ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN,

INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

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

TEMA:

"DETERMINACIÓN DE PATRONES PREDICTIVOS DE CONSUMO EN GRANDES VOLÚMENES DE DATOS, USANDO MÉTODOS DE DATA MINING"

AUTOR: ING. ROSERO CASA, DANIELA ELIZABETH

DIRECTOR: ING. MOLINA BUSTAMANTE, MARCO EDUARDO. PhD.

SANGOLQUI, 2019

VICERRECTORADO DE INVESTIGACIÓN,
INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA
CENTRO DE POSGRADO

CERTIFICACIÓN

Certifico que el trabajo de titulación, ***"DETERMINACIÓN DE PATRONES PREDICTIVOS DE CONSUMO EN GRANDES VOLÚMENES DE DATOS, USANDO MÉTODOS DE DATA MINING"*** fue realizado por la señora **Rosero Casa, Daniela Elizabeth** 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 Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolqui, 16 de Julio del 2019

Firma:

Dr. Marco Eduardo Molina Bustamante
CC: 1705613014

VICERRECTORADO DE INVESTIGACIÓN,
INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA
CENTRO DE POSGRADO

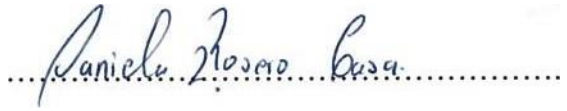
AUTORÍA DE RESPONSABILIDAD

Yo, ***Rosero Casa, Daniela Elizabeth*** con cédula de ciudadanía nº: 1716912892, declaró que el contenido, ideas y criterios del trabajo de titulación: ***"DETERMINACIÓN DE PATRONES PREDICTIVOS DE CONSUMO EN GRANDES VOLÚMENES DE DATOS, USANDO MÉTODOS DE DATA MINING"*** es de mi autoría y responsabilidad , cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolqui, 16 de Julio del 2019

Firma:



Ing. Daniela Elizabeth Rosero Casa

CC: 1716912892

VICERRECTORADO DE INVESTIGACIÓN,

INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

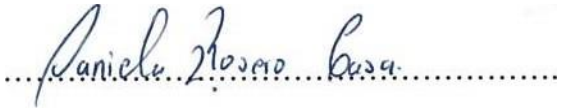
CENTRO DE POSGRADO

AUTORIZACIÓN

*Yo, **Rosero Casa, Daniela Elizabeth**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"DETERMINACIÓN DE PATRONES PREDICTIVOS DE CONSUMO EN GRANDES VOLÚMENES DE DATOS, USANDO MÉTODOS DE DATA MINING"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.*

Sangolqui, 16 de Julio del 2019

Firma:



Ing. Daniela Elizabeth Rosero Casa

CC: 1716912892

DEDICATORIA

El presente trabajo está dedicado a mi esposo Fernando, por inspirarme a cumplir mis sueños, por ser un gran padre, por permitirme ser madre de Matías y Martín y porque siempre estarás junto a mí.

Daniela

AGRADECIMIENTO

A Dios, por permitirme despertar cada día,

A mis padres, Santiago y Blanca por enseñarme valores que me han permitido alcanzar mis metas,

A mis hermanas, Gaby y Jaque, por estar presentes en cada alegría y tristeza,

A Matías, por siempre estar a mi lado, eres un hijo genial y estoy muy segura que llegarás a cumplir tus sueños,

A mi pequeño Martín, por regalarme los besos y abrazos que necesito para recargar energías, porque eres un hijo extraordinario y sé que siempre conseguirás los que te propones,

Al Ing. Marco Molina PhD, por guiarme y acompañarme en mi trabajo con todo su conocimiento,

A mis amigos de maestría, Taty y Edy, porque trabajamos juntos para cumplir nuestros objetivos.

Daniela

ÍNDICE DE CONTENIDO

CERTIFICACIÓN i

AUTORÍA DE RESPONSABILIDAD ii

AUTORIZACIÓN iii

DEDICATORIA iv

AGRADECIMIENTO v

ÍNDICE DE CONTENIDO vi

ÍNDICE DE TABLAS xi

ÍNDICE DE ILUSTRACIONES xiii

RESUMEN xiv

ABSTRACT xv

CAPITULO I

INTRODUCCION E INFORMACION GENERAL

1.1 Antecedentes 1

1.2 Contexto del Problema 1

1.3 Planteamiento del Problema 2

1.4 Justificación e Importancia 3

1.5 Objetivos 3

1.5.1 Objetivo General 3

1.5.2 Objetivos Específicos 3

1.6 Hipótesis 4

CAPITULO II

ESTADO DEL ARTE Y MARCO TEORICO

2.1 Estado del Arte 5

2.1.1 Definición de preguntas de investigación 6

2.1.2 Estrategia de búsqueda para los estudios primarios 6

2.1.3 Selección de documentos para su inclusión y exclusión (documentos relevantes) 6

2.1.4 Palabras clave de los resúmenes (Esquema de clasificación) 6

2.1.5 Extracción de datos y mapeo de estudios (mapa sistemático) 6

2.2 Aplicación del Estado del Arte 7

2.2.1 Definición de preguntas de investigación 7

2.2.2 Estrategia de búsqueda 7

2.2.3 Criterios de inclusión y exclusión 14

2.2.4 Proceso de selección 14

2.2.5 Extracción de Datos 15

2.2.6 Conclusión del Estado del Arte 19

2.3 Marco Teórico 20

2.3.1 Marketing 20

2.3.2 Tipos de marketing 20

2.3.3 Marketing Directo utilizando Minería de Datos 21

2.3.4 Inteligencia de Negocios 22

2.3.4.1 Data mart 23

2.3.4.2 Data Warehouse 23

2.3.4.3 Metodología Kimball 24

2.3.4.4 Metodología Bill Inmon 25

2.3.4.5 Exploración de datos 26

2.3.4.6 Balanced Scored Card 27

2.3.5 Minería de Datos 27

2.3.5.1 Técnicas de Minería de datos 27

2.3.5.2 Ciclo de Vida en Minería de Datos 37

CAPITULO III

METODOLOGIAS DE MINERIA DE DATOS

3.1 KDD 38

3.2 CRISP – DM 39

3.3 SEMMA 41

3.4 Comparación cualitativa de metodologías minería de datos 42

3.4.1 Metodología seleccionada 43

3.5 Herramientas de minería de datos....43

3.5.1 Selección de la herramienta de datos....45

CAPITULO IV

PROPUESTA DEL MODELO

4.1 Entendimiento del Negocio....47

4.1.2 Evaluación de la Situación....51

4.2 Comprensión de los datos....53

4.2.1 Fuentes de Información....53

4.2.2 Definición de variables....53

4.2.3 Exploración de variables....61

4.2.4 Verificación de calidad de los datos....68

4.3 Preparación de los datos....70

4.3.1 Selección de los datos....70

4.3.2 Limpieza y construcción de los datos....71

4.3.3 Integración y Formateo de Datos....72

4.4 Modelado....73

4.4.1 Selección de la técnica de modelado....73

4.4.2 Construcción del modelado....73

4.5 Evaluación....75

4.6 Resultados y análisis 78

CAPITULO V

CONCLUSIONES Y TRABAJOS FUTURO

5.1 Conclusiones 79

5.2 Trabajos Futuros 80

BIBLIOGRAFIA 81

ÍNDICE DE TABLAS

Tabla 1: *Documentos iniciales* 8

Tabla 2: *Identificación de palabras claves* 10

Tabla 3: *Cadenas de búsqueda de información experimental* 13

Tabla 4: *Estudios Candidatos* 15

Tabla 5: *Selección de estudios primarios* 15

Tabla 6: *Estudios Primarios* 16

Tabla 7: *Ponderación de Metodologías* 43

Tabla 8: *Selección de Herramienta* 46

Tabla 9*: Bines* 48

Tabla 10: *Solicitudes de tarjetas* 49

Tabla 11: *Variable – Edad* 53

Tabla 12: *Variable – NroCargas* 53

Tabla 13: *Variable – Género* 54

Tabla 14: *Variable – Estado Civil* 54

Tabla 15: *Variable – Nivel Estudios* 55

Tabla 16: *Variable - Situación Laboral* 55

Tabla 17: *Variable - Tipo Vivienda* 56

Tabla 18: *Variable - Sub Marca* 57

Tabla 19: *Variable – Marca* 58

Tabla 20: *Variable - Monto Transacción* 58

Tabla 21: *Variable - Número Transacción* 59

Tabla 22: *Variable - Línea de Negocio* 59

Tabla 23: *Calidad de los Datos* 69

Tabla 24: *Selección de las submarcas para el modelo* 71

ÍNDICE DE ILUSTRACIONES

Figura 1: Fases del SMS 5

Figura 2: Modelo Kimball 24

Figura 3: Modelo Bill Inmon 26

Figura 4: Modelo Predictivo 28

Figura 5: Árbol de decisión 29

Figura 6: Red Neuronal 33

Figura 7: Etapas de KDD 38

Figura 8: Fases de CRISP-DM 40

Figura 9: Fases de SEMMA 41

Figura 10: Proceso del funcionamiento de Tarjetas de Crédito 50

Figura 11: Proceso Masivo de Publicidad 52

Figura 12: Exploración de género 61

Figura 13: Exploración de estudios 62

Figura 14: Exploración de Estado Civil 63

Figura 15: Exploración de descripción laboral 64

Figura 16: Exploración de tipo de Vivienda 65

Figura 17: Exploración de marca 66

Figura 18: Exploración de SubMarca 67

Figura 19: Exploración de línea de negocio 68

Figura 20: Fechas de los datos 70

Figura 21: DataSet del Modelo 72

RESUMEN

El presente estudio tiene como objetivo generar una propuesta de patrones de consumo que permitan mejorar la entrega de fuentes de información para una publicidad focalizada a las necesidades de los clientes, mejorar el procesamiento de grandes volúmenes de datos y mejorar los tiempos de ejecución del proceso, la propuesta será materializada a través de la consolidación de las transacciones de los consumos, mediante un gestor de datos, realizar la limpieza y depuración de variables para aplicar técnicas de data mining, el resultado será disponibilizado en un repositorio que facilite el análisis y acceso para la obtención de los resultados. Para lograr el propósito se seguirá la **metodología de investigación experimental**, la cual consta de cinco fases, la fase **planteamiento de un problema de conocimiento** donde se realiza la elección del problema, continua con la **formulación de hipótesis** donde es la anticipación de un resultado, luego pasa a la **realización de un diseño adecuado a la hipótesis** donde el diseño refleja el plan o esquema de trabajo del investigador, es su organización formal. El diseño incluye diversos subprocesos, describe con detalle qué se debe hacer y cómo realizarlo, continúa con la fase de **recogida y análisis de datos**, y finalmente llega a la fase de **elaboración de conclusiones**. Al finalizar el estudio se espera tener un proceso documentado, que cumpla con el objetivo propuesto y que aporte a la gestión de Marketing en la mejora de su publicidad asertiva focalizada a las necesidades de cliente.

PALABRAS CLAVE:

- **DATA MINIG**
- **MARKETING**
- **PATRONES DE CONSUMO**

ABSTRACT

This study aims to generate a proposal of consumption patterns that allow improving the delivery of information sources for targeted advertising to the needs of customers, improve the processing of large volumes of data and improve the execution times of the process, The proposal will be materialized through the consolidation of consumption transactions, through a data manager, perform the cleaning and debugging of variables to apply data mining techniques, the result will be available in a repository that facilitates the analysis and access to Obtaining the results. To achieve the purpose, the experimental research methodology will be followed, which consists of five phases, the approach phase of a knowledge problem where the choice of the problem is made, continues with the formulation of hypotheses where the anticipation of a result is, then It goes on to the realization of a design adapted to the hypothesis where the design reflects the plan or scheme of work of the researcher, is his formal organization. The design includes various threads, describes in detail what should be done and how to do it, continues with the phase of data collection and analysis, and finally reaches the stage of drawing conclusions. At the end of the study, it is expected to have a documented process that meets the proposed objective and that contributes to Marketing management in the improvement of its assertive advertising focused on customer needs.

KEY WORD:

- **DATA MINIG**
- **MARKETING**
- **CONSUMPTION PATTERNS**

CAPITULO I

INTRODUCCION E INFORMACION GENERAL

1.1 Antecedentes

El marketing tradicional no es suficiente para que se puedan vender los productos y/o servicios, debido al cambio generacional y el uso progresivo de internet. Las campañas publicitarias pueden ser más eficaces si se las enfoca a los clientes interesados en comprar los productos, con lo cual se garantizaría el establecimiento de lazos de fidelidad.

Netflix, Amazon, eBay y Wish, son ejemplos de empresas que buscan establecer vínculos con los clientes, esto lo realizan en base a patrones de búsqueda, transacciones anteriores, compras de otros clientes similares y la información recopilada de las aplicaciones en tiempo real; aprovechándose del hecho de que los clientes, en su mayoría, realizan sus compras haciendo uso de tarjetas de crédito u otros medios de pago electrónico.

Tales entidades financieras actualmente tienen gran cantidad de datos, que seguirán creciendo a medida que se impulse el uso de medios de pago electrónico, pero no se aprovecha de su potencial, pues la publicidad que realizan estas entidades para que el cliente use su tarjeta no está siendo dirigida a las necesidades de los clientes. Es necesario encontrar métodos que ayuden a las entidades financieras a aprovechar los datos proporcionados históricamente por sus clientes en la mejora de las relaciones cliente – empresa.

1.2 Contexto del Problema

La Superintendencia de Bancos indica que en el Ecuador los consumos realizados por medio de tarjetas de crédito representan el 22.80% de la cartera de consumo. Cada entidad financiera

regulada por la Superintendencia de Bancos, tiene el área de Marketing que se encarga de la publicidad, consistente en dar a conocer a sus clientes las ofertas de los diferentes establecimientos. El envío masivo de publicidad puede causar malestar en el cliente, pues al tratarse de publicidad no deseada, genera rechazo por parte de este, pudiendo llegar a tomar la decisión de cerrar su medio electrónico de pago.

El encargado de la selección de la muestra para envío de publicidad es el analista de marketing, quien hace uso de métodos que privilegian la subjetividad, puesto que no existe documentación ni sistema automatizado que respalde las decisiones del analista.

1.3 Planteamiento del Problema

En la actualidad, el analista de marketing hace uso de herramientas de ofimática para el envío masivo de publicidad, en base al producto entregado sin contar con datos de gustos y preferencias de los clientes. La decisión de enviar dicha publicidad se la toma desde una perspectiva intuitiva o subjetiva que puede depender de varios factores no necesariamente técnicos.

Las consecuencias de las decisiones del analista no pueden ser peores, pues en primer lugar se provoca la deserción de clientes; en segundo lugar, se desperdician recursos económicos y en tercer lugar se pierde la oportunidad de afianzar la fidelidad de los clientes.

Por lo tanto, es necesario que se realicen campañas de marketing dirigidas a los clientes, tomando en cuenta sus preferencias a la hora de realizar sus inversiones. El resultado de esta práctica deberá redundar en un incremento de la fidelidad de los clientes y seguramente, en el incremento del uso de las tarjetas de crédito.

1.4 Justificación e Importancia

El proyecto de tesis se justifica por su aporte al área de marketing, para que el proceso de envío de publicidad no se base únicamente en la subjetividad del analista, sino en el procesamiento de grandes cantidades de datos históricos, acumulados por la institución, con lo cual se logra optimizar tiempo y recursos. Adicionalmente, el marketing directo, basado en patrones de consumo, conllevará la aceptación por parte de los clientes e incrementará la fidelidad de los mismos.

Para la institución financiera el presente proyecto de tesis tiene gran importancia, pues se trata de optimizar recursos (tiempo, dinero) en el proceso de envío de publicidad directa, al tiempo que se incrementará el uso de los medios electrónicos de pago y disminuirá la deserción de los clientes.

1.5 Objetivos

1.5.1 Objetivo General

Generar una propuesta de patrones de consumo para establecer una base de información para el proceso de envío de publicidad, utilizando herramientas para manejo de grandes volúmenes de datos.

1.5.2 Objetivos Específicos

OE1: Realizar una revisión literaria de metodología, de técnicas de minería de datos, y herramientas analíticas para grandes volúmenes de datos que permitan encontrar patrones predictivos de consumo.

OE2: Implementar un modelo para determinar los patrones de consumo, que permitan construir la base de información para el envío de una publicidad asertiva.

OE3: Evaluar el modelo de patrones de consumo propuesto, determinando el nivel de confianza y comparando la efectividad de la publicidad antes utilizada por el analista y después de tener identificado los patrones de consumo.

1.6 Hipótesis

La utilización de técnicas de minería de datos para identificar patrones de consumo, permitirán construir una base de información para el envío de publicidad focalizada, de acuerdo a las preferencias del cliente, logrando el incremento de la fidelidad y de consumos con medio electrónico de pago.

Señalamiento de Variables. -Se usarán variables tanto cualitativas como cuantitativas, en los algoritmos de descubrimiento de patrones. Para la validación cuantitativa se medirá el tiempo de implementación y para la calidad estadística se realizará análisis estadísticos que miden la varianza generados por el modelo.

Variables dependientes: Base de información para el envío de publicidad asertiva de acuerdo a las preferencias del cliente. Incremento de fidelidad y de consumos con medio electrónico de pago.

Variable Independientes: Patrones de Consumo

CAPITULO II

ESTADO DEL ARTE Y MARCO TEORICO

2.1 Estado del Arte

La técnica del SMS (Systematic Mapping Studies) según sus siglas en inglés, tiene como objetivo, identificar los artículos que sirvan de apoyo para: resolver una problemática, responder preguntas de investigación, o estudiar un área temática. La técnica sigue ciertos pasos, que llegando a cumplirlos nos ayuda a obtener el discernimiento adecuado de los artículos, eliminando el mayor sesgo posible. Por ello, esta técnica es considerada como la más fiable.[2]

Se opta por el uso del SMS debido a la minuciosidad que se requiere para el desarrollo de cada una de las etapas, lo cual es necesario para mitigar la aparición de una investigación que no sea la adecuada y que tome mucho más tiempo de lo planificado.

La figura 1 muestra los pasos del SMS:

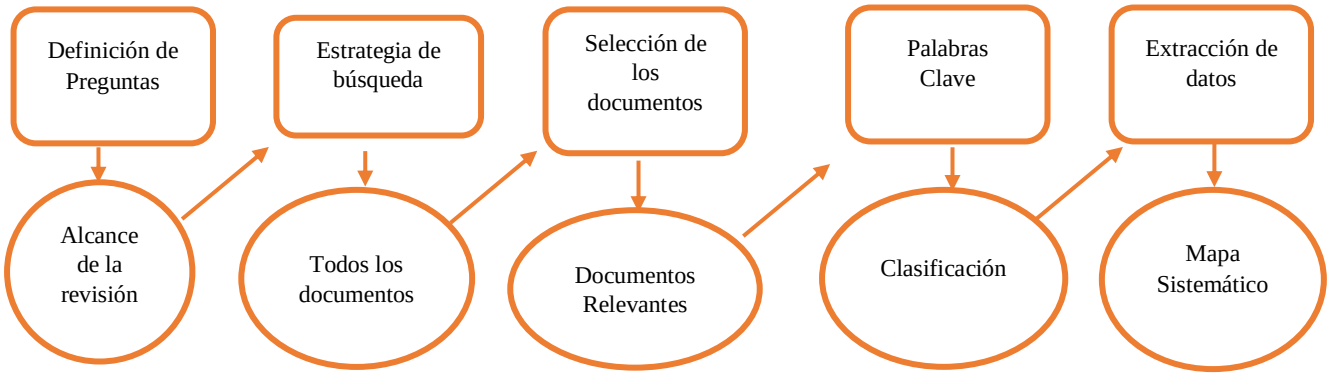


Figura 1: Fases del SMS

Fuente: Kai Petersen, Robert Feldt1, Shahid Mujtaba, Michael Mattsson

2.1.1 Definición de preguntas de investigación

Para definir las preguntas de investigación, es necesario; investigar los sitios donde los documentos han sido publicados, tener una lectura del resumen de cada documento y poder ir estructurando preguntas adecuadas para nuestra investigación. [2]

2.1.2 Estrategia de búsqueda para los estudios primarios

Los estudios primarios se identifican utilizando cadenas de búsqueda en bases de datos científicas o navegando manualmente a través de actas de congresos relevantes o publicaciones de revistas.

2.1.3 Selección de documentos para su inclusión y exclusión (documentos relevantes)

Los criterios se utilizan para discernir los documentos y poder utilizar los que son relevantes para responder a las preguntas de investigación.

2.1.4 Palabras clave de los resúmenes (Esquema de clasificación)

En la revisión de los documentos, se buscan palabras clave y conceptos que reflejen la contribución para la investigación, cuando se culmina el trabajo descrito el conjunto de palabras clave de diferentes artículos se combinan para obtener un conjunto de categorías representativo.

2.1.5 Extracción de datos y mapeo de estudios (mapa sistemático)

Los artículos relevantes se clasifican, y se colocan en una tabla explicando por qué debe ser considerado. [2]

2.2 Aplicación del Estado del Arte

2.2.1 Definición de preguntas de investigación

- OE1-RQ1 ¿Qué metodologías existen para encontrar patrones de consumo?
- OE1-RQ2 ¿Qué técnicas de minería de datos existen para encontrar patrones de consumos?
- OE1-RQ3 ¿Cuáles son las herramientas analíticas adecuadas para un modelo de patrones de consumo?
- OE2-RQ1: ¿Cuáles son los algoritmos de Data mining para determinar patrones de consumos, aplicado a un marketing digital?
- OE2-RQ2: ¿Cómo seleccionar el conjunto de datos?
- OE2-RQ3: ¿Cómo identificar los datos ocultos y sin explotar para su utilización en el modelo?
- OE3-RQ1: ¿Cómo determinar el asertividad del modelo?

2.2.2 Estrategia de búsqueda

Para llevar a cabo la estrategia de búsqueda se utilizaron los siguientes repositorios digitales:

- GOOGLE ACADEMICO - https://scholar.google.com.ec/
- SCIENCEDIRECT - https://www.sciencedirect.com/

Al realizar la búsqueda en los repositorios, se identificaron 30 documentos como se indica en la tabla 1:

Tabla 1:
Documentos iniciales

CÓDIGO	TÍTULO	REPOSITORIO DIGITAL
EC1	Mining Location-Based Service Data for Feature Construction in Retail Store Recommendation	SCIENCEDIRECT
EC2	Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes	GOOGLE ACADEMICO
EC3	Digitalisation and Big Data Mining in Banking	GOOGLE ACADEMICO
EC4	Data Mining for Supermarket Sale Analysis Using Association Rule	GOOGLE ACADEMICO
EC5	Comparative Study of Effective Performance of Association Rule Mining in Different Databases	GOOGLE ACADEMICO
EC6	Data Mining in Supermarket: A Survey	GOOGLE ACADEMICO
EC7	Applications of data mining in retail business	SCIENCEDIRECT
EC8	Application of Data Mining in Banking Sector	GOOGLE ACADEMICO
EC9	Knowledge management and data mining for marketing	GOOGLE ACADEMICO
EC10	Data mining techniques for customer relationship management	SCIENCEDIRECT
EC11	A data mining framework for targeted category promotions	GOOGLE ACADEMICO
EC12	Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision Making	SCIENCEDIRECT
EC13	An Online Mall CRM Model Based on Data Mining	GOOGLE ACADEMICO
EC14	The Impact of Knowledge Management and Data Mining on CRM in the Service Industry	GOOGLE ACADEMICO
EC15	Data mining application in credit card fraud detection system	GOOGLE ACADEMICO

CONTINÚA →

EC16	CRISP-DM: Towards a Standard Process Model for Data Mining	GOOGLE ACADEMICO
EC17	KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW	GOOGLE ACADEMICO
EC18	Knowledge Discovery and Data Mining: Towards a Unifying Framework	GOOGLE ACADEMICO
EC19	A Big Data Analytical Architecture for the Asset Management	SCIENCEDIRECT
EC20	Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan	SCIENCEDIRECT
EC21	A robust Datawarehouse as a requirement to the increasing quantity and complexity of travel survey data	SCIENCEDIRECT
EC22	A Comprehensive Method for Data Warehouse Design	GOOGLE ACADEMICO
EC23	A Decision Tree Analysis on J48 Algorithm for Data Mining	GOOGLE ACADEMICO
EC24	Non-sequential partitioning approaches to decision tree classifier	SCIENCEDIRECT
EC25	Linear regression model using Bayesian approach for energy performance of residential building	SCIENCEDIRECT
EC26	Prediction of kidney disease stages using data mining algorithms	SCIENCEDIRECT
EC27	A survey of neural network-based cancer prediction models from microarray data	SCIENCEDIRECT
EC28	Descriptive Modeling of Social Networks	SCIENCEDIRECT
EC29	Density-based clustering methods for unsupervised separation of partial discharge sources	SCIENCEDIRECT
EC30	Exploring Data Sets for Clusters and Validating Single Clusters	SCIENCEDIRECT

Se realiza la lectura de; título, resumen y palabras claves, que nos permita discernir sobre la utilización en nuestra investigación, luego de la revisión se estableció que 18 documentos son los primarios como se detalla en la tabla 2:

Tabla 2:
Identificación de palabras claves

CÓDIGO	TÍTULO	PALABRAS CLAVES	PRIMARIOS
EC1	Mining Location-Based Service Data for Feature Construction in Retail Store Recommendation	Urban mining Spatial and temporal data mining Location-based service Retail store recommendation	SI
EC2	Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes	Big data initiatives, Retail stores, Emerging service processes, Technology enablers, Privacy concerns, Shopping outcomes	SI
EC3	Digitalisation and Big Data Mining in Banking	big data analytics; data mining; banking; survey	SI
EC4	Data Mining for Supermarket Sale Analysis Using Association Rule	Data mining, Associations, supermarket	SI
EC5	Comparative Study of Effective Performance of Association Rule Mining in Different Databases	Data Mining, Association Rule Mining, Spatial Data Mining, RDBMS, Medical Database, Large Database, Distributed Database.	SI
EC6	Data Mining in Supermarket: A Survey	Data Mining, Supermarket, Association Rule, Cluster Analysis	SI
EC7	Applications of data mining in retail business	Data mining, Transaction databases, Organizational aspects, Application software, Educational institutions, Raw materials, Software algorithms, Merchandise, Throughput, Credit cards	SI
EC8	Application of Data Mining in Banking Sector	Data Mining, Banking Sector, Risk Management, CRM	SI

CONTINÚA →

EC9	Knowledge management and data mining for marketing	Data mining, Knowledge management Marketing decision support, Customer relationship management	SI
EC10	Data mining techniques for customer relationship management	Customer relationship management (CRM), Relationship marketing, Data mining, Neural networks Chi-square automated interaction detection (CHAID)Privacy rights	SI
EC11	A data mining framework for targeted category promotions	Cross-category purchases Target marketing Customized coupons Clustering Association rule mining	SI
EC12	Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision Making	Marketing Decision Support, Business Value, Profit-Analytics, Machine Learning	SI
EC13	An Online Mall CRM Model Based on Data Mining	Data mining CRM Online Mall	SI
EC14	The Impact of Knowledge Management and Data Mining on CRM in the Service Industry	Customer relationship management Relationship marketing Knowledge management Data mining	SI
EC15	Data mining application in credit card fraud detection system	Neural network, data mining, Clusters	SI
EC16	CRISP-DM: Towards a Standard Process Model for Data Mining	CRISP-DM (Cross Industry Standard Process for Data Mining)	SI
EC17	KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW	Data Mining Standards, Knowledge Discovery in Databases, Data Mining.	SI
EC18	Knowledge Discovery and Data Mining: Towards a Unifying Framework	Data mining, kdd	SI
EC19	A Big Data Analytical Architecture for the Asset Management	Asset Management Big data Big data analytics Data mining	NO

CONTINÚA →

EC20	Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan	Data mining Multiple dimensional constraints Ant colony system Apriori	NO
EC21	A robust Datawarehouse as a requirement to the increasing quantity and complexity of travel survey data	travel survey Datawarehouse fact dimension Kimball Origin-Destination software household-based individual-based person-based highlights animated maps trip	NO
EC22	A Comprehensive Method for Data Warehouse Design	data warehouse, multidimensional modeling, design methods, UML	NO
EC23	A Decision Tree Analysis on J48 Algorithm for Data Mining	Data Mining; Classification Techniques; J48; Decision Trees; Univariate algorithm; Multivariate algorithm; Pruning	NO
EC24	Non-sequential partitioning approaches to decision tree classifier	Decision tree Correlation Ferrer diagram Bell triangle Partitioning	NO
EC25	Linear regression model using Bayesian approach for energy performance of residential building	Linear regression bayesian frequentist	NO
EC26	Prediction of kidney disease stages using data mining algorithms	Prediction of kidney disease stages Data mining techniques Probabilistic neural networks Multilayer perceptron Support vector machine Radial basis function	NO
EC27	A survey of neural network-based cancer prediction models from microarray data	Cancer prediction models Neural networks Classification Clustering Filtering	NO

CONTINÚA →

EC28	Descriptive Modeling of Social Networks	complex network data mining social network mining descriptive modeling clustering searching for patterns	NO
EC29	Density-based clustering methods for unsupervised separation of partial discharge sources	Partial discharge DPC Clustering Spatial density	NO
EC30	Exploring Data Sets for Clusters and Validating Single Clusters	Cluster analysis cluster visualisation	NO

La identificación de los documentos primarios, son los que nos ayudarán a construir: las cadenas de búsqueda de información experimental y las cadenas de búsqueda de herramientas de soporte. A continuación, en la tabla 3, se presenta los resultados de la cadena de búsqueda de información experimental, donde se muestra que la cadena adecuada es la CD1 ya que nos devuelve el mayor número de documentos primarios, como se indica en la tabla a continuación:

Tabla 3:
Cadenas de búsqueda de información experimental

NRO	CADENA	ESTUDIOS	ESTUDIOS DEL GRUPO DE CONTROL
CD1	("KDD") AND ("Data mining") AND (("Customer Relationship management") OR ("CRM")) AND ("Decision support") OR ("Association rule") or ("marketing") OR ("Relationship marketing")	202	EC3, EC9, EC10, EC11, EC12, EC14, EC17, EC18
CD2	("Data mining") AND (("Customer Relationship management") OR ("CRM")) AND ("Decision support") AND ("Association rule")	236	EC9, EC15,

CONTINÚA →

CD3	("Data mining") AND (("Customer Relationship management") OR "CRM") AND ("Decision support")	1960	EC9, EC10,
CD4	("Data mining") AND (("Customer Relationship management") OR "CRM") AND ("TARGET")	3630	EC10

2.2.3 Criterios de inclusión y exclusión

Criterios de Inclusión

- El artículo explica técnicas sobre la minería de datos.
- El artículo expone el uso de gran cantidad de datos y como la minería de datos es aplicada al marketing.
- El artículo analiza los comportamientos de los clientes, que faciliten el análisis predictivo.

Criterios de Exclusión

- Artículos enfocados al análisis de pequeños volúmenes de datos.
- Artículos que presenten propuestas para la realización de análisis descriptivo de datos.
- Artículos que solo traten datos estructurados.
- Artículos publicados antes del 2010 no son considerados debido que la transformación tecnológica avanza constantemente.

2.2.4 Proceso de selección

De acuerdo al criterio de inclusión y exclusión, se realiza la búsqueda en los diferentes repositorios digitales, como una revisión preliminar de los resultados nos permitió constatar que las publicaciones relevantes para nuestro estudio se encontraban concentradas en los primeros puestos de la búsqueda. Las cadenas de búsqueda se conformaron a partir de los criterios de

inclusión generando varias posibilidades es así que el piloto y afinamiento se lo realizo en GOOGLE ACADEMICO y SCIENCEDIRECT. La ejecución de la búsqueda en estos motores específicos arrojó los siguientes resultados que se detallan en la tabla 4:

Tabla 4:

Estudios Candidatos

Repositorio	Nro Estudios con la Cadena	Nro Estudios seleccionados	Nro Estudios descartados	Porcentaje Acuerdo	Porcentaje Desacuerdo
GOOGLE ACADEMICO	202.00	16	186	92.08	7.92
SCIENCEDIRECT	185.00	14	171	92.43	7.57
TOTAL	**387.00**	**30.00**	**357**	**92.25**	**7.75**

Continuando con el SMS se realizó el análisis de los estudios primarios (ver tabla 5):

Tabla 5:

Selección de estudios primarios

Estudios	Cantidad	Porcentaje
Seleccionados	30	100
Descartados	12	40.00
Primarios	18	60.00
Sin Analizar	0	0

2.2.5 Extracción de Datos

Realizado el análisis de las cadenas de búsqueda, se procede a verificar cuál de ellas retorno un número de resultados manejable y en términos de criterios de inclusión más representativa. (ver tabla 6)

Tabla 6:
Estudios Primarios

Código	Título
EP1	Digitalisation and Big Data Mining in Banking
EP2	Knowledge management and data mining for marketing
EP3	Data mining techniques for customer relationship management
EP4	A data mining framework for targeted category promotions
EP5	Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision Making
EP6	The Impact of Knowledge Management and Data Mining on CRM in the Service Industry
EP7	KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW
EP8	Knowledge Discovery and Data Mining: Towards a Unifying Framework
EP9	Knowledge management and data mining for marketing
EP10	Data mining techniques for customer relationship management
EP11	A data mining framework for targeted category promotions
EP12	Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision Making
EP13	An Online Mall CRM Model Based on Data Mining
EP14	The Impact of Knowledge Management and Data Mining on CRM in the Service Industry
EP15	Data Mining Techniques for Customer Relationship Management
EP16	CRISP-DM: Towards a Standard Process Model for Data Mining
EP17	KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW
EP18	Knowledge Discovery and Data Mining: Towards a Unifying Framework

A continuación, se presentan los resultados de las preguntas realizadas en un inicio, las cuales apalancan esta investigación.

OE1-RQ1 ¿Qué metodologías existen para encontrar patrones de consumo?

La pregunta OE1-RQ1 encuentran su respuesta en los estudios EP7, EP8, EP16, EP17 en los cuales se detalla sobre SEMMA, CRISP-DM y KDD que tienen el objetivo de guiar la implementación de aplicaciones de minería de datos, y realizan una compresión de las similitudes entre ellos.

Los artículos describen los elementos finales entre la recopilación de datos, el descubrimiento de conocimientos y otros campos relacionados, definen el proceso de KDD y los algoritmos básicos de extracción de datos.

OE1-RQ2 ¿Qué técnicas de minería de datos existen para encontrar patrones de consumos?

En el artículo EP4 Y EP6 explica sobre el uso de la cesta de mercado y los algoritmos que pueden ser utilizados para encontrar patrones de consumos, se explica que el descubrimiento de reglas interesantes puede aportar a distintos escenarios para el desarrollo de nuevos productos, identificación de fraude e identificar el potencial de ventas cruzadas de artículos de productos.

OE1-RQ3 ¿Cuáles son las herramientas analíticas adecuadas para un modelo de patrones de consumo?

El artículo EP3, EP10 Y EP11, nos indiaca sobre los avances en las tecnologías: como el almacenamiento de datos, la extracción de datos y el software de gestión de campañas indica sobre la existencia de varias técnicas entre el software de minería de datos, cada una con sus propias ventajas y desafíos para diferentes tipos de aplicaciones. Se encuentra información sobre las redes

neuronales y la detección de interacción automatizada de chi-cuadrado (CHAID), indica que, si bien los diferentes enfoques abundan en el ámbito de la minería de datos, el uso de algún tipo de minería de datos es necesario para lograr los objetivos de la filosofía actual de gestión de relaciones con los clientes.

OE2-RQ1: ¿Cuáles son los algoritmos de Data mining para determinar patrones de consumos, aplicado a un marketing digital?

La pregunta OE2-RQ1 encuentra su respuesta en los estudios EP2, EP13 Y EP14 donde indica que parte de la información de mercadotecnia útil sobre las características del cliente y sus patrones de compra están en gran parte ocultos y sin explotar indica que actualmente que las relaciones con los clientes hacen que la función de marketing sea un área de aplicación ideal para beneficiarse enormemente del uso de herramientas de minería de datos para el soporte de decisiones. Propone una metodología sistemática que utiliza la extracción de datos y técnicas de gestión de conocimiento.

OE2-RQ2: ¿Cómo seleccionar el conjunto de datos?

La pregunta OE2-RQ2 tiene su guía en el estudio EP5 donde indica que los mensajes de marketing son más efectivos si llegan a los clientes correctos, propone una selección de conjuntos preocupada por las ganancias, un marco de modelado que integra los principios de aprendizaje estadístico y los objetivos de negocios en la forma de la maximización de las ganancias de la campaña.

OE2-RQ3: ¿Cómo identificar los datos ocultos y sin explotar para su utilización en el modelo?

La pregunta OE2-RQ3 puede encontrar información en el artículo EP2, donde habla sobre la proliferación de sistemas y tecnología de información, provocando que las empresas tengan cada vez más la capacidad de acumular enormes cantidades de datos de clientes en grandes bases de datos generando información de mercadotecnia útil sobre las características del cliente y sus patrones de compra que están en gran parte ocultos y sin explotar, propone una metodología sistemática que utiliza la extracción de datos y técnicas de gestión de conocimiento para administrar el conocimiento y apoyar las decisiones de marketing.

OE3-RQ1: ¿Cómo determinar el asertividad del modelo?

La pregunta OE3-RQ1 puede encontrar información en el artículo EP4, donde para demostrar el rendimiento del modelo propuesto, examina las transacciones de un programa de lealtad del mundo real de un importante minorista de comestibles. Realizando un análisis basado en escenarios utilizando la capacidad de respuesta de la promoción y la experiencia previa de los expertos en dominios sugiere que las promociones dirigidas podrían aumentar la rentabilidad entre un 15% y un 128% en relación con una campaña estándar no diferenciada.

2.2.6 Conclusión del Estado del Arte

Se concluye que la aplicación de la técnica del SMS aportó con la obtención de artículos científicos que servirán de apoyo para resolver la problemática propuesta, ya que estos abordan los temas de metodologías, herramientas y técnicas de minería de datos. En algunos artículos aplican temas de marketing focalizado lo cual puede servir como antecedentes prácticos para apoyar a nuestro proyecto.

2.3 Marco Teórico

2.3.1 Marketing

El marketing analiza el comportamiento de los mercados y de los consumidores con el objetivo de captar, retener y fidelizar a los clientes a través de la satisfacción de sus necesidades.

El marketing está presente en la mayor parte de nuestras actividades cotidianas, como consumidores estamos expuestos a miles de estímulos externos procedentes de la publicidad y de los comentarios de los vendedores, que nos aportan información sobre una gran variedad de productos y marcas.[3]

2.3.2 Tipos de marketing

A continuación, se detallan los tipos de marketing más conocidos:

Marketing estratégico. – Se enfoca en acciones tendientes a cumplir con las estrategias institucionales, su implantación requerirá una modificación de los procesos existentes.

Marketing mix. – Se conoce como el marketing de las 4P (producto, precio, promoción y distribución) las cuatro variables deben ser necesariamente definidas por parte de la empresa, para adaptarse a las necesidades del cliente.

Marketing operativo. – A diferencia del marketing estratégico que tiene un enfoque a largo plazo, este tipo de marketing, define unos objetivos y acciones a corto/medio plazo.

Marketing directo. –Intenta establecer una comunicación personalizada con el cliente, a través de publicidad más relacionada con aquello que el consumidor requiere o necesita.

Marketing relacional. – Este tipo de marketing es similar al marketing directo con la particularidad de que se enfoca en los clientes más rentables, priorizándolos y manteniendo con ellos una relación más cercana.

Marketing digital. – Se enfoca al mundo online, donde los usuarios hacen uso de internet, haciendo uso de herramientas especializadas de marketing digital y datos generados en la web.

2.3.3 Marketing Directo utilizando Minería de Datos

Según Piatetsky-Shapiro [38], la minería de datos es un proceso no trivial de descubrimiento de conocimiento novedoso, implícito, útil e integral, a partir de una gran cantidad de datos. En marketing directo, este conocimiento es una descripción de probables compradores o respondedores, y es útil para obtener mayor rendimiento que un marketing masivo.

Las empresas disponen de una gran cantidad de datos de los productos que ofertaron, como, por ejemplo, datos sobre el comportamiento transaccional de los clientes, productos adquiridos, opiniones vertidas y otros datos demográficos. A partir de esos datos, es posible realizar minería de datos y obtener perfiles de clientes que necesiten determinados productos y realizar un marketing directo.

A continuación, se describen los pasos que se deben seguir para realizar una campaña de marketing directo con minería de datos:

1. Obtener la base de datos de todos los clientes actuales.
2. Realizar Minería de datos en el conjunto de datos.

- Superposición: añadir información geo-demográfica a la base de datos.

- Preprocesamiento de datos: transformar dirección y área en códigos, tratar con valores faltantes, etc.
- Dividir la base de datos en un conjunto de entrenamiento y un conjunto de pruebas.
- Aplicar algoritmos de aprendizaje al conjunto de entrenamiento.

3. Evaluar los patrones encontrados en el conjunto de pruebas. Si el resultado no es satisfactorio se deben repetir los pasos anteriores.
4. Utilizar los patrones encontrados para predecir posibles compradores entre los actuales no compradores.
5. Promocionar productos entre los probables compradores mediante marketing directo.[4]

2.3.4 Inteligencia de Negocios

¿Qué es la inteligencia de Negocios?

La inteligencia empresarial (BI), como un término general que incluye; las aplicaciones, la infraestructura, las herramientas, y las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y el rendimiento.[5]

El manejo de la administración, la gestión y control de la información, como un arma estratégica, forma parte de la inteligencia del negocio, con apoyo de herramientas informáticas y analíticas que ayudan a las organizaciones a maximizar su rendimiento, generando eficacia operativa. Así mismo, la gestión del conocimiento ayuda a obtener mayor comprensión y entendimiento del entorno y de los procesos desde la propia experiencia de las personas y organizaciones. [6]

2.3.4.1 Data mart

Los datamarts son bases de datos orientadas a temas de cada área de negocios. Es un objeto de procesamiento analítico para el usuario final. Puede ser realizado por cada área de la empresa; como finanzas, marketing, etc. Son menos costosos y mucho más pequeños que un data warehouse.

Un datamart tiene datos específicos de un área / departamento de negocios el cual contiene solo un subconjunto de datos empresariales que son valiosos para una unidad o departamento de negocios específico el mismo que puede estar al detalle o resumido. [7]

2.3.4.2 Data Warehouse

Es el proceso de extraer y filtrar los datos asociados al desempeño del negocio, los cuales proceden de distintos sistemas de información, con el objetivo de conservarlos en un almacén de datos (datawarehouse). Se podrá acceder a este almacén de acuerdo con las necesidades que existan para apoyar la toma de decisiones de forma oportuna, eficaz y eficiente, por parte de los ejecutivos. [8]

Los cinco componentes de un almacén de datos son:

- Fuentes de datos de producción
- Extracción de datos y conversión
- El sistema de gestión de la base de datos del almacén de datos.
- Administración de almacenamiento de datos
- Herramientas de inteligencia de negocios (BI)

2.3.4.3 Metodología Kimball

Kimball sugiere utilizar una metodología Bottom-Up (ver ilustración 2), la información se extrae de los sistemas transaccionales para ser cargada en diferentes Data Marts cada uno de los cuales son independientes y están modelados de forma dimensional, la unión de los data marts formaran el data warehouse.[9]

El modelo dimensional fue diseñado para facilitar el análisis e interpretación de los datos, a partir de cierta redundancia de datos, se obtiene mayor espacio de almacenamiento y mayor procesamiento inicial.

Las consultas de un data warehouse basado en el modelo dimensional son mucho más rápidas porque las tablas ya han sido parcialmente desnormalizadas y en algunos casos están resumidas para algunos análisis. [10]

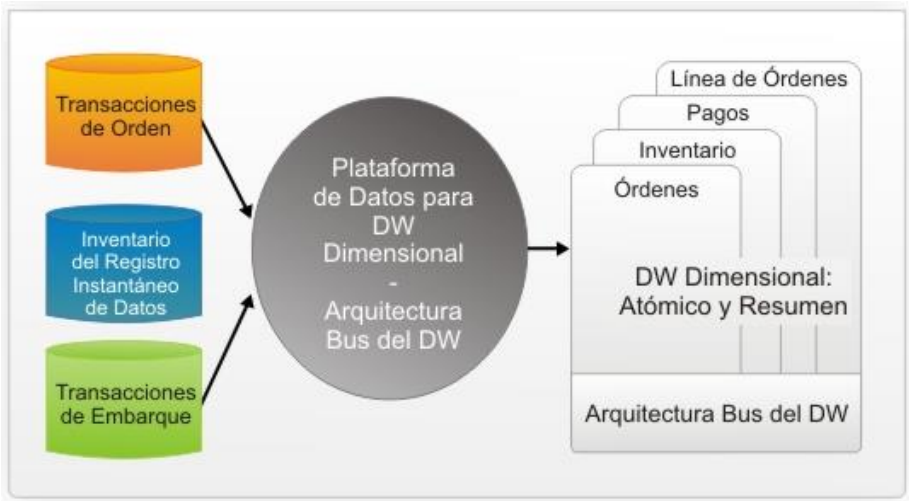


Figura 2: Modelo Kimball

Fuente: El rincón del BI

Las cuatro decisiones clave tomadas durante el diseño de un modelo dimensional incluyen:

1. Seleccionar el proceso de negocio.

2. Identificar el objetivo.

3. Identificar las dimensiones.

4. Identificar los hechos.

Las respuestas a estas preguntas se determinan considerando las necesidades del negocio junto con la realidad de los datos fuente subyacentes durante las sesiones de modelado colaborativo. Siguiendo el proceso de negocio, objetivo, dimensión y declaraciones de hechos, el equipo de diseño determina la tabla y Nombres de columnas, valores de dominio de muestra y reglas de negocio.[11]

2.3.4.4 Metodología Bill Inmon

Inmon sugiere una metodología top-down (ver ilustración 3), se nutrirá de ETL y se encontrará en 3FN, indica que una vez que se tiene el date warehouse cargado, se pueden realizar los data mart para las diferentes áreas del negocio y ser utilizado para minería de datos.

A principios de los años noventa, Inmon acuñó el término almacén de datos (DW) que es una colección no volátil, orientada al tema, integrada, variable en el tiempo de datos en apoyo de las decisiones de la dirección. Un DW está integrado porque los datos se recopilan a partir de una variedad de fuentes y se fusionan en un todo coherente.

Los procesos (Extraction-Transformation-Loading) son responsables de la extracción de datos de fuentes de datos operacionales heterogéneas, su transformación.[12]

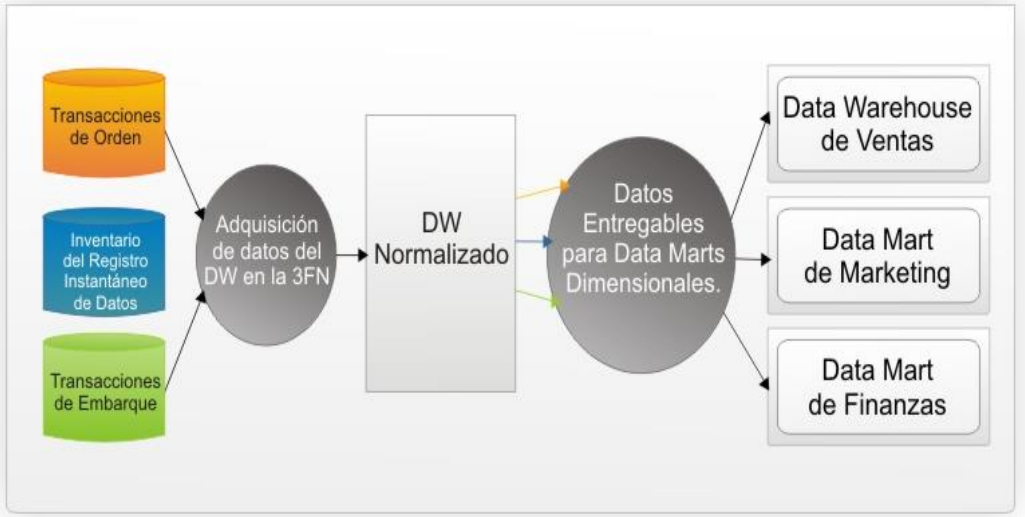


Figura 3: Modelo Bill Inmon

Fuente: El rincón del BI

2.3.4.5 Exploración de datos

En la empresa existe la necesidad de realizar un procesamiento estadístico de forma regular, para este proceso es necesario construir un almacén de datos de exploración. A continuación, se detallan las diferencias entre almacén de datos y de exploración de datos.

- El almacén de datos es una estructura persistente, mientras que el almacén de exploración se basa en un proceso analítico recurrente.
- El almacén de datos está diseñado para adaptarse al software inteligente de negocios, mientras que el almacén de exploración está diseñado para alojar el software de análisis estadístico.
- El almacén de datos contiene datos que están muy normalizados, mientras que el almacén de exploración a menudo contiene datos que se modifican, en previsión del análisis estadístico. [13]

2.3.4.6 Balanced Scored Card

Es la parte de visualización de cualquier sistema/herramienta que sirve de apoyo a la toma de decisiones es muy importante ya que proporciona una imagen completa del negocio, las cuales necesita conocer el usuario experto para poder explorar sus datos durante el proceso de toma de decisiones. El área de visualización, es parte del análisis exploratorio de datos (EDA), y fue creado durante la década de 1970 por el famoso estadístico John Tukey.[14]

2.3.5 Minería de Datos

¿Qué es la minería de Datos?

Es el proceso de la extracción de información oculta y predecible de grandes bases de datos, es una poderosa tecnología que aporta a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de minería de datos predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven).[15]

2.3.5.1 Técnicas de Minería de datos

Modelo Predictivo

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como valido. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión y redes neuronales. Pero, tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez técnicas de clasificación que pueden extraer perfiles de

comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato.[16]

Los modelos predictivos pueden ser: de ***clasificación*** en donde se dividen los datos en subconjuntos y de ***predicción*** donde los comportamientos del pasado ayudan a predecir situaciones actuales o del futuro. En la ilustración 4 se indica como es el proceso de un modelo predictivo.

Figura 4: Modelo Predictivo

Técnicas de análisis predictivo

Entre las principales técnicas de análisis predictivo se tienen: árboles de decisión, regresión lineal o logística, y redes neuronales.

Árboles de decisión

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Los árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales.

Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas.[17] Permiten identificar grupos homogéneos con alto o bajo riesgo y facilita la creación de reglas para realizar predicciones sobre casos individuales, es utilizado para:

- Segmentación
- Estratificación
- Predicción
- Reducción de datos y clasificación de variables
- Identificación de Interacción
- Fusión de categorías y discretización de variables continuas

El árbol incluye, un nodo raíz, nodos hoja que representan cualquier clase, nodos internos que representan condiciones de prueba (ver ilustración 5).

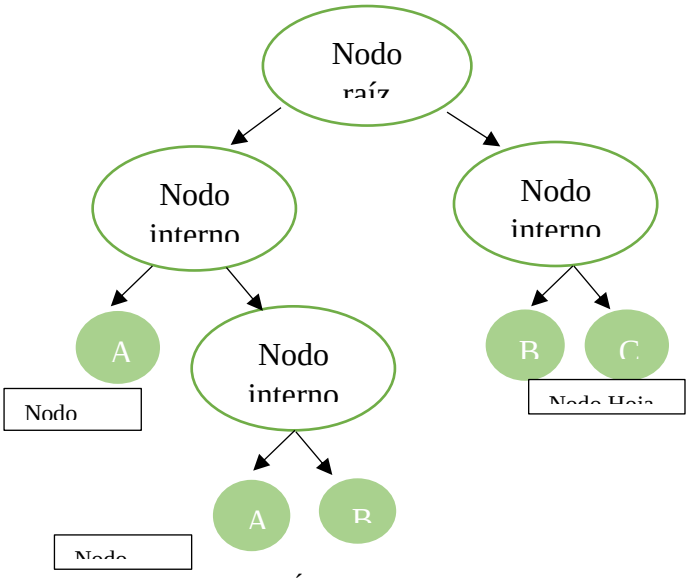


Figura 5: Árbol de decisión

***Tamaño del árbol.* -** La complejidad del árbol se puede medir mediante una métrica que contiene: el número total de nodos, número total de hojas, profundidad del árbol y número de atributos utilizados en la construcción del árbol. El tamaño del árbol debe ser relativamente pequeño, lo cual se logra mediante el uso de una técnica llamada poda.

***Regla de inducción en árboles.* -** La inducción del árbol de decisión está estrechamente relacionada con la inducción de la regla. Cada ruta que comienza desde la raíz de un árbol de decisión y termina en una de sus hojas que representa una regla.[18]

Existen diferentes tipos de árbol:

Árbol de clasificación y regresión (CART). - CART construye un árbol binario y divide un conjunto de datos basado en índice de *gini*. Se poda el árbol utilizando la mínima complejidad de costos, que se calcula utilizando el número de hojas y el porcentaje de instancias de datos mal clasificadas por el árbol. CART construye el árbol de regresión y predice la etiqueta de clase basada en la media ponderada del nodo.

C4.5.- Es un algoritmo popular y una extensión de ID3, selecciona las características en función de la ganancia de información. Maneja los datos de entrenamiento incompletos con valores perdidos y tiene la capacidad de usar características tanto continuas como discretas.

C5.0.- Es una versión avanzada de C4.5 con características adicionales como el aumento y costos desiguales para diferentes tipos de errores lo que genera un número de árboles más pequeños y realiza una poda global a fin de eliminar los subárboles que no son necesarios.

Árbol de Bayes ingenuo (NBTree). - Kohavi propone un clasificador híbrido llamado NBTree, que selecciona un nodo basado en el valor de utilidad más alto. Utiliza 5 veces el valor de validación cruzada de Naive Bayes para el cálculo del valor de utilidad en un nodo. NBTree es como un árbol de decisión clásico que incluye un clasificador de Bayes.

BFTree. - Haijian Shi sugiere el aprendiz del árbol de decisiones Best-First, (el primero que mejor expande el nodo) es decir, se selecciona el nodo basado en la máxima reducción de impurezas entre todos los disponibles.[19]

Regresión lineal y logística

Es un modelo estadístico que nos permite explicar los datos mediante el fenómeno de causa - efecto.

Antes de realizar un modelo de regresión lineal, se debe asegurar que existe una relación o correlación entre cada uno es decir variable independiente a la variable dependiente y la relación entre la variable dependiente con todos.

Hay varios métodos de prueba que se pueden utilizar para la prueba de linealidad, que son prueba de White, prueba de Terasvirta y prueba de Reset.[20]

Si hay una variable dependiente (Y) con *p* variables independientes (X1, X2,…, Xp), entonces, la regresión lineal múltiple de los modelos se pueden escribir de la siguiente manera.

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon \quad (1)$$

$$Y = X\beta + \varepsilon \quad (2)$$

Dónde

Y = variable dependiente (n × 1)

X = la matriz de la variable independiente (n × (p +1))

β = vector de parámetros del modelo de regresión ((p +1) × 1)

ε = vector de error (n × 1)

$$y = \begin{bmatrix} Y1 \\ Y2 \\ . \\ . \\ . \\ Yn \end{bmatrix} \; y = \begin{bmatrix} 1 & X11 \dots X1P \\ 1 & X21 \dots X2P \\ . & . \quad . \\ . & . \quad . \\ . & . \quad . \\ 1 & XN1 \dots XNP \end{bmatrix} \beta = \begin{bmatrix} \beta 0 \\ \beta 1 \\ . \\ . \\ . \\ \beta n \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon 1 \\ \varepsilon 2 \\ . \\ . \\ . \\ \varepsilon n \end{bmatrix}$$

El parámetro β se estima para obtener el modelo de regresión, a continuación, la fórmula para estimar el parámetro:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

Redes Neuronales

Fue introducido por Donald F. Specht en 1990 como una red basada en memoria que proporciona estimaciones de variables categóricas.

El algoritmo proporciona una aproximación suave de una función de destino, incluso con datos dispersos en un espacio multidimensional. Las ventajas de una red neuronal probabilística, es que son de rápido aprendizaje.

En la ilustración 6 se indica la red neuronal compuesta por cuatro capas: entrada, patrón (función del kernel RBF), suma y salida.[21]

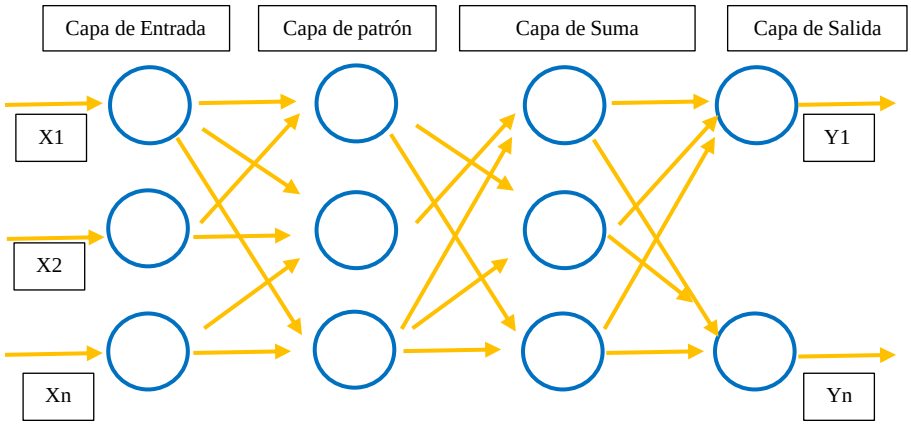


Figura 6: Red Neuronal

Modelo Descriptivo

Los modelos descriptivos, cubren un conjunto de técnicas que tienen como objetivo resumir los datos mediante la identificación de algunas características relevantes con el fin de describir cómo las cosas se organizan y realmente funcionan.[22]

Técnicas de Análisis Descriptivo

Entre las principales técnicas de análisis descriptivo se tiene: asociacion y clustering.

Reglas de asociación

Agrawal abordó por primera vez las reglas de la asociación en 1993. Señalando que hay algunas relaciones ocultas entre los artículos comprados en bases de datos transaccionales. Por ejemplo, hay asociaciones o relaciones entre artículos como el pan y la leche, que a menudo se compran juntos en una sola transacción de canasta.

Los resultados de la minería pueden ayudar a comprender el comportamiento de compra del cliente, que podría no haber sido previamente percibido.[23]

Una regla de asociación es de la forma X ⇒ Y, donde X e Y son conjuntos de elementos frecuentes en la base de datos dada y la intersección de X e Y es un conjunto vacío, es decir, X ∩ Y = ∅.

El soporte de la regla X ⇒ Y es el porcentaje de transacciones en la base de datos dada que contienen tanto X como Y, es decir, P (X ∪ Y).

La confianza de la regla X ⇒ Y es el porcentaje de transacciones en la base de datos que contiene X que también contiene Y, es decir, P (Y | X).

Por lo tanto, la minería de reglas de asociación se usa para encontrar todas las reglas de asociación entre los conjuntos de elementos en una base de datos determinada, donde el soporte y la confianza de estas reglas de asociación debe satisfacer el soporte mínimo especificado por el usuario y la confianza mínima.

El problema de la minería de reglas de asociación se puede dividir en dos subproblemas:

- Encontrar elementos frecuentes con sus soportes por encima del umbral mínimo de soporte.

- El uso de elementos frecuentes que se encuentran en el paso 1 para generar reglas de asociación que tienen un nivel de confianza superior al umbral mínimo de confianza.[23]

Clustering

Es una técnica de data mining no supervisada que consiste en encontrar conjuntos de datos en forma de agrupaciones, este método se aplica para exploración de datos.[24]

Los propósitos principales para aplicar clustering son los siguientes:

- Búsqueda de grupos inherentes en el conjunto de datos bajo los supuestos de que estos grupos realmente existen.

- Partición del conjunto de datos en subconjuntos. Aquí no es muy importante que los grupos estén más o menos bien separados. La partición del conjunto de datos es necesaria para reducir la complejidad del conjunto de datos y manejar los clústeres por separado. Se debe cumplir el criterio de homogeneidad.

- Búsqueda de agrupaciones individuales en el cual es suficiente encontrar uno o unos pocos clústeres bien separados y la mayoría de los datos podrían no estar asignado a cualquier cluster.[24]

Entre las principales técnicas de agrupamiento tenemos: algoritmos basados en la distribución, algoritmos basados en la jerarquía, algoritmos basados en la densidad y algoritmos basados en cuadrícula, la elección depende de la forma, densidad, anomalías y conocimiento a priori de los conjuntos de datos.[25]

- ***Métodos de partición.*** **-** dado un conjunto de n objetos, un método de partición construye k particiones de los datos, donde cada partición representa un clúster y k

≤ n. Es decir, divide los datos en k grupos tales que cada grupo debe contener al menos un objeto.

- ***Métodos jerárquicos.*** **-** un método jerárquico crea una descomposición jerárquica de los conjuntos de datos. Un método jerárquico puede clasificarse como cualquiera

    - Hay dos enfoques básicos para generar un agrupamiento jerárquico:

    - *Aglomerativo.* - Comienza con los puntos como grupos individuales y, en cada paso, fusionar el par de grupos más similares o más cercanos. Esto requiere una definición de similitud o distancia de racimo.

    - *Divisivo.* - Comienza con un grupo, todo incluido y, en cada paso, divide un grupo hasta que solo quedan grupos de puntos individuales. En este caso, tenemos que decidir, en cada paso que grupo dividir y cómo realizar la división.

- ***Métodos de densidad.*** **-** pueden dividir un conjunto de objetos en múltiples agrupaciones exclusivas, o una jerarquía de clusters. Por lo general, los métodos basados en densidad consideran solo clusters exclusivos, y no consideran los grupos difusos.

- ***Métodos basados en cuadrícula.*** **-** los métodos basados en cuadrícula cuantizan el espacio de objetos en un número finito de celdas que forman una estructura de rejilla. Todas las operaciones de clustering se realizan en la estructura (es decir, en el espacio cuantizado). La principal ventaja de este enfoque es su rapidez.[26]

2.3.5.2 Ciclo de Vida en Minería de Datos

Minería de datos como cualquier proceso tiene un ciclo de vida, que consiste en varios pasos que deben ser cumplidos secuencialmente, con la posibilidad de recurrencia entre ellos. Este ciclo de vida se explica con mayor detalle en el capítulo III.

CAPITULO III

METODOLOGIAS DE MINERIA DE DATOS

3.1 KDD

Según Fayyad [39] indica que es el proceso de usar métodos de minería de datos para extraer lo que se considera conocimiento según la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación requeridos de la base de datos.

Se consideran cinco pasos, representados en la ilustración 7:

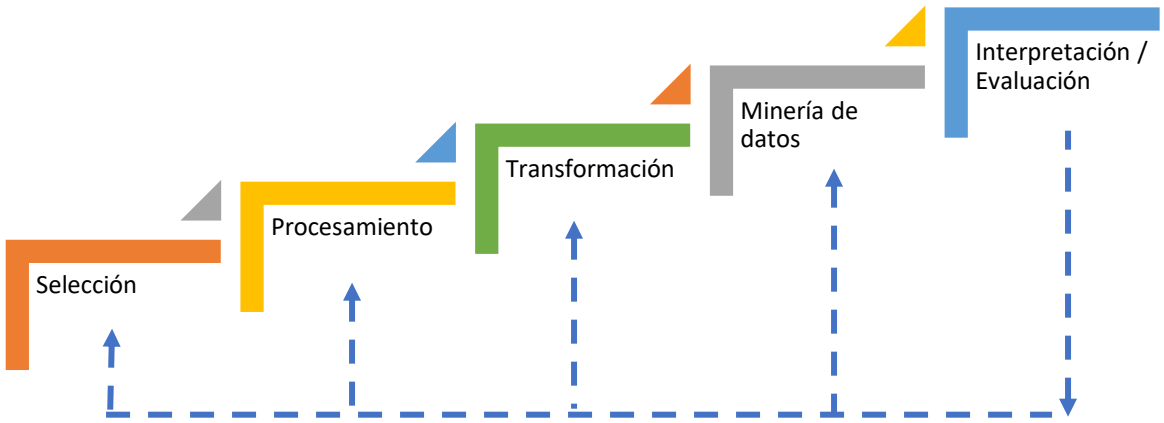


Figura 7: Etapas de KDD

Fuente: Fayyad, 1996

Selección. - Esta etapa consiste en crear un conjunto de datos de destino o centrarse en un subconjunto de variables o muestras de datos, en las que se realizará el descubrimiento.

Procesamiento. - Previo: esta etapa consiste en la limpieza y el procesamiento previo de los datos de destino para obtener datos consistentes.

Transformación. - Esta etapa consiste en la transformación de los datos utilizando la dimensionalidad, métodos de reducción o transformación.

Minería de datos. - Esta etapa consiste en la búsqueda de patrones de interés.

Interpretación / Evaluación. - Esta etapa consiste en la interpretación y evaluación de los patrones.

El proceso de KDD es interactivo e iterativo, involucra numerosos pasos y se toman muchas decisiones por el usuario.

El proceso de KDD debe ir precedido por el desarrollo de un entendimiento del dominio de la aplicación, el conocimiento previo relevante y los objetivos del usuario final. También debe ser continuado por la consolidación del conocimiento mediante la incorporación de este conocimiento en el sistema.[27]

3.2 CRISP – DM

El proceso CRISP-DM significa proceso estándar de datos de CRoss-Industry, consiste en un ciclo de seis etapas como se indica en la ilustración 8.

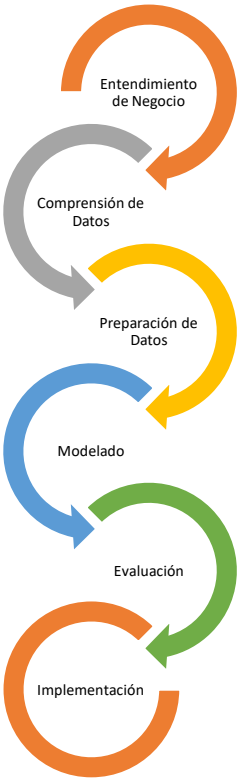


***Figura 8*:** Fases de CRISP-DM

Fuente: KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW

Entendimiento de negocio. - Esta fase inicial se enfoca en entender los objetivos del proyecto y los requisitos desde una perspectiva empresarial, y luego convertir este conocimiento en una minería de datos que nos permita definir el problema y diseñar los objetivos.

Comprensión de los datos. - Comienza con una recopilación de los datos con el objetivo de poder familiarizarse y conocer los problemas de calidad de los datos que nos permita descubrir las primeras ideas sobre los datos o para detectar subconjuntos interesantes para formular hipótesis.

Preparación de los datos. - La fase de preparación de los datos cubre todas las actividades para construir el conjunto de datos final a partir de los datos en bruto iniciales.

Modelado. - En esta fase, se seleccionan y aplican diversas técnicas de modelado y los parámetros se calibran a valores óptimos.

Evaluación. - En esta etapa, el modelo (o modelos) obtenido se evalúa más a fondo para asegurarse de que el negocio logre sus objetivos correctamente.

Implementación. - La creación del modelo generalmente no es el final del proyecto, incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá ser organizado y presentado de forma que el cliente pueda utilizarlo.[28]

3.3 SEMMA

El proceso SEMMA fue desarrollado por el Instituto SAS. El acrónimo SEMMA significa Muestra, Explore, modifique, modele y evalúe, se considera un ciclo con 5 etapas para el proceso indicado en la ilustración 9:

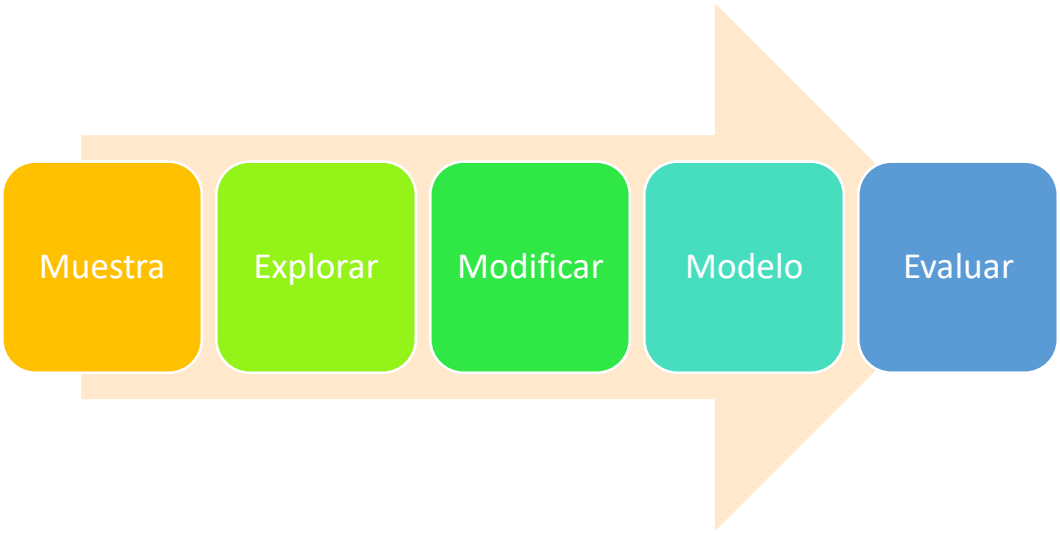


Figura 9: Fases de SEMMA

Fuente: KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW

Muestra. - Esta etapa consiste en muestrear los datos mediante la extracción de una parte de un gran conjunto de datos, suficiente para contener la información significativa, pero lo suficientemente pequeña para manipular rápidamente.

Explorar. - Esta etapa consiste en la exploración de los datos mediante la búsqueda de tendencias no anticipadas y anomalías para ganar comprensión e ideas.

Modificar. - Esta etapa consiste en la modificación de los datos mediante la creación, selección y transformación de las variables para enfocar el proceso de selección del modelo.

Modelo. - Esta etapa consiste en modelar los datos permitiendo que el software busque automáticamente para una combinación de datos que predice de manera confiable un resultado deseado.

Evaluar. - Esta etapa consiste en evaluar los datos mediante la evaluación de la utilidad y confiabilidad de los hallazgos del proceso de extracción de datos y estimar qué tan bien se realiza.[29]

3.4 Comparación cualitativa de metodologías minería de datos

Se asignará un punto por cada característica cumplida en las metodologías y se procederá a la tabulación como se indica en la tabla 7, los parámetros hacer evaluados son:

- Entendimiento del negocio
- Muestra de los datos
- Preparación de los datos
- Modelado

- Evaluación
- Implementación

Tabla 7:
Ponderación de Metodologías

Característica \ Metodología	KDD	CRISP-DM	SEMMA
Entendimiento del Negocio	0	1	0
Muestra de los datos	1	1	1
Preparación de los datos	1	1	1
Modelado	1	1	1
Evaluación	1	1	1
Implementación	1	1	1
Total, Ponderación	**5**	**6**	**5**

3.4.1 Metodología seleccionada

SEMMA y CRISP -DM son metodologías basadas en KDD, de acuerdo al análisis realizado se concluye que CRISP – DM es una metodología que abarca una etapa muy importante que es el entendimiento del negocio; si no se comprende el negocio, no se puede formular hipótesis adecuadas, que luego sean discernidas mediante la aplicación de modelos. En consecuencia, en el presente proyecto se utilizará la metodología CRISP – DM.

3.5 Herramientas de minería de datos

Weka. - Es un software de código abierto y fue desarrollado por la Universidad de Waikato, se basa en Java y puede usarse con Windows, MacOS y Linux. Conocido por sus amplias capacidades de aprendizaje automático, admite todas las tareas principales de minería de datos, como agrupamiento, asociación, regresión y clasificación.

WEKA puede experimentar problemas con el procesamiento si la cantidad de datos es demasiado grande. Esto se debe a que la herramienta de minería de datos intenta cargar todo en la

memoria. Para evitar esto, WEKA ofrece una línea de comando simple (CLI) que facilita el manejo de grandes cantidades de datos.[30]

RapidMiner. - Fue escrito en Java y contiene más de 500 operadores con diferentes enfoques para señalar las conexiones en los datos: hay opciones para la minería de datos, la minería de textos, la minería web y también para el análisis del estado de ánimo (análisis de sentimientos, minería de opinión), entre otras cosas. El programa también importa tablas de Excel, archivos SPSS y conjuntos de datos de muchas bases de datos, e integra las herramientas de extracción de datos WEKA y R.

RapidMiner admite todos los pasos del proceso de extracción de datos, incluida la presentación de resultados. La herramienta consta de tres módulos principales: RapidMiner Studio, RapidMiner Server y RapidMiner Radoop, cada uno de los cuales ejecuta diferentes técnicas de extracción de datos. Una fortaleza particular de RapidMiner es el análisis predictivo, que es el nombre que se le da a la predicción de futuros desarrollos basados en los datos recopilados.[31]

Anaconda. - Directamente desde la plataforma y sin la participación de DevOps, los científicos de datos pueden desarrollar e implementar modelos de aprendizaje automático e inteligencia artificial en producción.

Anaconda proporciona las herramientas necesarias para:

- Recopilar datos de archivos, bases de datos y lagos de datos.
- Administre los entornos con Conda (todas las dependencias de paquetes se cuidan en el momento de la descarga)

- Comparte, colabora y reproduce proyectos.

- Despliegue proyectos en producción con un solo clic de un botón.[32]

3.5.1 Selección de la herramienta de datos

Se realiza un cuadro comparativo como se detalla en la tabla 8, con características que deba cumplir la herramienta.

Se evalúa que:

- **Grandes Volúmenes de Datos.** - La herramienta debe ser capaz de procesar grandes volúmenes de datos.

- **Cumple ciclo de minería de Datos. -** Es importante contar con una herramienta que permita cumplir con el ciclo de minería de datos para no utilizar más herramientas que apoyen al proceso.

- **Varios Algoritmos. -** La herramienta debe estar en la capacidad de la ejecución de cualquier tipo de algoritmos sin tener restricción de procesamiento.

- **Multiplataforma. -** Se debe poder instalar en los diferentes sistemas operativos.

- **Multilenguaje. -** Debe tener el ambiente para poder ejecutar en otras lenguas de programación como Python y R.

Tabla 8:

Selección de Herramienta

HERRAMIENTA	RAPIDMINER	WEKA	ANACONDA
Grandes volúmenes de datos	1	0	1
Cumple el ciclo de minería de datos	1	1	1
Varios algoritmos	1	0	1
Multiplataforma	1	1	1
Multilenguaje	0	0	1
Total	**4**	**2**	**5**

De acuerdo a la tabla de puntuación y la investigación realizada, la herramienta Weka no es óptima para grandes volúmenes de datos, ya que todo el procesamiento lo realiza en memoria además es especialista en los algoritmos de clasificación: como redes neuronales artificiales, árboles de decisión, ID3 y algoritmos C4.5 sin embargo, WEKA tiene su debilidad cuando se trata de otras técnicas, como el análisis de conglomerados.

Rapidminer es una herramienta muy fuerte, tiene la parte gráfica que ayuda a realizar los modelos de una manera más ágil, sin necesidad de programación, sin embargo, se necesita especialistas que conozcan y aprender la herramienta.

La herramienta mejor puntuada es Anaconda, que tiene varios lenguajes de programación como con R y Python que en la actualidad son los lenguajes que abarcan la mayor parte de algoritmos y procesamiento de grandes volúmenes de datos.

CAPITULO IV

PROPUESTA DEL MODELO

4.1 Entendimiento del Negocio

El factor clave en el desarrollo de una estrategia competitiva, es la comprensión y el análisis del comportamiento del cliente y esto ayuda a adquirir y retener clientes potenciales. La minería de datos ayuda a las organizaciones a identificar clientes valiosos y predecir su comportamiento futuro. Lo que puede ser soportado por varios modelos de minería de datos.[33]

4.1.1 Tarjetas de Crédito

La Superintendencia de Bancos del Ecuador (https://www.superbancos.gob.ec), define a la tarjeta de crédito como:
"Es un documento que permite a su titular – o beneficiario de la tarjeta – adquirir bienes o servicios en establecimientos afiliados al sistema, difiriendo su pago o a crédito. Estos créditos pueden o no incluir intereses. Su uso incluye algunas tarifas, costos de emisión, costo de estado de cuenta, intereses y comisiones. Las tarjetas de crédito son intransferibles y deben emitirse a nombre de su titular. El pago mensual puede efectuarse del monto total de la obligación o del monto mínimo; sin embargo; el realizar solo un abono mínimo implica el pago de intereses." [34]

4.1.1.1 Marcas de Tarjetas de Crédito

Las entidades emisoras de tarjetas de crédito/débito mediante los procesos internos asocian electrónicamente a sus clientes y a su vez con la cuenta bancaria.

BIN. **-** Son los primeros dígitos de una tarjeta que identifica la marca de tarjeta y el banco emisor de la tarjeta. La tabla 9 muestra los nombres de las tarjetas con su bin respectivo en el Ecuador.

Tabla 9:
Bines

Red de emisión	IIN rangos
American Express	34, 371
Diners Club International	300-305, 309, 36, 38-396
Discover Card	6011, 622126-622925, 644-649, 65
MasterCard	51-55
Visa	4
Visa Electron	4026, 417500, 4405, 4508, 4844, 4913, 4917

Dentro de cada marca de tarjeta existen subclasificación de las tarjetas, la misma que dependen de los beneficios que se desea ofertar al cliente.[40][41][42]

- Discover
- Mastercard Classic
- Mastercard Elite Black Signature
- Mastercard Gold
- Mastercard Platinum
- Mastercard Signature
- Visa Classic
- Visa Elite Black Signature
- Visa Elite Infinite
- Visa Gold
- Visa Platinum
- Visa Signature

4.1.1.2 Colocación de Tarjetas

***Colocación Individual.* –** Es un procedimiento que inicia con la solicitud de una tarjeta por parte del cliente, a la entidad financiera, para ello remite información que será analizada por la institución y, de acuerdo a los parámetros definidos, se otorgará o rechazará la solicitud.

En uno de los 10 Bancos con mayor captación financiera existe un promedio de 236 solicitudes al año como se indica en la tabla 10.

Tabla 10:
Solicitudes de tarjetas

Mes	Nro Solicitudes	Porcentaje %
10-Ene	201	0.07
10-Feb	159	0.06
10-Mar	246	0.09
10-Abr	243	0.09
10-May	277	0.1
10-Jun	316	0.11
10-Jul	261	0.09
10-Ago	265	0.09
10-Sep	255	0.09
10-Oct	266	0.09
10-Nov	185	0.07
10-Dic	158	0.06
Total	**2832**	

***Colocación Masiva.* -**La colocación masiva de tarjetas se realiza mediante la utilización de bases de posibles clientes que cumplen con los requisitos de crédito y filtro de control.[35]

4.1.1.3 Proceso de las tarjetas de Crédito

En la ilustración 10 se indica el proceso de registro de transacción con una tarjeta de crédito.

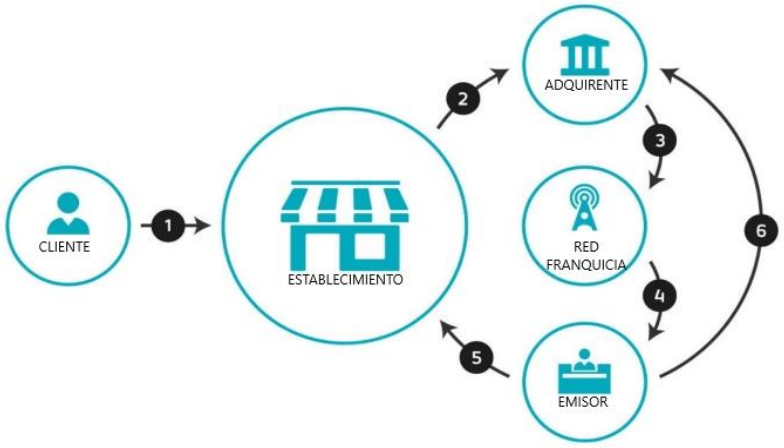


Figura 10: Proceso del funcionamiento de Tarjetas de Crédito

Fuente: Mastercard

Paso 1: El cliente paga con Tarjeta.

El cliente adquiere productos o servicios del establecimiento.

Paso 2: Autenticación del pago.

El sistema de punto de venta del establecimiento captura los datos de cuenta del cliente y los envía de forma segura al adquiriente.

Paso 3: Envío de transacción.

El adquiriente del establecimiento solicita a la franquicia la autorización del banco emisor del cliente.

Paso 4: Solicitud de autorización.

La franquicia envía la transacción al emisor para obtener la correspondiente autorización.

Step 5: Respuesta de autorización.

El banco emisor autoriza la transacción y reenvía la respuesta al establecimiento.

Step 6: Pago del establecimiento.

El banco emisor dirige el paso al adquiriente del establecimiento, que lo deposita a su vez en la cuenta del establecimiento comercial.[36]

Los actores que intervienen en el proceso de las tarjetas son:

- **Adquirente. -** También conocido como banco comercial, un adquirente es una entidad financiera con licencia de la franquicia, para ayudar a un establecimiento a receptar los pagos con las tarjetas.
- **Emisor. -** Una entidad emisora es el banco, entidad de crédito, entidad de ahorro y préstamos, entidad gubernativa (como una organización postal) o negocio minorista que ofrece una línea de crédito o una tarjeta de débito a un consumidor o empresa.
- **Red de Franquicia. –** La franquicia no es ni una entidad emisora ni un adquirente. Su papel consiste en proporcionar la tecnología y la red que hacen posibles las transacciones.
- **Establecimiento. -** Es el lugar donde el cliente realiza sus compras

4.1.2 Evaluación de la Situación

Fidelización de Clientes

Los programas de lealtad son una herramienta de diferenciación y alto impacto para crear relaciones a largo plazo con los clientes de una empresa, a fin de lograr una mejora en la cartera y los cobros, reducir los costos de servicio al cliente, realizar venta cruzada, entre otros beneficios.

El proceso de envío de publicidad o de venta cruzada en la actualidad, se establece mediante el envío masivo a los diferentes segmentos de los clientes como se indica en la ilustración 11, donde no se asegura de que esta publicidad surja efecto con la posibilidad de perder cercanía con los clientes.

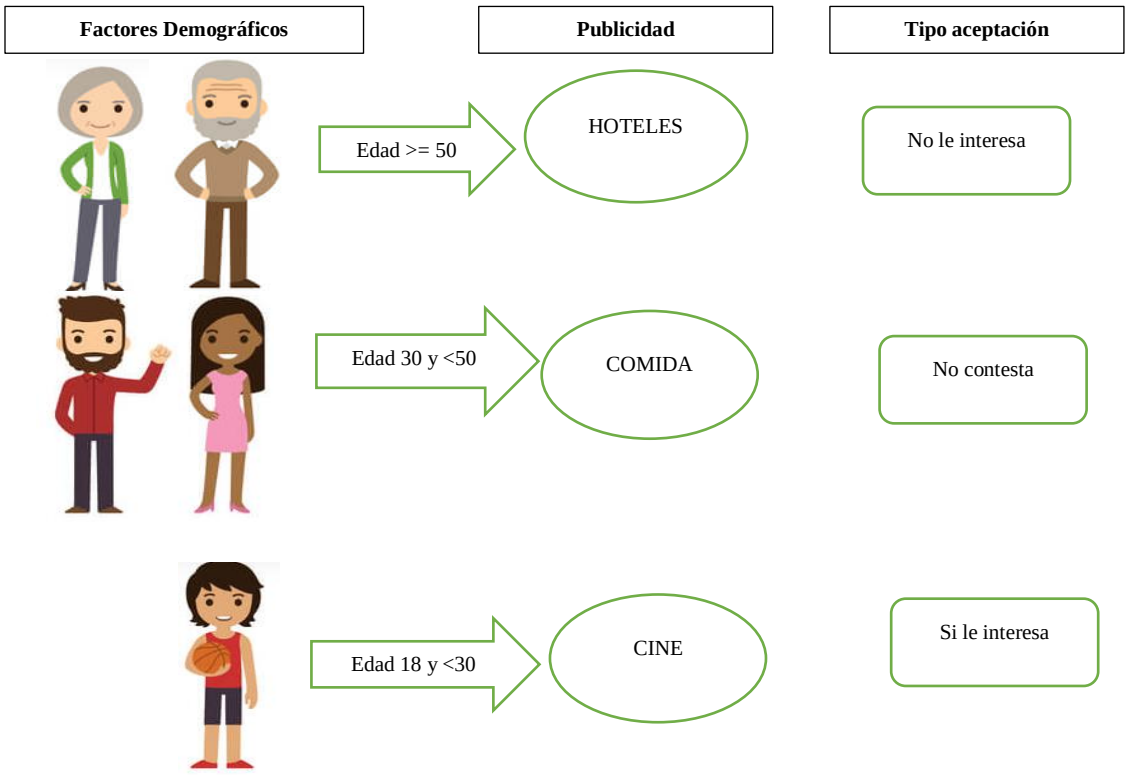


Figura 11: Proceso Masivo de Publicidad

En base a la transaccionalidad de las tarjetas de crédito que han sido colocadas de forma masiva o individual se pretende realizar campañas de las promociones de los establecimientos afiliados, con el fin de que el cliente pueda acceder a beneficios exclusivos de la marca y sentir más cercanía a la entidad financiera.

4.2 Comprensión de los datos

4.2.1 Fuentes de Información

Las fuentes de información utilizadas en el presente proyecto se describen a continuación:

- Base de información demográfica: es la fuente donde se almacena la información del cliente, donde se detallan variables demográficas, como: edad, genero, estado civil, e instrucción.

- Base de información financiera: es la fuente donde se tienen variables económicas como: tipo de vivienda, saldo promedio, etc.

- Base de información transaccional: es la fuente donde se encuentra el historial de transacciones del cliente, relativas a los pagos realizados con su medio electrónico.

4.2.2 Definición de variables

A continuación, se detallan las variables demográficas desde la tabla 11 a la tabla 16:

Tabla 11:
Variable – Edad

Nombre Variable	**Edad**
Fuente	Base Demográfica
Descripción	Es la edad del cliente.
Tipo de Datos	Int
Posibles Valores	0 – 100
Número de registros	1087431

Tabla 12:
Variable – NroCargas

Nombre Variable	**NroCargas**
Fuente	Base Demográfica
Descripción	Es la cantidad de dependientes del cliente.
Tipo de Datos	Int

CONTINÚA →

Posibles Valores	1 – 100
Número de registros	1087431

Tabla 13:
Variable – Género

Nombre Variable	Género
Fuente	Base Demográfica
Descripción	Es el género del cliente.
Tipo de Datos	varchar (10)
Posibles Valores	Femenino
	Masculino
Número de registros	Femenino: 488792
	Masculino: 598639

Tabla 14:
Variable – Estado Civil

Nombre Variable	EstadoCivil
Fuente	Base Demográfica
Descripción	Es el estado civil del cliente
Tipo de Datos	varchar (50)
Posibles Valores	Casado sin separación
	Unión libre < 2 años
	Unión libre > a 2 años
	Viudo
	Divorciado
	Casado con separación
	Soltero
Número de registros	Casado: 61238
	Casado con separación: 12765
	Casado sin separación: 474898
	Divorciado: 91883
	No definido: 21703
	Soltero: 395516
	Union libre < 2 anios: 8508
	Union libre > 2 anios: 7444
	Viudo: 13476

Tabla 15:
Variable – Nivel Estudios

Nombre Variable	NivelEstudios
Fuente	Base Demográfica
Descripción	Es el nivel de estudios del cliente
Tipo de Datos	varchar (30)
Posibles Valores	Sin Estudios
	Primarios/Básicos
	Medios/Secundarios
	Formación Intermedio o Técnica
	Universitarios
	Postgrado
Número de registros	null: 21035
	Formacion intermedia o tecnica: 20156
	Medios / secundarios: 494550
	Postgrado:25766
	Primarios/basicos: 50958
	Sin estudios: 653
	Universitarios: 474313

Tabla 16:
Variable - Situación Laboral

Nombre Variable	SituacionLaboral
Fuente	Base Demográfica
Descripción	Es la situación laboral del cliente
Tipo de Datos	varchar (100)
Posibles Valores	Jubilado
	Remesas del Exterior
	Otros
	Rentas
	Estudiante
	Misionero/Religioso
	Menor de Edad
	Ama de Casa
	Pensionista
	Empleado e independiente
	Independiente

CONTINÚA →

	Empleado
	No trabaja
Número de registros	Null: 24573
	Ama de casa: 22243
	Empleado: 726930
	Empleado e independiente: 29185
	Estudiante: 23875
	Independiente: 219561
	Jubilado: 19582
	Menor de edad: 639
	Misionero/religioso: 52
	No trabaja: 1248
	Otros: 16408
	Pensionista: 490
	Remesas del exterior: 280
	Rentas: 2365

A continuación, se detallan la variable financiera en la tabla 17:

Tabla 17:
Variable - Tipo Vivienda

Nombre Variable	TipoVivienda
Fuente	Base financiera
Descripción	Es el tipo de Vivienda del cliente
Tipo de Datos	varchar (100)
Posibles Valores	Propia no hipotecada
	Alquiler
	Anticresis
	Vive con Familiares
	Prestada
	Propia hipotecada
Número de registros	Null: 21878
	Alquiler: 146123
	Anticresis: 1237
	Prestada: 7773
	Propia hipotecada: 63399
	Propia no hipotecada: 324680
	Vive con familiares: 522341

CONTINÚA →

Tabla 18:
Variable Sub Marca

Nombre Variable	SubMarca
Fuente	Base transaccional
Descripción	Es la submarca del medio de pago electrónico
Tipo de Datos	varchar (100)
Posibles Valores	VISA TRADICIONAL INTERNACIONAL
	VISA TRADICIONAL GOLD
	MASTERCARD TRADICIONAL INTERNACIONAL
	VISA TRADICIONAL PLATINUM
	VISA ELITE BLACK SIGNATURE
	MASTERCARD TRADICIONAL GOLD
	MASTERCARD ELITE BLACK SIGNATURE
	MASTERCARD TRADICIONAL PLATINUM
	VISA ELITE INFINITE
	VISA RETAIL NACIONAL
	VISA COMERCIAL EMPRESARIAL
	VISA COMERCIAL CORPORATIVA
	MASTERCARD RETAIL NACIONAL
	VISA CONVENIO
	MASTERCARD COMERCIAL CORPORATIVA
	MASTERCARD COMERCIAL EMPRESARIAL
	DISCOVER TRADICIONAL INTERNACIONAL
	VISA TRADICIONAL INTERNACIONAL MICRO
	DISCOVER CONVENIO
	VISA RETAIL CREDIFACIL
	MASTERCARD TRADICIONAL INTERNACIONAL MICRO
	MASTERCARD CONVENIO
	DISCOVER TRADICIONAL INTERNACIONAL MICRO
Número de registros	VISA TRADICIONAL INTERNACIONAL: 268066
	VISA TRADICIONAL GOLD: 166576
	MASTERCARD TRADICIONAL INTERNACIONAL: 122712
	VISA TRADICIONAL PLATINUM: 118011
	VISA ELITE BLACK SIGNATURE: 98657
	MASTERCARD TRADICIONAL GOLD: 69199
	MASTERCARD ELITE BLACK SIGNATURE: 61786
	MASTERCARD TRADICIONAL PLATINUM: 45798

CONTINÚA →

VISA ELITE INFINITE: 44220
VISA RETAIL NACIONAL: 29585
VISA COMERCIAL EMPRESARIAL: 17086
VISA COMERCIAL CORPORATIVA: 12974
MASTERCARD RETAIL NACIONAL: 12554
VISA CONVENIO: 5858
MASTERCARD COMERCIAL CORPORATIVA: 4637
MASTERCARD COMERCIAL EMPRESARIAL: 2777
DISCOVER TRADICIONAL INTERNACIONAL: 2555
VISA TRADICIONAL INTERNACIONAL MICRO: 2354
DISCOVER CONVENIO: 1171
VISA RETAIL CREDIFACIL: 814
MASTERCARD TRADICIONAL INTERNACIONAL MICRO: 33
MASTERCARD CONVENIO: 6
DISCOVER TRADICIONAL INTERNACIONAL MICRO: 2

A continuación, se detallan las variables transaccionales desde la tabla 18 a la tabla 22:

Tabla 19:
Variable – Marca

Nombre Variable	Marca
Fuente	Base transaccional
Descripción	Es la marca del medio de pago electrónico
Tipo de Datos	varchar (100)
Posibles Valores	Visa
	Mastercard
	Discover
Número de registros	Discover: 3728
	Mastercard: 321401
	Visa: 762302

Tabla 20:
Variable - Monto Transacción

Nombre Variable	MontoTransaccion
Fuente	Base transaccional
Descripción	Es el monto que ha transaccionado el cliente.

CONTINÚA →

Tipo de Datos	float
Posibles Valores	>0
Número de registros	1529363

Tabla 21:
Variable - Número Transacción

Nombre Variable	NumeroTransaccion
Fuente	Base transaccional
Descripción	Es el número de transacciones del cliente.
Tipo de Datos	int
Posibles Valores	>0
Número de registros	1087431

Tabla 22:
Variable - Línea de Negocio

Nombre Variable	LineaNegocio
Fuente	Base transaccional
Descripción	Es la línea de negocio en la cual el cliente realizo la transacción.
Tipo de Datos	varchar (100)
Número de registros – Posibles valores	COMPRA EXTERIOR: 525777
	GASOLINERAS: 69218
	SUPERMAXI: 25807
	MEGAMAXI: 20940
	CNT: 19694
	AIG: 17390
	HYPERMARKET: 17113
	APL* ITUNES.COM/BILL 8: 16548
	FYBECA: 15126
	CLARO: 14395
	MOVISTAR: 14107
	DIRECTV: 14076
	AKI: 13901
	ALMACENES TIA: 10030
	FARMAENLACE CIA.LTDA.: 9658
	NETLIFE: 8432

CONTINÚA →

SUPERMERCADO SANTA MARIA: 8264
DIFARE: 7381
MI COMISARIATO: 6501
UNIVERSIDAD CATOLICA DE SANTIAGO DE: 5697
SANA: 5255
SUPERCINES: 4897
MULTICINES: 4890
JUGUETON: 4719
MARATHON SPORTS: 4651
SALUD S.A.: 4126
KYWI: 3697
LAN: 3435
SWEET & COFFEE: 2968
TVENTAS: 2417
ALMACEN GERARDO ORTIZ: 2329
MAYFLOWER: 2057
PAYLESS EC: 1938
CREPES Y WAFFLES: 1868
SEGUROS DEL PICHINCHA S.A: 1796
ALMACENES DE PRATI: 1696
ETAFASHION: 1600
CORAL: 1592
ECUASANITAS: 1554
MATRICULACION VEHICULAR: 1554
UBER: 1551
HUMANA S.A.: 1535
ZARA: 1521
PUNTO NET: 1500
FERRISARIATO: 1495
TIA S A ECUECU G: 1245
CINEMARK EC: 1212
#+NOMBRE?: 1211
Brussels Grill B: 1181
CRUZ AZUL: 1129
MUNICIPIO DE QUITO: 1064
GASOLINERA MOBIL: 1060
RM: 1048
DOMINOS: 1003

CONTINÚA →

4.2.3 Exploración de variables

***Género*.** -Se identifica que el 55% de las transacciones son realizadas por personad de sexo masculino. (ver ilustración 12)

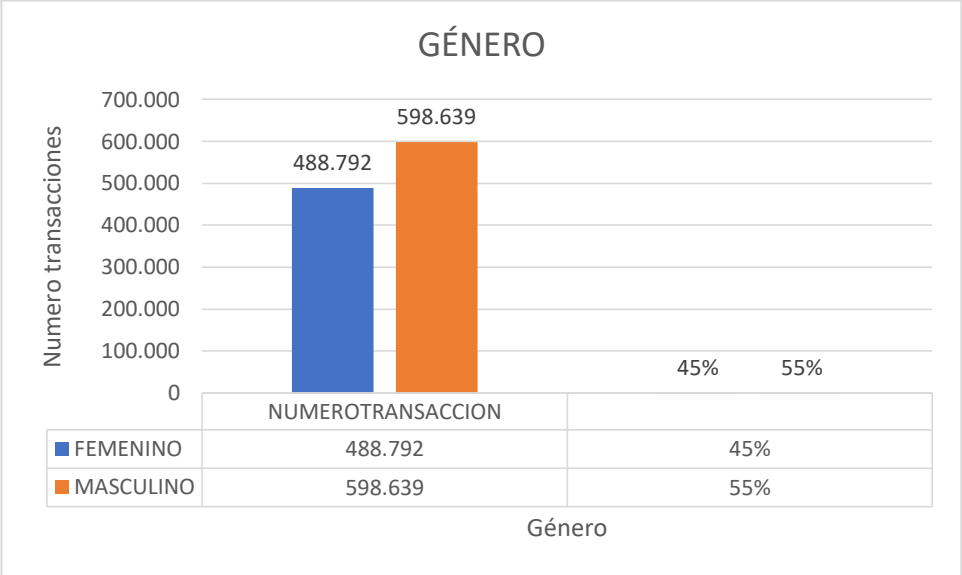


Figura 12: Exploración de género

Estudios. -Se identifica que el 45 % de las transacciones son ejecutados por medios/secundarios, seguido de los que tienen una instrucción universitaria. (ver ilustración 13)

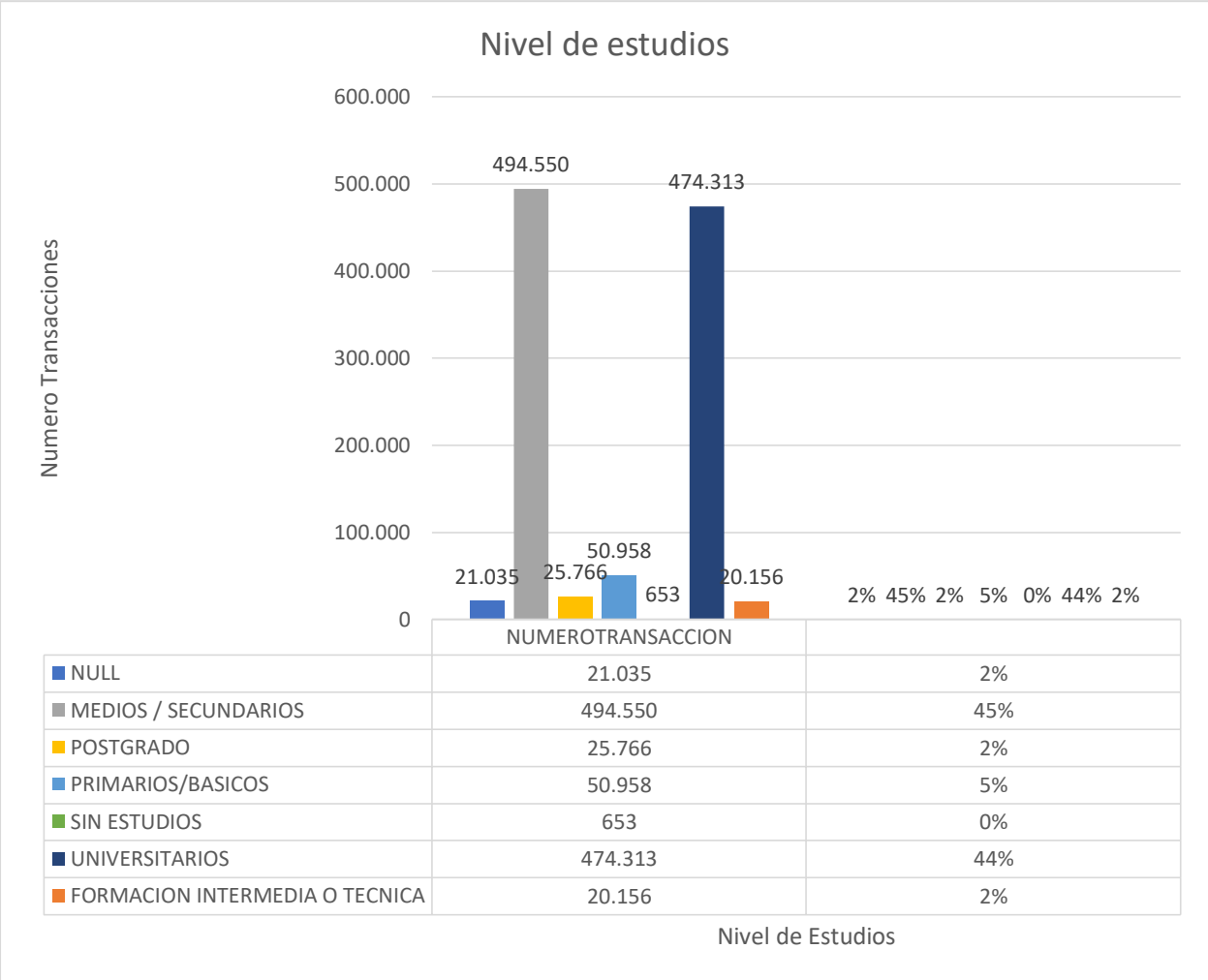
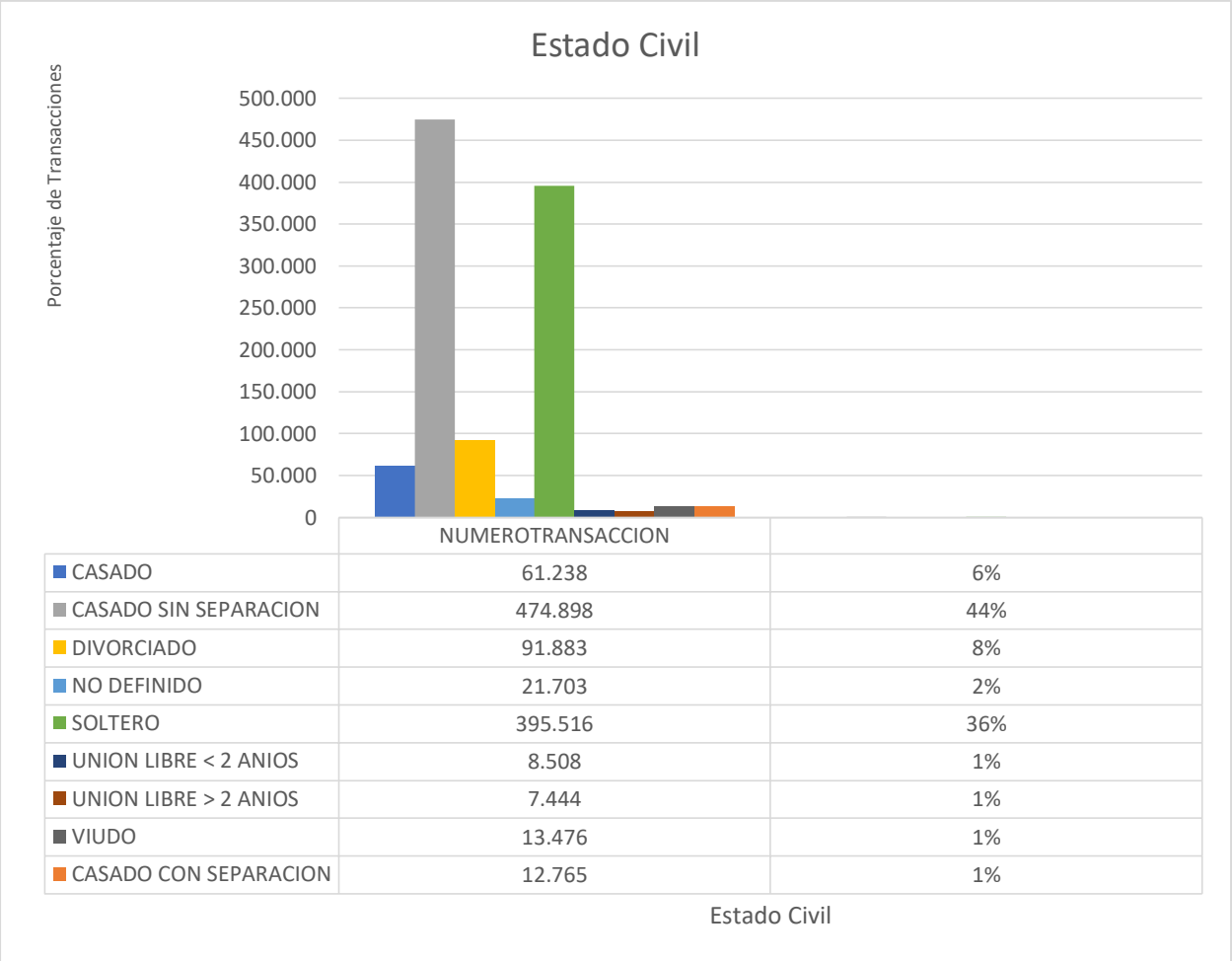


Figura 13: Exploración de estudios

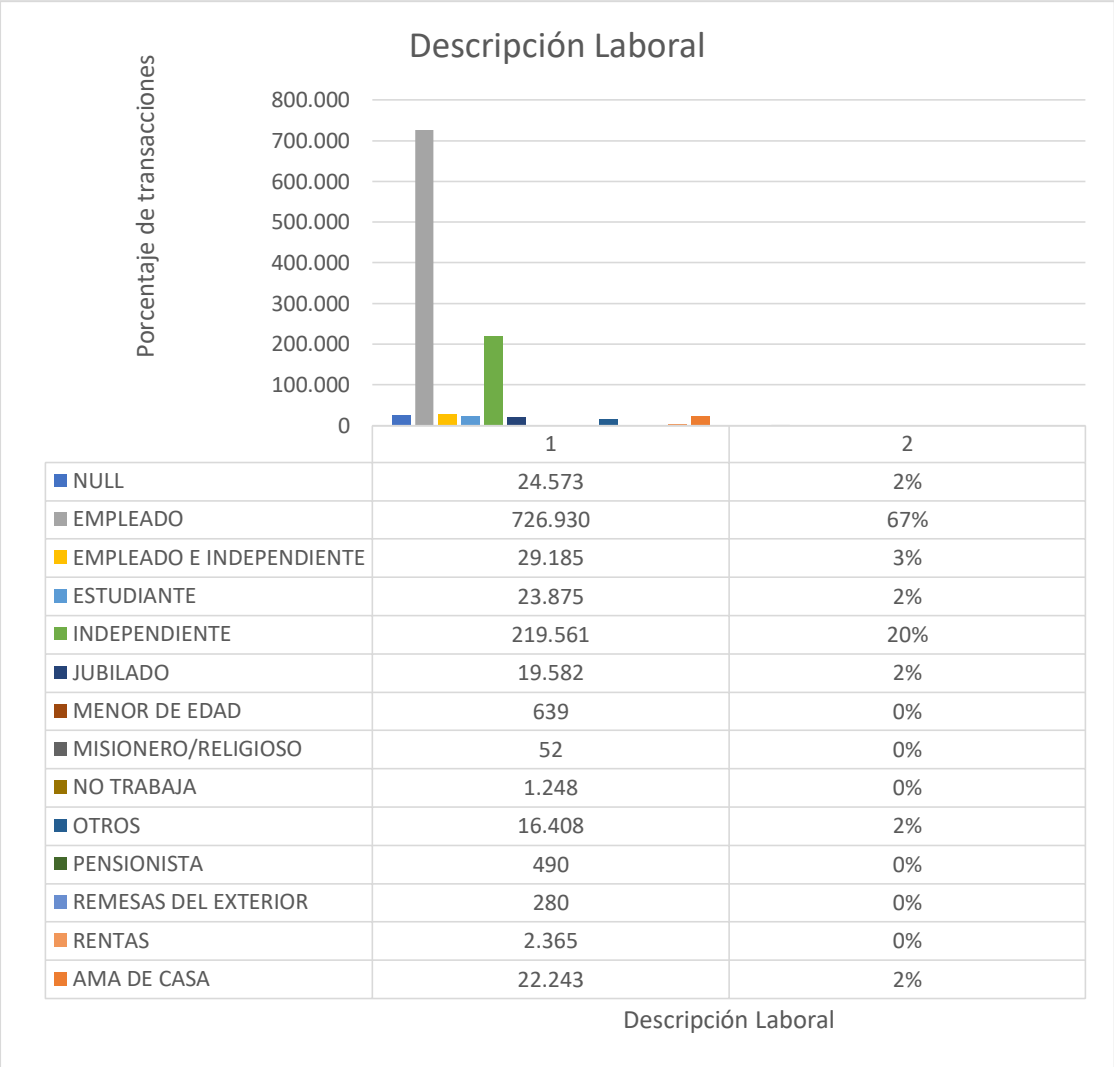
***Estado Civil*.** -Se identifica que el 44% de las transacciones son ejecutados por casados sin separación de bienes, mientras los solteros figuran en segundo lugar. (ver ilustración 14)



	NUMEROTRANSACCION	
CASADO	61.238	6%
CASADO SIN SEPARACION	474.898	44%
DIVORCIADO	91.883	8%
NO DEFINIDO	21.703	2%
SOLTERO	395.516	36%
UNION LIBRE < 2 ANIOS	8.508	1%
UNION LIBRE > 2 ANIOS	7.444	1%
VIUDO	13.476	1%
CASADO CON SEPARACION	12.765	1%

Figura 14: Exploración de Estado Civil

Descripción Laboral. **-**Se identifica que el 67% de las transacciones son realizadas por las personas que tiene relación de dependencia. (ver ilustración 15)



	1	2
NULL	24.573	2%
EMPLEADO	726.930	67%
EMPLEADO E INDEPENDIENTE	29.185	3%
ESTUDIANTE	23.875	2%
INDEPENDIENTE	219.561	20%
JUBILADO	19.582	2%
MENOR DE EDAD	639	0%
MISIONERO/RELIGIOSO	52	0%
NO TRABAJA	1.248	0%
OTROS	16.408	2%
PENSIONISTA	490	0%
REMESAS DEL EXTERIOR	280	0%
RENTAS	2.365	0%
AMA DE CASA	22.243	2%

Figura 15: Exploración de descripción laboral

***Tipo Vivienda*.** -Se identifica que el 48% de las transacciones son ejecutados por personas que viven con familiares, el 30% por las que tiene vivienda propia no hipotecada y finalmente el 13% por las que viven en vivienda alquilada. (ver ilustración 16)

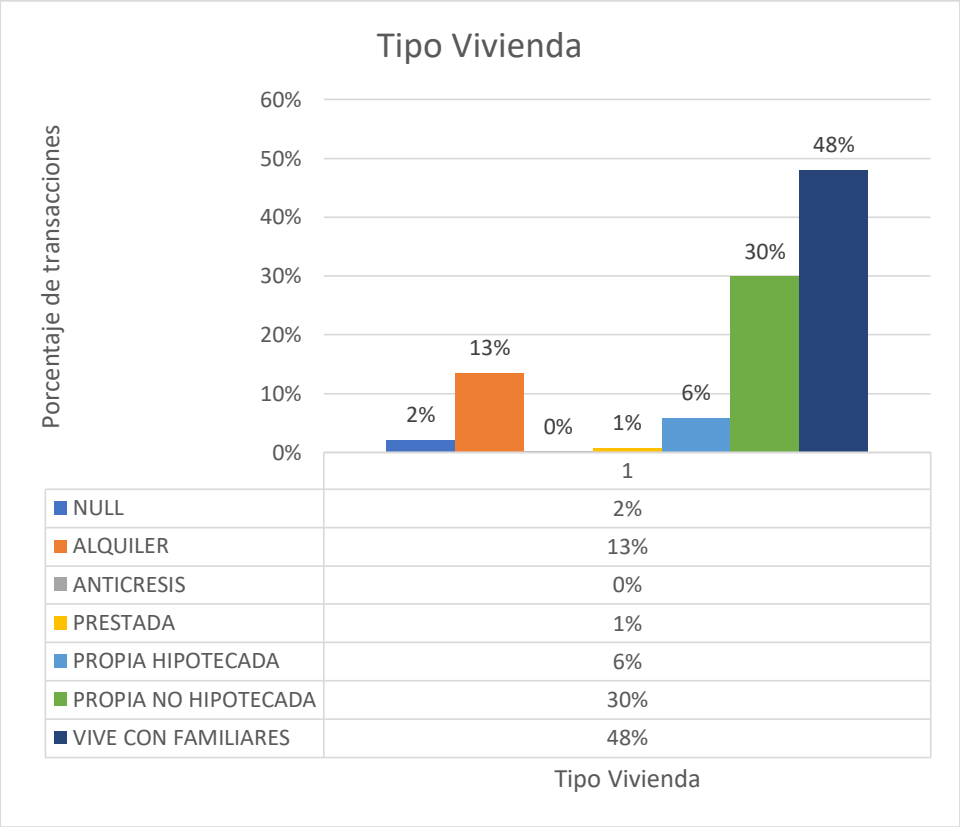


Figura 16: Exploración de tipo de Vivienda

Marca. **–** El 70% de las transacciones son ejecutadas con las tarjetas de la franquicia Visa. (ver ilustración 17)

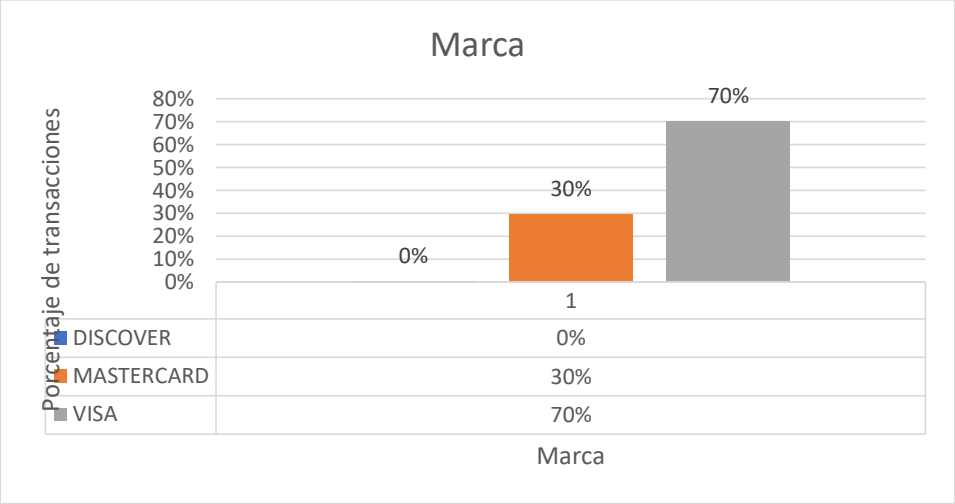


Figura 17: Exploración de marca

SubMarca. -Se identifica que la Visa Tradicional internacional ocupa el 25% de transaccionalidad. (ver ilustración 18)

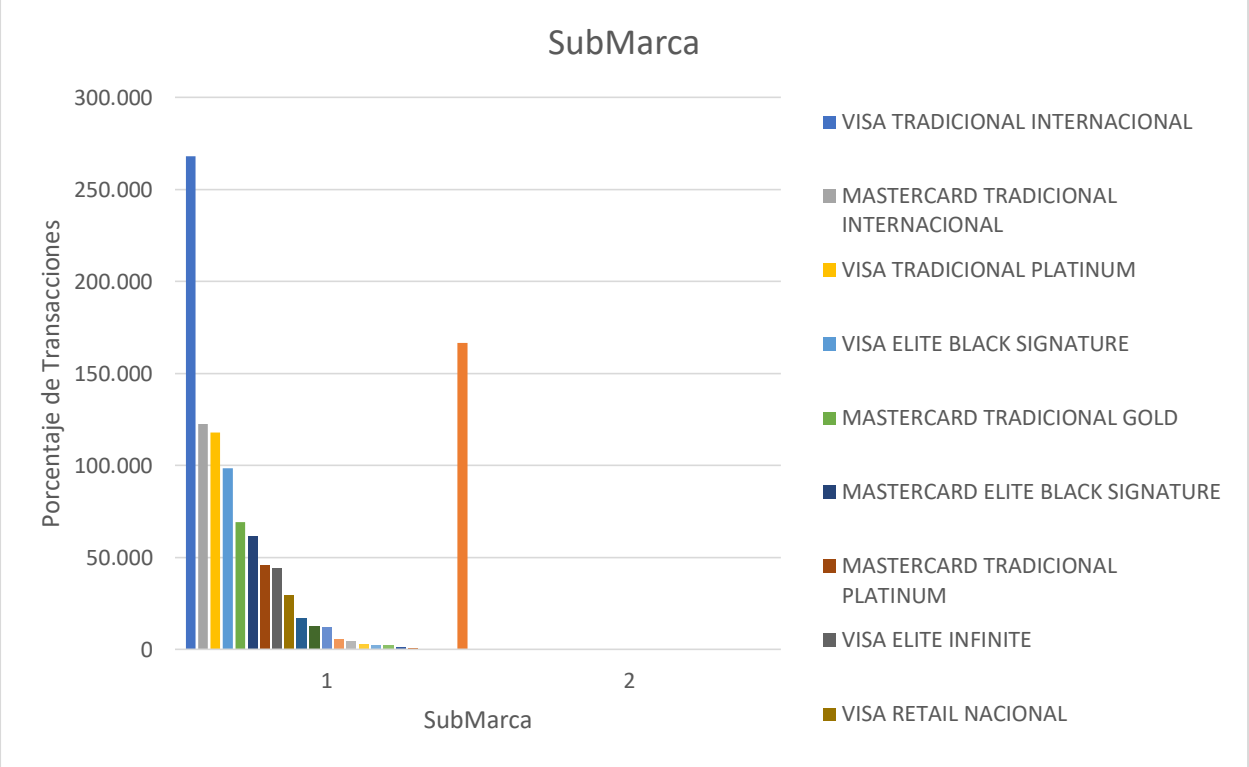


Figura 18: Exploración de SubMarca

Línea de Negocio. -El 48 % de las transacciones son ubicadas en la línea de negocio de compras del exterior. (ver ilustración 19)

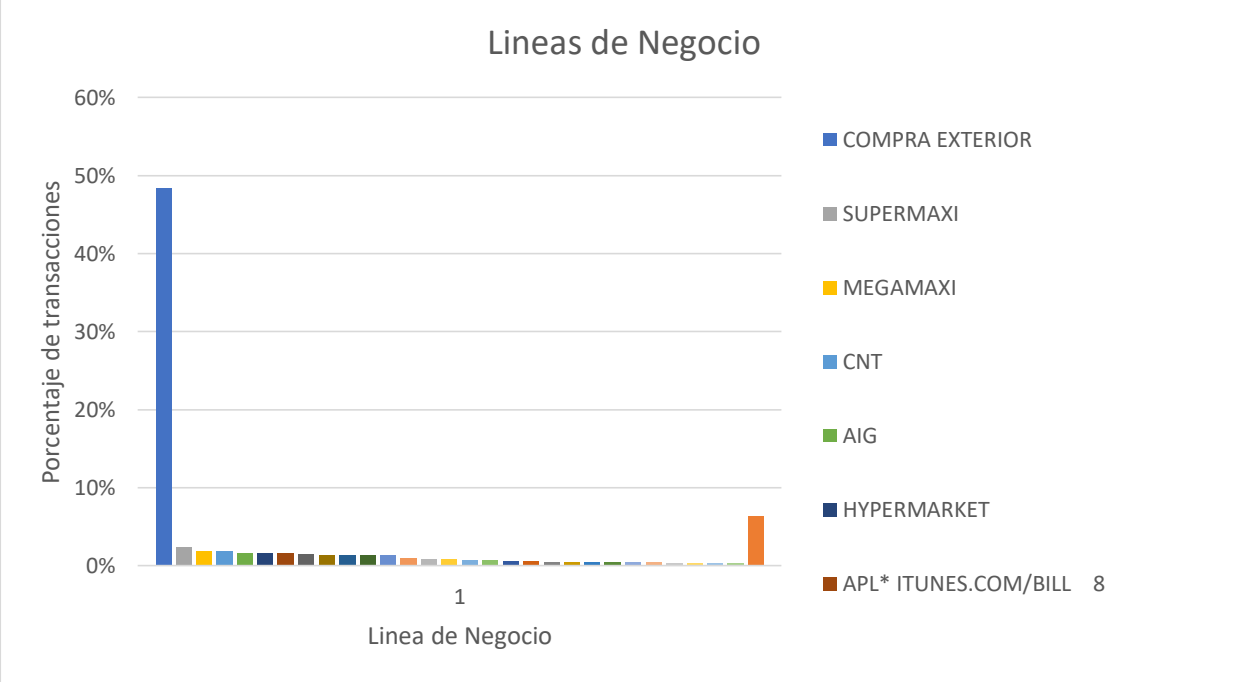


Figura 19: Exploración de línea de negocio

4.2.4 Verificación de calidad de los datos

Para validar la calidad de los datos, nos basaremos en la norma ISO/IEC20215 ya que es el estándar de seguridad de la información publicado por la Organización Internacional de Normalización, mediante el uso de una matriz, identificaremos que datos deben ser considerados para mejorar su calidad.[37]

Se analizan las características como se indica en la tabla 23 de las siguientes características:

Conformidad (CO). -Los datos deben cumplir estándares, convenciones o normativas vigentes y reglas similares referentes a la calidad de datos en un contexto de uso específico.

Confidencialidad (CN). -Los datos son accedidos e interpretados por usuarios autorizados en un contexto de uso específico.

Eficiencia (EF). **-** Los datos tienen atributos que pueden ser procesados y proporcionados con los niveles de rendimiento esperados mediante el uso de cantidades y tipos adecuados de recursos.

Precisión (PR). **-** Los datos tienen atributos que son exactos o proporcionan discernimiento en un contexto de uso específicos.

Trazabilidad (TR). **-**Los datos tienen atributos que proporcionan un camino de acceso auditado a los datos o cualquier otro cambio realizado sobre los datos en un contexto de uso específico.

Comprensibilidad (CM). **-** Los datos tienen atributos que permiten ser leídos e interpretados por los usuarios y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.

Tabla 23:
Calidad de los Datos

Datos\Características	CO	CN	EF	PR	TR	CM	Total
Género	1	1	1	1	1	1	6
Nivel de estudio	1	1	1	0	1	1	5
Estado Civil	1	1	1	0	1	1	5
Situación Laboral	1	1	1	0	1	1	5
Tipo de Vivienda	1	1	1	1	1	1	6
Marca	1	1	1	1	1	1	6
SubMarca	1	1	1	1	1	1	6
Línea de Negocio	1	1	1	1	1	1	6

Mediante la puntuación realizada se identifica que debemos mejorar la calidad de los datos de:

Nivel de Estudio. - Se identifica que el 2% de los valores están en null.

Estado Civil. - Se identifica que el 6% esta con un valor que puede ser considerado y unificado con los otros valores de la muestra.

Situación Laboral. - Se identifica que el 2% de los valores están en null.

4.3 Preparación de los datos

4.3.1 Selección de los datos

Se realiza una agrupación por fecha de consumo de las transacciones, donde se identifica que las fechas a considerar son: 2019-06-03 / 2019-06-04/ 2019-06-05. (ver ilustración 20)

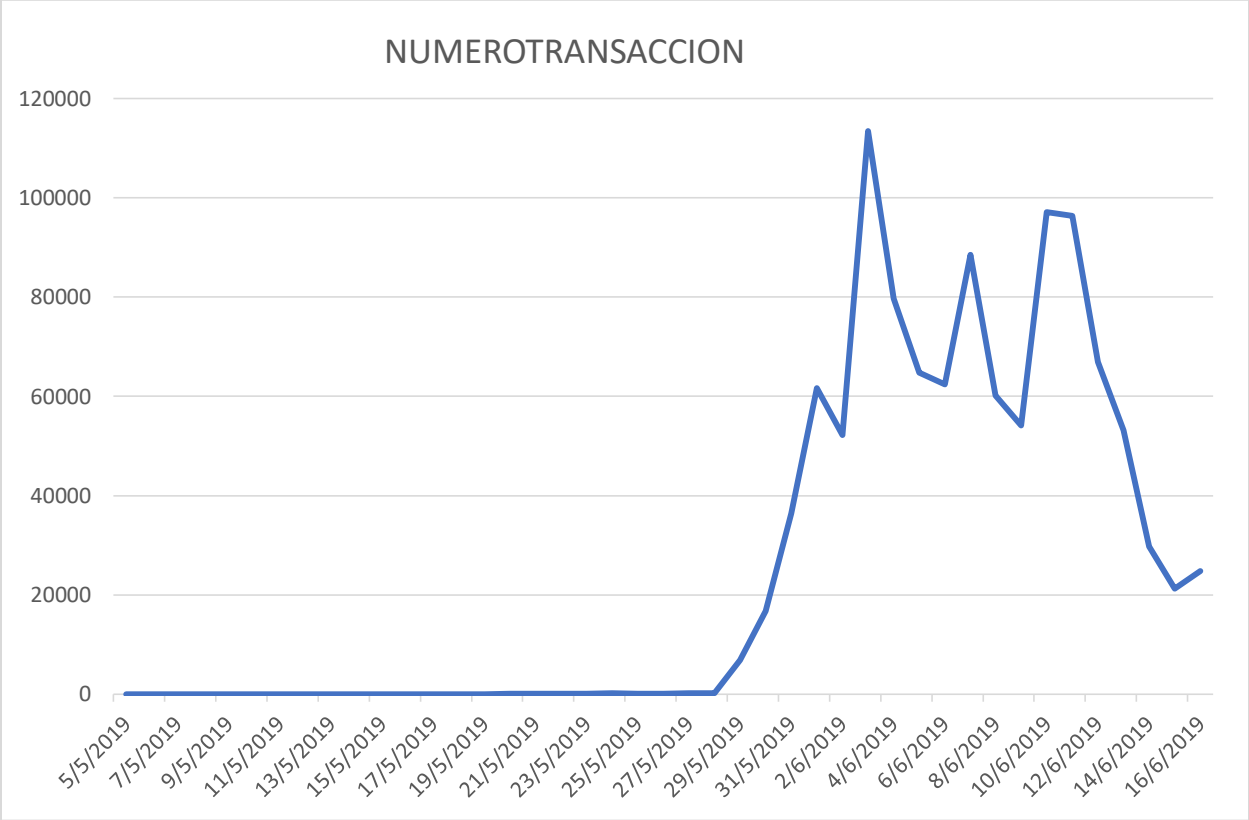


Figura 20: Fechas de los datos

Luego de la selección de las fechas a considerar se establece que no todas las submarcas deben ser consideradas, se analiza que submarcas serán útiles en nuestra muestra de acuerdo al porcentaje

de participación. Se seleccionan las submarcas que representan el 94% de la transaccionalidad como se indica en la tabla 24.

Tabla 24:
Selección de las submarcas para el modelo

SUBMARCA	NroTRX	%
VISA TRADICIONAL INTERNACIONAL	46079	0.24
VISA TRADICIONAL GOLD	28644	0.15
MASTERCARD TRADICIONAL INTERNACIONAL	22381	0.12
VISA TRADICIONAL PLATINUM	21011	0.11
VISA ELITE BLACK SIGNATURE	17003	0.09
MASTERCARD TRADICIONAL GOLD	12205	0.06
MASTERCARD ELITE BLACK SIGNATURE	10412	0.05
MASTERCARD TRADICIONAL PLATINUM	8226	0.04
VISA RETAIL NACIONAL	7937	0.04
VISA ELITE INFINITE	7135	0.04
MASTERCARD RETAIL NACIONAL	4634	0.02
VISA COMERCIAL EMPRESARIAL	2883	0.01
VISA COMERCIAL CORPORATIVA	2069	0.01
MASTERCARD COMERCIAL CORPORATIVA	735	0
DISCOVER TRADICIONAL INTERNACIONAL	536	0
VISA TRADICIONAL INTERNACIONAL MICRO	531	0
MASTERCARD COMERCIAL EMPRESARIAL	437	0
VISA RETAIL CREDIFACIL	151	0
DISCOVER CONVENIO	72	0
VISA CONVENIO	72	0
MASTERCARD TRADICIONAL INTERNACIONAL MICRO	5	0

Después de establecer que submarcas son útiles, se establece que deben ser clientes donde tenga por lo menos 3 transacciones en los distintos establecimientos.

4.3.2 Limpieza y construcción de los datos

Se procede a la limpieza de las variables de:

Estado Civil. - A los valores con Casado se les otorga el valor de Casado sin separación debido que es el valor más fuerte.

Situación Laboral. - Los valores con null se le dio el valor de Empleado.

Nivel de estudios. - Los valores de null se le otorga el valor de medios/secundarios.

Línea de Negocio. - En la variable se requieren realizar varias limpiezas ya que los Negocios pueden tener varias franquicias con distinto administrador y se guarda el nombre de la franquicia y el del administrador, se requiere para estos casos unificar los nombres. Los scripts de limpiezas están en el Anexo 1

4.3.3 Integración y Formateo de Datos

Se procede a la obtención del dataset de trabajo

Results Messages

	FechaConsumo	SubMarca	Genero	EstadoCivil	TipoVivenda	NivelEstudio	SituacionLaboral	FACTURA	LineaNegocio
1	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	FEMENINO	CASADO SIN SEPARACION	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO	F3021819	SEGUROS DEL PICHINCHA S.A
2	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	PRIMARIOS/BASICOS	INDEPENDIENTE	F3646458	SEGUROS DEL PICHINCHA S.A
3	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	FEMENINO	SOLTERO	PROPIA NO HIPOTECADA	PRIMARIOS/BASICOS	INDEPENDIENTE	F6682707	SEGUROS DEL PICHINCHA S.A
4	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	SOLTERO	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO	F2240848	SEGUROS DEL PICHINCHA S.A
5	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	SOLTERO	VIVE CON FAMILIARES	PRIMARIOS/BASICOS	EMPLEADO	F17281047	SEGUROS DEL PICHINCHA S.A
6	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	UNION LIBRE < 2 ANIOS	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO E INDEPENDIENTE	F4404583	SEGUROS DEL PICHINCHA S.A
7	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	CASADO SIN SEPARACION	ALQUILER	PRIMARIOS/BASICOS	INDEPENDIENTE	F3027144	SEGUROS DEL PICHINCHA S.A
8	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	INDEPENDIENTE	F4258427	SEGUROS DEL PICHINCHA S.A
9	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	CASADO SIN SEPARACION	VIVE CON FAMILIARES	UNIVERSITARIOS	EMPLEADO	F289905	SEGUROS DEL PICHINCHA S.A
10	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	DIVORCIADO	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F2246660	SEGUROS DEL PICHINCHA S.A
11	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	CASADO SIN SEPARACION	ALQUILER	MEDIOS / SECUNDARIOS	INDEPENDIENTE	F725941	SEGUROS DEL PICHINCHA S.A
12	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	CASADO SIN SEPARACION	PROPIA HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F741558	SEGUROS DEL PICHINCHA S.A
13	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	SOLTERO	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO E INDEPENDIENTE	F4874796	SEGUROS DEL PICHINCHA S.A
14	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	CASADO	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO E INDEPENDIENTE	F4872744	SEGUROS DEL PICHINCHA S.A
15	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	SOLTERO	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO E INDEPENDIENTE	F4877299	SEGUROS DEL PICHINCHA S.A
16	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	FEMENINO	SOLTERO	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO	F1446143	SEGUROS DEL PICHINCHA S.A
17	2019-06-03 00:00:00.000	MASTERCARD RETAIL NACIONAL	MASCULINO	CASADO SIN SEPARACION	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO	F3896035	SEGUROS DEL PICHINCHA S.A
18	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F1095742	SUPERMAXI
19	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F3068708	SUPERMAXI
20	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	FEMENINO	DIVORCIADO	VIVE CON FAMILIARES	UNIVERSITARIOS	EMPLEADO	F5158077	SUPERMAXI
21	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	MASCULINO	DIVORCIADO	PROPIA NO HIPOTECADA	UNIVERSITARIOS	EMPLEADO	F1562539	SUPERMAXI
22	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL INTERNACIONAL	FEMENINO	CASADO SIN SEPARACION	VIVE CON FAMILIARES	UNIVERSITARIOS	INDEPENDIENTE	F3424262	SUPERMAXI
23	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL GOLD	FEMENINO	SOLTERO	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F278673	SUPERMAXI
24	2019-06-04 00:00:00.000	MASTERCARD TRADICIONAL GOLD	FEMENINO	CASADO CON SEPARACI...	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	INDEPENDIENTE	F267934	SUPERMAXI
25	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL GOLD	MASCULINO	DIVORCIADO	PROPIA NO HIPOTECADA	PRIMARIOS/BASICOS	INDEPENDIENTE	F3342778	SUPERMAXI
26	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL GOLD	MASCULINO	SOLTERO	VIVE CON FAMILIARES	UNIVERSITARIOS	EMPLEADO	F2418085	SUPERMAXI
27	2019-06-04 00:00:00.000	MASTERCARD TRADICIONAL GOLD	FEMENINO	DIVORCIADO	PROPIA NO HIPOTECADA	UNIVERSITARIOS	NULL	F5041471	SUPERMAXI
28	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL GOLD	MASCULINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	FORMACION INTERME...	INDEPENDIENTE	F3071121	SUPERMAXI
29	2019-06-04 00:00:00.000	MASTERCARD TRADICIONAL GOLD	FEMENINO	CASADO SIN SEPARACION	VIVE CON FAMILIARES	MEDIOS / SECUNDARIOS	EMPLEADO	F5385516	SUPERMAXI
30	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL PLATINUM	MASCULINO	CASADO SIN SEPARACION	PROPIA HIPOTECADA	UNIVERSITARIOS	INDEPENDIENTE	F2617780	FYBECA
31	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL PLATINUM	FEMENINO	SOLTERO	VIVE CON FAMILIARES	UNIVERSITARIOS	INDEPENDIENTE	F3797841	FYBECA
32	2019-06-03 00:00:00.000	MASTERCARD TRADICIONAL PLATINUM	FEMENINO	CASADO SIN SEPARACION	PROPIA NO HIPOTECADA	MEDIOS / SECUNDARIOS	EMPLEADO	F4434052	FYBECA

Figura 21: DataSet del Modelo

4.4 Modelado

4.4.1 Selección de la técnica de modelado

De acuerdo al objetivo planteado en el apartado 1.5 de los Objetivos, se plantea la utilización de los algoritmos de Reglas de Asociación con el nodo Apriori, ya que necesitamos encontrar patrones de consumos para establecer un marketing directo.

4.4.2 Construcción del modelado

Se procederá a cargar las librerías Mlxtend de Python, para poder cargar nuestro dataset.

```
In [1]:
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

```
In [2]:
df = pd.read_csv('documents/tesis/FACTURA7.csv',engine='python')
df.head()
```

Out[2]:

	FACTURA	COMERCIO	TRANSACCIONES
0	F4570692	TRK GALEREYA S	1
1	F2652977	#+NOMBRE?	1
2	F4036538	#+NOMBRE?	1
3	F16253161	#+NOMBRE?	1
4	F5637636	002 HOUSTON'S LENX A	1

Se realiza un matriz de los diferentes comercios por cada cliente y se ponen los valores que les corresponde.

```
In [3]:
basket = (df
          .groupby(['FACTURA', 'COMERCIO'])['TRANSACCIONES']
          .sum().unstack().reset_index().fillna(0)
          .set_index('FACTURA'))
```

```
In [4]:
basket.head()
```

Out[4]:

COMERCIO	#+NOMBRE?	#21 LUCILLE'S SMOK R	#22 BRIO MILLENIA O	#265 SMOOTHIE KING H	#3 HOF'S HUT L	#30 BRIO WEST PALM P	#35 BRIO WOODLANDS T	#41 OCEAN PRIME TAMPA T	#7 LUCILLE'S SMOKE O	& CAFE XELA Q	...	Z FOOD	ZAFIRO TOURS M	ZARA	ZERIN
FACTURA															
F1002632	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
F1002659	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
F1003453	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
F1005575	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	
F1005668	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	

5 rows × 2409 columns

Los valores que tienen cero se mantienen y los que tienen un valor distinto a cero se coloca el uno, con el objetivo de tener una matriz de ceros y unos.

```
In [5]:
# Show a subset of columns
basket.iloc[:,[0,1,2,3,4,5,6,7,8,9]].head()
```

Out[5]:

COMERCIO	#+NOMBRE?	#21 LUCILLE'S SMOK R	#22 BRIO MILLENIA O	#265 SMOOTHIE KING H	#3 HOF'S HUT L	#30 BRIO WEST PALM P	#35 BRIO WOODLANDS T	#41 OCEAN PRIME TAMPA T	#7 LUCILLE'S SMOKE O	& CAFE XELA Q
FACTURA										
F1002632	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1002659	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1003453	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1005575	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1005668	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
In [6]:
def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1
```

```
In [7]:
basket_sets = basket.applymap(encode_units)
```

```
In [8]:
basket_sets.head()
```

Out[8]:

COMERCIO	#+NOMBRE?	#21 LUCILLE'S SMOK R	#22 BRIO MILLENIA O	#265 SMOOTHIE KING H	#3 HOF'S HUT L	#30 BRIO WEST PALM P	#35 BRIO WOODLANDS T	#41 OCEAN PRIME TAMPA T	#7 LUCILLE'S SMOKE O	& CAFE XELA Q	...	Z FOOD	ZAFIRO TOURS M	ZARA	ZERIN
FACTURA															
F1002632	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
F1002659	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
F1003453	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
F1005575	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
F1005668	0	0	0	0	0	0	0	0	0	0	...	0	0	0	

5 rows × 2409 columns

Con del dataset listo se procede a la ejecución del algoritmo Apriori. Donde nos muestra el soporte de los comercios frente a todas las transacciones.

```
In [31]: frequent_itemsets = apriori(basket_sets, min_support=0.1, use_colnames=True)
```

```
In [32]: frequent_itemsets.head()
```

Out[32]:

	support	itemsets
0	0.743021	(COMPRA EXTERIOR)
1	0.161056	(DIRECTV)
2	0.184909	(FYBECA)
3	0.231602	(GASOLINERAS)
4	0.153443	(MEGAMAXI)

Finalmente se obtienen ocho reglas de asociacion que deben ser analizadas.

```
In [54]: rules = association_rules(frequent_itemsets, metric="lift", min_threshold=0.01)
         rules.sort_index()
```

Out[54]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(COMPRA EXTERIOR)	(DIRECTV)	0.743021	0.161056	0.113179	0.152322	0.945775	-0.006489	0.989697
1	(DIRECTV)	(COMPRA EXTERIOR)	0.161056	0.743021	0.113179	0.702731	0.945775	-0.006489	0.864465
2	(FYBECA)	(COMPRA EXTERIOR)	0.184909	0.743021	0.120453	0.651418	0.876715	-0.016938	0.737211
3	(COMPRA EXTERIOR)	(FYBECA)	0.743021	0.184909	0.120453	0.162113	0.876715	-0.016938	0.972793
4	(GASOLINERAS)	(COMPRA EXTERIOR)	0.231602	0.743021	0.155980	0.673484	0.906413	-0.016105	0.787033
5	(COMPRA EXTERIOR)	(GASOLINERAS)	0.743021	0.231602	0.155980	0.209927	0.906413	-0.016105	0.972566
6	(SUPERMAXI)	(COMPRA EXTERIOR)	0.190323	0.743021	0.123160	0.647111	0.870918	-0.018254	0.728214
7	(COMPRA EXTERIOR)	(SUPERMAXI)	0.743021	0.190323	0.123160	0.165756	0.870918	-0.018254	0.970552

4.5 Evaluación

Luego de aplicar las reglas de Asociación se obtuvo la siguiente salida.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(COMPRA EXTERIOR)	(DIRECTV)	0.743021	0.161056	0.113179	0.152322	0.945775	-0.006489	0.989697
1	(DIRECTV)	(COMPRA EXTERIOR)	0.161056	0.743021	0.113179	0.702731	0.945775	-0.006489	0.864465
2	(FYBECA)	(COMPRA EXTERIOR)	0.184909	0.743021	0.120453	0.651418	0.876715	-0.016938	0.737211
3	(COMPRA EXTERIOR)	(FYBECA)	0.743021	0.184909	0.120453	0.162113	0.876715	-0.016938	0.972793
4	(GASOLINERAS)	(COMPRA EXTERIOR)	0.231602	0.743021	0.155980	0.673484	0.906413	-0.016105	0.787033
5	(COMPRA EXTERIOR)	(GASOLINERAS)	0.743021	0.231602	0.155980	0.209927	0.906413	-0.016105	0.972566
6	(SUPERMAXI)	(COMPRA EXTERIOR)	0.190323	0.743021	0.123160	0.647111	0.870918	-0.018254	0.728214
7	(COMPRA EXTERIOR)	(SUPERMAXI)	0.743021	0.190323	0.123160	0.165756	0.870918	-0.018254	0.970552

Se puede observar que existen dos reglas con valores de lift cerca de 1, la teoría indica que si el lift < 1 están negativamente correlacionadas, si lift=1 son independientes y si lift >1 están positivamente correlacionadas.

El valor de la confianza nos indica el porcentaje en el que se cumple la regla en el dataset.

El valor del soporte nos indicará el porcentaje de presencia de los comercios en la probabilidad de visita en otro comercio.

Regla 0

Se tienen que las personas que consumen en el comercio Compra Exterior, van adquirir Directv, si bien el lift es 0,94 muy cercano a 1, se muestra que el soporte del comercio antecedente es alto por lo cual se deduce que es muy probable que las personas que realizan compras en el exterior adquieran Directv.

Evaluación de la regla: coherente

Regla 1

Se tienen que las personas que consumen en el comercio Directv realizan compras en el exterior, si bien el soporte del antecedente es bajo, este se apalanca en el soporte del consecuente que es alto y determinaría como se puede aplicar.

Evaluación de la regla: coherente

Regla 2

Se tienen que las personas que consumen en el comercio Fybeca realizan compras en el exterior, si bien el soporte del antecedente es bajo, este se apalanca en el soporte del consecuente que es alto y determinaría como se puede aplicar.

Evaluación de la regla: coherente

Regla 3

Se tienen que las personas que consumen en el comercio de compras en el exterior, adquieren el mismo día productos en Fybeca, ya que la presencia del soporte del antecede es alto, esto quiere decir que hay varias personas que realizan la acción descrita en la regla.

Evaluación de la regla: coherente

Regla 4

Se tienen que las personas que consumen en el comercio de Gasolineras, realizan compras en el exterior, en el mismo día, como se evidencia el soporte que tiene la presencia del comercio de Compras en el Exterior es muy fuerte comparada con otros comercios.

Evaluación de la regla: coherente

Regla 5

Se tienen que las personas que consumen Gasolineras, son muy pocas en comparación con las que consumen en comercios en el exterior

Evaluación de la regla: coherente

Regla 6

Se tienen que las personas que consumen en el comercio de Supermaxi, consumen en el comercio de Compras en el exterior, a pesar que son pocas las que realizan esta transacción el mismo día.

Evaluación de la regla: coherente

Regla 7

Se tienen que las personas que consumen en Compra en el exterior, muy seguro consumirán en Supermaxi.

Evaluación de la regla: coherente

4.6 Resultados y análisis

Después de la ejecución del modelo para encontrar las reglas de asociación se identificó que existe un alto soporte de Compra en el exterior que es del 0.7, lo que indica que para poder hacer un marketing directo de las otras líneas de negocio se pueden utilizar métodos de publicidad en páginas que tengan la opción de compras en el exterior, como puede ser Amazon o ebay.

De acuerdo a los resultados obtenidos también se puede realizar un marketing directo de los comercios en el exterior en Fybeca, Directv y Supermaxi, por ejemplo, se puede promocionar tarjetas de regalo de Amazon en los comercios mencionados.

CAPITULO V

CONCLUSIONES Y TRABAJOS FUTURO

5.1 Conclusiones

- Se acepta la hipótesis planteada.

- La utilización de la técnica del SMS aportó con la obtención de artículos científicos que se utilizaron para resolver la problemática propuesta, ya que estos abordan los temas de metodologías, herramientas y técnicas de minería de datos.

- La metodología CRISP – DM, ayudó al entendimiento del negocio para poder focalizar de una manera correcta la selección y discriminación de las variables en el modelo de asociacion Apriori.

- Las reglas de asociación encontradas no llegaron a ser mayores a un lift de 1, sin embargo, permitieron encontrar reglas que para la hipótesis planteada son útiles al momento de realizar un marketing directo.

- En la construcción del dataset y obtención del conjunto de datos para realizar el modelo, se determinó que las tarjetas no son utilizadas en la mayoría de compras como un medio de pago electrónico, esto quiere decir que la población en el Ecuador utiliza en la mayoría de sus compras el efectivo.

- La utilización de la herramienta Anaconda Python ayudo a procesar una gran cantidad de datos, y a establecer las reglas de asociación más relevantes.

5.2 Trabajos Futuros

- Se debe realizar un proceso de automatización de la extracción de las variables, su transformación y carga con el objetivo de disponibilizar el dataset que se requiere para que el modelo pueda ser ejecutado continuamente de acuerdo al periodo de fechas que se desee realizar.

- El dataset construido permite aplicar otro tipo de modelos que permitan realizar diferentes estrategias de marketing directo, por lo cual a partir de la automatización de la generación del dataset, este proceso sería más rápido.

- Se deberá realizar una medición de las campañas de marketing directo realizadas a partir de las reglas encontradas con el fin de establecer la confianza del modelo.

- Publicar los resultados de la tesis.

BIBLIOGRAFIA

Bhargava, N., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. 1115.

ANACONDA. (n.d.). *https://www.anaconda.com/why-anaconda/*. Retrieved from https://www.anaconda.com/why-anaconda/.

Azevedo, A., & Santos, M. (2008). KDD, SEMMA AND CRISP-DM: A parallel overview. 2.

Bahari, F., & Elayidom, S. (2015). An Efficient CRM-Data Mining Framework for the Prediction of. *Procedia Computer Science*, 725-731.

Berlanga, V., Rubio Hurtado, M., & Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. 65.

Castillo Vicente, E. G., & Montalvo Rosero, E. M. (2012). *Plan Estratégico Tendiente a Mejorar los Niveles de Servicio.* UNIVERSIDAD CENTRAL DEL ECUADOR.

Castro Heredia, L. C. (2019). Density-based clustering methods for unsupervised separation of partial. *International Journal of Electrical Power & Energy Systems*, 224-230.

CX Ling, C. L.-K. (1998). Data mining for direct marketing: Problems and solutions. 1'2.

El-Houssainy A, R., & Ayman, A. (2019). Prediction of kidney disease stages using data mining algorithms. *Artificial Intelligence in Medicine*, 2-3.

Gartner. (2019). Retrieved from https://www.gartner.com/it-glossary/business-intelligence-bi/

iso25000. (n.d.). *https://iso25000.com/index.php/normas-iso-25000/iso-25012?limit=5&start=10*. Retrieved from https://iso25000.com/index.php/normas-iso-25000/iso-25012?limit=5&start=10.

Jaime Camposa, P. S. (2017). A big data analytical architecture for the Asset Management. 371.

Kamber, M., & Pei, J. (2006). *Data Mining: Concepts and techniques.* Illinois: Elsevier.

KIMBALL. (2016). Kimball Dimensional. 1.

Klawonn, F. (2016). Exploring data sets for clusters and validating single clusters. *Procedia Computer Science*, 1381-1390.

López, C. P. (2007). Minería de datos: técnicas y herramientas. 497.

Louis Davidson, . M. (2016). Pro SQL Server Relational Database Design and Implementation. 706.

Mastercard. (2019). Retrieved from Mastercard: https://www.mastercard.es/es-es/establecimientos/comience-a-aceptar/proceso-pago.html

Odun-Ayo, I., Goddy-Worlu, R., & Yahaya, J. (2019). A systematic mapping study of cloud policy languages and programming. *Journal of King Saud University - Computer and Information Sciences*, 3.

Parrilla, J. M. (2014). Cómo hacer inteligente su negocio: business intelligence a su alcance. 110.

Petersen, Feldt, Mujtaba, & Mattsson. (2008). Systematic Mapping Studies in. 1-10.

Pierre-Léo Bourbonnais, C. M. (2018). A robust datawarehouse as a requirement to the increasing quantity. 439.

Presser Carne, C. (2009). Data Data Mining Mining. 4.

R.J. Kuo, C. S. (2007). Association rule mining through the ant colony system for National. 1306-1307.

RAPIDMINER. (n.d.). *https://rapidminer.com/*. Retrieved from https://rapidminer.com/.

Rodríguez, I. (2006). *Principios y estrategias de marketing: (incluye web).* UOC.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). 220.

Shankru, G., & Vijayakumar, K. (2018). Non-sequential partitioning approaches to decision tree classifier. 276.

Sohn, F. V. (2012). Data Warehousing: The Ultimate Guide to Building Corporate Business Intelligent. 77.

Stattner, E., & Martine , C. (2015). Descriptive Modeling of Social Networks. *Procedia Computer Science*, 227.

Syarifah, D., & Heruna, T. (2018). Linear regression model using bayesian approach for energy performance of residential building. *Procedia Computer Science*, 672-673.

Trujillo, S. L.-M. (2003). A Comprehensive Method for Data Warehouse Desing. 2.

W.H. Inmon, . L. (2014). Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse. 131.

WEKA. (n.d.). *https://www.cs.waikato.ac.nz/ml/weka/*. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/.

WEKA. (n.d.). *https://www.cs.waikato.ac.nz/ml/weka/*. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data. 3.