UNIVERSIDAD TÉCNICA DEL NORTE

Facultad de Ingeniería en Ciencias Aplicadas

Carrera de Software

IMPLEMENTACIÓN DE TÉCNICAS DE MINERÍA DE DATOS Y VISUALIZACIONES UTILIZANDO INTELIGENCIA DE NEGOCIOS PARA LA TOMA DE DECISIONES EN EL COMERCIAL CADENA CASANOVA

Trabajo de Grado previo a la obtención del título de Ingeniero de Software

Autor:

Leslie Mishell Armas Uvidia

Director:

PhD. Iván Danilo García Santillán

Ibarra – Ecuador 2023

UNIVERSIDAD TÉCNICA DEL NORTE

BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1004198782		
APELLIDOS Y NOMBRES:	ARMAS UVIDIA LESLIE MISHELL		
DIRECCIÓN:	IBARRA, ARTURO HIDALGO Y QUITO		
EMAIL:	lmarmasu@utn.edu.ec		
TELÉFONO FIJO:	(06) 2547 203	**TELÉFONO MÓVIL:**	0991671290

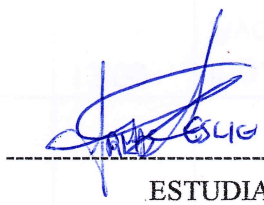
DATOS DE LA OBRA	
TÍTULO:	IMPLEMENTACIÓN DE TÉCNICAS DE MINERÍA DE DATOS Y VISUALIZACIONES UTILIZANDO INTELIGENCIA DE NEGOCIOS PARA LA TOMA DE DECISIONES EN EL COMERCIAL CADENA CASANOVA
AUTOR(ES):	LESLIE MISHELL ARMAS UVIDIA
FECHA:	19 de Abril del 2023
PROGRAMA:	PREGRADO
TÍTULO POR EL QUE OPTA:	INGENIERA DE SOFTWARE
DIRECTOR:	PhD. Iván García
ASESOR 1:	Msc. Marco Pusdá
ASESOR 2:	Msc. Pedro Granda

2. CONSTANCIAS

El autor (es) manifiesta (n) que la obra objeto de la presente autorización es original y se la desarrolló sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es (son) el (los) titular (es) de los derechos patrimoniales, por lo que asume (n) la responsabilidad sobre el contenido de esta y saldrá (n) en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, 10 de mayo de 2023

EL AUTOR:



ESTUDIANTE
Leslie Mishell Armas Uvidia
C.I: 1004198782

CERTIFICACIÓN DIRECTOR

Ibarra, 19 de abril del 2023

CERTIFICACIÓN DIRECTOR DEL TRABAJO DE TITULACIÓN

Por medio del presente yo García Santillán Iván Danilo, certifico que el Srta. Leslie Mishell Armas Uvidia portador de la cédula de ciudadanía número 1004198782, ha trabajado en el desarrollo del proyecto de grado **"IMPLEMENTACIÓN DE TÉCNICAS DE MINERÍA DE DATOS Y VISUALIZACIONESUTILIZANDO INTELIGENCIA DE NEGOCIOS PARA LA TOMA DE DECISIONES EN EL COMERCIAL CADENA CASANOVA"**, previo a la obtención del Título de Ingeniera en Software, esté trabajo se ha realizado con interés profesional y responsabilidad que certifico con honor de verdad.

Atentamente

PhD. García Iván
DIRECTOR DE TRABAJO DE GRADO

DEDICATORIA

Este trabajo lo dedico a mi Madre Elvia Vizcaíno, quién me brindó amor, calidez y apoyo en la vida que compartí con ella.
A mi abuelita Rosa Armas quién en vida dedicó tantos años a mi cuidado.

Han sido mi impulso
Leslie

AGRADECIMIENTO

A mis padres, Viviana Uvidia y Patricio Armas por brindarme la oportunidad de estudiar y el ejemplo de seguir adelante. A mi tío William García por enseñarme y enfocarme en el desarrollo de la carrera.

A mi abuelita Fanny Reascos que a pesar de estar lejos estuve pendiente de mí en todo aspecto.

A mi pareja Cristian Checa por darme aliento para alcanzar esta meta y ser mi soporte en el proceso de este trabajo.

A mis compañeros Francisco, Yamilex, Stiphen, Stalin y Brandon por haber hecho este camino más fácil y darme el apoyo necesario para seguir estos años.

Al docente Msc. Mauricio Rea, por todo lo aprendido en sus asignaturas y sobre todo por su amistad.

Al docente Msc. Marco Pusdá por brindarme la accesibilidad al desarrollo de este proyecto, como también su tolerancia.

Al docente PhD. Iván García, quiero agradecerle por su paciencia y conocimientos durante todo el proceso de elaboración de mi trabajo de titulación.

Muchas gracias a todos,
Leslie

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

RESUMEN

El presente proyecto se centra en el uso de técnicas de minería de datos y visualizaciones mediante inteligencia de negocios para la toma de decisiones en la empresa Comercial Cadena Casanova. Esto implica aplicar métodos para analizar grandes cantidades de datos y extraer información útil para la empresa, que puede utilizarse para mejorar la toma de decisiones en diferentes áreas del negocio, como el incremento de ventas, segmentación de clientes, marketing para ofertas.

El Comercial Cadena Casanova, que se dedica a la venta de productos al por mayor y menor, ha proporcionado los datos en lo que respecta a ventas, clientes y productos. Los datos receptados provienen del sistema llamado SICOF por medio de su base de datos en MS Access.

El progreso de este proyecto busca brindar información adicional que responda a las necesidades más relevantes de información y que pueda ser considerado como un recurso estratégico para la empresa. Como también, su propósito es proveer los indicadores necesarios para apoyar la toma de decisiones informadas.

Para el progreso de este trabajo se llevaron a cabo revisiones bibliográficas de los principales conceptos relacionados con la Inteligencia de Negocios, Técnicas de minería de datos con sus respectivas métricas cuantitativas, metodologías para llevar a cabo proyectos y estándares para garantizar la calidad de los datos, como la norma ISO/IEC 25012. Se utilizaron diversas herramientas y sistemas como MS Access, WEKA, Power BI, Python, Excel y Google Colab, los cuales fueron fundamentales para el resultado final. La metodología empleada para la realización en los respecta a Inteligencia de Negocios y Minería de datos fue CRISP-DM

Palabras Claves: Minería de Datos, Inteligencia de Negocios, Norma ISO/IEC 25012, Metodología CRISP-DM, Power BI, WEKA

ABSTRACT

This project focuses on the use of data mining techniques and visualizations through business intelligence for decision-making in the Comercial Cadena Casanova company. This involves applying methods to analyze massive amounts of data and extract useful information for the company, which can be used to improve decision-making in different areas of the business, such as increasing sales, customer segmentation, and marketing for offers.

Comercial Cadena Casanova, which sells products wholesale and retail, has provided data regarding sales, customers, and products. The data was collected from the SICOF system through its MS Access database.

The goal of this project is to provide additional information that responds to the most relevant information needs and can be considered a strategic resource for the company. Its purpose is also to provide the necessary indicators to support informed decision-making.

For the progress of this work, bibliographic reviews of the key concepts related to Business Intelligence, data mining techniques with their respective quantitative metrics, methodologies for project management, and standards to ensure data quality were conducted, such as the ISO/IEC 25012 standard. Various tools and systems were used, such as MS Access, WEKA, Power BI, Python, Excel, and Google Colab, which were fundamental for the final result. The methodology used for the project was CRISP-DM, which is a commonly used methodology in Business Intelligence and Data Mining.

Keywords: Data Mining, Business Intelligence, ISO/IEC 25012 Standard, CRISP-DM Methodology, Power BI, WEKA."

INTRODUCCIÓN

Tema

Implementación de Técnicas de Minería de Datos y visualizaciones utilizando Inteligencia de Negocios para la toma de decisiones en el Comercial Cadena Casanova.

Problema

Antecedentes

Usualmente en Ecuador, las microempresas o pequeñas empresas no hacen (o muy limitado) un estudio adecuado del negocio que les permita aumentar las ventas, así como también no realizan un presupuesto adecuado que permita el crecimiento de esta. Por lo que, la mejor opción para evitar pérdidas económicas, incluso hasta llegar a la quiebra, es realizar estrategias con la aplicación de Minería de Datos e Inteligencia de negocios para lograr la toma de decisiones correctas principalmente para mejorar los rendimientos financieros.

El mercado ecuatoriano está despertando, hacia mejorar su entendimiento y adopción de técnicas y metodologías más refinadas de analíticas de datos. Las organizaciones están en la necesidad de comprender cómo descomponer y extraer el valor monetario escondido en los datos de manera sostenible (Vargas, 2020).

Situación Actual

El comercial Cadena Casanova se dedica a la venta de víveres de consumo masivo, aseo, limpieza, confitería entre otros; también distribuye productos de marcas reconocidas en el Ecuador como son: NESTLÉ, COCA-COLA, FAMILIA, LA FABRIL, QUISUR, PRODUMAR. (Lema, 2016)

Esta empresa maneja un sistema de gestión empresarial SICOF, que proporciona información primordial sobre la situación actual del negocio. Sin embargo, estos datos son desaprovechados ya que no les dan su respectiva importancia para su análisis y explotación, lo que contribuiría a elevar el promedio de ventas.

Prospectiva

Se puede establecer que la empresa Cadena Casanova no realiza un análisis de datos correspondiente, por lo tanto, requiere de implantar el uso de análisis de datos a través de inteligencia de negocios y minería de datos para identificar patrones, establecer nivel de venta de productos, que contribuya a la toma de decisiones y futuras acciones.

Problema

El Comercial Cadena Casanova, tiene una escasa aplicabilidad de la cultura de datos en su modelo organizacional. Actualmente, no utiliza métodos que pueden contribuir a incrementar sus ventas, atracción de nuevos clientes, promociones y publicidad a medida, etc. Esto puede influir en la competitividad de la empresa y la expansión del negocio, conllevando a un estancamiento económico. A continuación, se puede observar en la Fig.1.

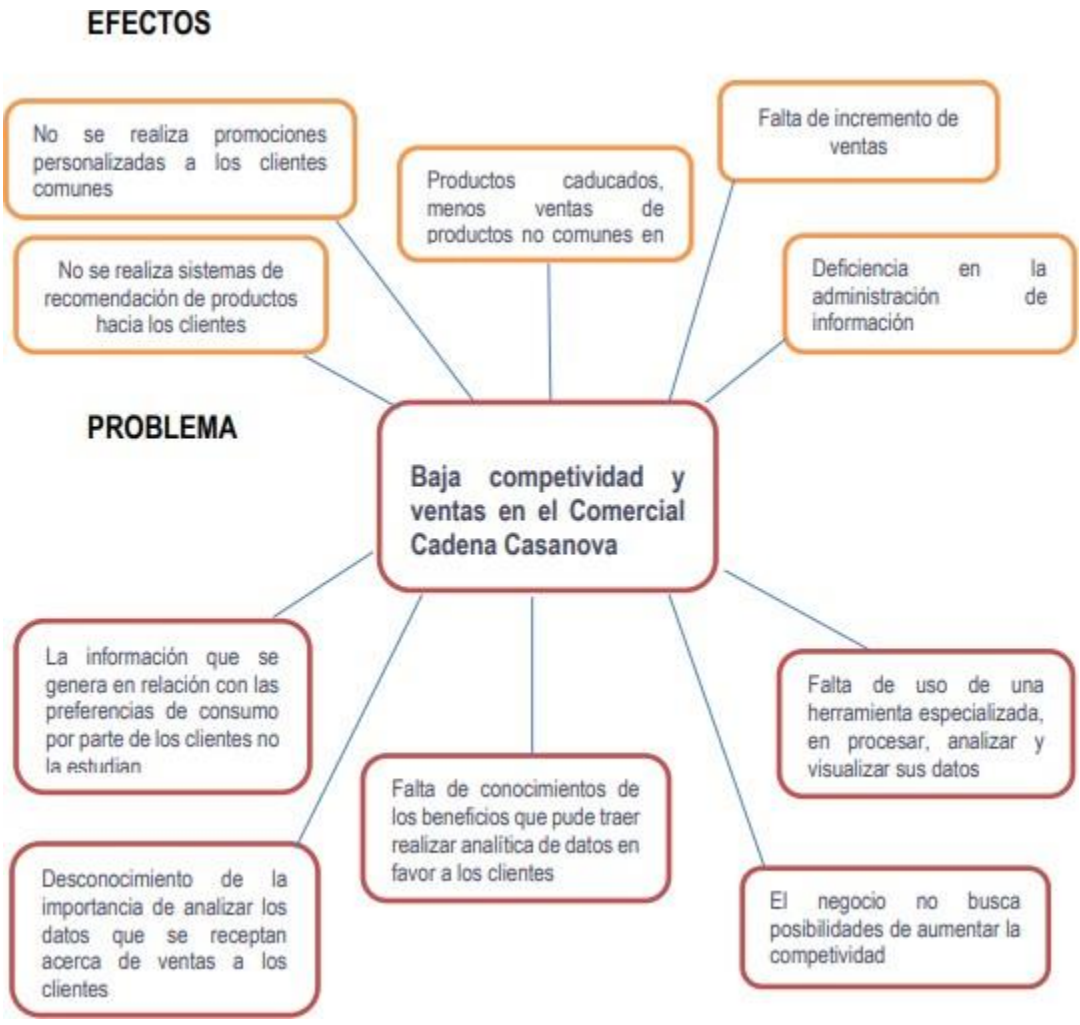


Fig. 1.Árbol de problemas (Elaboración propia)

Objetivos

Objetivo General

Implementar Técnicas de Minería de Datos y visualizaciones utilizando inteligencia de negocios para la toma de decisiones en el Comercial Cadena Casanova

Objetivos Específicos

- Revisar técnicas análisis de datos de visualización, agrupamiento y/o regresión como sustento teórico para el estudio.
- Implementar algoritmos de regresión y/o agrupamiento para segmentación de clientes y estimación de ventas utilizando Weka, y creación de visualizaciones en Power BI.
- Validar los modelos de regresión y/o agrupamiento utilizando métricas cuantitativas de rendimiento y las preguntas del negocio.

Justificación

El Covid-19 trajo consigo la quiebra de varios negocios, pero también el uso más usual de la tecnología. Es por ello, que las empresas han optado por usar la información que pueden generar los datos, algo que es primordial para saber las altas y bajas que tiene el negocio, conocer sus clientes más usuales, en si el estado en el que se encuentra la empresa.

Aplicar Minería de datos en esta empresa brindará una mayor organización de datos, estadísticas reales y predicciones que hará posible la toma de decisiones mucho más rápida. De esta manera hacemos efecto en el objetivo 8 “Promover el crecimiento económico inclusivo y sostenible, el empleo y el trabajo decente para todos”. (ONU,2015) Este objetivo representa el propósito del proyecto que es contribuir a incrementar los ingresos por medio de las ventas realizadas

La implementación de este proyecto también hace referencia al Objetivo 9, “Construir infraestructuras resilientes, promover la industrialización sostenible y fomentar la innovación”. (ONU,2015), al aplicar las nuevas tecnologías, el uso de BI y Data Mining. Se vincula con este objetivo ya que se proporciona una solución a la empresa para contribuir a aumentar el volumen de ventas.

Justificación Tecnológica. - La idea primordial para la toma de decisiones correctas en el negocio planteado es contar con información de datos históricos y realizar su respectivo análisis con las técnicas de Data Mining, las cuales generará patrones y dará a conocer la situación, y tendencias de los distintos años antes del Covid-19 y después, para así marcar una diferencia y lograr el cambio que se espera.

Justificación Teórica. - El presente proyecto considera los conceptos de Minería de datos, Business Intelligence, toma de decisiones, los cuales serán utilizados para que los datos receptados se conviertan en información importante para la empresa. La solución que se propone crear en este proyecto permitirá analizar y explotar la información creando reportes de acuerdo con las necesidades presentes y futuras.

Justificación Metodológica. – Se utilizará métricas cuantitativas (accuracy, precisión, recall, F1-score, AUC, etc.), para analizar el rendimiento, precisión y confiabilidad de los resultados obtenidos, reprimiendo así la gran cantidad de fuentes de potenciales errores

Alcance

El presente proyecto tiene como propósito implementar, analizar la información y datos que genera la empresa Comercial Cadena Casanova, estos datos servirán para crear la conciencia del uso debido de análisis de datos.

Esto permitirá desarrollar reportes, Dashboards y aplicar minería de datos, todo esto recopilando los datos de los últimos 2 años. Indagar sobre las ventas de artículos, los costos, ganancias, fiabilidad de los clientes y contestar preguntas del negocio para comprender la situación actual y tomar las decisiones correctas para el desempeño futuro.

El sistema SICOF provee los datos, ya que es el sistema que usan actualmente en la empresa, en conjunto con su base de datos en MS Access, de los cuales se obtienen archivos en formato .mdb.

Un paso primordial es la preparación de datos, utilizando el proceso de ETL (extracción, transformación y carga de datos). Esto generará un DataWarehouse limpio para realizar las visualizaciones y aplicar ciertos algoritmos predictivos y descriptivos (regresión, agrupamiento y asociación) los cuales serán útiles para acceder a datos claves de forma sencilla, organizando así la información para realizar el análisis propuesto.

En lo que conlleva al aspecto minería de datos se realizará basado en CRISP-DM (Cross-Industry Standard Process for Data Mining, 2000) en conjunto con las técnicas necesarias para ejecutar el modelo de estimación de ventas y la segmentación de clientes. Estos algoritmos se realizan en la herramienta Python, con las librerías respectivas para realización de patrones y predicciones.

Otra herramienta para utilizar es Power BI para realizar las visualizaciones y reportes necesarios, que muestren de una manera didáctica los resultados a obtener y ayuden a contestar preguntas del negocio.

Se utilizará el estándar ISO/IEC 25012 para la calidad de los datos obtenidos en los que se basan la toma de decisiones y de esta manera evitar datos erróneos que impedirán el desarrollo correcto del proyecto.

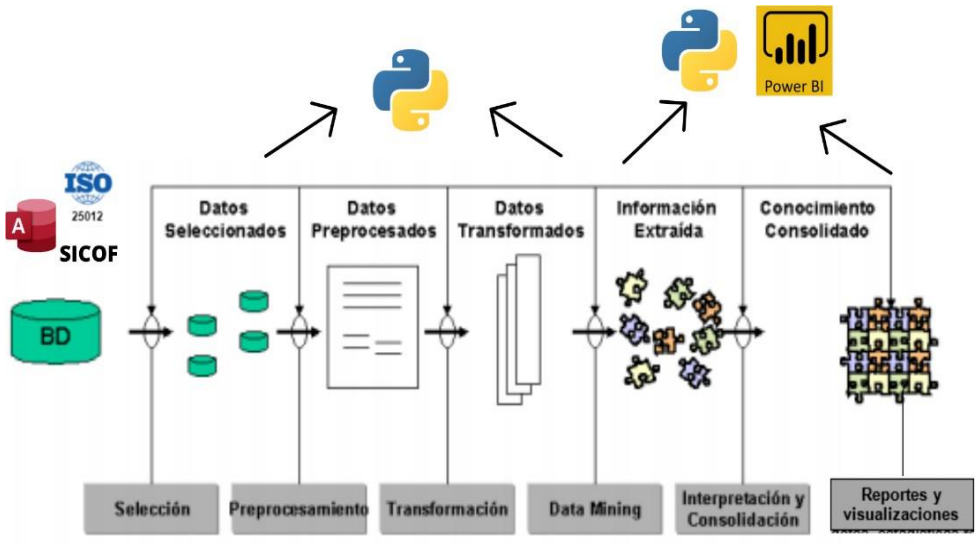


Fig. 2. Adaptación Técnicas de análisis de datos (J.M, 2018)

Fig. 3. Metodología CRISP-DM (Álvarez. 2020)

CAPÍTULO 1

Marco Teórico

1.1. Comercial Cadena Casanova

Los comienzos del Comercial Cadena Casanova se pronuncian brevemente por (Lema, 2016), donde comenta

Comercial Cadena Casanova inicia sus actividades económicas el 24 de agosto del 2012 como persona natural no obligado a llevar contabilidad, en la venta al por mayor y menor de productos diversos ubicada en la Provincia de Imbabura en la ciudad de Ibarra (Lema, 2016).

La empresa en el transcurso del año 2012 incrementó su capital de trabajo por el aumento de las ventas y el crecimiento de la misma, dejando de ser una persona natural no obligada a llevar contabilidad y pasando a ser una persona obligada a llevar contabilidad, cumpliendo con ciertas obligaciones tributarias establecidas por el SRI., la empresa se caracteriza por mantener un lazo comercial con sus clientes ofreciéndoles productos adecuados a su entera satisfacción y necesidad como pueden ser víveres de consumo masivo, aseo, limpieza, confitería entre otros, también distribuye productos de marcas reconocidas en el Ecuador como son: NESTLÉ, COCA-COLA, FAMILIA, LA FABRIL, QUISUR, PRODUMAR (Lema, 2016).

La organización empresarial es crucial para establecer la estructura jerárquica y definir los procesos necesarios para alcanzar las metas planificadas a largo plazo. Existen diferentes tipos de estructuras organizativas, cada una con sus propias ventajas y desventajas, que permiten a las empresas adaptarse a sus necesidades específicas y maximizar su eficiencia y productividad, se detalla la estructura a través del diagrama siguiente.

Fig. 4. Estructura organizacional (Elaboración Propia)

1.2. Introducción a Data Mining y Business Intelligence

1.2.1. Minería de Datos

La generación masiva de datos tanto que sean científicos, comerciales, económicos y de salud, debido a los avances tecnológicos, hoy en día la exploración y análisis de estos datos proporcionan una gran fuente de información y conocimiento.

La minería de datos es el campo en donde se descubren nuevas y sobresalientes relaciones, tendencias, como también patrones al momento de examinar un gran Dataset por medio de métodos y algoritmos que se enfocan en la visualización, análisis y modelización de información.

1.2.1. Relación con otras áreas

La relación de la minería de datos con otras áreas, en conjunto logran una mayor eficacia logrando que este sea un campo multidisciplinar

- **Base de Datos**: Interviene almacenes de datos como también el análisis en línea (OLAP), por medio de esto se adquiere información útil y comprensible.
- **Estadística:** Se considera el eje fundamental para la minería de datos, ya que a partir de esto se proporcionó varios algoritmos y técnicas.

- **Visualización de datos:** Este campo ayuda al usuario a descubrir, intuir y entender patrones que en algunas ocasiones serían difíciles de comprender.
- **Sistemas para la toma de decisiones:** Constituye un conjunto de sistemas y herramientas informatizados en donde su objetivo es proveer información necesaria para que el usuario tome una decisión correcta.
- **Aprendizaje Automático:** Es un subcampo de la Inteligencia Artificial en donde se desarrolla algoritmos capaces de aprender, este en conjunto con la estadística proporciona un análisis inteligente de los datos
- **Otras disciplinas:** Hoy en día con el gran alcance que ha tenido la tecnología Data Mining interviene en varias áreas o disciplinas, las cuales son útiles para el análisis de imagen, procesamiento de señales, etc.

1.2.2. Inteligencia de Negocios

Inteligencia de Negocios es el recurso que utiliza las organizaciones o empresas que están modernizadas, ya que con este puede contar para sacar provecho a un nivel superior de la información que se ha recopilado de sus clientes, proveedores, ventas en algunos casos y también de sus competidores, esto con el objetivo de lograr grandes ventajas en un mercado hostil y dinámico.

Utilizar BI es primordial ya que con esta las organizaciones logran tomar buenas decisiones, gracias a la conversión de la información en conocimiento, logrando así una mejora en el desarrollo organizacional de la empresa. Y trayendo grandes beneficios:

- Tangibles: Estos beneficios pueden traer captación de clientes, mayor ingreso de ventas, generación de ingresos, eliminar la sobreproducción de productos y también aumentar el desempeño de los empleados.
- Intangibles: Logra la toma de decisiones adecuadas por lo que puede ocasionar que la satisfacción de clientes aumente, se controla de manera más efectiva la información adquirida.
- Estratégicos: Impulsa a las empresas hacía que tipo de mercado debería enfocarse como también es posible analizar los clientes con mayor potencial

1.2.2.1. Elementos

Estos elementos son factores clave en la eficacia de los procesos de negocio

- **Los datos:** Es un conjunto de valores o elementos que por sí solos no tienen importancia y no son orientativos para alguna acción.
- **La información:** La información se refiere a un grupo de datos que han sido analizados y comprendidos, otorgándoles un significado concreto en términos de relevancia, propósito y contexto, logrando así que sean útiles e importantes para la toma de decisiones
- **El conocimiento:** Este es el conjunto de valores e información por lo que sería útil para las acciones futuras de la empresa.

1.2.3. Calidad de Datos - ISO/IEC 25012

En un proyecto de minería de datos o inteligencia de negocios, la calidad de los datos es de suma importancia. Por esta razón, se utiliza la norma ISO/IEC 25012 (Modelo de Calidad de Datos), que proporciona un formato estructurado para detallar requisitos, establecer medidas y llevar a cabo evaluaciones de calidad de datos. Esta norma se divide en características que dependen de los puntos de vista: inherentes y dependientes del sistema, y su importancia y uso dependen de las necesidades específicas del proyecto. En resumen, la norma ISO/IEC 25012 es una herramienta valiosa para garantizar la calidad de los datos en proyectos de minería de datos o inteligencia de negocios, y se divide en características que se adaptan a diferentes perspectivas y necesidades.

1.2.3.1. Calidad de datos Inherente

Se refiere al nivel de calidad que se encuentra intrínseco en los datos mismos. Se determina por el grado en que las características que miden la calidad de los datos cumplen con las necesidades explícitas e implícitas de los usuarios en un contexto específico. Se toman en cuenta los valores de dominio, posibles restricciones y relaciones de valores de datos y metadatos en este proceso (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).

1.2.3.2. Calidad de Datos Dependiente del Sistema

Se refiere al nivel de calidad de los datos que se logra y mantiene a través del uso de componentes tecnológicos específicos en los sistemas informáticos. En este caso, la calidad de los datos depende de la capacidad de los componentes del sistema para cumplir con los requisitos tecnológicos necesarios para procesar, almacenar y gestionar los datos de manera adecuado (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

En la siguiente te figura se muestra las características y su división de acuerdo con los puntos de vista

Característica	**Definición**	**Inherente**	**Dependiente del Sistema**
Exactitud	Los datos representan correctamente el valor deseado en un contexto específico	X	
Completitud	Los datos obligatorios estén completos	X	
Consistencia	Hace referencia al nivel en que los atributos de los datos no presentan contradicciones y son compatibles con otros datos dentro de un contexto concreto de uso	X	
Credibilidad	Es el contexto de autenticidad, en donde se observa si los datos son correctos y ciertos	X	
Actualidad	Los datos deben estar actualizados	X	
Accesibilidad	Es la medida de la facilidad de acceso de los datos en una situación determinada		X
Conformidad	Adecuación de los datos a los requerimientos y reglas establecidos para su uso y manejo en un contexto específico.	X	X
Confidencialidad	Aseguran que solamente es accesible e interpretable por usuarios autorizados	X	X
Eficiencia	Nivel de desempeño que se espera en el procesamiento y suministro de datos.	X	X
Precisión	Los datos tienen atributos que son exactos	X	X
Trazabilidad	Proveen una auditoría de intento de acceso y de cualquier cambio hecho	X	X
Comprensibilidad	Los datos tienen atributos que permiten ser leídos e interpretados por los usuarios	X	X
Disponibilidad	Los datos tienen atributos que le permiten ser recuperados por usuarios		X

Portabilidad	Analiza si los datos pueden ser copiados, reemplazados o eliminados al realizar un cambio de un sistema a otro, preservando el nivel de calidad		X
Recuperabilidad	Se comprueba que los datos mantienen y preservan un nivel de operaciones en caso de Fallos		X

Tabla 1. Calidad de datos (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

El proceso para llevar a cabo la evaluación de las características seleccionadas consta de cinco actividades (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).

- **Establecer los requisitos de la evaluación**: Se trata de identificar claramente qué se espera lograr con la evaluación y qué información se necesita recopilar para alcanzar esos objetivos (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).
- **Especificar la evaluación:** Definir en detalle los aspectos técnicos y metodológicos de la evaluación, incluyendo los módulos de evaluación, las métricas, las herramientas y las técnicas que se utilizarán para recopilar y analizar la información relevante (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).
- **Diseñar la evaluación:** Utiliza un enfoque metodológico adecuado y planificado cuidadosamente las tareas y los recursos necesarios (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).
- **Ejecutar la evaluación:** Implica llevar a cabo el plan diseñado previamente, utilizando las técnicas y herramientas especificadas para recopilar y analizar la información relevante (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).
- **Concluir la evaluación:** Concluir la evaluación implica el análisis y la interpretación de los resultados obtenidos en la etapa anterior, con el objetivo de llegar a conclusiones sobre el desempeño o el éxito en relación con los objetivos establecidos (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019).

1.3. Metodologías de Data Mining: CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) es la metodología más empleada en los últimos años para el desarrollo de proyectos de minería de datos en donde describe la manera en cómo se aborda el problema.

Esta metodología incluye un modelo y guía, en donde están constituidos de seis fases, aquí se debe tomar en cuenta que la sucesión no puede ser consecutiva por lo que se puede volver a una anterior en caso de que se desee revisar

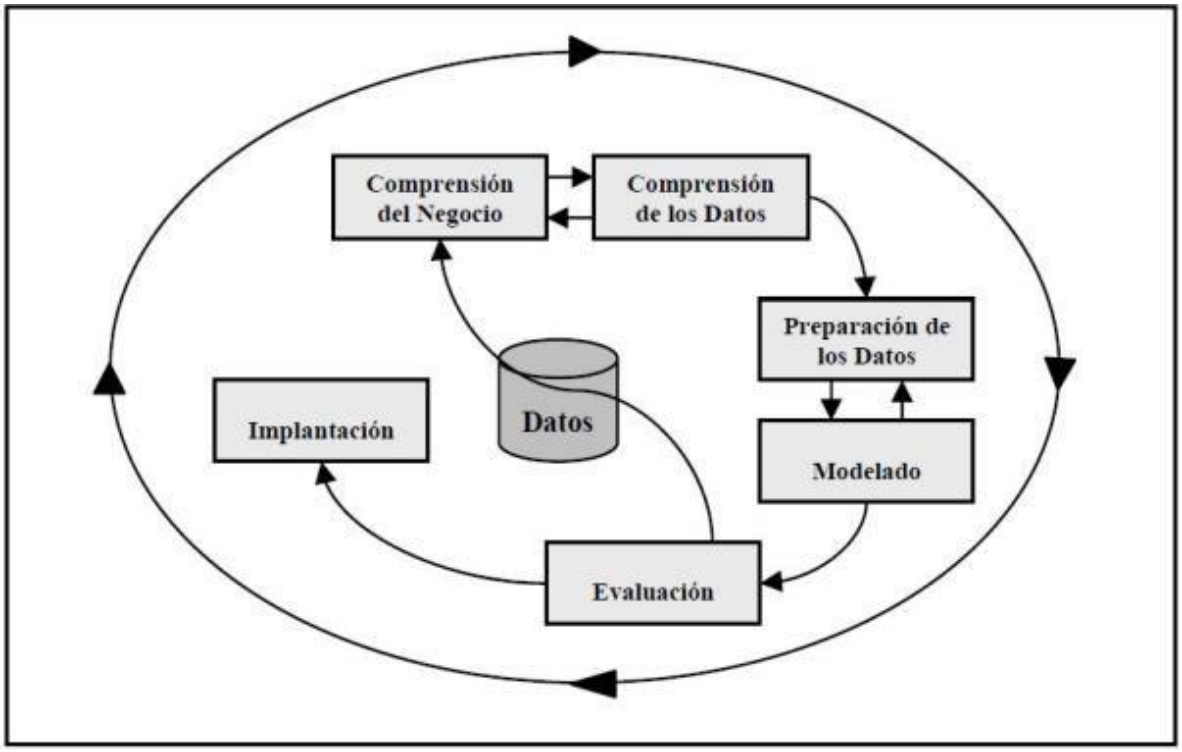


Fig. 1 Modelo de proceso CRISP–DM ([CRISP-DM, 2000)

1.3.1. Comprensión del negocio

Es la primera fase de CRISP-DM la cual podría considerar las más importante, en donde se comprende los objetivos y requisitos del proyecto a realizar todo esto desde una perspectiva de negocio, empresarial o institucional todo para conllevar a la creación de objetivos técnicos y un plan de proyecto. La comprensión del negocio es esencial ya que, si no se logra esta primera fase, aunque se desarrolle el algoritmo más robusto, los resultados no serán fiables (Arancibia, 2009)

Es por ello por lo que se debe entender el problema que se desea resolver, para así realizar una correcta recolección de datos e interpretar con éxito los resultados. (Arancibia, 2009)

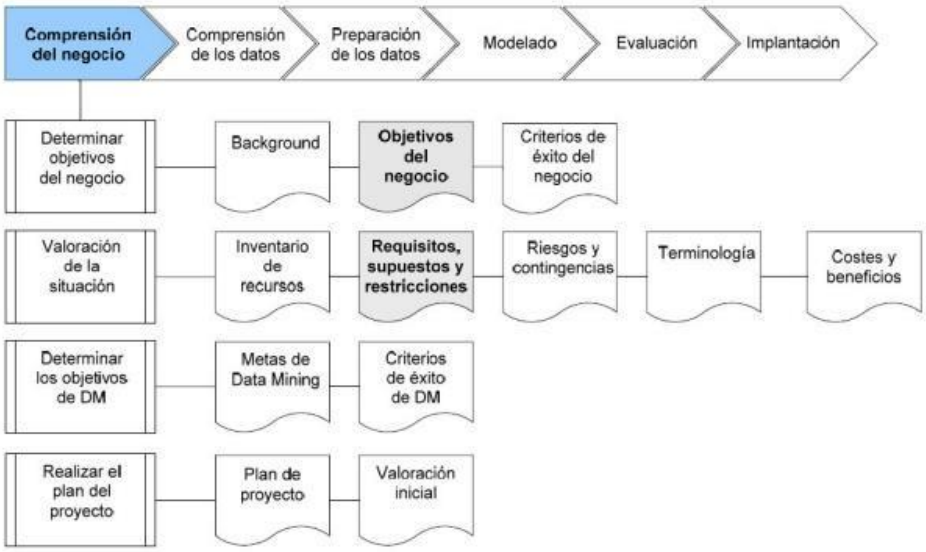


Fig. 6. Fase de comprensión del negocio (CRISP-DM,2000)

- **Determinar los objetivos del negocio:** La tarea inicial de un proyecto es establecer metas que reflejen los objetivos del negocio, definir el problema a resolver y los criterios de éxito para evaluar el impacto del uso de la minería de datos (Arancibia, 2009).
- **Evaluación o Valoración de la situación:** En esta tarea se especifica el estado actual antes de implementar el procedimiento propuesto de minería de datos, en donde se considera aspectos como: ¿Qué información previa se tiene sobre el problema en cuestión? ¿Existe la cantidad suficiente de datos necesarios para resolver el problema?, etc. (Arancibia, 2009).

  También proponer los requisitos del problema, todo esto con el objetivo de poder comparar con la situación antigua y la posterior después de la aplicación de DM para así medir el grado de éxito del proyecto (Arancibia, 2009).
- **Determinar los objetivos de la minería de datos:** El fin de esta tarea es representar los objetivos de la empresa, negocio o institución como metas del proyecto de minería de datos (Arancibia, 2009).

- **Realizar el plan del proyecto:** Esta es la última tarea de la primera fase de CRISP-DM el objetivo es desarrollar un plan del proyecto, en donde represente los pasos y técnicas que se van a implementar (Arancibia, 2009).

1.3.2. Compresión de los datos

Es la segunda fase de la metodología en donde se realiza la recolección inicial de los datos con el fin de relacionar con el problema y así conocer a profundidad la calidad y relaciones que permitan limitar las primeras hipótesis (Arancibia, 2009).

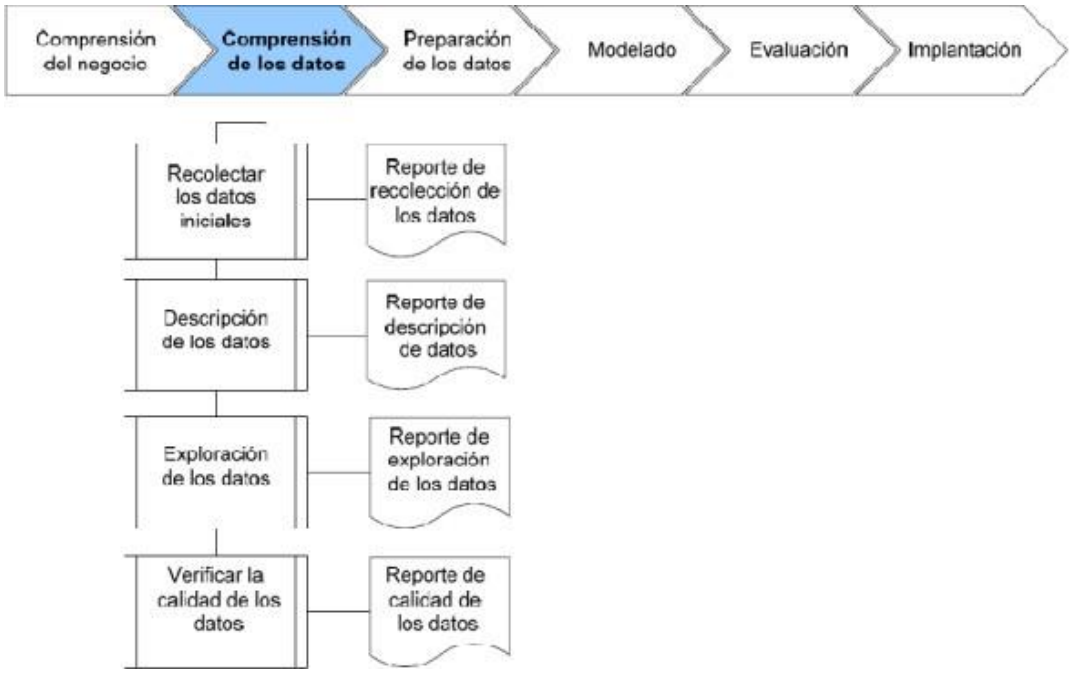


Fig. 2. Fase de comprensión de los datos (CRISP-DM,2000)

- **Recolección de datos iniciales:** Se recopilan los datos iniciales y se los prepara para su procesamiento posterior. Se lleva a cabo una identificación constante de las fuentes de datos, su ubicación, las técnicas empleadas en su recolección (Arancibia, 2009).
- **Descripción de los datos:** Una vez obtenido los datos iniciales, estos deben ser descritos, de esta manera se identifica el tipo, volúmenes de datos, formato, el significado de cada campo (Arancibia, 2009).

- **Exploración de los datos:** Continuamente de realizar la descripción de los datos, se procede a su exploración en donde se aplica pruebas estadísticas básicas que permiten saber sus propiedades de los datos en donde es necesario crear un informe de exploración de los datos con el fin de entenderlos de la mejor manera posible (Arancibia, 2009).
- **Verificar la calidad de los datos:** El fin de la última tarea de esta fase es la corrección de los datos por medio de verificaciones para así encontrar valores fuera de rango, la cantidad, la consistencia de valores individuales de los campos y así evitar problemas en el proceso (Arancibia, 2009).

1.3.3. Preparación de los datos

Consecutivamente de la comprensión de datos se efectúa la tercera fase de CRISP-DM, en donde se realiza la preparación de los datos para ser utilizados en las técnicas de minería de datos, se transforma los datos con el fin de utilizarlos en el modelado, mediante limpieza de datos, cambios de formato, adicionar valores, etc. Esto se realiza de acuerdo con la técnica de modelado a aplicar (Arancibia, 2009).

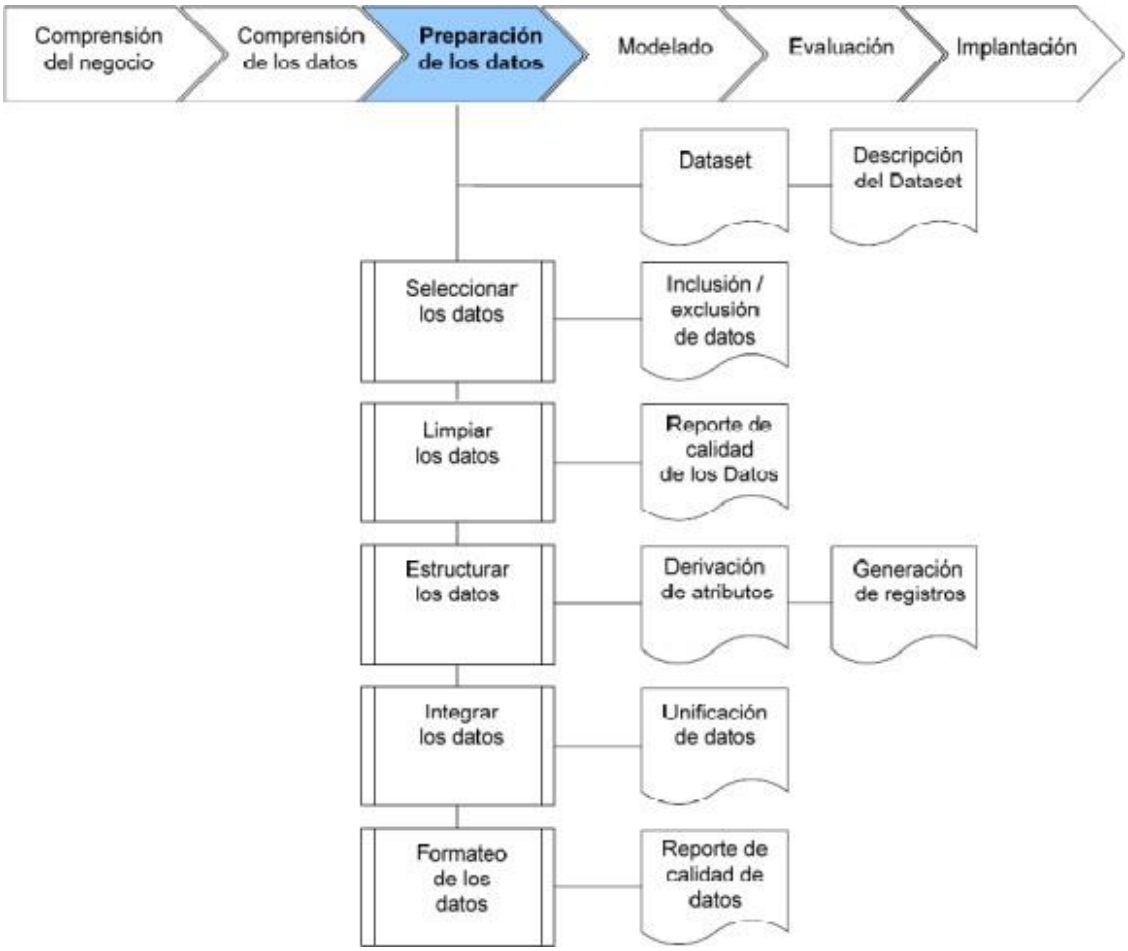


Fig. 8 Fase de preparación de los datos (CRISP-DM,*2000)*

- **Seleccionar los datos**

  En esta tarea se realiza un análisis con respecto a la calidad de valores, la corrección, tipo o volumen para así realizar una selección y formar un subconjunto con los datos necesarios. (Arancibia, 2009).

- **Limpiar los datos**

  Esta etapa puede consumir demasiado tiempo, debido a la aplicación de varias técnicas como la normalización de datos, tratamiento de valores, reducción de volumen, tratamiento de elementos duplicados; esto con el objeto de sean útiles al implementar en el modelado (Arancibia, 2009).

- **Construir los datos**

Generar datos a partir de los ya existentes, nuevos registros y también nuevas tablas o fusiones para que potencie la capacidad predictiva y la detección de comportamientos (Arancibia, 2009).

- **Integrar los datos**

  Desarrolla nuevas estructuras a partir de los datos a partir de los datos seleccionados

- **Formateo de los datos**

  Realizar cambios en la estructura o formato en los datos sin que altere su significado para así sea de manera más sencilla aplicar alguna técnica de minería de datos (Arancibia, 2009).

1.3.4. Modelado

En esta etapa de CRISP-DM se elige las técnicas de modelado según los criterios

- Estar acorde al problema
- Tener datos adecuados
- Acatar con los requisitos que se obtuvieron
- Tiempo correcto para lograr un modelo
- Comprensión de la técnica

Antes de realizar el modelado se debe realizar un método para evaluar los modelos para establecer el grado de adecuación de cada uno, a partir de esto se procese a la generación y evaluación del modelo dependiendo de las características de los datos de precisión (Arancibia, 2009).

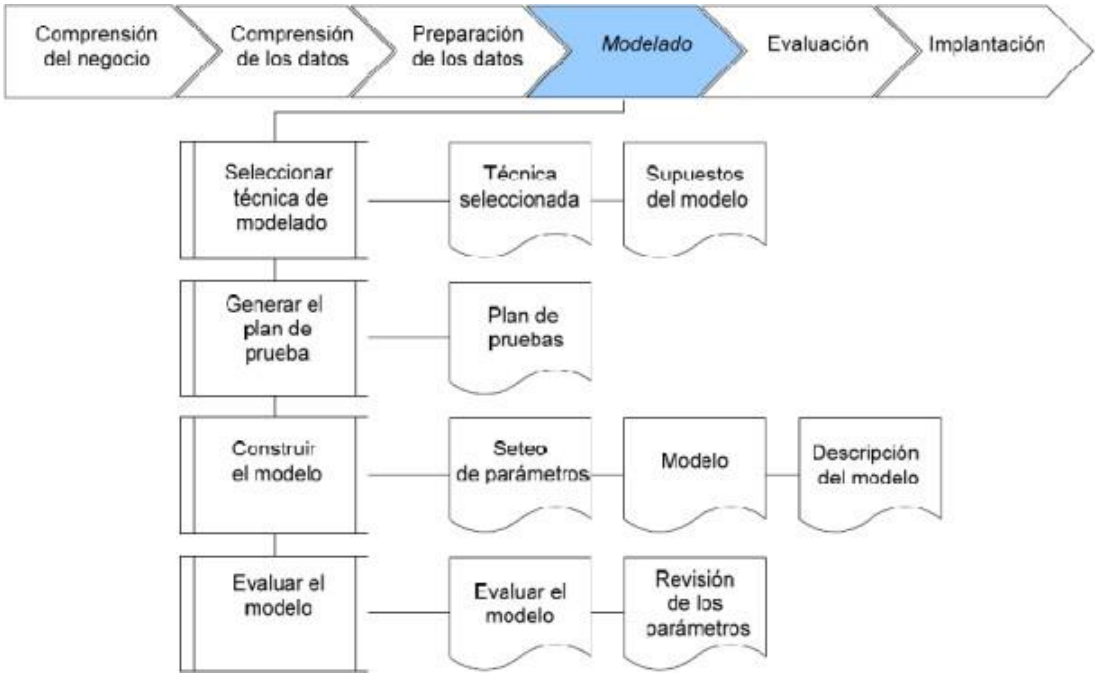


Fig. 9 Fase de modelado (CRISP-DM, 2000)

- **Escoger la técnica de modelado:** En esta etapa se debe seleccionar la técnica de minería de datos que se va a utilizar, Para ello, se considera el objetivo y meta del proyecto, también las herramientas de minería de datos que están a nuestro alcance, así mismo el dominio de la técnica elegida (Arancibia, 2009).
- **Generar el plan de prueba:** Se desarrolla un procedimiento para verificar la calidad y validez del modelo a realizar, aquí se divide la muestra entre datos de entrenamiento y validación, así se usa la razón de error para medir la calidad (Arancibia, 2009).
- **Construir el modelo:** Después de seleccionar la técnica con sus respectivas características, ejecutamos sobre los datos preprocesados para así generar uno o más modelos mediante un proceso iterativo de modificación de parámetros. Una vez realizado se debe interpretar los resultados y su rendimiento (Arancibia, 2009).

- **Evaluar el modelo:** Es la última tarea de esta fase en donde se analiza e interpretan los modelos, en base al dominio de las técnicas de Data Mining y criterios de éxito (Arancibia, 2009).

1.3.5. Evaluación.

Es la quinta fase de la metodología CRISP-DM, evalúa la calidad del modelo considerando si cumple con los criterios de éxito establecidos, es importante revisar el proceso, para corregir errores en caso de que existan. Para realizar esta fase se puede utilizar varias herramientas para interpretación de resultados (Arancibia, 2009).

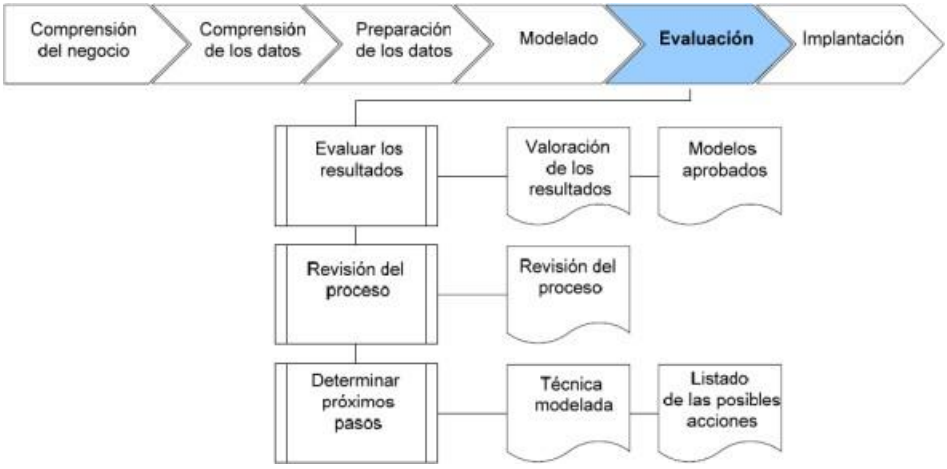


Fig. 10. Fase de evaluación (CRISP-DM, 2000)

- **Evaluar los resultados:** En esta tarea se evalúa el modelo en relación con los objetivos propuestos inicialmente (Arancibia, 2009).
- **Revisar el proceso:** Examina el proceso de minería de datos para así encontrar elementos o parámetros que pueden ser mejorados (Arancibia, 2009).
- **Determinar los próximos pasos:** Si en algún caso no se ha obtenido resultados satisfactorios, es posible hacer otra interacción desde la preparación de los datos o la modificación del modelo con distintos parámetros (Arancibia, 2009).

1.3.5. Despliegue o implantación.

Es la última fase de la metodología CRISP-DM en donde el modelo ha sido desarrollado y sus validaciones se han realizado, este se transforma en conocimiento. Es importante documentar

los resultados y etapas de todas las fases de la metodología, de manera correcta y perspicaz para que el usuario final, entienda rápidamente (Arancibia, 2009).

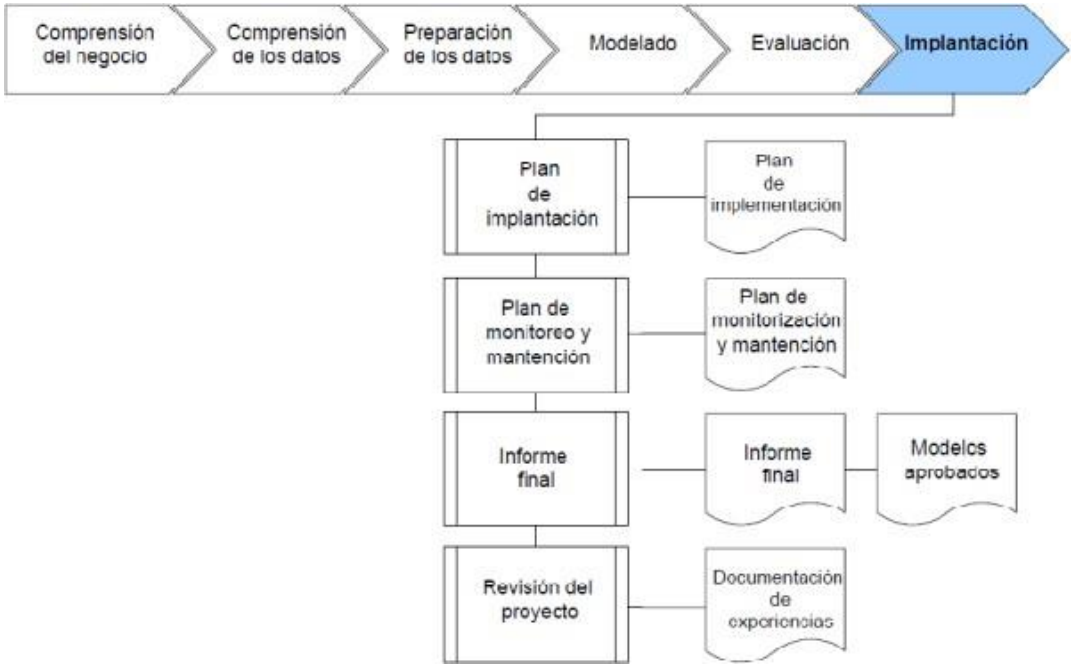


Fig. 11. Fase de implementación (CRISP-DM,2000)

- **Plan de implantación:** Se realiza una estrategia en base a los resultados de la evaluación, con el fin de implementar el resultado de la minería de datos (Arancibia, 2009).
- **Plan la monitorización y mantenimiento:** Con este proceso se genera una retroalimentación para verificar si el modelo está siendo manejado correctamente
- **Producir el informe final:** Propicia una conclusión del proyecto realizado, este depende del plan que se realizó de implementación, este documento puede ser un resumen de los puntos importantes de cada etapa del proyecto, y los resultados que se ha generado (Arancibia, 2009).
- **Revisión del proyecto:** En esta última tarea se evalúa los aspectos correctos y erróneos, como también lo que posiblemente se puede mejorar en lo que respecta al proyecto realizado (Arancibia, 2009).

1.4. Herramientas de Inteligencia de Negocios y Minería de datos

1.4.1. Herramientas de Business Intelligence

- **Microsoft Dynamics NAV:** Es una solución presentada por Microsoft en donde mejora el rendimiento de las pequeñas y medianas empresas, ayuda a automatizar y conectar compras, operaciones, contabilidad y administración de inventario (Vila, 2019).
- **Microstrategy Intelligence:** Creada por la empresa Microstrategy, transforma volúmenes robustos de datos en informes intuitivos orientados a sectores empresariales (Vila, 2019).
- **Microsoft Power BI:** Microsoft proporciona este software de manera gratuita, siendo un conjunto de aplicaciones en donde se transforman los datos en objetos visuales para todo tipo de usuarios sin depender de un experto (Rea D. C., 2021).
- **Microsoft Excel:** Esta herramienta facilitada por Microsoft brinda la posibilidad de procesar datos con el fin de adquirir cálculos matemáticos y así generar reportes estadísticos (Rea D. C., 2021).
- **Oracle BI:** Oracle Business Intelligence, herramienta presentada por Oracle, proporciona análisis descriptivos, transformando los datos en información de valor (Vila, 2019).
- **IBM Cognos Analytics :** Esta Solución está dirigida para expertos y no expertos, es donde permite creación de informes, análisis modelado de escenarios, analísticas predictivas, etc.; para la toma de decisiones correctas (Vila, 2019).
- **SAP Business Objects:** Esta herramienta de Inteligencia de Negocios permite explorar los datos y realizar análisis sólidos para que así los usuarios tomen decisiones eficaces (Rea K. G., 2020).
- **Pentaho:** Es un software libre, su ambiente de programación es Java, dispone de una variedad cantidad de gráficos para el análisis de información, minería de datos como también análisis multidimensional OLAP (Vila, 2019).
- **Jaspersoft para BI:** Este entorno es de código abierto, dirigido a usuarios que utilizan Java de Eclipse, este software permite diseñar informes, visualización de datos, en algunos casos a través del análisis OLAP, estos pueden par la web o dispositivos móviles (Vila, 2019).

1.4.2. Herramientas Data Mining

- **Orange:** Su principal función es el análisis predictivo y aplicación en la minería de datos, este software de código abierto provee herramientas para visualización de datos y machine learning. El manejo es para usuarios novatos, como expertos Orange presenta análisis con datos en flujo, posee indicadores de ubicación, analizar flujos de población en tiempo real, estas características pueden lograr un mejor resultado en el proyecto a realizar (Vila, 2019).
- **SAS Enterprise Miner:** El software SAS trabaja sobre grandes volúmenes de datos, logrando modelos predictivos y descriptivos. Su interfaz es intuitiva por lo que es fácil para los usuarios sin ninguna experiencia (Mancero, 2020).
- **Weka:** Esta herramienta de minería de datos y aprendizaje automático fue desarrollada en Java, está compuesta por un conjunto de algoritmos para realizar los análisis descriptivos con modelo predictivo, con técnica de preprocesamiento, métodos de entrenamiento (Mancero, 2020).
- **Knime:** Al igual que las herramientas anteriores está desarrollada en Java sobre la plataforma Eclipse, este software libre propicia la manera de crear de forma visual flujos de datos, modelos y vistas interactivas (Vila, 2019).
- **RapidMiner:** Este software desarrollado en Java, permite utilizar los algoritmos incluidos en Weka, incluye técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos (Mancero, 2020).

1.5. Técnicas descriptivas y predictivas, algoritmos.

La minería de datos se centra en explorar y descubrir el conocimiento a partir de los datos, lo que representa un cambio significativo en la forma en que se aborda el análisis de datos, la aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos (Beltrán, 2016).

La elección del modelo de Minería de Datos a utilizar se ve influenciada por los objetivos del negocio, que se logran al diseñar y probar diversas combinaciones de algoritmos (Beltrán, 2016).

Los modelos de Data Mining se clasifican como predictivos y descriptivos, se tiene una variable con valor desconocido, y la finalidad es determinar, esta variable se llama respuesta,

variable dependiente u objetivo, mientras que aquellas utilizadas para hacer la predicción son los predictores o variables independientes (Beltrán, 2016).

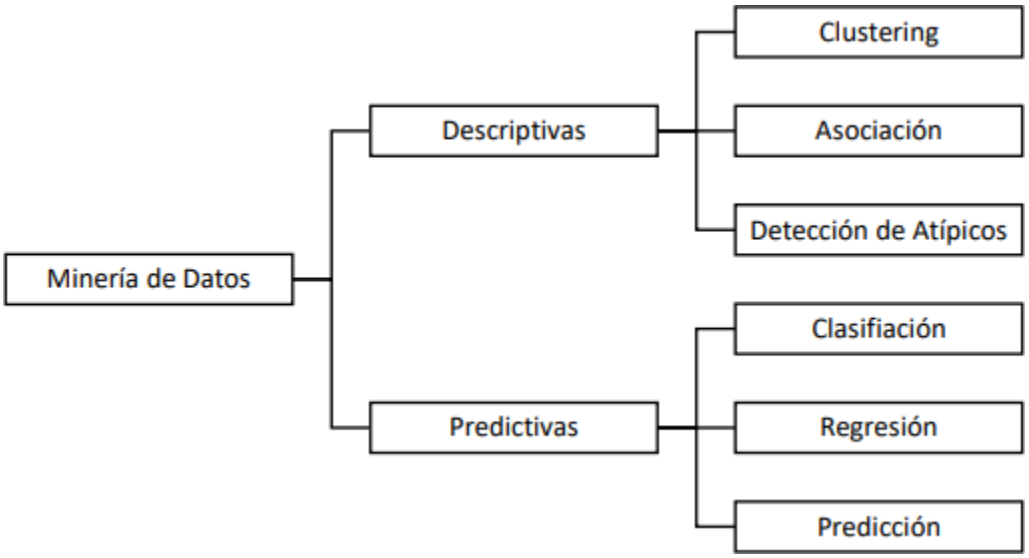


Fig. 12. Técnicas de minería de datos (Hernández Valadez, 2006)

1.5.1. Técnicas Descriptivas

Cada técnica da solución a distintos problemas, en donde se destacan los siguientes:

1.5.1.1. Clustering (Agrupamiento)

Este proceso consiste en agrupar datos en donde los valores de sus atributos sean similares, este método descriptivo se basa en predecir basándose en datos del pasado. Los métodos básicos para llevar a cabo dos tipos de clustering numérico: el clustering particional y el clustering ascendente jerárquico. (Larranaga, Inza, & Moujahid)

- **Clustering Particional:** Su fin es realizar una partición en donde todos los objetos se encuentren en grupos o clusters, dividiéndolos en k clusters posibles y cluster disjuntos (Larranaga, Inza, & Moujahid).
- **Clustering Ascendente Jerárquico:** Este tipo de agrupación se basa en construir un dendrograma que es la agrupación de cada dos objetos más cercanos, este

método es superior al clustering particional. Dentro de la técnica clustering, contiene algunos algoritmos, estos son los más frecuentes (Larranaga, Inza, & Moujahid):

K-medias: Este procedimiento utiliza la distancia euclídea, sus parámetros son: el número de grupo, K, y sus respectivos centroides, en donde estos son la media de un conjunto de datos y se aplica a objetos en espacios continuos n-dimensionales debiéndose especificar el número de clúster a encontrar (Larranaga, Inza, & Moujahid).

Algoritmo DBSCAN: Este algoritmo basado en densidad es una de los más rápidos en el campo de clustering, se puede encontrar un cluster separando del ruido más disperso (Larranaga, Inza, & Moujahid).

1.5.1.2. Asociación

Esta técnica de asociación es importante en la Minería de Datos, consiste en encontrar las asociaciones interesantes en forma de relaciones de implicación entre los valores de los atributos de los objetos de un conjunto de datos (Hernandez, 2008).

- **Algoritmo A priori:** Se trata de una técnica que permite ahorrar tiempo y recursos al procesar grandes conjuntos de datos, su principal principio "a priori" se basa en buscar iterativamente conjuntos frecuentes con cardinalidad 1 hasta k, posteriormente se genera las reglas de asociación (Hernandez, 2008)
- **FP-Growth:** Utiliza una estructura de árbol en donde a partir de itemsets frecuentes que no necesitan la generación de candidatos para cada tamaño, genera las reglas de asociación, esta técnica permite reducir el espacio de almacenamiento necesario para procesar grandes conjuntos de datos, al mismo tiempo que simplifica el proceso de cálculo del soporte para todos los elementos (Hernandez, 2008).

1.5.1.3. Detección de atípicos

Para el correcto desarrollo de un proyecto de minería de datos, la calidad del dataset es primordial, para que al momento de su análisis se toma las correctas decisiones empresariales, es una tarea difícil detectar valores que se desvían significativamente de la norma, los valores atípicos

o también denominados anómalos tienen propiedades diferentes con respecto a la generalidad (Rea D. C., 2021).

Para detectar valores atípicos, se utiliza varios métodos entre los cuales están:

- **Detección valores atípicos usando clustering:** Las técnicas de agrupamiento o clustering ordena los objetos en grupos, en donde el grado de asociación es máximo si los objetos corresponden al mismo grupo (Rea D. C., 2021).
- **Detección de valores atípicos basados en la distancia:** A través de la distancia se puede detectar los valores atípicos, en donde si los puntos están lejos del vecindario local, son datos inusuales caso contrario si se encuentran cerca están correctos (Rea D. C., 2021).
- **Detección de valores atípicos basados en la densidad:** Se basa en la densidad de regiones en los datos, si se observa que existe baja densidad en las regiones se considera valores atípicos (Rea D. C., 2021).
- **Detección de valores atípicos basados en la distancia y densidad:** Consta del algoritmo RODHA: Robust Outlier Detection using Hybrid Approach que consiste en la detección de valores atípicos en conjunto con la distancia, la densidad y entropía, volviendo la detección más robusta (Rea D. C., 2021).

1.5.2. Técnicas predictivas

Esta área de la minería de datos consiste en obtener datos, procesarlos y generar conocimiento para predecir patrones de comportamiento y tendencias, para la aplicación de esta técnica se debe disponer de una gran cantidad de datos, actuales como pasados (Vila, 2019).

Existen diversos tipos de modelos entre los cuales están:

1.5.2.1. Clasificación

Estos modelos pueden ser descriptivos o predictivos, se identifica si un elemento del Dataset pertenece a una de las clases predefinidas

- **Árboles de Decisión:** Esta técnica es muy popular debido a su efectividad en el análisis de grandes conjuntos de datos, y es ampliamente utilizada para identificar patrones y relaciones útiles entre variables (Vila, 2019).

- **Redes Neuronales Artificiales:** Cuando se proporciona un patrón de entrada a una red neuronal, esta genera un patrón de salida específico. La arquitectura de la red neuronal se basa en una distribución en paralelo de un gran número de nodos (neuronas) y conexiones que permiten su funcionamiento (Vila, 2019).
- **Vecino más próximo:** Es una técnica sencilla que detecta datos atípicos mediante la identificación del vecino más cercano entre los datos previamente conocidos y clasificados a través del sistema de votación (Vila, 2019).

1.5.2.1. Regresión

La salida de este algoritmo es un valor numérico en donde forma relaciones entre los elementos del Dataset

- **Regresión lineal:** Es un algoritmo que tiene la función de "dibujar una recta" en el cual nos indicará la tendencia de un conjunto de datos continuos (Mancero, 2020)
- **Regresión múltiple:** Modelo estadístico que se utiliza para analizar y evaluar las relaciones entre un resultado o variable dependiente continuo y varios predictores o variables independientes (Mancero, 2020).
- **Árboles de regresión:** Un árbol de regresión es básicamente un árbol de decisión que se usa para la tarea de regresión que se puede usar para predecir salidas de valor continuo en lugar de salidas discretas (Mancero, 2020).

1.6. Trabajos existentes

Para la realización del proyecto, se ha tomado en cuenta algunos trabajos ya realizados en donde se aplica técnicas de minería de datos como sus metodologías, que se describirán a continuación

(Mancero, 2020) utilizó los datos históricos (2015-2020) provenientes de la entidad aduanera del Ecuador, en donde su principal objetivo es detectar cuales son los principales factores de contrabando en el Ecuador, por medio de patrones y técnicas predictivas (clasificación y regresión). Mediante la Metodología KDD eso se aplicó para obtener una vista minables. Para el

desarrollo de los modelos utilizaron KNIME Y WEKA, basándose en la técnica de agrupamiento (clustering) donde se identificó información estratégica para la toma de decisiones dentro de la entidad de control concluyendo que existe contrabando por Santa Rosa y por la frontera con Colombia, logrando así una competencia con la producción local.

(Rea K. G., 2020) en este proyecto se utilizaron datos históricos de las ventas en unidades a nivel de subcategoría de producto, de los años 2018 y 2019. A través de la fuente de información de la base de datos MySQL. Para el entrenamiento de datos en la herramienta analítica Rapidminer, en este proyecto, se utiliza la metodología Design Science Research y la metodología KDD para encontrar el modelo analítico más adecuado para predecir las ventas de la línea OTC de un laboratorio farmacéutico. Los modelos para utilizar son: Arimax, Redes Neuronales y Holt Winters y con la métrica MAPE se obtuvo una precisión de error del 10% mínimo.

(Cortina, 2015) en este trabajo se sigue rigurosamente cada una de las fases de la metodología CRISP-DM para analizar los datos académicos almacenados por la universidad en sus sistemas informáticos, además, se realizan consultas SQL significativas para obtener muestras representativas de los datos y obtener conclusiones adicionales que no se contemplaban inicialmente en los objetivos de la minería de datos. Todo el proceso se lleva a cabo utilizando la herramienta Rapid Miner.

(Vila, 2019) en este proyecto se aplicó técnicas predictivas de minería de datos (clasificación y regresión), para procesar datos históricos de los estudiantes desde del año 2017 a 2018., como también se realizó el proceso KDD para obtener una vista minable, Los principales resultados demostraron que los mejores algoritmos fueron RandomTree y Logistic, para obtener el conocimiento se tomó en cuenta la intersección de los resultados obtenidos de ambos algoritmos.

(Paspuel, 2022) en este trabajo aplicaron las técnicas de Data son las reglas de asociación y el clustering, y se utilizan herramientas informáticas como el lenguaje de programación Python, para descubrir hábitos de compra de productos y accesorios de bicicletas, concluyendo que reglas de asociación tienes las probabilidades más altas en la adquisición de productos.

(Nazate, 2022) este proyecto se realizó recopilando las transacciones de la Papelería Sari Popular con datos desde abril del 2021 hasta 2022, el proyecto se llevará a cabo mediante el uso

de herramientas y software específicos, tales como Power BI, Python, PostgreSQL, así como hojas de cálculo como Excel y Google Sheets. La metodología elegida para la ejecución del proyecto es Hefesto. Y aplica el algoritmo de asociación A priori el cual dio una confianza cercana a 1 que indica que es óptima

CAPÍTULO 2

DESARROLLO

2.1. Definición de Requerimientos y preguntas del negocio

2.1.1. Requerimientos

La determinación de requerimientos es una de las etapas fundamentales en el desarrollo del proyecto en donde el cliente y proveedor definen el alcance y las limitaciones durante el proyecto. Es primordial realizar un previo análisis a la base de datos en donde se conocerá las circunstancias en las que se debe desempeñar la inteligencia de negocios y cómo es que debe estar funcionando.

2.1.2. Indicadores

Según (NEIRA, 2015) Los indicadores se relacionan con tres aspectos: medida, medición y métricas en donde se calcula estas y se fija límites de control, lo que también significa un aspecto clave en donde se desencadena el control y el seguimiento en los proyectos y las organizaciones de manera general. En donde se divide en distintos tipos:

- Indicadores de Entrada: Miden la cantidad, calidad y oportunidad de los recursos humanos, financieros, materiales, tecnológicos y de información de un proyecto (NEIRA, 2015).
- Indicadores de Salida: miden bienes y servicios entregados por el proyecto (Neira & Romero, 2015).
- Indicadores de Impacto: miden la calidad y cantidad de resultados a largo plazo generados por el programa (Neira & Romero, 2015).
- Indicadores de Resultado: miden los resultados intermedios o de corto plazo generados por los productos del programa (Neira & Romero, 2015).

Los Indicadores sé utilizan como las claves de desempeño para evaluar la actuación de los proyectos y de aquellos involucrados en su desarrollo, se trata de una parte importante de la programación de rendimiento, que proviene de la descomposición del objetivo estratégico general

de la organización y son métricas para el control de la evolución de las acciones en los proyectos (Neira & Romero, 2015)

2.1.3. Requerimientos e indicadores en Comercial Cadena Casanova

La recepción de requerimientos se realizó haciendo una propuesta sobre "Las preguntas del negocio" presentando a la gerente del Comercial en donde ella aprobó y añadió algunas más. Las preguntas están basadas principalmente para conocer información de los productos, clientes, categorías, con sus respectivos indicadores que permiten medir (Nazate, 2022). Se puede observar en la siguiente tabla:

Requerimientos	Indicadores
1) Monto total de ventas de cada categoría en el año 2021 2) Monto total de ventas de cada categoría en el año 2022	
3) Cantidad de facturas de más $500 que se emitieron en 2021 4) Cantidad de facturas de más $500 que se emitieron en 2022 5) Periodo de tiempo con mayor número de ventas en el año 2021 6) Periodo de tiempo con mayor número de ventas en el año 2022	
7) Clientes con mayor número de productos comprados 8) Ventas por cliente en un periodo de tiempo 9) El producto más vendido de cada categoría	

Tabla 2. Requerimiento e indicadores

Con esta representación de los requerimientos e indicadores alcanzados se obtiene una visión clara del proyecto, en el desarrollo del proyecto se podrá ver con más precisión el uso de este esquema.

2.2. Calidad de datos – Estándar ISO/IEC 25012

Los datos obtenidos deberán ser evaluado bajo la ISO/IEC 2501, tomando como referencia a (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019) en donde genera diferentes rangos de evaluación basándose en las características del estándar

Para realizar la evaluación de calidad de datos se tomó como guía la estructura de ISO/IEC 25040, en relación con la base de datos adquiridos, para medir la calidad de datos se realizan los siguientes pasos:

Establecer los requisitos de la evaluación: El objetivo principal es comprobar si los datos que se van a utilizar se consideran como necesarios, como también verificar si están establecidos con los formatos esperados, basándose en la ISO/IEC 25012 se estimara las características: Exactitud (Semántica y Sintáctica) y Completitud. (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Especificar la evaluación: En esta actividad se establecen las métricas de aceptación de las características anteriormente mencionadas que son Exactitud y Completitud. De esta manera se muestra los requisitos de evaluación en el siguiente cuadro

Característica	**Variables**	**Escala**
X = Porcentaje de datos del atributo que cumplen con las reglas definidas.		
Característica	Inadecuado	X<=10%
	Poco idóneo	X>10% & X<=45%
	Rango Suficiente	X>45% & X<=85%
	Idóneo en su totalidad	X>85

Tabla 3. Métricas de aceptación (Calabrese, Esponda, Pasini, Boracchia, & Pesado, 2019)

Resultados

Tabla Clientes

Atributos	**Exactitud semántica**	**Exactitud sintáctica**	**Integridad**
RUC	100%	99,9%	100%
CLIENTE	99,85%	93,63%	100%

Tabla 4. Resultados de métricas en tabla clientes

Analizada la tabla Clientes con los atributos principales dio como resultados que todos están en el rango de Idóneo en su totalidad superando el 85% que cumplen con las características

Tabla Producto

Atributos	**Exactitud semántica**	**Exactitud sintáctica**	**Integridad**
NOMBRE	100%	63,26%	100%

Tabla 5. Resultados de métricas en tabla Producto

La tabla Productos muestra que los atributos principales dieron como resultados que están en el rango de Idóneo en su totalidad superando el 85% que cumplen con las características sin embargo en exactitud sintáctica está en la escala de rango suficiente.

Tabla Facturas

Atributos	**Exactitud semántica**	**Exactitud sintáctica**	**Integridad**
NROFACTURA	99,9%	100%	100%

Tabla 6. Resultados de métricas en tabla Facturas

Analizada la tabla Facturas con los atributos principales dio como resultados que todos están en el rango de Idóneo en su totalidad superando el 85% que cumplen con las características

2.3. Proceso ETL y creación de DataWarehouse usando la metodología CRISP-DM.

En el desarrollo de este proyecto se emplea la Metodología CRISP-DM a los datos recolectados del Comercial Cadena Casanova, en donde se realizan varios procesos como Extracción, Transformación y Carga, Para ello, se implementarán cada una de las etapas propuestas.

2.3.1. Comprensión del Negocio

El Comercial Cadena Casanova ofrece productos de primera necesidad y venta al por mayor y menor, en la provincia de Imbabura ciudad de Ibarra. A partir de esto se seguirá con las etapas de la metodología

2.3.1.1. Objetivos del Negocio

Existe tres objetivos claros para lograr una mejoría de ingresos y evitar la caducidad de productos en el Comercial Cadena Casanova, uno de los principales es lograr una fidelización de clientes por medio de la segmentación a partir de los datos adquiridos, personalización de ofertas o combos para lograr la venta de productos que no han tenido ventas, realizar predicciones de los productos por categoría para adquirir la cantidad necesaria y evitar la caducidad (Cortina, 2015)

2.3.1.2. Valoración de la situación

La empresa emplea un sistema llamado SICOF el cual provee los datos, en conjunto con su base de datos en MS Access, de los cuales se obtienen archivos en formato .mdb. Con un análisis breve se muestra que datos necesarios para realizar el proyecto y el desarrollo de los objetivos propuestos, en los cuales están las ventas realizadas, inventario, clientes, facturas, asientos contables, proveedores (Cortina, 2015).

2.3.1.3. Plan de proyecto

El plan de proyecto consiste en varias etapas:

- Recolección de datos desde MS Access, en donde se incluye una depuración inicial y la importación a Power BI
- Exploración y Verificación de la calidad de datos
- Realización de consultas con el objetivo de obtener una muestra que represente adecuadamente los datos (Cortina, 2015).
- Preparación previa de los datos (selección, limpieza, conversión y formateo) con el fin de facilitar la aplicación de minería de datos en ellos (Cortina, 2015).

- Escoger las técnicas y herramientas adecuadas para realizar el modelo y aplicación sobre los datos
- Análisis y evaluación de los resultados obtenidos en la fase anterior (Cortina, 2015).
- Creación y exposición de informes que reflejen los resultados alcanzados en concordancia con los objetivos previamente establecidos (Cortina, 2015).

2.3.2. Comprensión de los Datos

La siguiente etapa involucra la obtención, descripción, exploración y verificación de la calidad de los datos (Cortina, 2015).

2.3.2.1. Recolectar los datos iniciales

A simple vista las relaciones de la base de datos no están correctamente estructuradas, las relaciones de las tablas que se realiza el análisis están erróneas, por lo que se realiza un modelo aparte con los datos adquiridos que son tablas de productos, categorías, clientes haciendo referencia a las facturas que incluyen Nro. de ingreso, fecha, Nro. factura, total. Estas tablas se han elegido en base a los objetivos que se van a realizar (Cortina, 2015).

Otra falla encontrada es que en la tabla productos no existe un campo de categoría por lo que se realizó manualmente en base al inventario.

2.3.2.2. Descripción de los Datos

El nuevo modelo de base datos se realizará con esquema dimensional en estrella esto se realizó en la Herramienta de Power BI

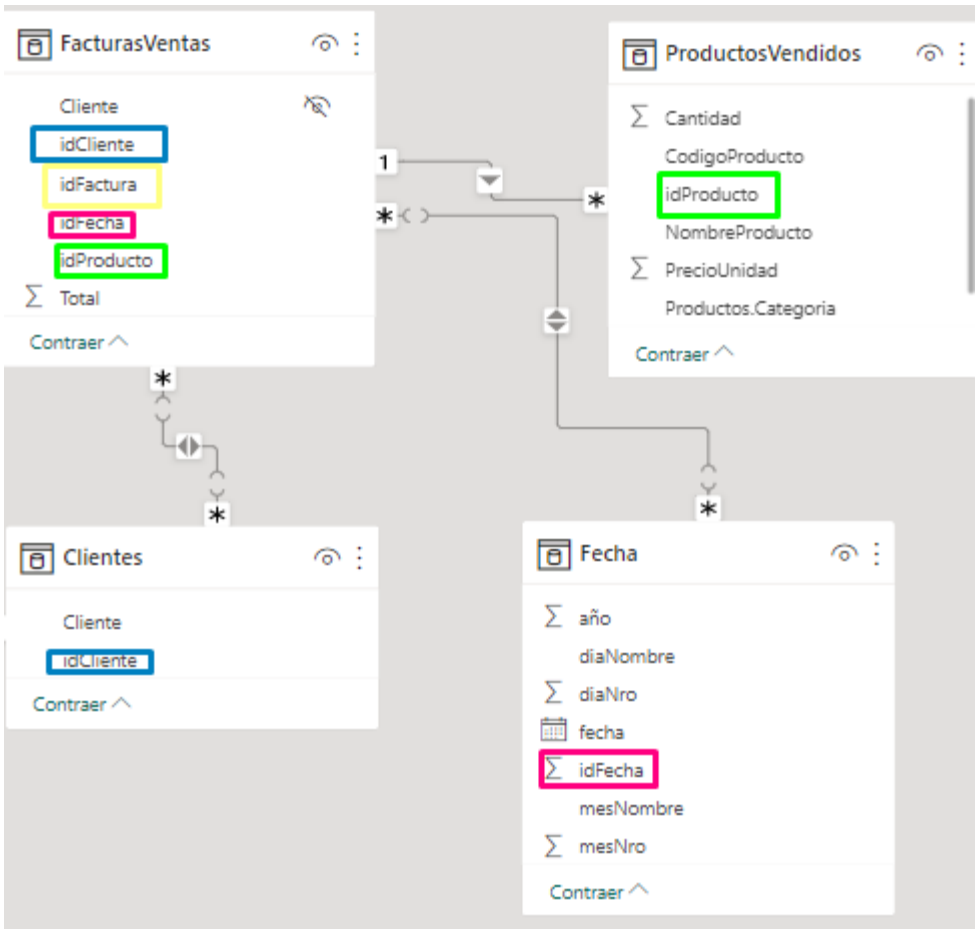


Fig. 13. Esquema dimensional inicial

Podemos observar que utilizamos simplemente cuatro tablas FacturasVentas, ProductosVendidos, Clientes y Fecha, a continuación, describiremos cada una de ellas detallando cada uno de sus campos.

Tabla FacturasVentas

La tabla que se presenta es la tabla principal del depósito de datos, que también se conoce como "tabla de hechos". Es en esta tabla donde se registra toda la información relacionada con las transacciones comerciales. Al ser la tabla central, su clave primaria es una combinación de las claves principales de las otras tablas, conocidas como "tablas dimensionales" (Cortina, 2015).

Las claves principales que utiliza son el identificador de cliente (idCliente), el identificador de fecha (idFecha) y el identificador de producto (idProducto), los cuales a su vez son claves

foráneas (foreign keys). Esta tabla contiene un total de 119,361 registros desde enero de 2021 hasta diciembre de 2022

- **idFactura**: Tipo texto. Este campo posee el código que identifica a la factura y es único.
- **idCliente:** Tipo numérico. Esta variable es el conjunto de varios números que identifica a las persona u organización, a la misma vez es único
- **idFecha**: Es de tipo numérico y se trata de un número único que distingue cada fecha que ha sido insertada en la tabla de fechas
- **idProducto**: Es de tipo numérico y se trata de un número único que asocia cada una de las ventas de los productos vendidos y que es único
- **Total:** Tipo decimal, Esta variable muestra el monto total de la venta realizada

Tabla Clientes

La tabla contiene información sobre las fechas y su clave principal es el campo idFecha. La tabla tiene un total de 21.842 registros.

- **idCliente**: Tipo numérico. Esta variable es el conjunto de varios números que identifica a las persona u organización, a la misma vez es único
- **cliente:** Tipo texto. Este campo muestra el nombre del cliente u organización

Tabla Fecha

Esta tabla contiene la información acerca de las fechas. Su clave primaria es el campo idFecha, y tiene un total de 119.361 registros desde enero del 2021 hasta diciembre del 2022

- **idFecha:** Tipo numérico Este campo es un número que identifica a cada fecha insertada en la tabla de fechas y que es único para cada fecha
- **diaNro:** Tipo numérico, donde se muestra el número de día que se realizó la venta
- **diaNombre**: Tipo texto, donde se muestra el nombre de día que se realizó la venta
- **mesNro:** Tipo numérico, donde se muestra el número de mes que se realizó la venta
- **mesNombre:** Tipo texto, donde se muestra el nombre de mes que se realizó la venta
- **año:** Tipo numérico, donde se muestra el año que se realizó la venta

Tabla ProductosVendidos

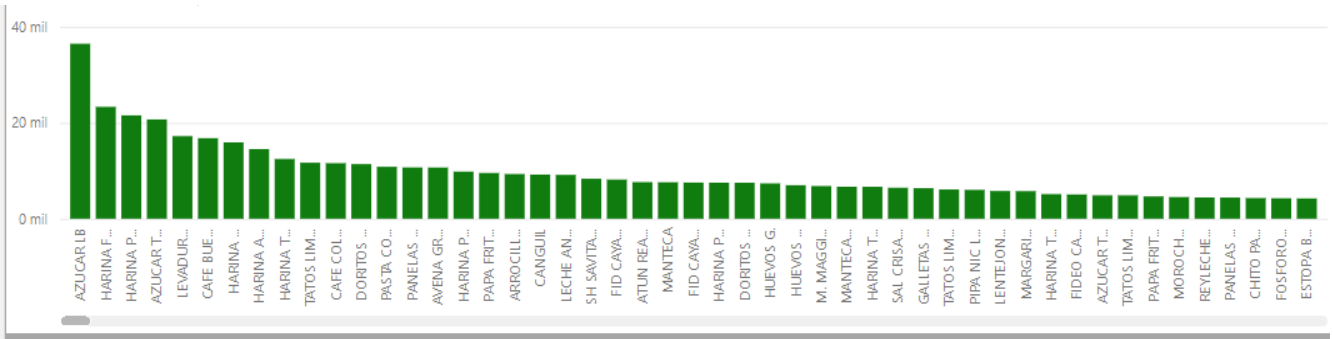
Esta tabla contiene la información acerca de los productos que se han vendido. Su clave primaria es el campo idProducto, tiene un total de 397101 registros.

- **idProducto:** Es de tipo numérico y se trata de un número único que distingue a cada una de las ventas de los productos vendidos y que es único
- **codigoProducto**: Tipo Texto. Este campo es un conjunto de letras y números que identifica a los productos en su inventario y es único
- **nombreProducto:** Tipo texto. Este campo muestra el nombre de los productos que se encuentra en el inventario
- **categoría**: Tipo texto. Esta variable indica a la categoría a la cual pertenece el producto de acuerdo con el inventario
- **precioUnidad:** Tipo decimal. Indica el precio unitario de cada uno de los productos

2.3.2.3. Exploración de los Datos

Ya realizada la descripción de los datos a utilizar, podemos explorarlos, que implica realizar pruebas estadísticas en donde proporciona información de los datos, y así a la misma vez crear sus respectivos gráficos y frecuencias de distribución (Cortina, 2015)

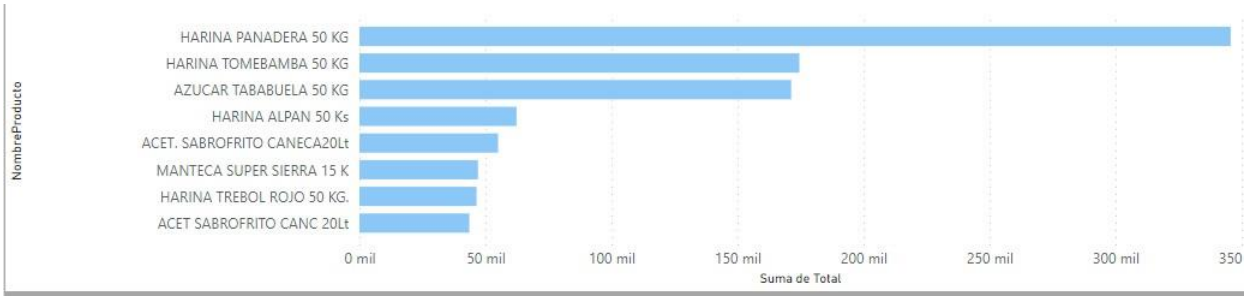
En el siguiente gráfico se muestra la distribución de los productos vendidos con respecto a la cantidad en el intervalo de enero 2021 a Diciembre 2022



NombreProducto	Suma de Cantidad
GALLETAS SALTICAS INTEGRAL	60,00
GOMAS GRISSLY GUSANO 12 U	60,00
JB S/ALES BLA F/ROJOS*2U	60,00
LENTEJA LIBRA	60,00
MANTECA REINA FUNDA 1LB	60,00
MOROCHO MOLIDO Lbr	60,00
PAN PASCUA BUEN PAN 420G	60,00
PAPEL ROSAL SUPER ROLLO DE 12	60,00
PH ELITE PREM-EX/G X 18R	60,00
T/C SCOTT MULTI ABSORB 55H	60,00
TALLARIN 400G	60,00

Fig. 14. Productos vendidos por cantidad

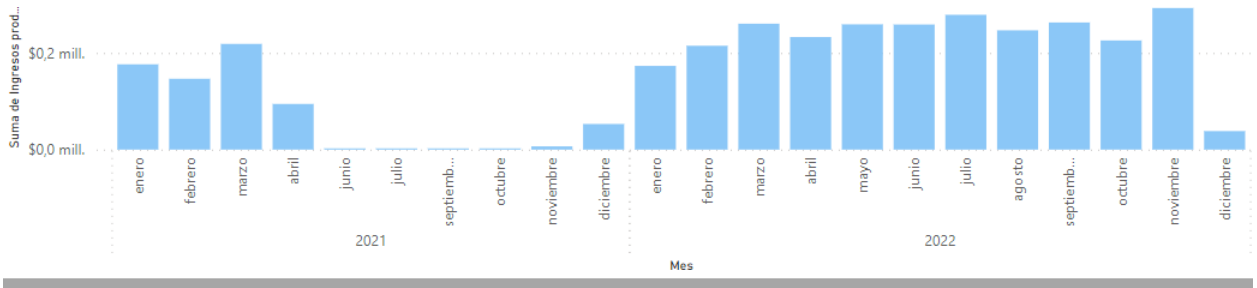
En el siguiente gráfico muestra los ingresos que lograron cada uno de los productos



NombreProducto	Suma de Total
HARINA PANADERA 50 KG	345.656,58
HARINA TOMEBAMBA 50 KG	174.527,73
AZUCAR TABABUELA 50 KG	171.297,12
HARINA ALPAN 50 Ks	62.320,40
ACET. SABROFRITO CANECA20Lt	54.991,00
MANTECA SUPER SIERRA 15 K	47.021,64
HARINA TREBOL ROJO 50 KG.	46.457,53
ACET SABROFRITO CANC 20Lt	43.583,37
HARINA PANIFICACION 50 KG.	41.715,24
LARK ROJO x 10 DE 10 s	35.424,21

Fig. 15. Productos vendidos por ingresos

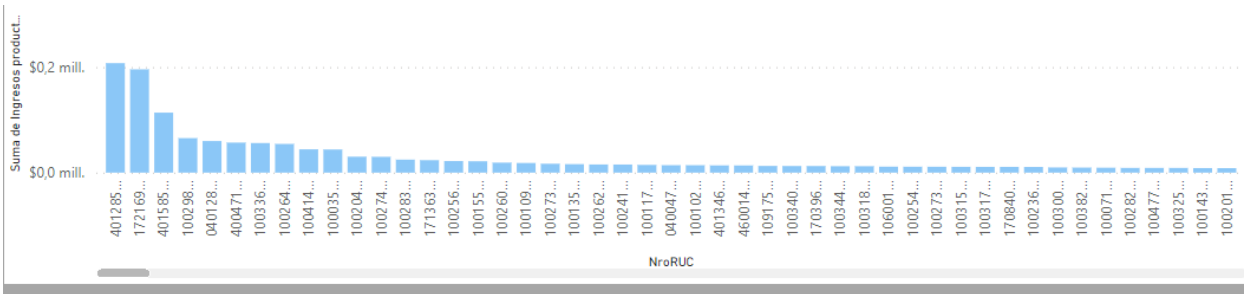
Se observa los ingresos obtenidos en los distintos meses de los dos años



Año, Mes	Suma de Ingresos productos
2021, enero	$176.432,73
2021, febrero	$146.561,48
2021, marzo	$218.698,04
2021, abril	$94.447,92
2021, junio	$1.902,72
2021, julio	$2.035,55
2021, septiembre	$568,21
2021, octubre	$2.437,79
2021, noviembre	$6.615,35
2021, diciembre	$52.940,07

Fig. 16. Ingresos por tiempo

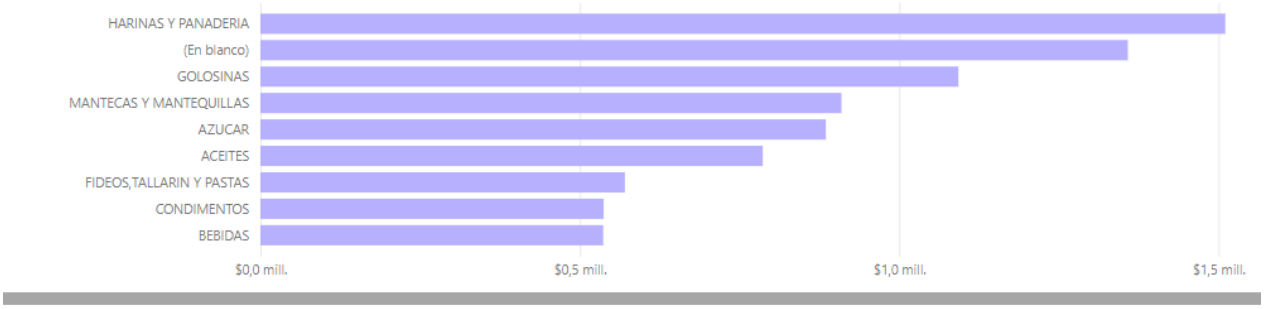
En el gráfico se puede observar los ingresos obtenidos por clientes



NroRUC	Suma de Ingresos productos
401285218001	$207.635,90
1721695821001	$195.861,86
401585716	$113.755,14
1002985578001	$65.403,99
0401285218001	$59.877,23
400471504001	$56.912,10
1003369921	$56.080,62
1002642708001	$54.422,41
1004149181	$44.187,31
1000353555	$43.849,24

Fig. 17. Ingresos por cliente

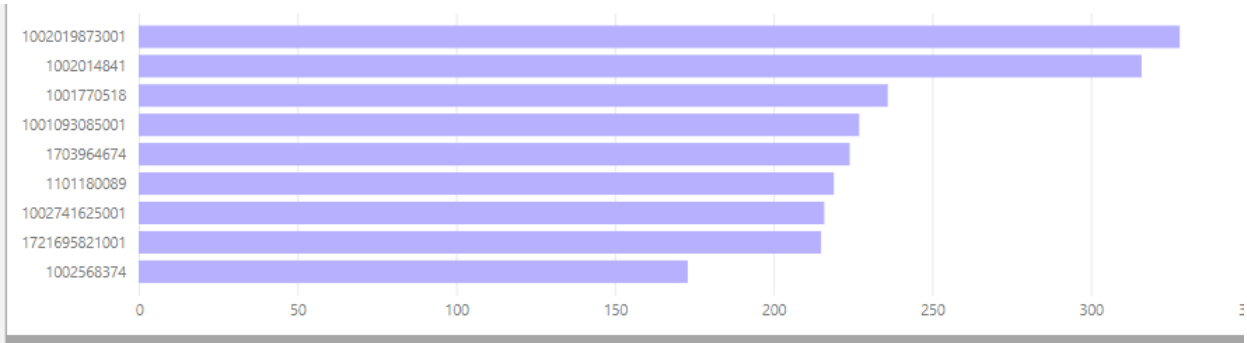
En la siguiente figura se observa los ingresos por categoría



Categoría	Suma de Ingresos productos
HARINAS Y PANADERIA	$1.510.524,61
	$1.357.924,80
GOLOSINAS	$1.092.646,61
MANTECAS Y MANTEQUILLAS	$909.581,11
AZUCAR	$885.215,12
ACEITES	$786.440,59
FIDEOS,TALLARIN Y PASTAS	$570.536,69
CONDIMENTOS	$537.385,98
BEBIDAS	$536.722,66
DETERGENTES	$495.371,21
ARROCES	$492.039,20

Fig. 18. Ingresos por cliente

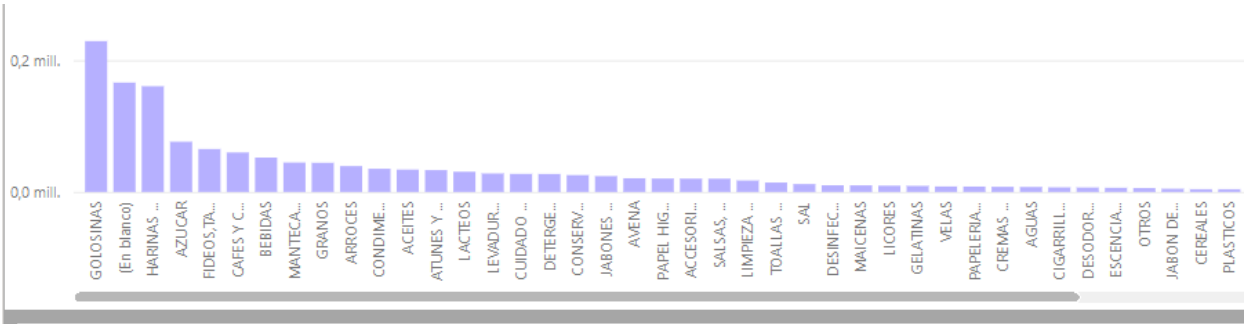
A continuación, la siguiente figura indica la frecuencia de los clientes en el 2021-2022



NroRUC	Recuento de Fecha
1002019873001	328
1002014841	316
1001770518	236
1001093085001	227
1703964674	224
1101180089	219
1002741625001	216
1721695821001	215
1002568374	173
401585716	170
1704101813	168

Fig. 19. Clientes por tiempo

Se observa en la siguiente imagen las cantidades vendidas por categoría



Categoria	Suma de Cantidad
GOLOSINAS	228.965,00
	165.961,00
HARINAS Y PANADERIA	160.435,00
AZUCAR	76.228,00
FIDEOS,TALLARIN Y PASTAS	65.107,00
CAFES Y CHOCOLATES EN POLVO	59.869,00
BEBIDAS	52.104,00
MANTECAS Y MANTEQUILLAS	44.581,00
GRANOS	44.209,00
ARROCES	39.162,00
CONDIMENTOS	35.114,00

Fig. 20. Categoría por cantidad

2.3.2.4. Verificar la calidad de los datos

Una vez realizada la exploración inicial de los datos y un análisis breve se llega a la conclusión que están completos. Estos datos cubren los requerimientos necesarios para la obtención de los resultados y objetivos definidos. Tampoco se encuentran valores fuera de rango por lo que no había demasiadas dificultades en el proceso de minería de datos, sin embargo, se debe realizar una limpieza ya que en la exploración se observó datos en blanco que es en el campo de categorías en la tabla ProductosVendidos esto se ha visto a simple vista por lo que podría a haber en las otras tablas más valores en blanco. (Cortina, 2015)

2.3.3. Preparación de los Datos

En esta parte de la metodología CRISP-DM prepara los datos para ajustarlos a las técnicas de minería de datos que se aplican en ellos. Por lo que se limpiara para así eliminar errores y mejorar su calidad como también añadir nuevos datos si es necesario (Cortina, 2015).

2.3.3.1. Seleccionar los Datos

Se seleccionan los campos necesarios para cumplir objetivos del proceso de minería de datos. Por lo cual se seleccionó:

Tabla Fecha

- idFecha

Tabla Cliente

- idCliente

Tabla ProductosVendidos

- idProducto
- CodigoProducto
- NombreProducto

Cantidad

- PrecioUnidad

Tabla FacturasVentas

- idFacturas
- idFecha
- idCliente
- idProducto
- Total

El motivo para la inclusión o exclusión de algunos campos es, como se ha mencionado antes, la importancia de dichos campos en relación con los objetivos de la minería de datos que se definieron en la fase 1 (comprensión del negocio) de la metodología (Cortina, 2015).

2.3.3.2. Limpiar los Datos

La base de datos del proyecto contiene la información necesaria para lograr los objetivos de la minería de datos. Sin embargo, se realizó la limpieza a los mismo, eliminando nulos, blancos y campos innecesarios para el desarrollo del modelo (Cortina, 2015).

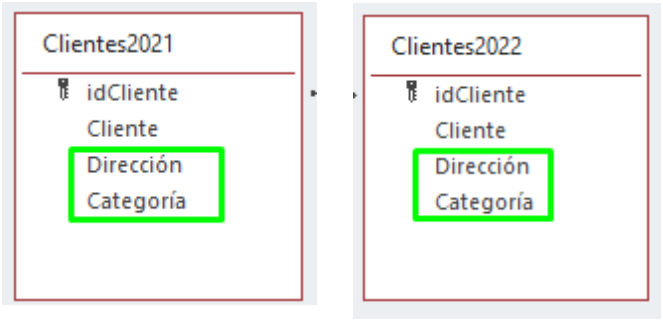
Tablas Clientes

Limpieza: Eliminación de los campos dirección y categoría Limpieza de campos nulos y blancos

Tabla 7. Limpieza de Clientes

La tabla Clientes2021 inició con un registro de 19780 datos, al final de la limpieza terminó con 19780, consecutivamente la tabla Clientes2022 inicio con 21873 y en la limpieza solo se encontró dos datos en blanco, dando un total de 21896 datos.

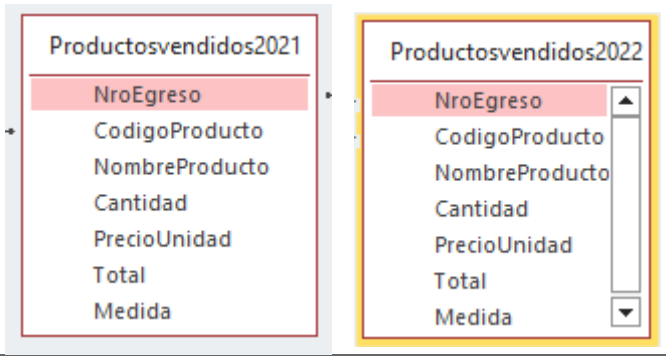
Tablas ProductosVendidos

Limpieza: Eliminación del campo medida Limpieza de nulos y blancos

Tabla 8. Limpieza de ProductosVendidos

La tabla ProductosVendidos2021 comenzó con 89256 y ProductosVendidos2022 con 304148, consecutivamente se terminó con el mismo número de datos.

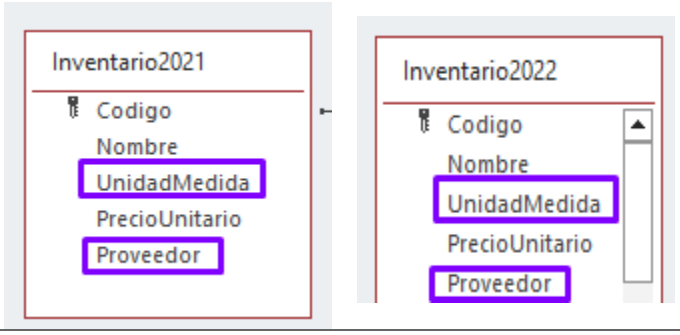
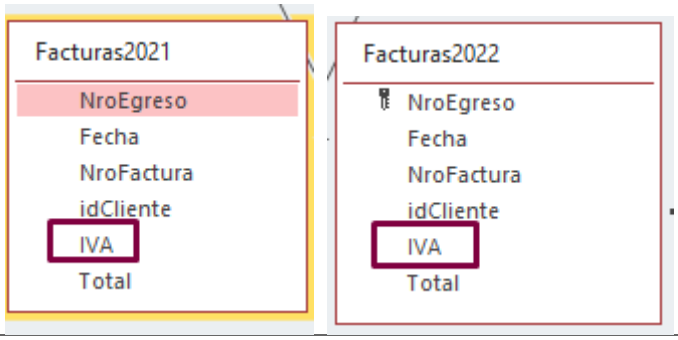
Tablas Inventarios

Limpieza: Eliminación del campo medida Limpieza de nulos y blancos

Tabla 9. Limpieza de Inventarios

La tabla Inventarios2021 empezó con 10757, después de la limpieza resultó el mismo número de datos, de esta misma manera Inventarios2022 se registró 12993.

Tablas Facturas



Limpieza:

Eliminación de los campos IVA

Limpieza de campos nulos y blancos

Tabla 10. Limpieza de Facturas

El número de facturas del 2021 registradas en un comienzo fue 25250 y facturas del 2022 hubo un total 94689, después de la limpieza de datos se obtuvo 24719 y 94641 respectivamente.

2.3.3.3. Construir los DatosRegistros generados

- Se anexo las tablas Clientes2021 y Clientes2022 para crear una llamada Clientes
- Se anexo las tablas ProductosVendidos2021 y ProductosVendidos2022 para crear una sola, consecutivamente esta misma tabla se unificó con Inventario en donde se toma como clave principal al código del producto para llamar al campo categoría que es necesario.
- Como también se unificó las tablas Facturas2021 y Facturas 2022 como resultado generó la tabla FacturasVentas.

Atributos

- Únicamente se renombró a los campos necesarios como idFactura, idProducto.
- En la tabla FacturasVentas existía compras por los consumidores finales por lo que se colocó un id para ese campo que es 01789 para este tipo de clientes

2.3.3.4. Integrar los Datos

Se generó la tabla Fecha en la cual se tomó las fechas de compras de la tabla factura, en esta existe los campos idFecha, Fecha, diaNro, diaNombre, mesNro, mesNombre, año. Y se eliminó las fechas repetidas.

idFecha	Fecha	diaNro	diaNombre	mesNro	mesNombre	año	trimestre
212021	2/1/2021	2	sábado	1	enero	2021	1
312021	3/1/2021	3	domingo	1	enero	2021	1
412021	4/1/2021	4	lunes	1	enero	2021	1
512021	5/1/2021	5	martes	1	enero	2021	1
612021	6/1/2021	6	miércoles	1	enero	2021	1
712021	7/1/2021	7	jueves	1	enero	2021	1
812021	8/1/2021	8	viernes	1	enero	2021	1
912021	9/1/2021	9	sábado	1	enero	2021	1
1012021	10/1/2021	10	domingo	1	enero	2021	1
1112021	11/1/2021	11	lunes	1	enero	2021	1

Fig. 21. Tabla Fecha

2.3.3.4. Formateo de los datos

La tabla FacturasVentas es el núcleo del almacén de datos, conocida también como la "tabla de hechos", ya que es donde se almacena toda la información relacionada con la venta de productos y clientes. En donde los campos finales son: idFactura, idProducto, idCliente, Cantidad, Total, NroEgreso

idFactura	Total	idProducto	idFecha	idCliente	NroEgreso	Cantidad
5F343380	$1,78	LF1.3	1812022	01789	577971	1
5F343392	$1,78	LF1.3	1812022	01789	577985	1
5F343248	$1,78	LF1.3	1712022	01789	577786	1
5F343275	$1,78	LF1.3	1712022	01789	577827	1
5F343324	$1,78	LF1.3	1712022	01789	577891	1
5F343336	$1,78	LF1.3	1812022	01789	577908	1
5F343474	$1,78	LF1.3	1812022	01789	578092	1
5F343511	$1,78	LF1.3	1812022	01789	578134	1
5F343541	$1,78	LF1.3	1812022	01789	578168	1
5F343562	$1,78	LF1.3	1812022	01789	578193	1

Fig. 22. Tabla FacturasVentas

Finalmente, la relación de las tablas se muestra de esta manera, el campo Egreso no se usa, pero tampoco se puede eliminar, ya que tiene una relación con ProductosVendidos

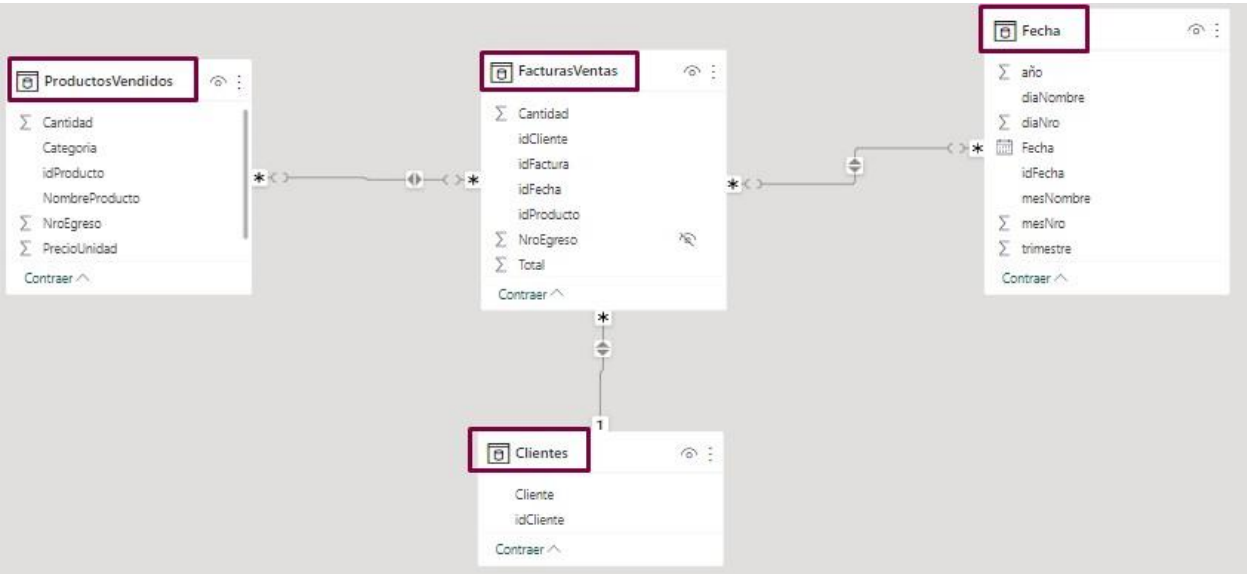


Fig. 23. Modelo relacional

2.3.4. Modelado

Durante esta fase de la metodología se seleccionó la técnica o técnicas más adecuadas para cumplir con los objetivos establecidos de la minería de datos, a continuación, y una vez realizado un plan de prueba para los modelos escogidos, se procederá a aplicar dichas técnicas sobre los

datos para generar el modelo y por último se tendrá que evaluar si dicho modelo ha cumplido los criterios de éxito o no (Cortina, 2015)

2.3.4.1. Escoger la Técnica de Modelado

Debido a que se va a utilizar el software WEKA para realizar la minería de datos, deberemos utilizar alguna de las técnicas de modelado que nos ofrece esta herramienta de acuerdo con los objetivos de nuestro proyecto (Cortina, 2015)

De los modelos disponibles que ofrece esta herramienta, se ha seleccionado el modelo de asociación, con el algoritmo de A priori, ya que este permitirá encontrar relaciones dentro un conjunto de transacciones, como también para la segmentación de clientes según ciertos aspectos modelo de Cluster como K-means y Clasificación que se utilizará Random Tree y Naive Bayes

2.3.4.2. Generar el Plan de Prueba

Para medir la calidad, exactitud y validez del modelo se utilizó ciertas métricas entre estas están: Accuracy, Sensibilidad, Especificidad, Coeficiente Kappa, Precisión, etc. Esto se calcula automáticamente en la herramienta Weka el ejecutar el modelo de clasificación por otra parte para el modelo de asociación se mide con la confianza; por último, en la aplicación de cluster con el algoritmo K-means se utilizará Silueta, Índice de Calinski-Harabasz, Índice de Dunn; a continuación, se describe los parámetros para calcular las métricas (Cortina, 2015)

- **TP**: (Verdaderos positivos): Son aquellos valores que el algoritmo clasifica como positivos y que realmente son positivos (Diaz, s.f.).
- **TN**: (Verdaderos negativos): Son aquellos valores que el algoritmo clasifica como negativos (en este caso 0) y que realmente son negativos (Diaz, s.f.).
- **FP**: (Falsos positivos): Son aquellos valores que el algoritmo clasifica como positivos cuando en realidad son negativos (Diaz, s.f.).
- **FN:** (Falsos negativos): Son aquellos valores que el algoritmo clasifica como negativos cuando en realidad son positivos (Diaz, s.f.).
- **VP:** número de verdaderos positivos.
- **Co**: Es la medida de la concordancia observada, expresada como una proporción.
- **Ca**: Es la medida de la concordancia que se esperaría por azar, también expresada como una proporción.

Con las características descritas se calcula las siguientes métricas:

- **Tasa de error:** Es la probabilidad de que la prueba pase por alto un verdadero positivo. Se calcula como FN / FN + VP
- Sensibilidad: Es una medida de cuán bien un modelo puede identificar casos positivos. También se conoce como Tasa de Verdaderos Positivos (TPR) y se calcula dividiendo los casos verdaderamente positivos identificados por el modelo entre el total de casos positivos en el conjunto de datos: VP/(VP+FN) (BERRÍOS, 2015).
- Especificidad: Es una medida de cuán bien un modelo puede identificar casos negativos. También se conoce como Tasa de Verdaderos Negativos (TNR) y se calcula dividiendo los casos verdaderamente negativos identificados por el modelo entre el total de casos negativos en el conjunto de datos: VN/ (VN+FP (BERRÍOS, 2015)).
- Exactitud (Accuracy): Se refiere a la proporción de predicciones correctas que el modelo realiza, es decir, cuán bien el modelo puede predecir los resultados correctos. Se calcula dividiendo el número total de predicciones correctas (verdaderos positivos y verdaderos negativos) entre el número total de predicciones realizadas: (VP+VN) /(VP+FP+FN+VN) (BERRÍOS, 2015).
- Coeficiente Kappa: Es una medida que compara la concordancia observada en un conjunto de datos con la que podría esperarse por mero azar. Se calcula mediante la fórmula K= (Co – Ca) / (1- Ca), donde Co es la proporción de la concordancia observada y Ca es la proporción de concordancia que se esperaría por mero azar (BERRÍOS, 2015).
- Curva ROC: Se representa la sensibilidad del modelo frente al valor obtenido de restar la especificidad a la unidad se calcula 1-especificidad (BERRÍOS, 2015).
- **Precisión:** Esta métrica se refiere a la proporción de verdaderos positivos respecto a todos los resultados positivos (verdaderos positivos y falsos positivos). Se calcula como VP/(VP+FP) (BERRÍOS, 2015).
- Recall: También conocido como la métrica de exhaustividad, se refiere a la proporción de verdaderos positivos respecto a todos los casos positivos presentes en los datos (verdaderos positivos y falsos negativos). Se calcula como TP/(TP+FN) (BERRÍOS, 2015).
- **F – Measure:** Resume la precisión y sensibilidad en una sola métrica. Se calcula: 2*Precisión*Sensibilidad/Precisión + Sensibilidad

- **Confianza:** se refiere a la cantidad de veces que una regla dada resulta ser verdadera en la práctica. Se calcula Soporte (X U Y) / Soporte (X) d
- **Lift:** Compara la frecuencia de un patrón observado con respecto a lo que se esperaría ver ese patrón sólo por azar. Se calcula conf (X→Y) / Soporte(Y)
- **Índice de Calinski-Harabasz:** Es una herramienta importante para garantizar la precisión y significado de los resultados del análisis de conglomerados.
- **Silueta:** La métrica de silueta es una técnica que mide la calidad de un análisis de conglomerados o clustering al evaluar qué tan bien se agrupan los datos en diferentes clusters
- **Índice de Davies-Bouldin:** Este es un método de evaluación interna en el que se determina la calidad de la agrupación utilizando medidas y características que son propias del conjunto de datos.

2.3.4.3. Construir el Modelo

En esta parte se ejecutará el modelo que se ha seleccionado, a continuación, se representa el ajuste a los parámetros que se ha realizado, como los descubrimientos de la aplicación del modelo y la descripción de estos (Cortina, 2015)

Ajuste de parámetros

Los atributos seleccionados se han definido para cumplir con los objetivos planteados, ya que varían según el modelo.

- **Objetivo 1:** Implementar algoritmos para segmentación de clientes
  Dentro de este objetivo se realizó la creación de nuevos atributos, se contiene los siguientes campos

Atributos
idCliente
Nombre
Fecha_nacimiento
Ciudad

Provincia
Tipo _de_ vivienda
Estado
Ocupación
Total

Tabla 11. Atributos para creación de primer modelo

De primera instancia se agrega el campo edad con una función DAX tomando como valor la Fecha_nacimiento

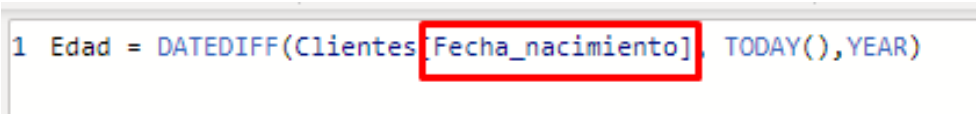


Fig. 24. Función Fecha

Como también se realiza la categorización a cada uno de los clientes en base al Monto total de compras realizadas, en donde

A→≥ 5000 B→ ≥ 500 C→ ≥ 0.01

Para realizar la categorización, realizamos una función DAX tomando en base al campo Total

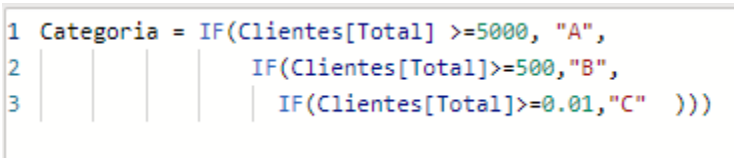


Fig. 25. Función para categorización de clientes

Para finalizar excluimos los campos idCliente, Nombre y Fecha_nacimiento, consecutivamente se transforma a archivo csv (comma-separated values) ya que es compatible con

la herramienta Weka la cual se utilizó para la minería de datos

```
Vivienda;Estado;Ciudad;Provincia;Ocupacion;Edad;Categoria
PROPIA NO HIPOTECADA;CASADO;ANTONIO ANTE;IMBABURA;Venta al por menor de bebidas no alcohólicas (no destinadas al consumo en el lugar de venta) en esta;57;A
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Venta al por menor de gran variedad de productos en tiendas, entre los que predominan, los productos;47;A
ARRENDADA;SOLTERO;IBARRA;IMBABURA;Venta al por menor de gran variedad de productos en tiendas, entre los que predominan, los productos;33;A
PROPIA NO HIPOTECADA;SOLTERO;IBARRA;IMBABURA;Otros cultivos de frutas tropicales y subtropicales: papayas, babacos, chamburos, aguacates, higos,;42;A
PROPIA NO HIPOTECADA;SOLTERO;IBARRA;IMBABURA;EMPLEADO PRIVADO;40;A
ARRENDADA;CASADO;IBARRA;IMBABURA;Elaboración de pan y otros productos de panadería secos: pan de todo tipo, panecillos, bizcochos, to;56;A
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;77;A
VIVE CON FAMILIARES;CASADO;IBARRA;IMBABURA;Actividades de los hogares como empleadores de personal doméstico, como sirvientes, cocineros, camar;33;A
PROPIA NO HIPOTECADA;DIVORCIADO;IBARRA;IMBABURA;Venta al por menor de gran variedad de productos en tiendas, entre los que predominan, los productos;66;B
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Cría y reproducción de cerdos.;53;B
VIVE CON FAMILIARES;SOLTERO;ANTONIO ANTE;IMBABURA;Venta al por menor de productos textiles, prendas de vestir y calzado en puestos de venta y mercados;37;B
VIVE CON FAMILIARES;SOLTERO;IBARRA;IMBABURA;Venta al por menor de frutas, legumbres y hortalizas frescas o en conserva en establecimientos espec;27;B
VIVE CON FAMILIARES;SOLTERO;ESPEJO;CARCHI;Cría y reproducción de cerdos.;23;B
VIVE CON FAMILIARES;CASADO;IBARRA;IMBABURA;Venta al por menor de equipos de telecomunicaciones: celulares, tubos electrónicos, etcétera. Incluy;43;B
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;49;B
ARRENDADA;CASADO;IBARRA;IMBABURA;EMPLEADO PRIVADO;38;B
ARRENDADA;SOLTERO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;38;B
ARRENDADA;SOLTERO;IBARRA;IMBABURA;Fabricación de artículos de bisutería: anillos, brazaletes, collares y artículos de bisuterías simil;32;B
PROPIA NO HIPOTECADA;CASADO;QUITO;PICHINCHA;Venta al por menor de bebidas no alcohólicas (no destinadas al consumo en el lugar de venta) en esta;60;B
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;53;B
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;67;B
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Elaboración de pan y otros productos de panadería secos: pan de todo tipo, panecillos, bizcochos, to;51;B
PROPIA NO HIPOTECADA;VIUDO;IBARRA;IMBABURA;Venta al por menor de gran variedad de productos en tiendas, entre los que predominan, los productos;74;B
ARRENDADA;SEPARADO;IBARRA;IMBABURA;Venta al por menor de prendas de vestir y peletería en establecimientos especializados.;38;B
PROPIA NO HIPOTECADA;UNION DE HECHO;ANTONIO ANTE;IMBABURA;Venta al por menor de gran variedad de productos en tiendas, entre los que predominan, los productos;
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Elaboración de pan y otros productos de panadería secos: pan de todo tipo, panecillos, bizcochos, to;66;B
ARRENDADA;SOLTERO;IBARRA;IMBABURA;Elaboración de pan y otros productos de panadería secos: pan de todo tipo, panecillos, bizcochos, to;33;B
VIVE CON FAMILIARES;SOLTERO;IBARRA;IMBABURA;Venta al por menor de prendas de vestir y peletería en establecimientos especializados.;33;B
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Cultivo de mangos.;51;B
ARRENDADA;CASADO;OTAVALO;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;40;B
VIVE CON FAMILIARES;SOLTERO;IBARRA;IMBABURA;Restaurantes de comida rápida, puestos de refrigerio y establecimientos que ofrecen comida para llev;30;B
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Actividades de los hogares como empleadores de personal doméstico, como sirvientes, cocineros, camar;43;B
PROPIA NO HIPOTECADA;SOLTERO;IBARRA;IMBABURA;Restaurantes, cevicherías, picanterías, cafeterías, etcétera, incluido comida para llevar.;46;B
PROPIA NO HIPOTECADA;CASADO;IBARRA;IMBABURA;Venta al por menor de otros artículos en puestos de venta o mercado como: alfombras, tapices, libros;63;B
PROPIA NO HIPOTECADA;CASADO;OTAVALO;IMBABURA;Venta al por menor de frutas, legumbres y hortalizas frescas o en conserva en establecimientos espec;55;B
ARRENDADA;SOLTERO;SAN MIGUEL DE URCUQUÍ;IMBABURA;Venta al por menor de otros artículos en puestos de venta o mercado como: alfombras, tapices, libros;33;B
```

Fig. 26. Modelo minable para segmentación de clientes

Por otra parte, este mismo modelo lo usaremos para el agrupamiento de clientes con productos, en base a la categorización de clientes y sus características.

```
Vivienda;Estado;Ciudad;Provincia;Ocupacion;Productos;Categoria_Cliente;Rango Edad
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;GALL_FESTIVAL_VAI_X12_6U;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;GALL_FESTIVAL_VAI_12/6U;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;GALLETAS_SALTICAS_TUBO_70G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;DORITOS_QUESO_23G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;DORITOS_QUESO_26G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_35G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_LIMON_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_QUESO_37G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;TATOS_QUESO_43G;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;BOMBRIL_ESTRELLA_DE_6;C;Mayores a 60
ARRENDADA;CASADO;IBARRA;IMBABURA;Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre;CHEETOS_DE_QUESO_21G;C;Mayores a 60
```

Fig. 27. Modelo minable para segmentación de clientes específicos con productos

El siguiente modelo se realizó con los atributos Cliente, Producto, Categoría, en donde el objetivo es agrupar clientes específicos con los productos en base a sus preferencias

```
Cliente;Productos;Categoria
MARIA V CANDO AREQUIP;AMBIENTALES_TIPS_POTPOURRI;DESINFECTANTE
GUDIÑO MEJIA DANILO FRANCISCO;AMB_GLADE_AERSL_PLUM_400ML;DESINFECTANTE
JOE GARCIA;FRESKLIN_DOYPACK_LAVANDA_200ML;DESINFECTANTE
CONSUELO ORDONEZ MINDA;INSECTICIDA_SAPOLIO_MATACUCARACHAS_400CM;DESINFECTANTE
ALVAREZ SILVAARLIVYS;FRESKLIN_DOYPACK_LAVANDA_200ML;DESINFECTANTE
ALVAREZ SILVAARLIVYS;FULL_CLORO_500ML;DESINFECTANTE
LORENA CASANOVA IMBAQUINGO;FULL_CLORO_500ML;DESINFECTANTE
NANCY DE JESUS CUENCA HERRERA;INSECTICIDA_SAPOLIO_MATACUCARACHAS_400CM;DESINFECTANTE
EDUARDO LOPEZ HERRER;PINOKLIN_MANZANA_500_CC;DESINFECTANTE
CRISTINA DE LA TORRE;FRESKLIN_DOYPACK_FLORAL_200ML;DESINFECTANTE
DANIELA TITO;AMBIENTALES_TIPS_MANZANA;DESINFECTANTE
ALINA ANDRADE JIMENEZ;FRESKLIN_MANZANA__1000_CC;DESINFECTANTE
LUIS RECALDE;AMB_TIPS_BANO_EUCALIPTO_90G;DESINFECTANTE
LETICIA MINA ARAUJO;INSECTICIDA_RAID_2ACC_ZANC__DE_360CC;DESINFECTANTE
LETICIA MINA ARAUJO;I_RAID_DOBLE_ACCION_Z/M_400CM;DESINFECTANTE
JOSE ALVAREZ;AMBIENTALES_TIPS_CANELA;DESINFECTANTE
JOSE ALVAREZ;AMBIENTALES_TIPS_MANZANA;DESINFECTANTE
MARIA MOREIRA ALCIVAR;AMBIENTALES_TIPS_CANELA;DESINFECTANTE
ALEXANDRA RODRIGUEZ REVEL;FRESKLIN_LAVANDA_1000_CC;DESINFECTANTE
ALEXANDRA RODRIGUEZ REVEL;AMB_TIPS_BANO_LAVANDA_90G;DESINFECTANTE
MONICA GUERRA;DESIN_OLIMPIA_EUCA_450ML;DESINFECTANTE
GLORIA ESTELA CAICEDO;FULL_CLORO_1000CC;DESINFECTANTE
MARIA JOSE SALAS CHALA;DESIN_OLIMPIA_EUCA_450ML;DESINFECTANTE
CECILIA PINEDA CIFUENTES;DESINFEC_OLIMPIA_EUCAL,_900_CC;DESINFECTANTE
VERONICA ERAZO;GEL_SIDANNE_PLANCHA_X_32U;DESINFECTANTE
BLANCA CASANOVA SALAZAR;DESINF_WINPLUS_CITRU_GL;DESINFECTANTE
BLANCA CASANOVA SALAZAR;DESINF_WINPLUS_CITRU_GALON;DESINFECTANTE
CECILIA PINEDA CIFUENTES;FRESKLIN_DOYPACK_FLORAL_200ML;DESINFECTANTE
GALO RODRIGUEZ;AMBIENTALES_TIPS_CANELA;DESINFECTANTE
DERMA  VERONICA FREIRE PATOJA;INSECTICIDA_SAPOLIO_MATACUCARACHAS_400CM;DESINFECTANTE
```

Fig. 28. Modelo minable para segmentación de clientes categorizados con productos

- **Objetivo 2**: Algoritmos para estimación de ventas

Para el modelo enfocado en este objetivo, se tomó en cuenta las 12477 facturas con más productos, como también los 35 productos más vendidos. Para la realización de la vista minable se realizó una tabla dinámica con la función Buscar, la cual busca los productos que se encuentra en las cada una de las facturas, la cual si hay coincidencia dará como resultado 1, a continuación, se muestra la función y la tabla dinámica

```
=BUSCARV([@NombreProducto];$E$2:$F$40;2;FALSO)
```

Fig. 29. Función buscar en Excel

Etiquetas de fila	ARROCILLO_lbr	ATUN_REAL_A/F_ACEITE_160GR	AVENA_GRANEL	AZUCAR_LB	AZUCAR_TABABUELA	CAFE_BUENDIA_10g	CAFE_COLCAFE_SOB_10gr	CANGUIL	DORITOS_DE_LIMOI
5F250607				1					
5F250647								1	
5F250904						1			
5F250990								1	
5F251459				1		1			
5F251498		1				1		1	
5F251603								1	
5F251649								1	
5F251965						1			
5F252145						1		1	
5F252436								1	
5F253300								1	
5F253344						1			
5F253946		1							
5F254828						1			
5F255367				1					
5F255615		1		1				1	
5F255693				1		1			
5F256308		1		1				1	

Fig. 30. Tabla dinámica de transacciones

A partir de esto se reemplazó las casillas vacías como N y las que contiene 1 como V, también se eliminó la Columna en donde se encuentran las Facturas, dando como resultado ya nuestro archivo para realizar la aplicación del algoritmo de asociación

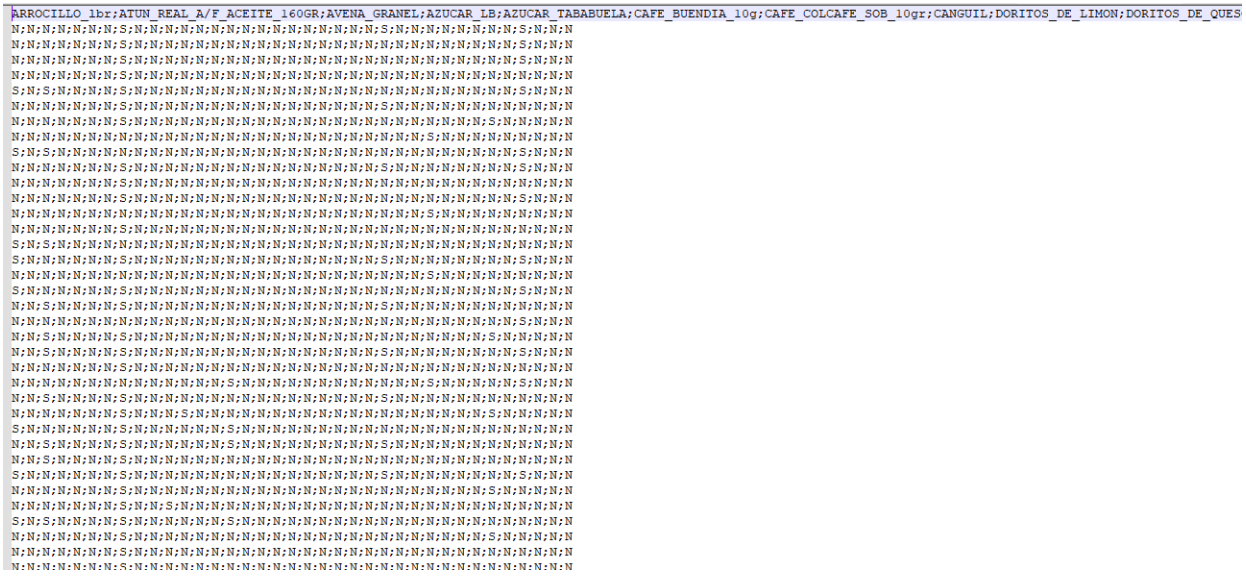


Fig. 31.Modelo minable para asociación de productos

2.3.5. Evaluación

Las vistas minables se desarrolla en la herramienta Weka en donde se aplicara los algoritmos de clasificación (Random Tree y Naive Bayes), agrupación (k-means) y asociación (A priori y FPGrowth), las cuales se describe a continuación y los resultados se observan en el siguiente capítulo

a) Aplicación de Algoritmo de Clasificación

Esta se basa en la segmentación de clientes en donde se encontró características similares de una persona perteneciente a cada categoría, se utilizó los Algoritmos Random Tree y Naive Bayes

- **Random Tree**

  Se utilizó una gran cantidad de datos para el entrenamiento debido a la simplicidad del modelo, la facilidad de interpretación y su velocidad en la clasificación de nuevos datos (Vila, 2019).

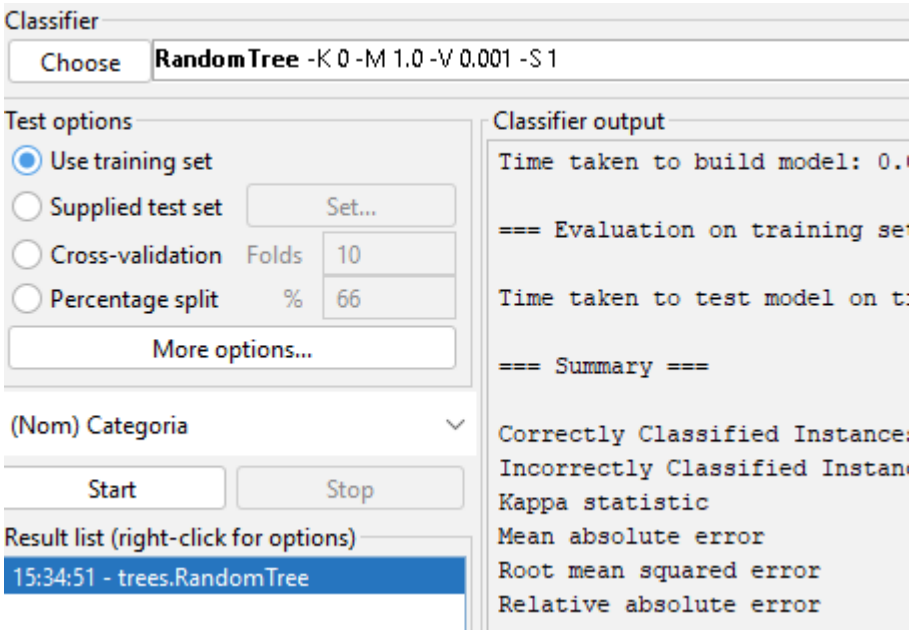


Fig. 32. Aplicación Random Tree

- **Naive Bayes**

  El clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características

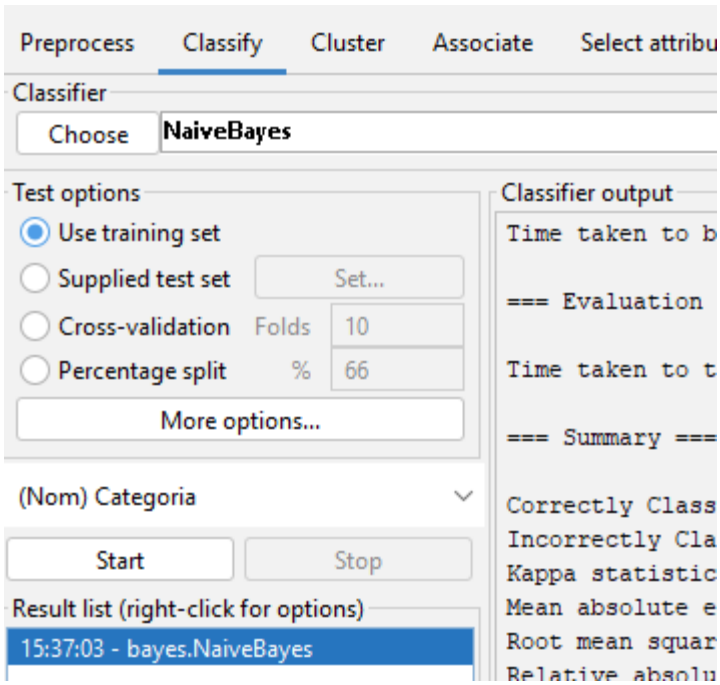


Fig. 33. Aplicación de Naive Bayes

b) Aplicación de Algoritmo de Agrupamiento

Esta se basa en el agrupamiento de clientes categorizados y clientes en específico con los productos que tienen mayor preferencia, se utilizó el algoritmo K-means, para conocer el número óptimo de clusters, se realizó "método del codo" o "elbow method", que implica trazar la suma de las distancias cuadradas intra-cluster en función del número de clusters. El número óptimo de clusters es donde la suma de las distancias intra-cluster deja de disminuir significativamente, formando una curva con forma de codo

- **Clientes en específico**

  Aplicando el método del codo en conjunto con inercia, entre un rango de 2 a 20 se concluyó que utilizaremos 6 clusters

```
# Graficar la inercia en función del número de clústeres
plt.plot(range(2, 20), inertias)
plt.xlabel("Número de clústeres")
plt.ylabel("Inercia")
plt.title("Gráfico de codo")
plt.show()
```

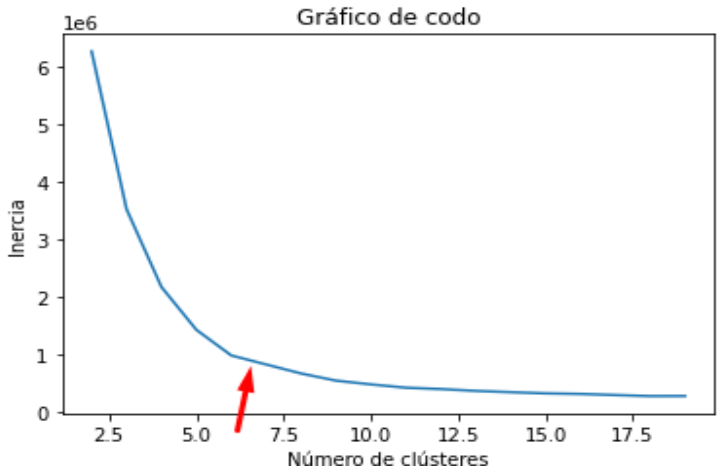


Fig. 34. Aplicación del método codo e inercia

Con el número de clusters óptimo aplicamos el algoritmo K-means el cual agrupó el producto de mayor preferencia con el nombre del Cliente

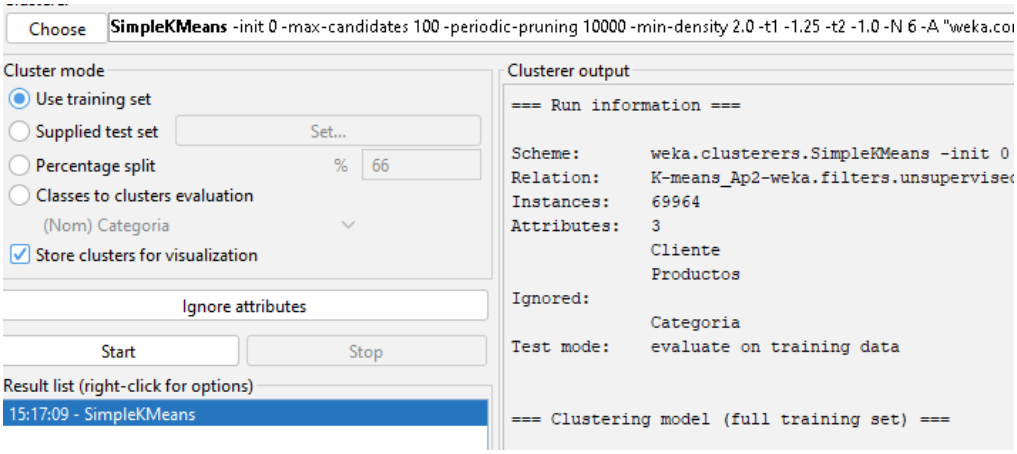


Fig. 35. Aplicación de algoritmo K-Means para clientes específicos

- **Clientes por categorización**

Aplicando el método del codo en conjunto con inercia, entre un rango de 2 a 20 se concluyó que utilizaremos 5 clusters

```
# Graficar la inercia en función del número de clústeres
plt.plot(range(2, 20), inertias)
plt.xlabel("Número de clústeres")
plt.ylabel("Inercia")
plt.title("Gráfico de codo")
plt.show()
```

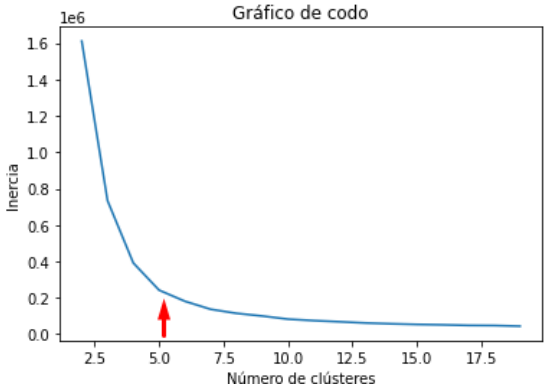


Fig. 36. Aplicación del método codo e inercia

El algoritmo K-means se aplicará con 5 clusters, en donde se obtuvo la agrupación de características de clientes similares con un cierto producto

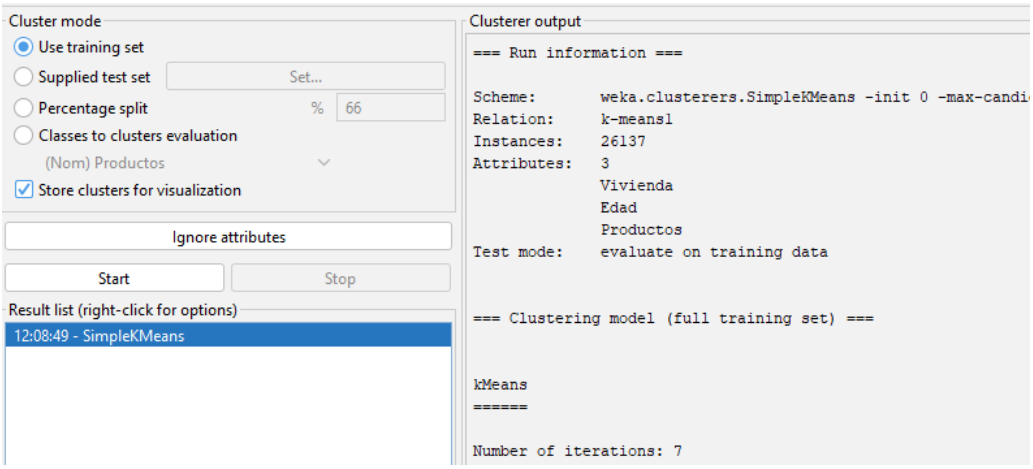


Fig. 37. Aplicación de algoritmo K-Means para segmentación de clientes

a) Aplicación de Algoritmo de Asociación

Esta se basa en la asociación de productos en donde se encontró características similares de los productos vendidos en base a las facturas realizadas, se utilizó el algoritmo A priori y FPGrowth

- A priori

Este algoritmo toma cada parte de un conjunto de datos más grande y lo "puntúa" o lo contrasta con otros conjuntos de alguna manera ordenada

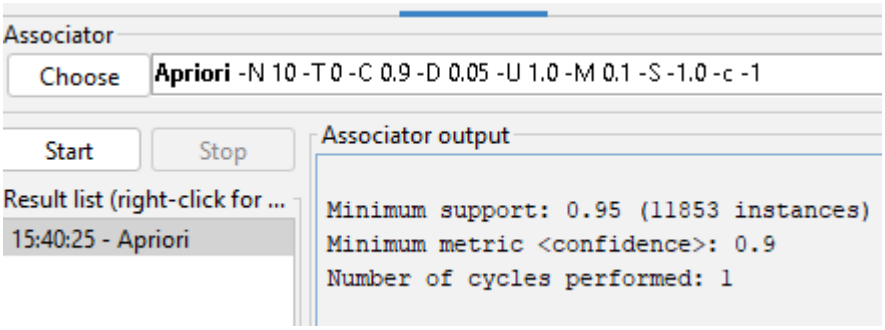


Fig. 38. Aplicación de algoritmo A priori

- FPGrowth

Deriva del "a-priori" y es un algoritmo que se caracteriza por ser muy eficiente y además escalable, es decir, se puede usar para grandes volúmenes de datos con un orden razonablemente bajo

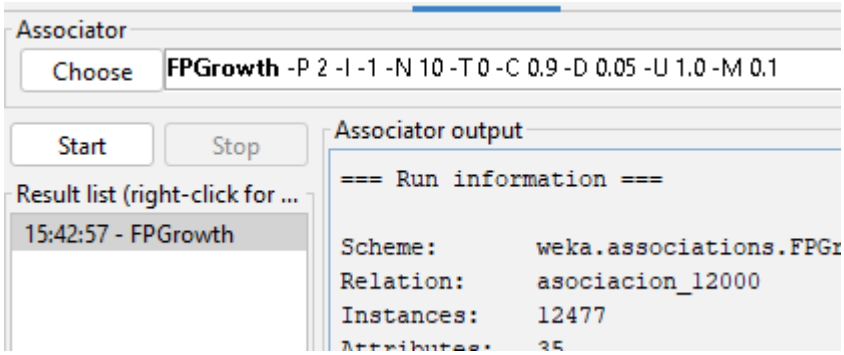


Fig. 39.Aplicación de algoritmo FPGrowth

CAPÍTULO 3

RESULTADOS

3.1. Evaluación de algoritmos mediante métricas de rendimiento

Para evaluar los algoritmos aplicados (clasificación y regresión) se emplearon métricas cuantitativas de calidad

3.1.1. Evaluación de Tareas de clasificación

- Random Tree

En la siguiente figura se muestra el desarrollo del algoritmo con 511 instancias y 6 atributos

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     Clasificacion_Final
Instances:    511
Attributes:   6
              Vivienda
              Estado
              Provincia
              Ocupacion
              Categoria
              Rango Edad
Test mode:    evaluate on training data

=== Classifier model (full training set) ===


RandomTree
==========

Ocupacion = Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, entre
|   Estado = CASADO
|   |   Vivienda = ARRENDADA
|   |   |   Rango Edad = Mayores a 60 : C (37/4)
|   |   |   Rango Edad = Entre 40 y 59
|   |   |   |   Provincia = IMBABURA : C (37/4)
|   |   |   |   Provincia = CARCHI : C (1/0)
|   |   |   |   Provincia = ESMERALDAS : C (0/0)
|   |   |   |   Provincia = PICHINCHA : C (0/0)
|   |   |   |   Provincia = SUCUMBIOS : C (0/0)
|   |   |   |   Provincia = NAPO : C (0/0)
|   |   |   Rango Edad = Menores de 40 : C (8/0)
```

Fig. 40.Primera parte de resultados de aplicación de Random Tree

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         482               94.3249 %
Incorrectly Classified Instances        29                5.6751 %
Kappa statistic                          0.6533
Mean absolute error                      0.0545
Root mean squared error                  0.165
Relative absolute error                 39.4593 %
Root relative squared error             63.2277 %
Total Number of Instances              511

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,998    0,466    0,944      0,998    0,970      0,697    0,962     0,993     C
                 0,500    0,002    0,800      0,500    0,615      0,628    0,990     0,702     A
                 0,520    0,002    0,963      0,520    0,675      0,688    0,963     0,797     B
Weighted Avg.    0,943    0,413    0,943      0,943    0,936      0,695    0,963     0,970

=== Confusion Matrix ===

   a   b   c   <-- classified as
 452   0   1 |   a = C
   4   4   0 |   b = A
  23   1  26 |   c = B
```

Fig. 41.Segunda parte de resultados de aplicación de Random Tree

En la Tabla 12. se observa la matriz de confusión que se obtiene después de ejecutar el algoritmo Random Tree, usando que se entrene todo el dataset

		ClASE PREDICHA		
		C	B	A
	C	452	0	1
Clase verdadera	B	4	4	0
	A	23	1	26

Tabla 12. Matriz de confusión de Random Tree

En la aplicación de esta técnica existen otras métricas para medir la calidad como se puede observar en la Tabla 13.

Medida	Valor
Accuracy	94.32%
Tasa de error	5.67%
Sensibilidad	99.8%
Coeficiente Kappa	0.653
Curva ROC	0,96
Precisión	94.3%
Recall	94.3%
TP Rate	94.3%
FP Rate	41.3%
F – Measure	93.6%

Tabla 13. Métricas de Random Tree

- **Naive Bayes**

  En la siguiente Fig. 41. se muestra el desarrollo del algoritmo con 511 instancias y 6 atributos

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Clasificacion_Final
Instances:    511
Attributes:   6
              Vivienda
              Estado
              Provincia
              Ocupacion
              Categoria
              Rango Edad
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

                                                                                                       Class
Attribute                                                                                                  C      A      B
                                                                                                      (0.88) (0.02)  (0.1)
==============================================================================================================================
Vivienda
  ARRENDADA                                                                                            185.0    4.0   19.0
  PROPIA NO HIPOTECADA                                                                                 144.0    5.0   20.0
  VIVE CON FAMILIARES                                                                                  115.0    2.0   12.0
  PRESTADA                                                                                               7.0    1.0    2.0
  PROPIA HIPOTECADA                                                                                      7.0    1.0    2.0
  [total]                                                                                              458.0   13.0   55.0

Estado
  CASADO                                                                                               249.0    6.0   25.0
```

Fig. 42.Primera parte de resultados de aplicación de Naive Bayes

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         454               88.8454 %
Incorrectly Classified Instances        57               11.1546 %
Kappa statistic                          0.057
Mean absolute error                      0.1172
Root mean squared error                  0.2511
Relative absolute error                 84.9229 %
Root relative squared error             96.218  %
Total Number of Instances              511

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,998    0,966    0,890      0,998    0,941      0,134    0,645     0,917     C
                 0,000    0,002    0,000      0,000    0,000      -0,006   0,628     0,027     A
                 0,040    0,000    1,000      0,040    0,077      0,190    0,666     0,340     B
Weighted Avg.    0,888    0,856    0,887      0,888    0,841      0,137    0,647     0,847

=== Confusion Matrix ===

   a   b   c   <-- classified as
 452   1   0 |   a = C
   8   0   0 |   b = A
  48   0   2 |   c = B
```

Fig. 43.Segunda parte de resultados de aplicación de Naive Bayes

En la Tabla 14. se observa la matriz de confusión que se obtiene después de ejecutar el algoritmo Naive Bayes, usando que se entrene todo el dataset

		Clase predicha		
		A	B	C
	C	452	1	0
Clase verdadera	A	8	0	0
	B	48	0	2

Tabla 14. Matriz de confusión de Naive Bayes

En la aplicación de esta técnica existen otras métricas para medir la calidad como se observa en la Tabla 15.

Medida	**Valor**
Accuracy	88.84%
Tasa de error	11.15%
Sensibilidad	99.8%
Coeficiente Kappa	0.057
Curva ROC	0.647
Precisión	88.7%
Recall	88.8%
TP Rate	88.8%
FP Rate	85.6%
F – Measure	84.1%

Tabla 15. Métricas de evaluación de Naive Bayes

3.1.2. Evaluación de Tareas de agrupamiento

a) Clientes en específico

```
Clusterer output
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 118589.00000000001

Initial starting points (random):

Cluster 0: 'WASHINTON DAVID GUEVARA CASTILLO',DEJA_X_36_DE_360_Gr
Cluster 1: 'AIDA FRAGA',TALCO_REXONA_55G
Cluster 2: 'BETTY MARISOL TIRIRA',ENERG_V220_12U/600ML
Cluster 3: 'MARIA ROMELIA CAMPUES',ACEITE_PAL/ORO_F_15U_100_280ML
Cluster 4: 'JUSTO FELIX LOMAS',SH_HYS_375ML_PROT/CAID
Cluster 5: 'LOPEZ GUERRERO  OSCAR ALEXANDER',CARA_LECHE_MIEL_UNIV_450G

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                                          Cluster#
Attribute                                                             Full Data                  0
                                                                      (69964.0)          (57150.0)
====================================================================================================
Cliente                                                      JESUS PORTILLA MELO    JESUS PORTILLA MELO
Productos                                                  AZUCAR_TABABUELA_50_KG  AZUCAR_TABABUELA_50_KG      AC
```

Fig. 44.Primera parte de resultados de aplicación de K-means para clientes específicos

```
                                                                                          Cluster#
Attribute                                                             Full Data                  0
                                                                      (69964.0)          (57150.0)
====================================================================================================
Cliente                                                      JESUS PORTILLA MELO    JESUS PORTILLA MELO
Productos                                                  AZUCAR_TABABUELA_50_KG  AZUCAR_TABABUELA_50_KG      A




Time taken to build model (full training data) : 0.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      57150 ( 82%)
1       1483 (  2%)
2       5885 (  8%)
3        531 (  1%)
4       3253 (  5%)
5       1662 (  2%)
```

Fig. 45.Segunda parte de resultados de aplicación de K-means para clientes específicos

En la Tabla 16. Se puede observar las métricas aplicadas a este modelo y el resultado de cada una de ellas

Métrica	Valor	Descripción
Índice de Calinski-Harabasz	156599.72	Aceptable

Índice de Davies-Bouldin	0.489	Aceptable
Medida de silueta	0.604	Aceptable

Tabla 16. Evaluación de modelo de clientes específicos

Para describir si es aceptable o no depende del número de datos y los clusters que se han realizado. Con el Índice de Calinski-Harabasz se ha obtenido el valor de 156599.72 en donde, sugiere que los clusters tienen una buena separación entre sí y están bien definidos en términos de su distribución en el espacio de características. Índice de Davies-Bouldin se ha alcanzado la cifra de 0.489 lo que indica una buena calidad de la agrupación, ya que es relativamente bajo, esto sugiere que los clusters están bien separados y tienen características distintas. Y la medida silueta sugiere que mientras más cerca esté el valor de 1 por lo que 0.60 es un valor moderado y sugiere que los clusters tienen cierta cohesión y separación entre sí.

b) Clientes por características

```
Clusterer output
              Productos
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 31787.84779436819

Initial starting points (random):

Cluster 0: ARRENDADA,77,PIMIE_CHINI_DIS_50_SOB
Cluster 1: 'PROPIA NO HIPOTECADA',66,PASTA_COLGATE_T/A_60ML
Cluster 2: 'PROPIA NO HIPOTECADA',47,SAL_CRISAL_GRANDE_X_25
Cluster 3: 'PROPIA NO HIPOTECADA',66,BARBERA_BIC_CONFORT_3
Cluster 4: 'PROPIA NO HIPOTECADA',74,ACEITE_FAVORITA_DE_365_ML

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                                   Cluster#
Attribute                                           Full Data             0                     1
                                                    (26137.0)      (9527.0)              (2078.0)
=====================================================================================================
Vivienda                                  PROPIA NO HIPOTECADA      ARRENDADA  PROPIA NO HIPOTECADA   PROPIA N
Edad                                                   49.0555        52.6697               63.6165
Productos                                 HARINA_PANADERA_50_KG  HARINA_PANADERA_50_KG PASTA_COLGATE_T/A_60ML
```

Fig. 46.Primera parte de resultados de aplicación de K- means para clientes por características

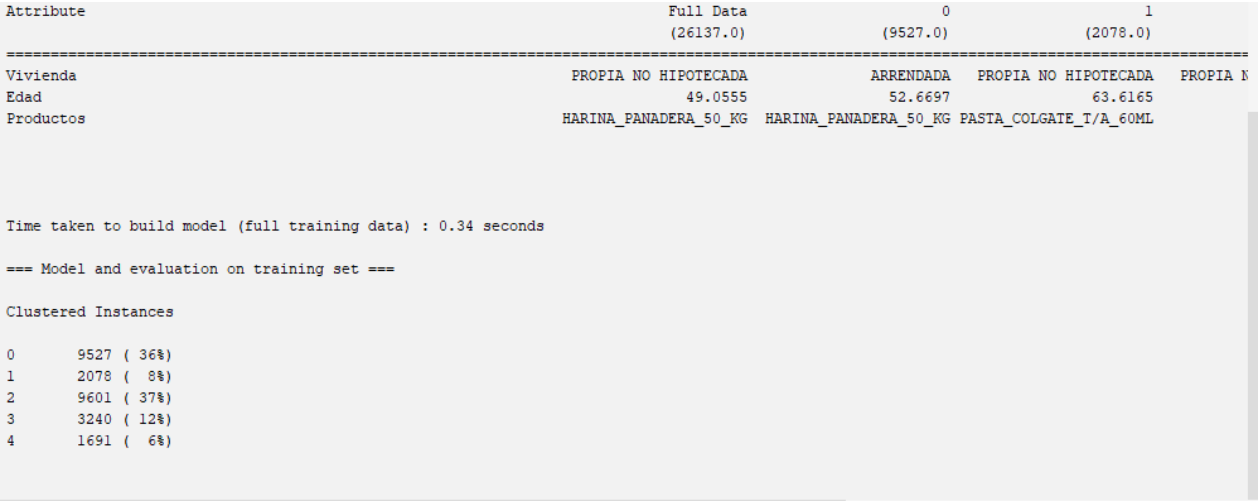


Fig. 47.Segunda parte de resultados de aplicación de K- means para clientes por características

En la Tabla 17. Se puede observar las métricas aplicadas a este modelo y el resultado de cada una de ellas

Métrica	**Valor**	**Descripción**
Índice de Calinski-Harabasz	166781.26	Aceptable
Índice de Davies-Bouldin	0.509	Aceptable
Medida de silueta	0.603	Aceptable

Tabla 17. Evaluación de modelos de clientes con características

Para describir si es aceptable o no depende del número de datos y los clusters que se han realizado. Ejecutando el Índice de Calinski-Harabasz y obteniendo el valor 166781.26 sugiere que los clusters tienen una buena separación y cohesión entre sí. El Índice de Davies-Bouldin con un valor de 0.509 indica que los clusters están bien definidos y no se superponen significativamente. Medida de silueta adquirió el valor de 0.603 muestra que el valor de silueta es mayor a 0.5 lo que indica que los clusters están bien definidos

3.1.3. Evaluación de Tareas de asociación

- **A priori**

En la siguiente Fig. 47. se muestra el desarrollo del algoritmo con 12477 instancias y 35 atributos

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     asociacion_12000
Instances:    12477
Attributes:   35
              ARROCILLO_1br
              ATUN_REAL_A/F_ACEITE_160GR
              AVENA_GRANEL
              AZUCAR_LB
              AZUCAR_TABABUELA
              CAFE_BUENDIA_10g
              CAFE_COLCAFE_SOB_10gr
              CANGUIL
              DORITOS_DE_LIMON
              DORITOS_DE_QUESO
              FIDEO_CAYAMBE_TLL_87_400G
              GALLETAS_SALTICAS_TUBO_70G
              HARINA_ALPAN_LBR
              HARINA_FLOR_1LBR
              HARINA_MAIZ_CRUDO_LBR
              HARINA_PANADERA_1LB
              HARINA_PANADERA_50_KG
              HARINA_PANIFICACION_LBR
              HARINA_TOMEBAMBA_Lbr
              HARINA_TRIGO_LBR
              HUEVOS_G
              HUEVOS_G_X_30
              LECHE_ANDINA_LARVIDA_ENT
              LENTEJON_CANADIENSE__Lbr
              LEVADURA_LEVAPAN_500_G
              M_MAGGI_DE_30g
```

Fig. 48. Primera parte de resultados de aplicación de A priori

```
Minimum support: 0.95 (11853 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 83

Size of set of large itemsets L(3): 140

Size of set of large itemsets L(4): 110

Size of set of large itemsets L(5): 35

Size of set of large itemsets L(6): 3

Best rules found:

 1. PAPA_FRITA_NATURAL=F 12474 ==> SH_SAVITAL_KERAT_25ML=F 12473    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
 2. FID_CAYAMBE_LAZO_CHI_400G=F 12441 ==> SH_SAVITAL_KERAT_25ML=F 12440    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
 3. FID_CAYAMBE_LAZO_CHI_400G=F PAPA_FRITA_NATURAL=F 12439 ==> SH_SAVITAL_KERAT_25ML=F 12438    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
 4. HUEVOS_G=F 12387 ==> SH_SAVITAL_KERAT_25ML=F 12386    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
 5. HUEVOS_G=F PAPA_FRITA_NATURAL=F 12385 ==> SH_SAVITAL_KERAT_25ML=F 12384    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
 6. HUEVOS_G=F FID_CAYAMBE_LAZO_CHI_400G=F 12352 ==> SH_SAVITAL_KERAT_25ML=F 12351    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
 7. HUEVOS_G=F FID_CAYAMBE_LAZO_CHI_400G=F PAPA_FRITA_NATURAL=F 12350 ==> SH_SAVITAL_KERAT_25ML=F 12349    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
 8. PIPA_NIC_LIMON_25CVS=F 12334 ==> SH_SAVITAL_KERAT_25ML=F 12333    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
 9. PIPA_NIC_LIMON_25CVS=F PAPA_FRITA_NATURAL=F 12332 ==> SH_SAVITAL_KERAT_25ML=F 12331    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
10. PIPA_NIC_LIMON_25CVS=F FID_CAYAMBE_LAZO_CHI_400G=F 12299 ==> SH_SAVITAL_KERAT_25ML=F 12298    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.99)
```

Fig. 49. Segunda parte de resultados de aplicación de A priori

A continuación, en la Tabla 18. Se muestran las medidas con el valor promedio de todos los hallazgos con el algoritmo A priori, sin embargo, las medidas deben evaluarse individualmente.

Medida	Promedio valor
Confianza	**1**
Lift	**1**
Convicción	**0.99**

Tabla 18. Promedio de métricas de evaluación A priori

- **FP Growth**

  En la Fig. 49. se muestra el desarrollo del algoritmo con 12477 instancias y 35 atributos

```
Scheme:       weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
Relation:     asociacion_12000
Instances:    12477
Attributes:   35
              ARROCILLO_1br
              ATUN_REAL_A/F_ACEITE_160GR
              AVENA_GRANEL
              AZUCAR_LB
              AZUCAR_TABABUELA
              CAFE_BUENDIA_10g
              CAFE_COLCAFE_SOB_10gr
              CANGUIL
              DORITOS_DE_LIMON
              DORITOS_DE_QUESO
              FIDEO_CAYAMBE_TLL_87_400G
              GALLETAS_SALTICAS_TUBO_70G
              HARINA_ALPAN_LBR
              HARINA_FLOR_1LBR
              HARINA_MAIZ_CRUDO_LBR
              HARINA_PANADERA_1LB
              HARINA_PANADERA_50_KG
              HARINA_PANIFICACION_LBR
              HARINA_TOMEBAMBA_Lbr
              HARINA_TRIGO_LBR
              HUEVOS_G
              HUEVOS_G_X_30
              LECHE_ANDINA_LARVIDA_ENT
              LENTEJON_CANADIENSE__Lbr
              LEVADURA_LEVAPAN_500_G
              M_MAGGI_DE_30g
```

Fig. 50. Primera parte de resultados de aplicación de FP Growth

```
              LEVADURA_LEVAPAN_500_G
              M_MAGGI_DE_30g
              MANTECA
              PASTA_COLGATE_T/A_60ML
              PIPA_NIC_LIMON_25CVS
              SAL_CRISAL_2_KS
              TATOS_LIMON_35G
              FID_CAYAMBE_LAZO_CHI_400G
              SH_SAVITAL_KERAT_25ML
              HARINA_TOMEBAMBA_50_KG
              PAPA_FRITA_NATURAL
=== Associator model (full training set) ===

FPGrowth found 8 rules (displaying top 8)

1. [AZUCAR_TABABUELA=F]: 9475 ==> [AZUCAR_LB=F]: 9475   <conf:(1)> lift:(1.32) lev:(0.18) conv:(2278.95)
2. [HARINA_PANIFICACION_LBR=S]: 1377 ==> [HARINA_FLOR_1LBR=S]: 1377   <conf:(1)> lift:(9.06) lev:(0.1) conv:(1225.03)
3. [HARINA_FLOR_1LBR=S]: 1377 ==> [HARINA_PANIFICACION_LBR=S]: 1377   <conf:(1)> lift:(9.06) lev:(0.1) conv:(1225.03)
4. [AZUCAR_LB=F, SAL_CRISAL_2_KS=S]: 1524 ==> [AZUCAR_TABABUELA=F]: 1524   <conf:(1)> lift:(1.32) lev:(0.03) conv:(366.68)
5. [AZUCAR_TABABUELA=F, SAL_CRISAL_2_KS=S]: 1524 ==> [AZUCAR_LB=F]: 1524   <conf:(1)> lift:(1.32) lev:(0.03) conv:(366.56)
6. [AZUCAR_LB=F, ATUN_REAL_A/F_ACEITE_160GR=S]: 1466 ==> [AZUCAR_TABABUELA=F]: 1466   <conf:(1)> lift:(1.32) lev:(0.03) conv:(352.72)
7. [AZUCAR_TABABUELA=F, ATUN_REAL_A/F_ACEITE_160GR=S]: 1466 ==> [AZUCAR_LB=F]: 1466   <conf:(1)> lift:(1.32) lev:(0.03) conv:(352.61)
8. [AZUCAR_LB=F]: 9476 ==> [AZUCAR_TABABUELA=F]: 9475   <conf:(1)> lift:(1.32) lev:(0.18) conv:(1139.98)
```

Fig. 51. Segunda parte de resultados de aplicación de FP Growth

A continuación, en la Tabla 19. Se muestran las medidas con el valor promedio de todos los hallazgos con el algoritmo FP Growth sin embargo, las medidas deben evaluarse individualmente.

Medida	Promedio valor
Confianza	**0.93**
Lift	**27.52**
Convicción	**1027.93**

Tabla 19. Promedio de métricas de evaluación FP Growth

La exactitud de las reglas formadas se evaluará con las métricas donde:

a) **conf:** Mientras el valor esté más cerca de 1 es más preciso
b) **lift:** Si el valor del lift es igual a 1, esto significa que la aparición del conjunto en cuestión es la misma que la esperada en condiciones de independencia. Si el valor del lift es mayor que 1, indica que el conjunto aparece más veces de lo esperado bajo condiciones de independencia. Por otro lado, si el valor del lift es menor que 1, esto sugiere que el conjunto aparece menos veces de lo esperado bajo condiciones de independencia (Barrios, 2019).

- **A priori**

  En la Tabla 20. Indica las reglas con las respectivas métricas

Regla Nro	**Conf**	**Lift**
1	**1**	**1**
2	**1**	**1**
3	**1**	**1**
4	**1**	**1**
5	**1**	**1**
6	**1**	**1**
7	**1**	**1**
8	**1**	**1**
9	**1**	**1**
10	**1**	**1**

Tabla 20. Métricas de evaluación de reglas A priori

- **FP Growth**

Regla Nro	**Conf**	**Lift**
1	**1**	**1.32**
2	**1**	**9.06**
3	**1**	**9.06**
4	**1**	**1.32**
5	**1**	**1.32**
6	**1**	**1.32**
7	**1**	**1.32**
8	**1**	**1.32**

En la Tabla 21. Indica las reglas obtenidas con las métricas de evaluación

Regla Nro	**Conf**	**Lift**
1	**1**	**1.32**
2	**1**	**9.06**
3	**1**	**9.06**
4	**1**	**1.32**
5	**1**	**1.32**
6	**1**	**1.32**
7	**1**	**1.32**
8	**1**	**1.32**

Tabla 21. Métricas de evaluación de reglas FP Growth

Con la Aplicación de los Algoritmos Apriori y FP Growth sobre el modelo para encontrar la asociación de los productos vendidos muestra una confianza aproximada de 1 por lo que es un rango más que aceptable

3.2. Análisis e interpretación de los resultados

3.2.1. Análisis e interpretación de resultados de las tareas de clasificación

- **Análisis de resultados**

La evaluación cuantitativa a los resultados de la aplicación de técnicas de clasificación como son Random Tree y Naive Bayes, utilizando todos los datos para el entrenamiento de esta, nos mostró varias métricas para evaluar, como son la exactitud, precisión, tasa de error, coeficiente Kappa, tasa de Tp, F measure, tasa FT y la Curva ROC.

Como se puede observar , el trabajo realizado principalmente es para segmentar los clientes en base a la categoría que representa el monto invertido en compras en donde se ha trabajado con 511 instancias y 6 atributos, en donde se observa que la categoría predominante es la "C" que pertenece a compras realizadas menores a $500, con registro de 450 clientes, mientras tanto en la categoría "B" con 50 datos y por cantidad inferior de 8 registros la categoría "A" , al realizar la aplicación de los algoritmos sobre estos se muestra que los clasificadores cuentan con valores aceptables sin embargo indica un diferencia mostrando que el Random Tree ha logrado mejores resultados, los cuales se describirán a continuación:

El algoritmo Random Tree clasificó 482 instancias de manera correcta con llevando a 29 instancias mal clasificadas, por lo que obtiene lo siguiente

- La categoría A que son los valores de compras mayores a $5000 4 registros y 4 de forma errónea
- Para la categoría B que pertenecen a las comprar entre $500 y $4999 clasifica como 26 correctos y 24 incorrectos
- Para la categoría C que son los valores inferiores a $500 clasifica como 452 registros y un dato erróneo

El Algoritmo Naive Bayes clasificó 452 instancias correctamente por otro lado obtiene 59 instancias mal clasificadas

- La categoría A y B muestra que se han clasificado 0 valores
- Para la categoría C indica qué 452 registros se han clasificado correctamente

A simple en el momento de clasificar el Algoritmo de Random Tree es superior a Naive Bayes, como también basándose en las evaluaciones de las métricas, por lo tanto, la información que será utilizada será del algoritmo Random Tree donde proporciona un conjunto de reglas para la toma de decisiones (similar a un árbol de decisión) y muestra un procesamiento simple e interpretación de la información

- **Interpretación de resultados**

En el proceso de análisis de resultados y selección del mejor modelo de clasificación utilizando el algoritmo Random Tree, la interpretación de datos se lleva a cabo desde la raíz del árbol hacia las hojas, enfocándose específicamente en aquellas hojas cuyo atributo de clase es de interés, la interpretación de datos se realizó desde la raíz hacia las hojas cuyo atributo de clase sea Categoría = C identificando los siguientes patrones de clientes pertenecientes a esta, ya que la mayoría pertenecen a esta y son los clientes principales que debe enfocarse el comercial (Vila, 2019).

Pertenecen a la Categoría C

- Personas mayores a 60 años que son casadas cuya vivienda es arrendada y se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas.
- Personas mayores a entre 40 y 59 años que son casadas cuya vivienda es arrendada y se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas, proveniente de Imbabura
- Personas menores de 40 que son casadas cuya vivienda es arrendada y se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas.
- Personas menores de 40 que son divorciadas cuya vivienda es arrendada y se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas.
- Personas mayores a 60 y son solteros se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas.
- Personas que son menores de 60 que son de Imbabura y se dedican a la Fabricación de prendas de vestir de telas tejidas, de punto y ganchillo, de telas no tejidas.
- Personas casado que se dedica a Lavado y limpieza en seco
- Personas menores de 40 años que es soltera y vive en vivienda propia, que se dedica a Lavado y limpieza en seco

- Personas menores de 40 años que es soltera y vive con familiares y se dedica a Lavado y limpieza en seco
- Personas mayores a 60 años que son casadas y se dedican a la Venta al por menor de gran variedad de productos en tiendas
- Personas entre 40 y 59 años que son casadas, viven en vivienda propia y se dedican a la Venta al por menor de gran variedad de productos en tiendas
- Personas menores a 40 años y se dedican a la Venta al por menor de gran variedad de productos en tiendas
- Personas entre 40 y 59 años que son solteras y se dedican a la Venta al por menor de gran variedad de productos en tiendas
- Personas menores de 40 años que viven con los familiares y se dedican a la Venta al por menor de gran variedad de productos en tiendas
- Personas que son menores a 60 años, provenientes a Imbabura y se dedican a la Venta al por menor de bebidas no alcohólicas
- Personas que se dedican a todas las actividades de transporte de carga por carretera
- Personas mayores a 60 años y que viven arrendando, dedicándose a la venta al por menor de otros artículos en puestos de venta o mercado
- Personas entre 40 y 59 años, que se dedica a la venta al por menor de otros artículos en puestos de venta o mercado
- Personas menores a 40 años que viven con familiares y se dedican a la venta al por menor de otros artículos en puestos de venta o mercado
- Personas de Pichincha que se dedican a la producción de leche cruda de vaca
- Personas casadas que tienen vivienda propia y se dedican a la cría y reproducción de cerdos
- Personas solteras de Imbabura
- Personas que se dedican a explotación de criaderos de pollos y reproducción de aves de corral, pollos y gallinas
- Personas que se dedican a la venta al por menor de gran variedad de productos entre los que no predominan los productos alimenticios
- Personas dedicadas a la venta al por menor de carne y productos cárnicos (incluidos los de aves de corral)

- Personas casadas dedicadas a restaurantes, cevicherías, picanterías, cafeterías, etcétera, incluido comida para llevar.
- Personas dedicadas Venta al por menor de perfumes, artículos cosméticos y de uso personal
- Personas que trabajan en construcción de todo tipo de edificios residenciales: casas familiares individuales
- Personas que trabajan en transporte terrestre de pasajeros por sistemas de transporte urbano
- Personas menores de 40 años que se encuentran solteros y se dedican a Venta al por menor de prendas de vestir y peletería en establecimientos especializados
- Persona que trabajan en servicios de apoyo a la fabricación de prendas de vestir
- Personas que viven con familiares y se dedica a cultivos de frutas tropicales y subtropicales
- Personas casadas que viven arrendando y se dedican la elaboración de pan y otros productos de panadería secos
- Personas entre 40 y 59 años, proveniente de Imbabura que son casadas y trabajan en la elaboración de pan y otros productos de panadería secos
- Personas que viven con familiares y trabajan en elaboración de pan y otros productos de panadería secos
- Personas dedicadas a actividades de acondicionamiento y mantenimiento de terrenos para usos agrícolas
- Personas dedicadas a actividades de lavado, corte, recorte, peinado, teñido, coloración, ondulación y alisado del cabello
- Personas que trabajan en cultivo de papa
- Personas dedicadas a actividades de operaciones preparatorias de fibras textiles: devanado y lavado de seda, desengrase
- Personas que se ocupan en servicios sociales, de asesoramiento, de bienestar social, de remisión y servicios similares

3.2.2. Análisis e interpretación de resultados de las tareas de agrupación

- **Análisis de resultados**

La evaluación cuantitativa a los resultados de la aplicación de técnicas de agrupación que es K-means utilizando todos los datos para el entrenamiento respectivo, mostró la división de los clusters por lo que se analiza respecto a esto

a) Clientes en específico

Para la aplicación de este algoritmo se utilizó dos campos que son el nombre de cliente y productos, realizando así 69964 instancias con 3 atributos. A continuación, en la Tabla 22. Muestra los principales hallazgos

CLIENTE	PRODUCTO
JESUS P	Azúcar Tababuela 50 Kg
AIDA F	Aceite Alesol 12 Fd 900ml
BETTY T	Energizante V220 12u/600ml
MARIA C	Aceite Pal/Oro F 15u 100_280ml
JUSTO F	Sal Crisal Grande X25
OSCAR G	Cara Leche Miel Univ 450g

Tabla 22. Hallazgos de clientes específicos

b) Clientes por características

Como se puede observar en la tabla, el trabajo realizado para agrupar los productos con los clientes categorizados y sus diversas características, en donde se existe 26137 y 3 atributos. En la Tabla 23. Se observa los resultados obtenidos de este algoritmo

- **Interpretación de resultados**

Tipo de vivienda	**Edad**	**Sugerencia de Producto**
Arrendada	52 o más años	Harina Panadera 50 Kg
Propia no hipotecada	63 o más años	Pasta Colgate T/A 60ml
Propia no hipotecada	37 años	Azúcar Lb
Propia no hipotecada	51 años	Sal Crisal 2ks

Propia no hipotecada	73 o más años	Fósforos El Sol Pqt

Tabla 23. Hallazgos de clientes por características

3.2.3. Análisis e interpretación de resultados de las tareas de asociación

- **Análisis de resultados**

La evaluación cuantitativa a los resultados de la aplicación de técnicas de asociación como son A priori y FP Growth, utilizando todos los datos para el entrenamiento de esta, nos mostró varias métricas para evaluar, como son la confianza, lift y convicción.

Como se puede observar en la Tabla, el trabajo realizado principalmente es para asociar los productos en base a las facturas realizadas, con 12477 y 35 atributos, al realizar la aplicación de los algoritmos sobre estos se muestra que las asociaciones cuentan con valores aceptables los cuales se describirán a continuación:

- El algoritmo A priori asoció 12477 y 35 atributos de manera correcta, logrando 10 reglas
- El algoritmo FP Growth asoció 12477 y 35 atributos de manera correcta, obteniendo 8 reglas

A simple en el momento de asociar los productos los dos algoritmos han logrados reglas con un porcentaje alto de confianza por lo que se utilizará las dos técnicas que servirá para la toma de decisiones

- **Interpretación de resultados**

  A continuación, en la Tabla 24. Se indica las reglas que dieron como resultados en la aplicación de este algoritmo

Reglas
Si adquieren una funda de fideo Cayambe de 400g y una papa frita natural es probable que compren un Shampoo Savital de 25 ml
Si compran Huevos y una papa frita natural pueden comprar un Shampoo Savital de 25 ml

Si compran Huevos y funda de fideo Cayambe de 400g podrían adquirir un Shampoo Savital de 25 ml
Si obtienen Pipa limón pueden adquirir una papa frita natural
Si compran azúcar Tababuela pueden comprar azúcar de 1 libra
Si compran harina de panificación de 1 lbr es posible que adquieran harina flor de 1 lbr
Si compran azúcar de 1lb más sal Crisal hay la posibilidad que adquieran azúcar Tababuela
Si compran azúcar de 1 lbr más atún real de 160gr podrían obtener azúcar Tababuela

Tabla 24. Hallazgos de productos

3.2.3. Análisis al Sistema BI y preguntas del negocio

La construcción de Dashboards y reportes se realizaron con el fin de cumplir los indicadores planteados

c

Fig. 52. Pantalla principal de sistema BI

A continuación, en la Tabla 25., se resume las preguntas planteadas y la ubicación en las pantallas

Pantalla Sistema BI	Preguntas del negocio
1. Ventas	Monto total de ventas de cada categoría en el año 2021 Monto total de ventas de cada categoría en el año 2022 Cantidad de facturas de más $500 que se emitieron en 2021 Cantidad de facturas de más $500 que se emitieron en 2022 El producto más vendido de cada categoría Periodo de tiempo con mayor número de ventas en el año 2021 Periodo de tiempo con mayor número de ventas en el año 2022
2. Clientes	Clientes con mayor número de productos comprados Ventas por cliente en un periodo de tiempo

Tabla 25. Pantallas de sistema BI y Preguntas del negocio

Se puede observar las Fig. 52. Y Fig. 53. que es el Sistema BI con las preguntas y pantallas mencionadas anteriormente

Fig. 53. Pantalla de ventas

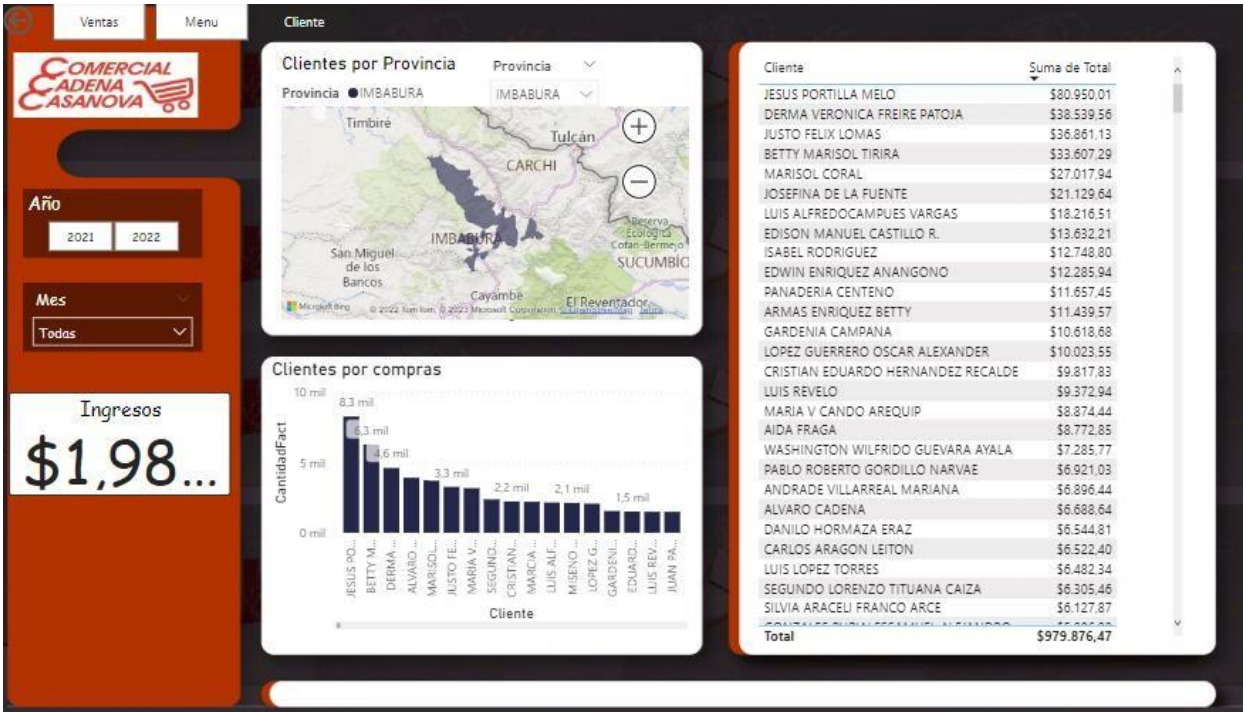


Fig. 54. Pantalla de Clientes

3.3. Discusión de resultados con trabajos relacionados

Con la revisión de trabajos previos enfocados en Inteligencia de negocios y minería de datos se observa que la mayoría de los proyectos no toman como base la aplicación de alguna norma para medir la calidad de los datos con lo que se está trabajando, aun así han obtenido excelentes resultados logrando el objetivo que es la mejora de la toma de decisiones para el avance de negocio.

En otra instancia la metodología más usada en trabajos anteriores que se pudo observar es de KDD, sin embargo a diferencia del presente proyecto que se realizó en base a CRISP-DM tomando como base la aplicación y conclusión de (Cortina, 2015) en donde afirma que es una metodología que funciona y que además es sencilla de usar, ya que solamente hay que seguir una serie de fases que están claramente delimitadas y está pensada para que cualquier persona con conocimientos de bases de datos y estadística pueda utilizarla.

Otro punto clave en este proyecto ha sido la aplicación de Inteligencia de negocios que tiene como objetivo primordial mejorar las ventas y conocer el movimiento de la empresa, el cual se asimila al trabajo de (Nazate, 2022) en donde se aplica esta rama y muestra la gran funcionalidad de la herramienta Power BI para la visualización óptima, operaciones DAX (Data Analysis Expressions) para realizar filtros o consultas en conjunto con componentes visuales que son de agrado para el usuario.

En lo que respecta a la aplicación de las técnicas de Minería de datos se basó en los resultados de (Nazate, 2022) en donde indica la eficacia de los algoritmos de asociación A priori permite visualizar de manera efectiva la probabilidad de que un producto sea adquirido conjuntamente con otros productos coincidiendo a la misma vez con (Paspuel, 2022) que muestra la aplicación de algoritmo de agrupamiento que se asemeja con el presente proyecto ya que el objetivo principal es conocer las tendencias de compra de los clientes

En base a (Vila, 2019) se muestra la importancia de los algoritmos de clasificación en donde se aplica para ciertos clientes con el fin de clasificarlos según características similares y comprender de mejor manera los datos con los que se está trabajando

Como parte de las limitaciones en el transcurso del desarrollo del proyecto

Inicialmente se recibió la base de datos en formato mdb en el cual en su primer análisis se observó campos incoherentes e incompletos, como también tablas y relaciones mal estructuradas, por otra parte, el campo categoría de los productos era algo esencial para el desarrollo del proyecto, pero no estaba vinculada con el inventario, por lo que se optó por etiquetar manualmente la categoría de cada producto

Los datos del año 2021 están incompletos ya que la base ha sido parcialmente borrada por lo que se trabajó con un poco cantidad de datos de este año.

De los clientes no existía mucha información para realizar la segmentación de estos

El equipo en el cual se desarrolló el proyecto es de poca capacidad, por lo que no permitió que los algoritmos trabajan con la totalidad de los datos

CONCLUSIONES

- El Data Warehouse se ha construido utilizando la metodología CRISP-DM, la cual ha sido fundamental debido a que cada uno de sus pasos ha sido minuciosamente diseñado y resulta fácil de ejecutar. Además, esta metodología proporciona una descripción detallada de las técnicas que se deben utilizar en cada fase del proceso

- La adopción de la Norma ISO/IEC 25015 ha sido fundamental para iniciar el proceso de comprensión de los datos, ya que su aplicación permite la evaluación de la calidad de estos. Al asegurarnos de que los datos cumplen con los estándares de calidad establecidos, podemos estar seguros de que los resultados obtenidos se alinean con los objetivos previamente determinados

- Los proyectos de Inteligencia de Negocios y la minería de datos son herramientas poderosas para lograr el éxito empresarial a través de la toma de decisiones precisas basadas en datos. Al analizar los datos, se pueden identificar patrones y tendencias que ayuden a optimizar el funcionamiento de la organización y a mejorar la eficiencia en las operaciones, así como crear estrategias de marketing efectivas para retener a los clientes existentes y aumentar las ventas.

.

RECOMENDACIONES

- La metodología debe elegirse en base a los objetivos planteados, es recomendable que se realice un análisis previo y revisión de trabajos anteriores para saber cuál es la ideal para aplicarla, como también tomar en cuenta los datos disponibles y su estructura, para lograr excelentes resultados

- La recopilación completa de datos es necesaria y mucho más con los datos históricos para comparar los diferentes comportamientos, se recomienda trabajar con mínimo dos años de recepción de datos y sobre todo tratar de que estén completos

- Sería muy importante desarrollar los algoritmos de agrupamiento y asociación con más cantidad de datos y realizar la comparación entre estos para así mejorar la exactitud y precisión de los diferentes modelos

- Se recomienda no enfocarse en una sola herramienta para la realización del Data Warehouse como también para la aplicación de algoritmos, ya que se podría experimentar en varias y de la misma manera escoger la que mejor resultados se ha obtenido

- Para un desarrollo a futuro más fructífero de aplicación de Data Mining se recomienda que el negocio automatice la base de datos, retirando campos que no son necesarios y mejorando la relaciones entre tablas

- Se recomienda la aplicación de normas de calidad de datos antes y después de la limpieza de estos, para verificar la autenticidad y normalización de los datos

REFERENCIAS

Alonso, J. F., & Geovany, M. Q. (2019). Minería de datos Streams Aplicada a Parámetros abióticos: caso práctico: Invernadero de rosas Espeiasa I. Obtenido de http://repositorio.espe.edu.ec/bitstream/21000/20790/1/T-ESPE-039650.pdf

Arancibia, J. A. (s.f.). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.

Barrios. (26 de Julio de 2019). HEALTH Big Data. Obtenido de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Barrueta, A., & Castillo, E. (2018). Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos. Obtenido de https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/626023/Barrueta%20MR.pdf?sequence=1&isAllowed=y

Beltrán, B. (s.f.). Minería De Datos. Obtenido de http://bbeltran.cs.buap.mx/NotasMD.pdf

Berríos, L. H. (Diciembre de 17 de 2015). Aplicación de un sistema de alerta temprana basada en la minería de datos para identificar patrones delictivos en la ciudad de chiclayo. Obtenido de https://tesis.usat.edu.pe/bitstream/20.500.12423/543/1/TL_Jimenez_Berrios_LeslyHaymet.pdf

Borbor, A. L., & Leal, G. K. (2018). Análisis comparativo de herramientas de inteligencia de negocios para la creación de tableros de control utilizando bases de datos de diferentes fuentes. Obtenido de http://repositorio.ug.edu.ec/bitstream/redug/32313/1/B-CISC-PTG-1516%20Rodr%c3%adguez%20Borbor%20Andrea%20Lisbeth%20.%20Romo%20Leal%20Ginger%20Katherine.pdf

Bravo, J. D., Rincón, C. M., & Marín, D. L. (2019). Inteligencia de negocios: Evolución del concepto, importancia y beneficios para las pequeñas y medianas empresas. Obtenido de https://repositorio.uniagustiniana.edu.co/bitstream/handle/123456789/925/PaezBravo-JuanDavid-2019.pdf?sequence=1&isAllowed=y

Buitrón, S. A. (2019). Detección de patrones de deserción estudiantil utilizando técnicas descriptivas de agrupamiento, asociación y atípicos en minería de datos para la gestión académica en la universidad técnica del norte. Ibarra.

Calabrese, J., Esponda, S., Pasini, A., Boracchia, M., & Pesado, P. (14 de Octubre de 2019). Guía para evaluar calidad de datos basada en ISO/IEC 25012. Obtenido de XXV Congreso Argentino de Ciencias de la Computación: http://sedici.unlp.edu.ar/bitstream/handle/10915/91086/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Chung, N. (2017). Análisis exploratorio del ewom mediante herramientas de data mining. Revista de Investigación en Modelos Financieros, 81-95.

Cortina, V. G. (Octubre de 2015). Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario. Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Diaz, R. (s.f.). The machine learning. Obtenido de https://www.themachinelearners.com/metricas-de-clasificacion/#:~:text=TP%20(True%20Positive)%20%E2%80%93%20Son,y%20que%20realmente%20son%20negativos.

Diego, R. F. (2019). Minerìa de datos y toma de decisiones en el supermercado “MEGA BODEGA 9:9. Obtenido de https://dspace.uniandes.edu.ec/bitstream/123456789/10717/1/ACTFMFG022-2019.pdf

Fassler, M. U. (2017). Minería de datos para la toma de decisiones en la unidad de nivelación y admisión. Chimborazo: Cumbres.

Gutiérrez Pacherres, J. J. (2017). Propuesta de una metodología de extracción de conocimientos a partir de datos de las prestaciones del seguro integral de salud en la región piura en el año 2016. Piura.

Hernández, H. M., Mass, R. C., & Zúñiga-Pérez, L. M. (2016). Inteligencia de los negocios. Revista Clío América , 194-211.

Hernandez, R. (2008). Descubrimiento de Conjuntos Frecuentesde´Items en Datos Est´aticos y Din´amicos. Obtenido de http://docplayer.es/139175729-Descubrimiento-de-conjuntos-frecuentes-de-items-en-datos-estaticos-y-dinamicos.html

Inocente, M. E. (2017). Implementación de business intelligence para mejorar la eficiencia de la toma de decisiones en la gestión de proyectos. Lima.

Jaramillo, A., & Paz-Arias, H. (2015). Aplicación de Técnicas de Minería de Datospara Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. Revista Tecnológica ESPOL , 64-90.

Jesús, E.-Z. J. (07 de Noviembre de 2019). Implementation of the CRISP-DM methodology for geographical segmentation using a public database. Ingeniería Investigación y Tecnología, 2-17.

JUKA, S. D. (Junio de 2019). “análisis comparativo de herramientas open source para data mining sobre datos públicos del Ministerio De Educación De La República Del Ecuador. Obtenido de http://repositorio.puce.edu.ec/bitstream/handle/22000/17060/Tesis%20Sergio%20P%c3%a1ez%20Doc.pdf?sequence=1&isAllowed=y

La inteligencia de negocios: una estrategia para la gestión de las empresas productivas . (2017). Revista Ciencia UNEMI, 40-48.

Larranaga, P., Inza, I., & Moujahid, A. (s.f.). Clustering. Obtenido de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t14clustering.pdf

Logreira, C. (2011). Minería de datos y su incidencia en la toma de decisiones empresariales en el contexto de crm. Ingeniería Solidaria, 68-71.

López, J. M. (2006). Aplicaciones prácticas utilizando microsoft excel y weka.

Mamani, Y. (s.f.). Business Intelligence: herramientas para la toma de decisiones en procesos de negocio. Obtenido de https://www.researchgate.net/profile/Yonatan-Mamani-Coaquira/publication/323993348_Business_Intelligence_herramientas_para_la_toma_de_decisiones_en_procesos_de_negocio/links/5ab6bc4ba6fdcc46d3b6b9ee/Business-Intelligence-herramientas-para-la-toma-de-decis

Mancero, T. B. (Abril de 2020). Detección de patrones de contrabando para la gestión de aprehensiones y retenciones, utilizando técnicas predictivas de clasificación y regresión de minería de datos. Obtenido de http://repositorio.utn.edu.ec/bitstream/123456789/10860/2/04%20ISC%20572%20TRABAJO%20GRADO.pdf

Maribel, L. S. (2016). Modelo de gestión de talento humano para la empresa "Comercial Cadena Casanova" De La Ciudad De Ibarra, Provincia De Imbabura. Obtenido de https://dspace.uniandes.edu.ec/bitstream/123456789/5837/1/PIUIADM008-2017.pdf

Marisol, L. C. (2015). Inteligencia de negocios para la toma de decisiones del departamento de cartera de la Cooperativa Finander. Ibarra.

Márquez, M. A. (2017). Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una Distribuidora De Suplementos Nutricionales. Obtenido de https://repositorio.usil.edu.pe/server/api/core/bitstreams/494c61fa-4b5c-4c00-9965-551a56369c57/content

Martínez, C. G. (2020). PUBs por RStudio. Obtenido de Reglas De Asociación: https://rpubs.com/Cristina_Gil/Reglas_Asociacion

MEJÍA, C. E. (2015). "Implementación De Un Sistema De Inteligencia De Negocios En El Departamento De Ventas De Una Empresa Nacional del. Quito.

Melillanca, E. (2018). Evaluación de modelos de clasificación: Matriz de Confusión y Curva ROC. Obtenido de http://www.ericmelillanca.cl/content/evaluaci-n-modelos-clasificaci-n-matriz-confusi-n-y-curva-roc

Merino, E. M., & Merino, M. J. (2017). Análisis de los Modelos de Inteligencia de Negocios basados en Big Data en las Pymes del Ecuador. Revista Ciencia & Tecnología, 46-57.

Miriam, A. P. (2019). Análisis de Impacto en el Desempeño de la Toma de Decisiones en. Guayaquil.

Molina Rea, K. G. (15 de Septiembre de 2020). Implementación de un modelo analítico para la predicción de la venta del portafolio de. Obtenido de http://repositorio.espe.edu.ec/xmlui/bitstream/handle/21000/22561/T-ESPE-043875.pdf?sequence=1&isAllowed=y

Monjas, Y. B. (s.f.). MINERÍA DE DATOS. Obtenido de https://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

Morales, M. R., & Cardoso, S. L. (2017). Inteligencia de negocios basada en Bases de Datos InMemory. Revista Publicando, 201-217.

Nazate, W. (2022). Implementación de inteligencia de negocios para apoyar en la toma de decisiones informadas de la empresa "SARI POPULAR. Obtenido de http://repositorio.utn.edu.ec/bitstream/123456789/13025/2/04%20ISC%20652%20TRABAJO%20DE%20GRADO.pdf

ONU. (25 de Septiembre de 2015). Obtenido de https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Orellana, M., & Cedillo, P. (2020). Detección de valores atípicos con técnicas de minería de datos y métodos estadísticos. Enfoque UTE, 56-67.

Orozco. (2017). Método de reglas de asociación para el análisis de afinidad entre objetos de tipo texto. Obtenido de https://repositorio.cuc.edu.co/bitstream/handle/11323/165/72208501.pdf?sequence=1&isAllowed=y

Panana, C. E., & Paredes, J. P. (2021). Inteligencia de Negocios y la Toma de Decisiones en la Hiperbodega Precio Uno, Huacho 2021. Obtenido de http://200.48.129.167/bitstream/handle/UNJFSC/5500/GUERRERO%20y%20MORALES.pdf?sequence=1&isAllowed=y

Paspuel, B. G. (2022). Análisis de datos aplicando las técnicas de data mining (reglas de asociación y clustering) para fortalecer el comercio electrónico descubriendo hábitos de compra de productos y accesorios de bicicletas en la ciudad de Tulcán. Obtenido de http://repositorio.utn.edu.ec/bitstream/123456789/12618/2/04%20ISC%20633%20TRABAJO%20DE%20GRADO.pdf

Peña, J. A., Ávila, A. E., Lugo, A. J., & Montelongo, D. L. (2018). El uso de herramientas tecnológicas de minería de datos en el análisis de datos climatológicos. Revista Iberoamericana de las Ciencias Computacionales e Informática.

Rea, D. C. (2021). Detección de patrones de contrabando para la gestión de información de aprehensiones y retenciones utilizando técnicas descriptivas de agrupamiento, asociación y atípicos en minería de datos. Ibarra.

Rea, K. G. (15 de Septiembre de 2020). Implementación de un modelo analítico para la predicción de la venta del portafolio de productos OTC de un Laboratorio Farmacéutico. Obtenido de Molina Rea, Karina Gabriela: http://repositorio.espe.edu.ec/bitstream/21000/22561/1/T-ESPE-043875.pdf

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias . Revista Iberoamericana de Inteligencia Artificial, 11-18.

Rivera, I. W. (2006). Minería de datos: herramienta de apoyo en la selección de equipos de proyectos informáticos. Obtenido de file:///C:/Users/Leslie/Downloads/Dialnet-MineriaDeDatos-4786672.pdf

Ronald, C. P. (2019). Aplicación de inteligencia de negocios para la toma de decisiones en el área comercial de la empresa computer. Huancavelica.

Rueda, J. F. (s.f.). Health Data Miner. Obtenido de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Tello, E. A., & Velasco, J. M. (2014). Inteligencia de negocios: estrategia para el desarrollo de competitividad en empresas de base tecnológica. Contaduría y Administración , 127-158.

Ulloa, P. A., Castillo, D. V., Mena, V. M., & Jácome, D. J. (2020). Inteligencia de negocios en la gestión administrativa de una empresa distribuidora del sector eléctrico. Cuadernos de desarrollo aplicados a las TIC, 43-67.

Urgiles, C. M., & Amoroso, M. S. (s.f.). Revisión de algoritmos para la detección de valores atípicos. Obtenido de https://killkana.ucacue.edu.ec/index.php/killkana_tecnico/article/view/287/353

Vargas, N. (23 de Marzo de 2020). DATTA. Obtenido de https://datta.com.ec/articulo/big-data-y-analitica-una-realidad-ecuatoriana

Vila, D. (Abril de 2019). Detección de patrones de deserción estudiantil utilizando técnicas predictivas de clasificación y regresión de minería de datos, para la gestión académica de la Universidad Técnica Del Norte. Obtenido de http://repositorio.utn.edu.ec/bitstream/123456789/9095/1/04%20ISC%20515%20TRABAJO%20DE%20GRADO.pdf

ANEXOS

ANEXO A: Código de aplicación de método codo y métricas en el algoritmo K-Means

https://colab.research.google.com/drive/1SKTqd2KXYDeD8g1dso2E42JxltZF_LYy?usp=sharing

https://colab.research.google.com/drive/1SrsJtfEqtRPpwOI2bObonosWEuzYZhy?usp=sharing

Ibarra, 04 de mayo de 2023

CERTIFICADO DE CULMINACIÓN DE PROYECTO

Mediante el presente certifico que la Srta. **LESLIE MISHELL ARMAS UVIDIA** con cédula de ciudadanía **1004198782**, estudiante de la Universidad Técnica del Norte, culminó de manera satisfactoria el estudio del proyecto **"Implementación de Técnicas de Minería de Datos y visualizaciones utilizando inteligencia de negocios para la toma de decisiones en el Comercial Cadena Casanova"**. Asimismo, deseo expresar mi más sincera gratitud tanto a la institución, al director de tesis y al tesista por el excelente trabajo que han llevado a cabo en el desarrollo de este proyecto.

Por consiguiente, tengo a bien informar que se han realizado satisfactoriamente la revisión de cumplimiento de los requerimientos funcionales, por lo que se recibe el proyecto con plena conformidad.

La Srta. LESLIE MISHELL ARMAS UVIDIA, puede hacer uso de este documento para los fines pertinentes.

Atentamente,

Firmado electrónicamente por:
VIVIANA ANDREA
CASANOVA HEREDIA

Dra. Viviana Casanova

PROPIETARIA DEL

COMERCIAL CADENA

CASANOVA

RUC: 0401317847001
TELF: 062 640 - 043