/eswa
- Literature Review Application of data mining techniques in customer relationship management (2012). E.W.T. Ngai, Li Xiu, D.C.K. Chau, www.elsevier.com/locate/esw
- Customer data mining for lifestyle segmentation (2012). V.L. Miguéis, A.S. Camanho, João Falcão e Cunha, www.elsevier.com/locate/esw
- Wirth, R., & Hipp, J. (2000). Crisp-dm: towards a standard process modell for data mining.
- Sales forecasting using extreme learning machine with applications in fashion retailing (2008). Zhan-Li Sun, Tsan-Ming Choi, Kin-Fan Au, Yong Yu, www.elsevier.com/locate/esw
- Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Citeseer.
- Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." John Wiley and Sons, Inc. WIREs Comp Stat 2: 433–59. http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf.
- Cristina Gil Martínez. 2018. "Análisis de Componentes Principales", https://github.com/CristinaGil/Ciencia-de-Datos-R
- Smith, L. I. (2002). A tutorial on principal components analysis.

- López, Á. J. C., Caicedo, C. G., & Oviedo, P. C. J. (2017). Análisis multivariado para segmentación de clientes basada en RFM. Tecnura, 21(54), 41-51.
- Bosch, M., Montoya, R., & Inostroza, R. (2001). Estudio de los Efectos de la Introducción de un Producto de Marca Propia en una Cadena de Retail. Technical report, Departamento de Ingeniería Industrial, Universidad de Chile.
- Pardo, A., Garrido, J., Ruiz, M. Á., & San Martín, R. (2007). La interacción entre factores en el análisis de varianza: errores de interpretación. Psicothema, 19(2), 343-349.
- Castro, M. C. B., Carrión, G. A. C., & Roldán, J. L. (2007). Investigar en economía de la empresa:¿ partial least squares o modelos basados en la covarianza?. In El comportamiento de la empresa ante entornos dinámicos: XIX Congreso anual y XV Congreso Hispano Francés de AEDEM (p. 63). Asociación Española de Dirección y Economía de la Empresa (AEDEM).
- Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.
- Mardia, K. V., J. T. Kent, and J. M. Bibby (1979) Multivariate Analysis, London: Academic Press.
- Venables, W. N. and B. D. Ripley (2002) Modern Applied Statistics with S, Springer-Verlag.
- Macas, M., Lagla, L., Fuertes, W., Guerrero, G., & Toulkeridis, T. (2017, April). Data Mining model in the discovery of trends and patterns of intruder attacks on the data network as a public-sector innovation. In 2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 55-62). IEEE.

- Guayasmín, A., Fuertes, W., Campaña, M., & Toulkeridis, T. (2018, November). Formalistic Modelling Based on Pattern Recognition Applied to the Knowledge and Human Talent Sector in Ecuador. In 2018 ICAI Workshops (ICAIW) (pp. 1-6). IEEE.
- Han, M. (2008, December). Customer segmentation model based on retail consumer behavior analysis. In 2008 International Symposium on Intelligent Information Technology Application Workshops (pp. 914-917). IEEE.
- Das, G. (2015). Impact of store attributes on consumer-based retailer equity: An exploratory study of department retail stores. Journal of Fashion Marketing and Management, 19(2), 188-204.