UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TITULO DEL TRABAJO DE TITULACIÓN

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DEL CRECIMIENTO DEL PORTAFOLIO DE CLIENTES DE LA EMPRESA “CUSTOM PLACE”

AUTOR

SUÁREZ ALVARADO JOSÉ MAXIMILIANO

PROYECTO DE UNIDAD INTEGRACIÓN CURRICULAR

Previo a la obtención del grado académico en
INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

TUTOR

Ing. Orozco Iguasnia Walter

Santa Elena, Ecuador

Año 2023

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

TRIBUNAL DE SUSTENTACIÓN

Ing. José Sánchez A., Mgtr
DIRECTOR DE LA CARRERA

Ing. Walter Orozco L., Mgtr
TUTOR

Ing. Eudice Haz L., Msi
DOCENTE ESPECIALISTA

Ing. Marjorie Coronel S. Mgti.
DOCENTE GUÍA UIC

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por José Maximiliano Suárez Alvarado, como requerimiento para la obtención del título de Ingeniero en Tecnologías de la Información.

La Libertad, a los 18 días del mes de febrero del año 2023

TUTOR

Ing. Walter Orozco I.

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **José Maximiliano Suárez Alvarado**

DECLARO QUE:

El trabajo de Titulación, Aplicación De Técnicas De Minería De Datos Para La Predicción Del Crecimiento Del Portafolio De Clientes De La Empresa "Custom Place" previo a la obtención del título en Ingeniero en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

La Libertad, a los 18 días del mes de febrero del año 2023

EL AUTOR

José M Suárez

José Maximiliano Suárez Alvarado

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN DE ANTIPLAGIO

Certifico que después de revisar el documento final del trabajo de titulación denominado Aplicación De Técnicas De Minería De Datos Para La Predicción Del Crecimiento Del Portafolio De Clientes De La Empresa "Custom Place", presentado por el estudiante, José Maximiliano Suárez Alvarado fue enviado al Sistema Antiplagio, presentando un porcentaje de similitud correspondiente al 3%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.

CERTIFICADO DE ANÁLISIS
magister

SUAREZ_JOSE_Trab_Titulacion

3% Similitudes

6% Texto entre comillas
< 1% similitudes entre comillas
< 1% Idioma no reconocido

Nombre del documento: SUAREZ_JOSE_Trab_Titulacion.docx
ID del documento: d988cd6e42dd3b44fb65b1bbdfabda811e5ccfab
Tamaño del documento original: 2,01 Mo

Depositante: WALTER ARMANDO OROZCO IGUASNIA
Fecha de depósito: 23/2/2023
Tipo de carga: interface
fecha de fin de análisis: 23/2/2023

Número de palabras: 11.703
Número de caracteres: 74.265

Ubicación de las similitudes en el documento:

TUTOR

Ing. Walter Orozco I.

UNIVERSIDAD ESTATAL PENÍNSULA DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **José Maximiliano Suárez Alvarado**

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales de artículo profesional de alto nivel con fines de difusión pública, además apruebo la reproducción de este artículo académico dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor

Santa Elena, a los 18 días del mes de febrero del año 2023

EL AUTOR

José M Suárez

José Maximiliano Suárez Alvarado

AGRADECIMIENTO

A mis padres que siempre estuvieron apoyándome a lo largo de mi carrera, por haberme inculcado buenos valores y estando allí en cada momento preguntando si ya entregué todo, si no tengo nada pendiente, si voy bien en las materias

A mi familia en general por estar pendientes de mi progreso en esta etapa.

A todos los docentes que me fueron guiando a lo largo de mi carrera, a mi tutor y mi docente de integración curricular por guiarme en mí proyecto

José Maximiliano, Suárez Alvarado

DEDICATORIA

Este trabajo va dedicado a mis padres, mis hermanos y mi familia en general, por estar apoyándome en cada momento durante mi trayecto educativo y que nunca me dé por vencido que luche por lo que deseo y por lo que aspiro.

José Maximiliano, Suárez Alvarado

ÍNDICE DE LOS CONTENIDOS

TRIBUNAL DE SUSTENTACIÓN I

CERTIFICACIÓN II

DECLARACIÓN DE RESPONSABILIDAD III

CERTIFICACIÓN DE ANTIPLAGIO IV

AUTORIZACIÓN V

AGRADECIMIENTO VI

DEDICATORIA VII

ÍNDICE DE LOS CONTENIDOS VIII

ÍNDICE DE LAS TABLAS X

ÍNDICE DE LOS GRÁFICOS XI

RESUMEN XII

ABSTRACT XIII

INTRODUCCIÓN 1

1. FUNDAMENTACIÓN 3

1.1. ANTECEDENTES 3

1.2. DESCRIPCIÓN DE PROYECTO 5

1.3. OBJETIVOS 7

1.3.1. OBJETIVO GENERAL 7

1.3.2. OBJETIVO ESPECIFICO 7

1.4. JUSTIFICACIÓN 8

1.5. ALCANCE 9

1.6. METODOLOGÍA DE PROYECTO 10

1.6.1. METODOLOGÍA DE LA INVESTIGACIÓN 10

1.6.2. VARIABLE DEL PROYECTO 10

1.6.3. TÉCNICA DE RECOLECCIÓN DE INFORMACIÓN 10

1.7. GRUPO POBLACIONAL INVOLUCRADO. 11

1.8. METODOLOGÍA DE DESARROLLO 12

2. PROPUESTA 14

2.1. MARCO CONTEXTUAL 14

2.1.1. EMPRESA CUSTOM PLACE 14

2.1.2. MISIÓN DE LA EMPRESA CUSTOM PLACE 14

2.1.3. MINERÍA DE DATOS COMO HERRAMIENTA PRINCIPAL COMO SOPORTE AL MARKETING. 14

2.1.4. BASE LEGAL 15

2.2. MARCO CONCEPTUAL 16

2.2.1. POWER BI DESKTOP 16
2.2.2. RSTUDIO 16
2.2.3. RAPIDMINER STUDIO 16
2.2.4. EXCEL 16
2.2.5. VISUAL STUDIO 2019 16
2.2.6. MICROSOFT SQL SERVER MANAGEMENT 17
2.2.7. REDES NEURONALES 17
2.2.8. ARBOLES DE DECISIÓN 17
2.2.9. MÁQUINA DE VECTORES DE SOPORTE 18
2.2.10. MACHINE LEARNING 19
2.2.11. MÉTRICAS DE RENDIMIENTO 19
2.2.12. DATA WAREHOUSE 19
2.2.13. BASE DE DATOS 20
2.3. MARCO TEÓRICO 20
2.3.1. DATA MINING Y EL DESCUBRIMIENTO DEL CONOCIMIENTO 21
2.3.2. DATA MINING: CONCEPT AND TECHNIQUES 21
2.3.3. PATTERN RECOGNITION AND MACHINE LEARNING 22
2.4. DESARROLLO DE LA PROPUESTA 23
2.4.1. FASE UNO: OBTENCIÓN DE LOS DATOS. 23
2.4.2. FASE DOS: PREPARACIÓN DE LOS DATOS 26
2.4.2.1. CREACIÓN DEL DATA WAREHOUSE 26
2.4.3. FASE TRES: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS. 30
2.4.3.1. EXPORTACIÓN DE LOS DATOS 30
2.4.3.2. IMPLEMENTACIÓN DE LA TÉCNICA ÁRBOL DE DECISIONES. 32
2.4.3.3. IMPLEMENTACIÓN DE LA TÉCNICA REDES NEURONALES 34
2.4.3.4. APLICACIÓN DE LA TÉCNICA DE REGRESIÓN LINEAL MÚLTIPLE 35
2.4.3.5. APLICACIÓN DE LA TÉCNICA DE MÁQUINA DE SOPORTE PARA VECTORES 37
2.4.4. FASE 4: EVALUCIÓN DE RESULTADOS 37
2.4.4.1. EVALUACIÓN DE RESULTADOS DE LOS MODELOS. 37
2.4.4.2. GRÁFICOS ESTADÍSTICOS DE LOS RESULTADOS 39
2.5. RESULTADOS OBTENIDOS 43
2.5.1. RESULTADOS DE LOS MODELOS 43
CONCLUSIONES 45
RECOMENDACIONES 47
REFERENCIAS 48

ÍNDICE DE LAS TABLAS

TABLA I CANTIDAD DE CLIENTES 11
TABLA II BENEFICIARIOS DIRECTOS E INDIRECTOS 11
TABLA III METODOLOGÍA DE DESARROLLO 12
TABLA IV CONTENIDO DE LA TABLA "PROVINCIA" 24
TABLA V CONTENIDO DE LA TABLA "CIUDAD" 24
TABLA VI CONTENIDO DE LA TABLA "GÉNERO" 24
TABLA VII CONTENIDO DE LA TABLA "CLIENTE" 25
TABLA VIII CONTENIDO DE LA TABLA "PRODUCTO" 25
TABLA IX CONTENIDO DE LA TABLA "VENTAS" 26
TABLA X LIBRERÍAS PARA LA TÉCNICA ÁRBOL DE DECISIONES 32
TABLA XI PORCENTAJE DE LOS NODOS 33
TABLA XII LIBRERÍAS UTILIZADAS PARA CREAR RED NEURONAL 34
TABLA XIII LIBRERÍAS PARA REGRESIÓN LINEAL MÚLTIPLE 36
TABLA XIV LIBRERÍAS PARA MÁQUINA DE SOPORTE PARA VECTORES. 37
TABLA XV MÉTRICAS DEL ÁRBOL DE DECISIONES 38
TABLA XVI MÉTRICAS DE LA RED NEURONAL 38
TABLA XVII MÉTRICAS DE REGRESIÓN LINEAL MÚLTIPLE 38
TABLA XVIII MÉTRICAS DE MÁQUINA DE SOPORTE DE VECTORES 39
TABLA XIX RESULTADOS DE LAS MÉTRICAS DE RENDIMIENTO 43
TABLA XX TIEMPO DE ESPERA DE LOS CLIENTES 44

ÍNDICE DE LOS GRÁFICOS

FIG. 1. METODOLOGÍA CRISP-DM [15] 12
FIG. 2. ESQUEMA DE LA METODOLOGÍA DEL PROYECTO 13
FIG. 3. CLASIFICACIÓN DE MINERÍA DE DATOS. [50] 15
FIG. 4 EJEMPLO DE UNA RED NEURONAL [51] 17
FIG. 5 EJEMPLO DE ÁRBOL DE DECISIONES [32]. 18
FIG. 6 EJEMPLO DE SVM (SUPPORT VECTOR MACHINE) [33]. 18
FIG. 7 MACHINE LEARNING Y SUS ALCANCES [52]. 19
FIG. 8 EJEMPLO DE UN DATA WAREHOUSE [53]. 20
FIG. 9 MINERÍA DE DATOS COMO UN PASO PARA EL PROCESO DE KDD [46]. 22
FIG. 10 BASE DE DATOS ORIGINAL 23
FIG. 11 NUEVA BASE DE DATOS 23
FIG. 12. DATAMART VENTAS 27
FIG. 13. PROCESO GENERAL ETL 28
FIG. 14. PROCESO ETL DE CLIENTES 29
FIG. 15. MAPPING ORIGEN-DESTINO CLIENTE 29
FIG. 16. ORIGEN DEL TIPO DE MODELO A EXPORTAR 30
FIG. 17. SELECCIÓN DE TABLAS A EXPORTAR 31
FIG. 18. DESTINO DEL TIPO DE MODELO A EXPORTAR 31
FIG. 19 ESQUEMA DEL ÁRBOL DE DECISIONES 33
FIG. 20 ESQUEMA DE LA RED NEURONAL 35
FIG. 21 ESQUEMA DE REGRESIÓN LINEAL MÚLTIPLE 36
FIG. 22 GRÁFICO DE DISPERSIÓN DEL MODELO DE REDES NEURONALES DESDE RAPIDMINER 40
FIG. 23 GRÁFICO DE DISPERSIÓN DEL MODELO DE REDES NEURONALES DESDE VISUAL STUDIO 40
FIG. 24 GRÁFICO DE DISPERSIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DESDE RAPIDMINER 41
FIG. 25 GRÁFICO DE DISPERSIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DESDE VISUAL STUDIO 41
FIG. 26 GRÁFICO DE DISPERSIÓN DEL MODELO DE ÁRBOLES DE DECISIONES DESDE VISUAL STUDIO 42
FIG. 27 GRÁFICO DE DISPERSIÓN DEL MODELO DE ÁRBOLES DE DECISIONES DESDE RAPIDMINER 42
FIG. 28 CORRELACIÓN DE VARIABLES 43

RESUMEN

Custom Place, una empresa de manufacturación sublimada enfrenta problemas de organización de datos que afectan su atención al cliente. Para solucionar esto, se propone utilizar técnicas de minería de datos como Redes Neuronales, Arboles de Decisiones, Maquina de Soporte de Vectores y Regresión Lineal Múltiple siguiendo la metodología CRISP-DM, con la ayuda de bibliotecas de R para crear modelos de árboles de decisión, redes neuronales y máquinas de soporte de vectores. Se utilizarán gráficos de dispersión y cuatro métricas de rendimiento MSE, RMSE, MAE y R^2 para evaluar los modelos. Con estas técnicas, la empresa Custom Place podrá mejorar su gestión de datos y ofrecer así una mejor atención al cliente, lo que se traducirá en una mayor eficiencia y competitividad en el mercado. La minería de datos es una herramienta poderosa que puede ayudar a las empresas a aprovechar al máximo sus datos y mejorar su rendimiento en general.

Palabras claves: CRISP-DM, Minería de datos, Inteligencia de negocio.

ABSTRACT

Custom Place, a sublimation manufacturing company, faces data organization problems that affect its customer service. To solve this, techniques such as Neural Networks, Decision Trees, Support Vector Machines, and Multiple Linear Regression are proposed using the CRISP-DM methodology, with the help of R libraries to create decision tree models, neural networks, and support vector machines. Scatterplots and four performance metrics MSE, RMSE, MAE, and R^2 will be used to evaluate the models. With these techniques, Custom Place can improve its data management and offer better customer service, resulting in increased efficiency and competitiveness in the market. Data mining is a powerful tool that can help companies make the most of their data and improve their overall performance.

Keywords: CRISP-DM, Data mining, Business intelligence.

INTRODUCCIÓN

La minería de datos tiene un papel muy importante en el análisis de datos, ya que con esto resulta de manera más sencilla encontrar patrones, relaciones, y conocimientos útiles que ayuden a la correcta toma de decisiones, es decir que sean más acertadas a la realidad de los mismos datos y del negocio, empresa u organización en el que se desee emplear. Muchas empresas en la actualidad utilizan procesos de minería de datos para descifrar valores en relación con lo necesitado, dando soporte a que estas empresas o compañías tengan un mejor desarrollo competitivo con otras franquicias que básicamente se dediquen a lo mismo.

Incluso la minería de datos actualmente cubre muchas otras áreas interdisciplinarias como por ejemplo en salud, el marketing, la educación, entre otras más. Por su eficacia y eficiencia al desarrollar estos modelos, a partir de la utilización de procesos estadísticos, de bases de datos, de inteligencia de negocios y de inteligencia artificial

Este proyecto se centra en conocer por medio de modelos de minería de datos el crecimiento del portafolio de los clientes de la empresa Custom Place empresa que se dedica a la manufacturación de sublimados.

Contará con un total de dos capítulos en donde el capítulo I se encuentra los antecedentes, técnicas de recolección de información, los objetivos de este proyecto, la justificación, la metodología que se llevará a cabo y la descripción en donde se narra lo que abarcará el proyecto.

El capítulo II contara con los marcos: teóricos, conceptuales y contextuales de esta manera se obtendrán los conceptos de varias herramientas utilizadas, bases legales de acuerdo a las leyes de Ecuador, bases teóricas con respecto a otros trabajos científicos e información relevante que realice el análisis teórico.

En este mismo capítulo se desarrollara la propuesta que está conformada de cuatro fases en donde la primera etapa se trata de recopilar información de la empresa es decir sus bases de datos, también se utilizó técnicas como la observación y la entrevista, en la segunda etapa se centra en hacer la correcta limpieza, transformación y extracción de los datos obtenidos de las bases de datos, en la tercera etapa se trata de realizar los modelos de minería de datos se desarrollaron 4 modelos: modelo de árbol de decisiones, redes neuronales, regresión lineal múltiples y máquina de soporte de vectores, en la cuarta etapa se trata de evaluar los modelos realizados en la etapa tres, se utilizaron las siguientes métricas de rendimiento para

corroborar los resultados obtenidos: error cuadrático medio (MSE), error absoluto medio (MAE), raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). También habrá el apartado de los resultados obtenidos donde se concluye cuál de los modelos dio mejores resultados.

CAPITULO I

1. FUNDAMENTACIÓN

1.1. ANTECEDENTES

Desde hace muchos años varias compañías de comercio de compra y venta de productos se han visto en la necesidad de administrar sus actividades comerciales con la finalidad de obtener un control de ingresos en tiempo real [1] dando como resultado una acumulación de una gran cantidad de datos que se vuelven difíciles de manejarlos para los gerentes o usuarios que están al frente de un negocio a la hora de conocer el progreso del negocio [2]. Sin embargo, muchas de estas instituciones han conseguido recolectar grandes volúmenes de datos sobre las actividades que éstas realizan día a día, pero muy pocos datos han logrado ser tratados mientras que la mayoría de estos solo se acumulan hasta perderse [3].

Bajo este contexto, la empresa "Custom Place" ubicada en la ciudadela puertas del sol en la ciudad de Salinas, es una empresa que se dedica a la manufacturación de sublimados en vasos, tazas, camisetas y demás objetos, comenzó sus funciones un 17 de mayo de 2019 y tiene como misión ofrecer a la distinguida clientela un producto de calidad y ajustado a las necesidades personalizadas resaltando la calidez y calidad. Hoy en día la empresa sigue ofreciendo aquellos servicios, uno de los encargados de la empresa manifestó en una entrevista que, (ver anexo 1) al tener un gran número de clientes, los procesos para cubrir las necesidades de cada cliente son muy lentos e ineficientes debido a que desconoce los servicios detallados que son más adquiridos por sus clientes, provocando así una fuerte disminución y pérdida económica.

Otro de los problemas que dio a conocer es que a pesar de haber perdido muchos clientes en ese periodo, este posee grandes cantidades de registros en su base de datos, que hoy en día se le dificulta clasificarlos, puesto que recalcó que el no conocer el tipo de servicio que los clientes adquieren, el gerente desconoce la manera más eficaz y eficiente de comprar materiales ya que muchas veces éste ha invertido en materiales que nunca se logró utilizar causando pérdidas económicas y de recursos pero esto no duro mucho ya que unos meses después la empresa comenzó a tener pérdida de clientes, ya que al tener competencia con respecto a otras empresas que brindaban los mismos servicios éstas si poseían una mejor organización y estrategias para ganar clientes de manera más fácil.

Adicionalmente menciona que no conoce las necesidades o intereses de las personas, y que conlleva a tener una ineficiente comprensión entre lo que ofrece la empresa y lo que el cliente

desea, haciendo que éste proporcione una información errónea de los servicios que ofrece tomando decisiones erróneas y poco acertadas dando como resultado la pérdida de clientes, disminución económica y una mala posición en el mercado por lo que la empresa se dio cuenta que cada proceso de venta y de un nuevo cliente que pide el servicio que brinda la empresa es llenado en una base de datos.

Para tener una mejor guía y comprensión del tema se ha seleccionado tres tesis relacionadas al tema de minería de datos, los cuales se detallarán a continuación con el lugar de donde se desarrollaron, a que rama está diseñada, que metodología utiliza, las diferencias y similitudes que tengan en común este trabajo con los citados.

A nivel de Latinoamérica sobresale el trabajo de investigación denominado "Análisis de la minería de datos aplicada en empresas del sector retail", desarrollado por Xiomara Silva en el año 2020, publicada en el repositorio de la Universidad Católica San Pablo [4], este trabajo es casi similar al que se presentara con la diferencia de que se usara una metodología diferente, y se realizaran otras técnicas de minería de datos, con esta tesis se tendrá una guía sobre cómo aplicar las técnicas de minería de datos por lo que será bastante útil para el desarrollo de este trabajo.

Otro proyecto realizado a nivel local, denominado "Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.'", publicada en el repositorio de la Universidad Península de Santa Elena [5], y presentada por Alex Villao, este trabajo obtuvo grandes resultados y se realizaron comparaciones y reportes de análisis de cada una de las técnicas que se realizaron para predecir qué estudiantes aprobaran y cuales obtuvieron un mejor rendimiento académico, también mediante la metodología que utilizo se permitió conocer cómo funciona el negocio y como se puede implementar las técnicas de minería de datos a este negocio, lo distinto que tendrá con respecto al proyecto en desarrollo es que la metodología a aplicar será basada en Crisp-DM, por lo que los datos que se obtendrán serán analizados e interpretados de una manera distinta.

En otro trabajo citado a nivel nacional es, Minería de datos para la gestión de compras de medicamentos en el Hospital Básico El Puyo, esta tesis fue extraída del repositorio institucional UNIANDES [6] , lo que trata de implementar este trabajo es predecir la cantidad de materiales hospitalarios que se adquieran para su uso eficaz para así tomar decisiones en los distintos campos dentro del hospital, posee la misma metodología que usare

en mi proyecto, por eso se tomó como referencia este trabajo para tener una guía de cómo implementar dicha metodología, el uso de esta técnica así sea de otra rama la idea y la interpretación es la parecida a la de este proyecto.

Este proyecto ayudara a tomar decisiones de manera efectiva y eficaz, que a su vez ayuda al incremento económico de la empresa conociendo los datos y realizando un correcto análisis de los mismos, esto conlleva a tener una mejor agilidad y automatización a la hora de dar la correcta información a los clientes específicos, para cada interés, para cada generación y para cada tipo de cliente, al utilizar minería de datos, la empresa entraría a un campo competitivo con empresas que ofrecen similares productos, al aplicar estas técnicas de minería de datos conllevara a la empresa a tomar decisiones relevantes y especificas a cada cliente.

1.2. DESCRIPCIÓN DE PROYECTO

Con el objetivo de solucionar la problemática, se propone aplicar la técnica de árbol de decisiones y la técnica de redes neuronales que ayuden a la toma de decisiones para una correcta gestión de los clientes que posee la institución y para que ésta tenga un crecimiento económico y competitivo con demás empresas similares, que a partir del análisis de datos contenidos dentro del gestor de base de datos que posee la empresa ayudara en la predicción de posibles clientes potenciales.

El presente trabajo constara de cuatro fases que están sujeto a la metodología CRISP-DM [7] las que se detallaran a continuación:

Fase de Obtención de los datos.

Esta fase constara en seleccionar y extraer los datos más relevantes y útiles suministrados por la empresa, los cuales serán presentados en hojas de cálculos de Excel, con registros de datos de los clientes y de las ventas realizadas desde mayo 2019 hasta mayo del 2022, los mismos que servirán para la siguiente fase.

Fase de Preparación de los datos.

Esta fase comprenderá de la creación de la data Warehouse, en la cual a partir de los datos obtenidos en la primera fase se hará una limpieza y depuración de los datos, para que éstos no contengan ningún dato duplicado o erróneo que provoque algún problema en el análisis final, puesto que, una vez obtenido la tabla de la data Warehouse con los datos correctos

permitirá la implementación de técnicas de minería de datos que se realizaran en la fase siguiente.

Fase de Aplicación de minería de datos.

En esta fase, después de haber desarrollado el data Warehouse, se procederá a aplicar técnicas de minería de datos, orientado a un análisis predictivo, en la cual se usarán cuatro técnicas que permitan obtener este tipo de análisis, los cuales se mencionan a continuación: **redes neuronales, árboles de decisión, Máquina de Vectores de Soporte (SVM) y Regresión Lineal Múltiple.**

Fase de Evaluación de resultados.

Luego de la aplicación de las técnicas utilizadas anteriormente se procederá con el análisis e interpretación de resultados generados a partir de los datos los cuales serán plasmados en gráficos estadísticos como: gráficos de líneas para predecir la línea o la tendencia del crecimiento de clientes para la empresa, gráficos de dispersión que servirá para conocer un interés común entre los clientes y así tener una mejor relación con los clientes y gráficos circulares para hacer comparativas entre los resultados de cada mes y un análisis general y por último se realizaran gráficos de mapas de calor.

Todas estas pautas permitirán conocer los futuros clientes potenciales, tendencias de lo que ofrece la empresa, así como también conocer si el negocio lograra la meta planteada como objetivo, para que ésta tenga un éxito al prestar sus servicios.

Toda esta información obtenida será presentada ante los encargados de la empresa, para que en función de estos análisis estos puedan tomar decisiones con respecto al negocio y así realizar los posibles cambios para optimizar la relación con los clientes fijos y consumidores futuros.

Adicional a eso al finalizar el trabajo se realizarán las debidas recomendaciones conseguidas a partir de los resultados.

Las herramientas que se usarán para el desarrollo de este trabajo se describirán a continuación:

- **Power BI Desktop:** programa para graficar resultados con grandes cantidades de datos.

- **RStudio:** Lenguaje para la realización de los modelos de minería de datos.

- **RapidMiner Studio:** programa para realización de modelos automáticos de minería de datos y graficar los resultados obtenidos.

- **Excel:** programa con el tipo de datos que se encontraban los datos iniciales de la empresa.

- **Visual Studio 2019:** Realización de los procesos ETL e implementación de técnicas de minería de datos.

- **Microsoft SQL Server Management:** Gestor de base de datos relacional.

- **Weka:** programa para comparar resultados obtenidos mediante R.

El presente trabajo tiene como línea de investigación principal Tecnologías y Sistemas de la información (TSI) [8] y como sub-línea de investigación se detalla en Inteligencia Computacional debido a que los métodos de predicción se realizan mediante algoritmos.

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Aplicar técnicas de minería de datos para predecir el crecimiento del portafolio de clientes de la empresa Custom Place mediante análisis exploratorios, predictivos y gráficos estadísticos.

1.3.2. OBJETIVO ESPECIFICO

- Crear data Warehouse con datos específicos obtenidos de las tablas principales dadas por la empresa.
- Aplicar técnicas de “árbol de decisión”, “redes neuronales”, “máquina de vectores de soporte” y “regresión lineal múltiple” para predecir clientes.

- Explicar los resultados obtenidos mediante gráficos estadísticos y representación visual.

1.4. JUSTIFICACIÓN

El uso de las tecnologías y los sistemas de información es fundamental para las medianas y pequeñas empresas que están en proceso de desarrollo [9], ya que con estas herramientas se pueden optimizar y mejorar los procesos que realizan las empresas u organizaciones, las cuales ayudan a obtener ventajas competitivas que les permitan obtener una mejor posición en el mercado y así atraer clientes y tener un mayor nivel de productividad y desempeño [10].

Actualmente las empresas deben tomar diversas decisiones lo cual es un proceso muy complejo ya que existen varias posibles soluciones de cómo afrontar o resolver un problema, por lo que dependiendo de cómo se tome una decisión se puede entender como es la lógica del negocio por ello se busca incrementar la eficiencia y la eficacia a la hora de tomar decisiones para así cumplir con los objetivos y/o metas planteadas inicialmente en una empresa [11].

Este proyecto permitirá a la empresa conocer los posibles clientes potenciales a partir del análisis predictivo de la información crucial contenida en la base de datos proporcionada por la empresa mediante la aplicación de técnicas de minería de datos como “redes neuronales” y “arboles de decisión”, el cual ayudará a la empresa a entender de mejor manera más acertada las necesidades y requerimientos de los clientes.

El presente trabajo no solo brindara soporte a la toma de decisiones sino también ayudara a obtener una mejor comunicación entre el cliente y la empresa, ya que con la creación y diseño de un data warehouse aplicando las técnicas ETL para los datos, el desarrollo de este proyecto permitirá analizar resultados muy fáciles de interpretar para que la empresa ofrezca el servicio que el cliente necesite.

Este trabajo está alineado a los objetivos del Plan Nacional Creación de Oportunidades que está vigente desde 2021-2025 [12], el cual se describirá a continuación:

Eje 3: Económico y generación de empleo.

Objetivo 3: Fomentar la productividad y competitividad en los sectores agrícolas, industrial, acuícola y pesquero, bajo el enfoque de la economía circular.

Política 3.1: Mejorar la competitividad y productividad agrícola, acuícola, pesquera e industrial incentivando el acceso a infraestructura adecuada, insumos y uso de tecnologías modernas y limpias.

Eje 3: Económico y generación de empleo.

Objetivo 4: Garantizar la gestión de las finanzas públicas de manera sostenible.

Política 4.3: Incrementar la eficiencia de las empresas públicas con un enfoque de calidad y rentabilidad económica y social.

1.5. ALCANCE

En vista de las necesidades y problemas que posee la empresa, el presente trabajo se enfoca en la aplicación de técnicas de minería de datos que a través del análisis adecuado de los datos que posee la empresa "Custom Place" se obtendrá información relevante y útil que ayudara en la gestión de compras y ventas de productos, relación con clientes, así como también en la toma decisiones acertadas y que para tener una mejor comprensión de los resultados obtenidos se realizaran las siguientes fases:

- **Fase de Obtención de los datos.**
- **Fase de Preparación de los datos.**
- **Fase de Aplicación de las técnicas de minería de datos.**
- **Fase de Evaluación de resultados.**

- En la **fase de obtención de datos** se limita a conseguir únicamente los datos sobre ventas y clientes, los cuales están contenidos dentro de una hoja de cálculo de Excel, Esta base de datos consta dos tablas; la de clientes con los siguientes datos: nombre del cliente u organización, cedula o identificación, la edad, el género, el teléfono, el correo electrónico y la dirección, y por otro lado la tabla de Ventas que muestra las cantidades y los valores de las ventas realizadas.
- En la **fase de la preparación de los datos** se creará el data Warehouse con datos específicos y haciendo depuración de los mismos para evitar datos duplicados, una vez obtenido aquello se utilizarán técnicas de extracción, carga y transformación de datos.

- En la **fase de Aplicación de las técnicas de minería de datos**, una vez que hemos obtenido el data set de la fase anterior se comenzara a aplicar las técnicas de árbol de decisiones y redes neuronales para minería de datos mediante un análisis exploratorito para verificar las variables que servirán para al final obtener un resultado predictivo.

- En la **fase de evaluación de resultados**, con la información obtenida en la fase anterior se crearán gráficos estadísticos con ayuda de la herramienta Power BI, el cual ayudará la creación de estos de manera rápida, entendible y eficaz, para tener una mejor interpretación de los datos, así como también se harán las conclusiones en base a los resultados que se obtuvieron en el análisis.

1.6. METODOLOGÍA DE PROYECTO

1.6.1. METODOLOGÍA DE LA INVESTIGACIÓN

Debido a la poca existencia de información acerca de los procesos y la importancia de los datos que el presente proyecto requiere para un análisis adecuado de minería de datos, el trabajo a realizar contara con un estudio de carácter exploratorio ya que con el fin de examinar un tema desconocido o poco estudiando [13], se indagarán trabajos relacionados a análisis predictivos y aplicación de técnicas de minería de datos para comparar semejanzas y diferencias frente al trabajo propuesto como "Aplicación de técnicas de minería de datos para predecir el portafolio de clientes de la empresa CUSTOM PLACE".

Con el propósito de obtener y recolectar información importante de la empresa se utilizó la metodología de tipo diagnostica [13], para conocer en que se basa el negocio, como es su organización estructural, los tipos de soluciones que aplican ante diversos problemas y la efectividad de la toma de decisiones que utilizan.

1.6.2. VARIABLE DEL PROYECTO

Lo que se plantea solucionar con este trabajo es mejorar el tiempo con el que se toman las decisiones a la hora de ofrecer un buen servicio al cliente en base a los resultados que se obtendrán, por este motivo la variable del proyecto es la que se describió en este párrafo.

1.6.3. TÉCNICA DE RECOLECCIÓN DE INFORMACIÓN

Para la recolección de información y determinación de las bases del negocio, este trabajo contara con la aplicación de dos técnicas de recolección de información: entrevista (ver anexo 1) y observación (ver anexo 2) [14].

En cuanto a la técnica de observación se limitó a analizar los datos obtenidos por parte de la empresa en formato .xls (extensión de hoja de cálculo de Excel), y en formato accedb (extensión base de datos de Access) en la cual estos documentos cuentan con una lista de clientes que la empresa posee con sus datos correspondientes, los mismos que están comprendidos entre los años 2019 a 2022. Acorde a ello se generó una tabla generalizada con la cantidad de clientes por periodo.

Tabla I

Cantidad de clientes

Periodo de Ventas	**Cantidad de Clientes**
2019-2020	7
2020-2021	27
2021-2022	21

A partir de los datos expuestos se describió lo siguiente:

- En base a la entrevista personal (ver anexo 1) realizada a los encargados de la empresa, entre el periodo 2019-2020 hubo poca concurrencia de clientes debido a la afectación de la pandemia ya que ésta influenció mucho en la economía de las personas haciendo que la empresa tenga menos clientela que la esperada.

- En el siguiente periodo aumento considerablemente la cantidad de clientes, pero no fue lo que la empresa deseaba, en consecuencia, a esto, se dedujo que este problema ya no era a causa la pandemia como factor principal, sino más bien problemas de gestión de negocios de la misma empresa.

1.7. GRUPO POBLACIONAL INVOLUCRADO.

Este proyecto beneficiará directamente a los dirigentes de la empresa, mientras que sus clientes se beneficiarán de forma indirecta, ya que las respuestas y los servicios que la empresa ofrece a sus clientes serán optimizadas.

Tabla II

Beneficiarios directos e indirectos

Beneficiarios directos	
Administración	3
Gerencia	2

Beneficiarios indirectos
Clientes

1.8. METODOLOGÍA DE DESARROLLO

La metodología de desarrollo que se implementará será CRISP-DM [7], esta metodología cuenta con 6 fases, pero en este trabajo se han fusionado varias fases que a continuación se describirán:

Tabla III

Metodología de desarrollo

Metodología Crisp-Dm	**Metodología de este trabajo**
Comprensión del negocio	Obtención de los datos
Comprensión de los datos	
Preparación de los datos	Preparación de los datos.
Modelado	Aplicación de las técnicas de minería de datos.
Evaluación	Evaluación de resultados.
Despliegue	

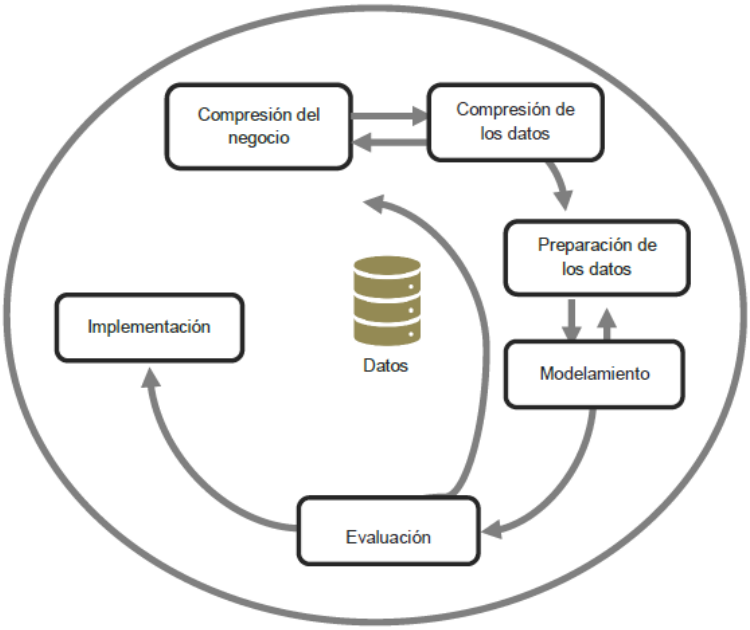


Fig. 1. Metodología CRISP-DM [15]

Fig. 2. Esquema de la metodología del proyecto

Fase de Obtención de los datos: en esta fase se realizará la recolección de los datos, en este caso la empresa facilito una base de datos que contiene las tablas de: clientes y ventas, el cual está dentro de un libro de Excel que cuenta con 2 hojas como corresponde a las tablas y además cuenta con una pequeña base de datos en Microsoft Access.

Fase de Preparación de los datos: en esta fase se creará el data warehouse a partir de los datos extraídos del Datamart para realizar el posterior análisis e implementación de la siguiente fase con datos limpios aplicando las técnicas ETL de base de datos. La creación de este data warehouse será basada en la metodología multidimensional, ya que con estos datos se procederá a aplicar la minería de datos.

Fase de aplicación de minería de datos: se aplicará las siguientes técnicas de predicción "Redes neuronales" y "Árbol de decisiones" utilizando los datos previamente estructurados y corregidos, por lo que se eliminaron los datos redundantes o que estuviesen incorrectos debido al formato de las bases de datos y así se contará con el data warehouse limpio y sin problemas.

Fase de Evaluación de Resultado: luego de los resultados obtenidos se aplicarán técnicas de validación de errores para evitar problemas e inconsistencias en los resultados, estos datos se analizarán y se presentaran en tablas y gráficos estadísticos, el cual se presentará los datos obtenidos a los representantes de la empresa, y así puedan conocer de mejor manera cuáles serán los posibles clientes potenciales.

CAPITULO II

2. PROPUESTA

2.1. MARCO CONTEXTUAL

2.1.1. EMPRESA CUSTOM PLACE

La empresa custom place fue creada el 17 de mayo de 2019 (ver anexo 1), es una empresa dedicada a la manufacturación de sublimados para todo tipo de productos, desde ropa hasta utensilios del hogar y de la cocina. La empresa se encuentra ubicada en la ciudadela puertas del sol del cantón Salinas de la provincia de Santa Elena a dos cuadras de la unidad educativa "Nuestro Mundo".

2.1.2. MISIÓN DE LA EMPRESA CUSTOM PLACE

Ofrecer a nuestra distinguida clientela un producto de calidad y ajustado a las necesidades personalizadas de sus requerimientos, resaltando atención de calidad y calidez (ver anexo 1).

2.1.3. MINERÍA DE DATOS COMO HERRAMIENTA PRINCIPAL COMO SOPORTE AL MARKETING.

La minería de datos es el campo que permite descubrir nueva información útil dentro de una gran cantidad de datos, estas a su vez se han empleado en diversos campos desde la medicina hasta los negocios de servicio, y estos se pueden clasificar en dos grandes grupos que son: por técnicas de verificación y por métodos de descubrimientos [16].Los de las técnicas de verificación se centran o se limitan a demostrar hipótesis proporcionadas por el usuario, mientras que los de métodos de descubrimientos se centran en encontrar patrones interesantes de forma automática, dentro de este grupo se reúnen todas las técnicas de predicción que pueden ser de carácter descriptivo o predictivo [17]

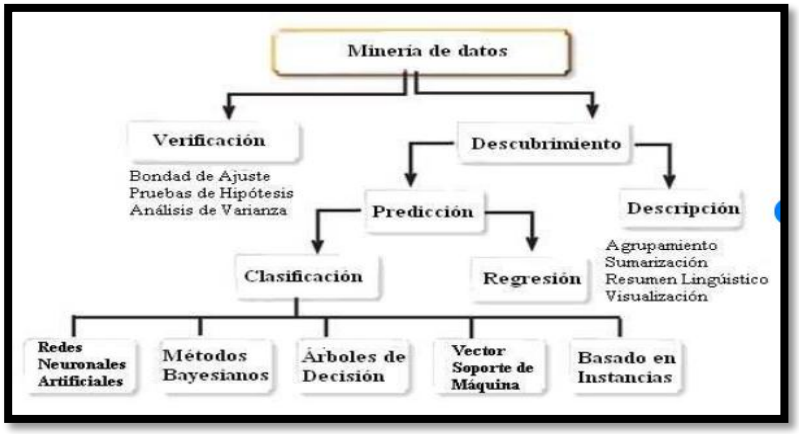


Fig. 3. Clasificación de minería de datos. [50]

.

2.1.4. BASE LEGAL

Las empresas deben cumplir con las regulaciones establecidas en la Constitución de la República del Ecuador, la Ley Orgánica de Protección de Datos Personales, su Reglamento, y la Ley de Protección de Datos Personales en el Sector Público [18]. Estas regulaciones establecen las normas para el tratamiento de datos personales, incluyendo la recolección, almacenamiento, uso, y divulgación de dicha información [19]. Algunas de las regulaciones más importantes que deben cumplir las empresas son las siguientes:

- Obtención de consentimiento previo del titular de los datos antes de recolectar, almacenar, usar o divulgar su información personal.
- Utilizar medidas de seguridad adecuadas para proteger los datos personales de accesos no autorizados, pérdida, destrucción o alteración.
- Proporcionar al titular de los datos acceso a su información personal, así como la posibilidad de rectificar o actualizarla.
- Informar a las autoridades competentes en caso de una violación a la seguridad de los datos personales.
- No utilizar los datos personales para fines diferentes a los informados al titular de los mismos.
- No divulgar los datos personales a terceros sin el consentimiento previo del titular.
- Establecer un sistema de responsabilidad interno y externo para garantizar el cumplimiento de estas regulaciones.

Es importante tener en cuenta que el incumplimiento de estas regulaciones puede resultar en sanciones administrativas y/o penales [20].

2.2. MARCO CONCEPTUAL

2.2.1. POWER BI DESKTOP

Es un programa de Microsoft, de visualización de datos el cual sirve para tomar decisiones de manera rápida y eficaz, ya que permite extraer datos de cualquier tipo de base de datos [21].

2.2.2. RSTUDIO

Es un lenguaje de programación interpretado, interactivo y orientado a objetos, cuenta con una sintaxis muy clara, aparte posee un sin número de librerías, es utilizado como un lenguaje de extensión para aplicaciones que soliciten de una interfaz [22].

2.2.3. RAPIDMINER STUDIO

RapidMiner es una plataforma de minería de datos y análisis predictivo que permite a los usuarios analizar, visualizar y predecir patrones en grandes conjuntos de datos. RapidMiner ofrece una variedad de herramientas para limpiar, transformar y preparar datos, así como para realizar tareas de análisis estadístico y aprendizaje automático. [23].

2.2.4. EXCEL

Es una herramienta muy eficaz para conseguir información con significado a partir de grandes cantidades de datos. También realiza cálculos sencillos y para ejecutar el seguimiento de cualquier tipo de información [24].

2.2.5. VISUAL STUDIO 2019

Visual Studio 2019 es un entorno de desarrollo integrado (IDE, por sus siglas en inglés) desarrollado por Microsoft. Es una plataforma completa para desarrollar aplicaciones para Windows, Linux, Android, iOS, web y nube. Incluye herramientas para desarrollo de código, depuración, pruebas, seguimiento de errores, control de versiones, integración con otros sistemas y mucho más [25]. Con Visual Studio 2019, los desarrolladores pueden utilizar una variedad de lenguajes de programación, incluyendo C#, C++, F#, Visual Basic, Python, JavaScript y TypeScript. También incluye soporte para frameworks y bibliotecas populares, como .NET, ASP.NET, Xamarin, React y Angular. [26].

2.2.6. MICROSOFT SQL SERVER MANAGEMENT

Microsoft SQL Server Management Studio (SSMS) es una herramienta integral para la administración de bases de datos de Microsoft SQL Server. Proporciona una interfaz gráfica fácil de usar para configurar, administrar y trabajar con bases de datos SQL Server. Es compatible con versiones anteriores y actuales de SQL Server, incluyendo SQL Server Express, LocalDB, SQL Azure, entre otros. [27].

2.2.7. REDES NEURONALES

Las redes neuronales son una clase de modelos de aprendizaje automático inspirados en la estructura y función del cerebro humano. Consisten en una gran cantidad de nodos interconectados llamados "neuronas" que trabajan juntos para procesar información y tomar decisiones [28]. En una red neuronal básica, cada neurona recibe información, realiza cálculos matemáticos y envía información a otras neuronas. Estas conexiones entre neuronas tienen pesos asociados que controlan la importancia de la entrada para la salida final [29]. Son modelos computacionales inspirados en las características neurofisiológicas del cerebro humano y están formadas por un gran número de neuronas dispuestas en varias capas e interconectadas entre sí mediante conexiones con pesos [30].

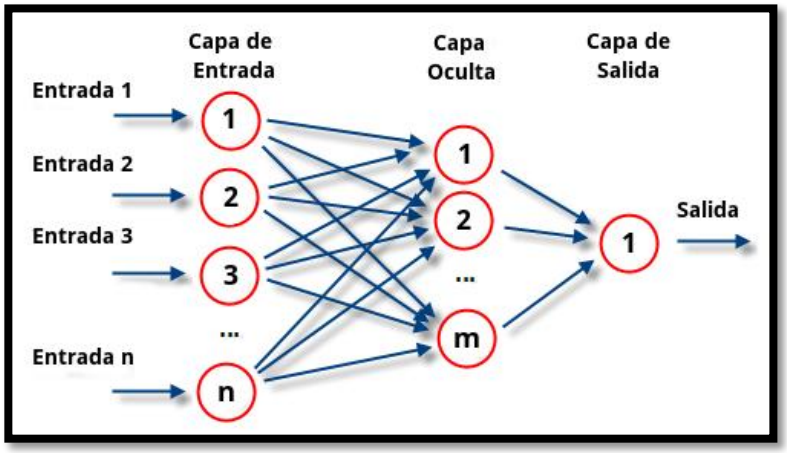


Fig. 4 Ejemplo de una red neuronal [51]

2.2.8. ARBOLES DE DECISIÓN

Son series de decisiones o condiciones organizadas de forma jerárquica, a modo de árbol, son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en calidad de los datos [30]. Los árboles de decisión son una técnica de aprendizaje automático utilizada para resolver problemas de clasificación y regresión. Cada nodo de árbol representa una prueba de atributo, cada rama representa un resultado de prueba y cada hoja representa una clase o valor de salida. La construcción de árboles de decisión se basa en la técnica de

partición recursiva [29], tiene un proceso iterativo con dos fases: la primera fase selecciona los atributos con mayor relevancia para particionar el conjunto de datos y de esta manera particionar el conjunto de datos en subconjuntos más pequeños [31]. El proceso continúa hasta que se cumplen ciertos criterios de parada, como cuando se alcanza un determinado nivel de clasificación o se alcanza un tamaño mínimo de subconjunto [32].

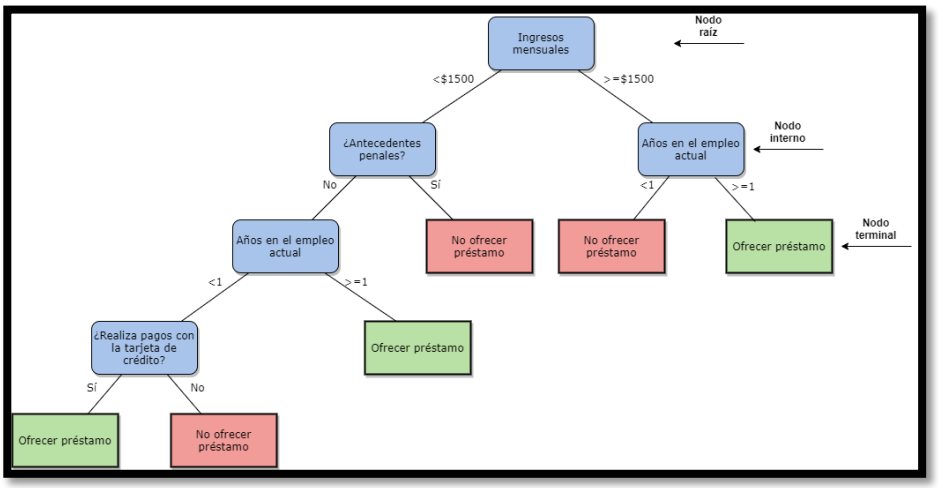


Fig. 5 Ejemplo de árbol de decisiones [32].

2.2.9. MÁQUINA DE VECTORES DE SOPORTE

Es un algoritmo de aprendizaje automático para clasificación y regresión [33]. El objetivo de esta técnica es encontrar un hiperplano que divida los datos en diferentes clases de la mejor manera posible. Un hiperplano es una línea o plano que divide el espacio en dos regiones [34]. En el caso de un problema de clasificación binaria se refiere a dos clases, el objetivo es encontrar el hiperplano que divide los datos en dos grupos, uno para cada grupo [34].

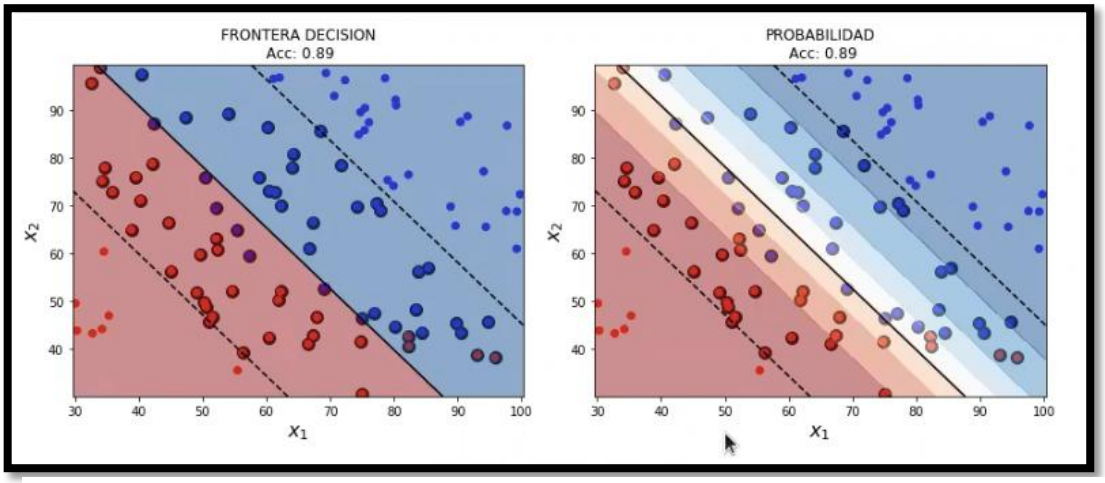


Fig. 6 Ejemplo de SVM (Support Vector Machine) [33].

2.2.10. MACHINE LEARNING

Machine learning o aprendizaje automático es una rama de la inteligencia artificial que se enfoca en desarrollar sistemas y algoritmos que aprenden y mejoran automáticamente a partir de los datos sin ser programados explícitamente [29]. Existen diferentes tipos de aprendizaje automático, como el aprendizaje supervisado en el que tiene un conjunto de datos etiquetados y el objetivo es aprender a hacer predicciones basadas en esos datos, el aprendizaje no supervisado en el que el objetivo es descubrir patrones o estructuras en los datos sin etiquetas y que aprenda por refuerzo, donde el objetivo es tomar acciones en el entorno para maximizar las recompensas [35].

Fig. 7 Machine learning y sus alcances [52].

2.2.11. MÉTRICAS DE RENDIMIENTO

El desarrollo de nuevos algoritmos de machine learning y Deep learning [36] que buscan ir mejorando el rendimiento de los problemas de clasificación multiclase con datos no balanceados, por ello es necesario optar por el algoritmo mejor adaptado mediante una o varias métricas de evaluación de rendimiento en varios algoritmos seleccionados que ayuden con la comprobación de la información obtenida [37].

2.2.12. DATA WAREHOUSE

Un data warehouse o también conocido como almacén de datos es un sistema de almacenamiento y recuperación de información diseñado para apoyar el análisis y la toma de decisiones empresariales [38]. Un data warehouse combina datos de diferentes fuentes, como sistemas transaccionales, hojas de cálculo y archivos de texto, y los organiza de manera coherente para que los usuarios puedan acceder y analizar la información de manera eficiente [39]. Las datas warehouses se utilizan para el análisis de datos a gran escala, como el análisis de ventas, el seguimiento de inventarios y el análisis de tendencias del mercado [40].

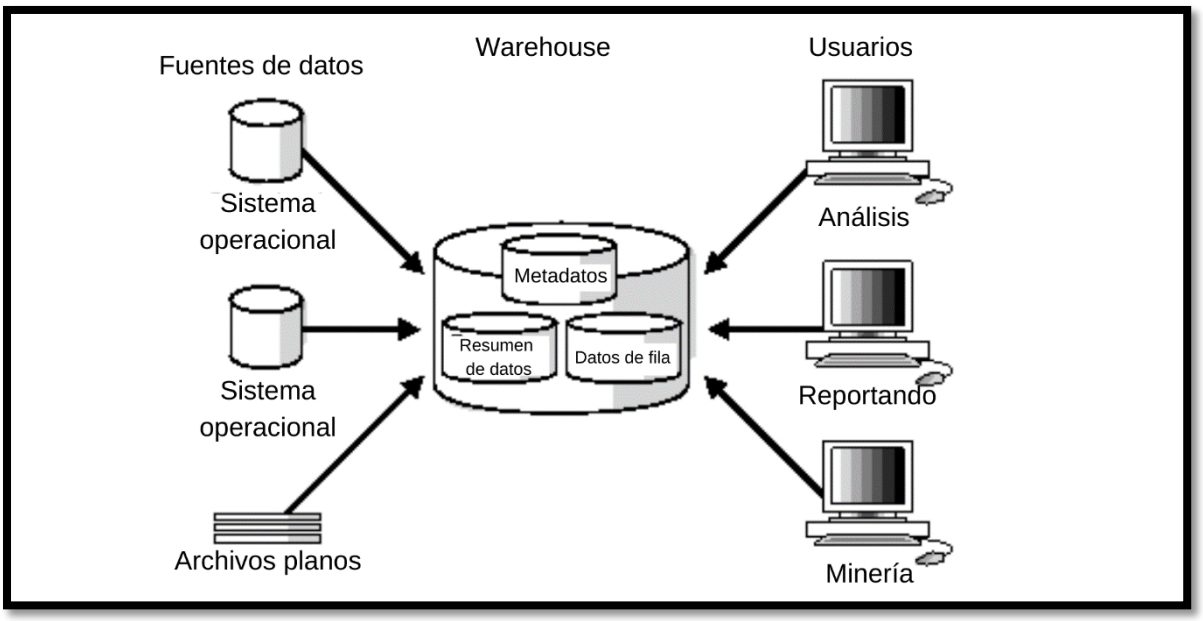


Fig. 8 Ejemplo de un data warehouse [53].

2.2.13. BASE DE DATOS

Una base de datos es un sistema de almacenamiento y recuperación de información que permite a los usuarios acceder y gestionar datos de manera eficiente [41]. Una base de datos está compuesta por un conjunto de datos organizados en tablas, registros y campos, que se relacionan entre sí para permitir la recuperación de información de manera rápida y precisa [42]. Las bases de datos se utilizan en una amplia variedad de aplicaciones, como el almacenamiento de información de clientes, el seguimiento de inventarios, el análisis de datos empresariales, la gestión de sistemas de información y la recopilación de datos en investigaciones científicas [43].

2.3. MARCO TEÓRICO

Para fundamentar este trabajo en el ámbito investigativo y teórico se consultó de varias fuentes de los cuales se detallará los trabajos de mayor impacto y de mejor relevancia para este proyecto.

2.3.1. DATA MINING Y EL DESCUBRIMIENTO DEL CONOCIMIENTO

En los últimos tiempos, ha existido un enorme incremento en nuestro entorno al generar y recolectar datos, debido al procesamiento de datos e información que realiza toda empresa u organización, sin embargo, de todos estos datos existe información de gran importancia que aplicando técnicas básicas simples no sería sencillo acceder [44].

El descubrimiento del conocimiento en base de datos o (KDD),"es la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" [45].

Por una parte, el Data Mining y el KDD contribuyen a la toma de decisiones tácticas y estrategias, aportando de manera automática para la generación de conocimiento y en consecuencia a la toma acertada de decisiones y su aplicación amplia en las diferentes ramas de la investigación

2.3.2. DATA MINING: CONCEPT AND TECHNIQUES

Muchas personas tratan el data mining como un sinónimo de otro termino popular, el descubrimiento del conocimiento de los datos o en sus siglas el KDD, mientras que otras personas ven el data mining como un paso esencial en el proceso del descubrimiento de conocimiento [46].

La técnica de descubrimiento de conocimiento de datos o KDD, posee una secuencia de pasos iterativa que es:

- Limpieza de datos.
- Integración de datos.
- Selección de datos.

- Transformación de los datos.
- Minería de datos.
- Evaluación de patrones.
- Presentación de conocimiento.

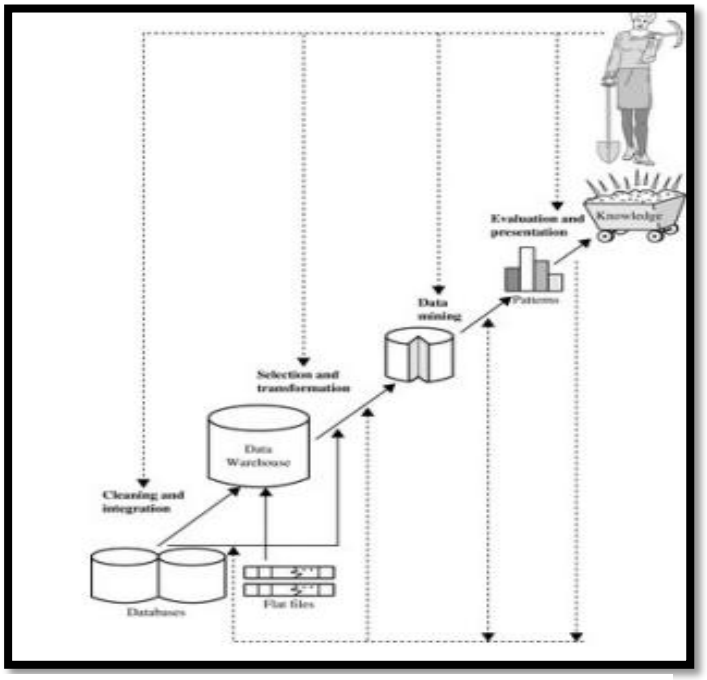


Fig. 9 Minería de datos como un paso para el proceso de KDD [46].

2.3.3. PATTERN RECOGNITION AND MACHINE LEARNING

El reconocimiento de patrones tiene su origen en la ingeniería, mientras que el aprendizaje automático surgió de la informática. Sin embargo, estos dos términos pueden considerarse dos facetas de un mismo campo, y juntas han experimentado un desarrollo sustancial en los últimos tiempos. En concreto, los métodos bayesianos han pasado de ser un nicho especializado a convertirse en la corriente principal, mientras que los modelos gráficos han surgido como una puesta general para describir y aplicar modelos probabilísticos [47].

Al mismo tiempo, la aplicación de la práctica de los métodos bayesianos ha ido mejorando enormemente gracias al desarrollo de una serie de algoritmos de inferencia aproximada, como la propagación de expectativas. Del mismo modo, los nuevos modelos basados en

kernels han tenido un impacto significativo tanto en los algoritmos como en las aplicaciones tanto en los algoritmos como en las aplicaciones [29].

2.4. DESARROLLO DE LA PROPUESTA

2.4.1. FASE UNO: OBTENCIÓN DE LOS DATOS.

La Base de datos original de la empresa Custom Place, cuenta con 3 tablas que son: Clientes, Ventas y Productos. Para la cardinalidad de la base de datos se observa que tiene relación uno a muchos: Clientes y Ventas, Productos y Ventas. La información fue extraída de un documento de Excel el cual consta de 3 hojas respectivamente con datos de los clientes, de los productos y del registro de las ventas.

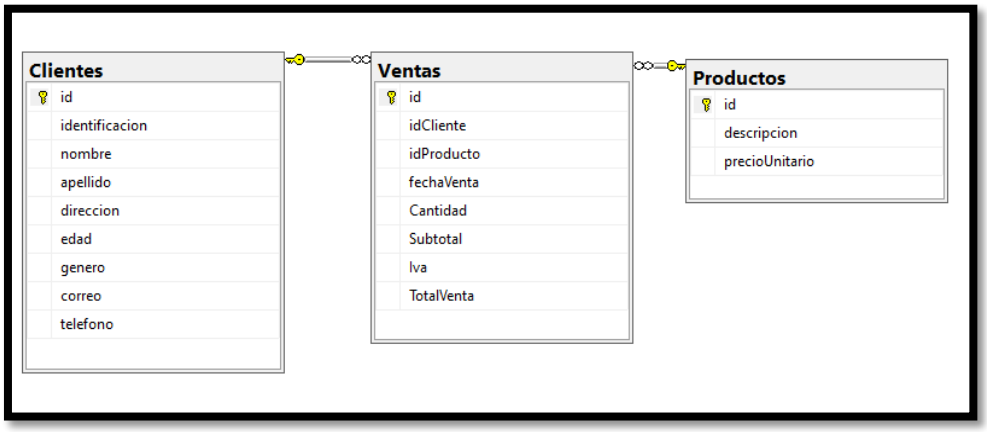


Fig. 10 Base de datos Original

Se realizó una nueva base de datos con la información antes descrita y esta nueva base de datos consta de 7 tablas que son: Provincia, Ciudad, Género, Clientes, Ventas y Productos, con esta nueva base de datos se tendrán datos mucho más organizados y estructurados.

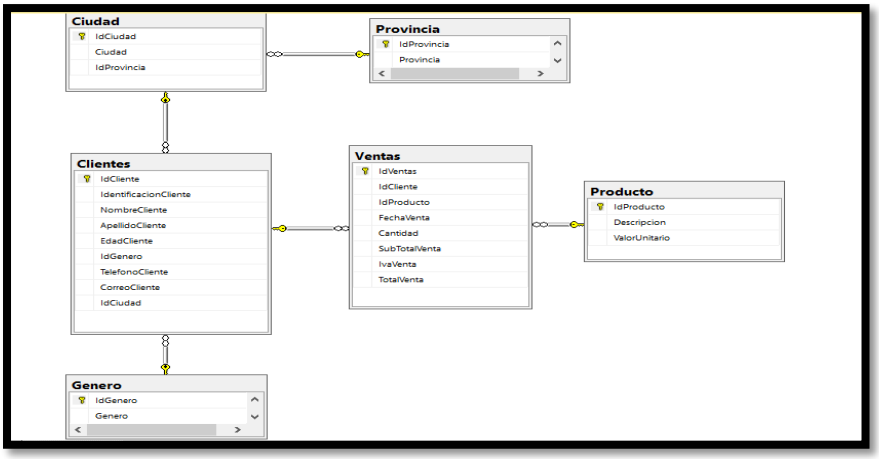


Fig. 11 Nueva base de datos

En cada tabla se detalla lo siguiente:

Tabla IV

Contenido de la tabla "Provincia"

Nombre de Columna	Tipo
idProvincia	int
Provincia	varchar(50)

La tabla provincia cuenta con dos columnas en donde el idProvincia es el indicador de la provincia y la columna provincia contiene el nombre como tal de la provincia.

Tabla V

Contenido de la tabla "Ciudad"

Nombre de Columna	Tipo
idCiudad	int
Ciudad	varchar(50)
idProvincia	int

La tabla ciudad cuenta con tres columnas en donde el idCiudad es el indicador de la ciudad, la columna ciudad contiene el nombre como tal de la ciudad y el idProvincia contiene el indicador de la provincia a la que corresponde la ciudad.

Tabla VI

Contenido de la tabla "Género"

Nombre de Columna	Tipo
idGenero	int
Genero	nchar(10)

La tabla género cuenta con dos columnas en donde el idGenero es el indicador del género y la columna género contiene el nombre como tal del género.

Tabla VII

Contenido de la tabla "Cliente"

Nombre de Columna	Tipo
idCliente	int
identificacionCliente	nchar(10)
NombreCliente	varchar(50)
ApellidoCliente	varchar(50)
EdadCliente	Int
IdGenero	Int
IdCiudad	Int
telefonoCliente	nvarchar(50)
CorreoCliente	varchar(50)

La tabla cliente cuenta con nueve columnas en donde el idCliente es el indicador del cliente, la columna identificaciónCliente contiene la cedula del cliente, el NombreCliente contiene el nombre del cliente, el ApellidoCliente contiene el apellido del cliente, la EdadCliente contiene la edad del cliente, el idGenero es el indicador que contiene el género del cliente, el idCiudad es el indicador que contiene la ciudad donde vive el cliente, el teléfonoCliente corresponde al teléfono del cliente y por último en la columna CorreoCliente corresponde al correo del cliente.

Tabla VIII

Contenido de la tabla "Producto"

Nombre de Columna	Tipo
idProducto	int
Descripción	nvarchar(50)
ValorUnitario	numeric(18,2)

La tabla producto cuenta con tres columnas en donde el idProducto es el indicador del producto, la columna descripción contiene el nombre del producto como tal y el valor unitario contiene el costo del producto.

Tabla IX

Contenido de la tabla "Ventas"

Nombre de Columna	Tipo
idVentas	int
idCliente	int
NombreCliente	Varchar(105)
GeneroCliente	Nchar(10)
EdadCliente	int
idProducto	Int
DescripcionProducto	nvarchar(50)
FechaVenta	Date
cantidad	Int
Subtotal	numeric(18,2)
IvaVenta	Numeric(18,2)
TotalVenta	Numeric(18,2)

La tabla ventas cuenta con doce columnas en donde el idVentas es el indicador de la venta, la columna idCliente contiene al indicador del cliente que corresponde a esa venta, la columna NombreCliente contiene la mezcla del nombre y apellido del cliente, la columna GeneroCliente contiene como tal el género en letras, en la columna EdadCliente consta de la edad como tal del cliente.

La columna idProducto contiene al indicador del producto que corresponde a esa venta, en la columna DescripcionProducto consta con el nombre del producto, la columna FechaVenta que contiene la fecha en donde se realizó la venta, la columna cantidad contiene la cantidad de productos que se vendieron, el campo subtotal contiene el valor a pagar sin IVA, el campo IvaVenta contiene el valor del IVA que se aumentara al subtotal y el campo TotalVenta contiene el total a pagar por la venta realizada.

2.4.2. FASE DOS: PREPARACIÓN DE LOS DATOS

2.4.2.1. CREACIÓN DEL DATA WAREHOUSE

Existen dos enfoques para el data warehouse o almacén de datos, los cuales son: el enfoque según Inmon y el enfoque según Kimball [48], el trabajo presentado contara con el enfoque

según Inmon debido a que gracias a su modelo es mucho más flexible a las organizaciones o empresas que ofrecen productos o servicios.

Este enfoque cuenta con una regresión descendente es decir de arriba abajo, lo que significa que primero se realiza el data warehouse mientras que internamente a este se desarrollaran los datamarts, este proceso ayuda a que los datos ingresen normalmente al data warehouse y dentro de mismo estos se distribuyan a cada tabla como corresponda.

Se procede a crear una nueva base de datos que servirá para el procesamiento de los datos, estos datos transformados serán a su vez depurados ya que datos nulos o inválidos no serán transportados a esta nueva base de datos.

Esta nueva base de datos contará con 4 tablas, 3 tablas de dimensiones que serán las tablas "D_Clientes", "D_Productos" y "D_Fecha" también contará con 1 tabla de hechos que constituirá la tabla de "H_Ventas".

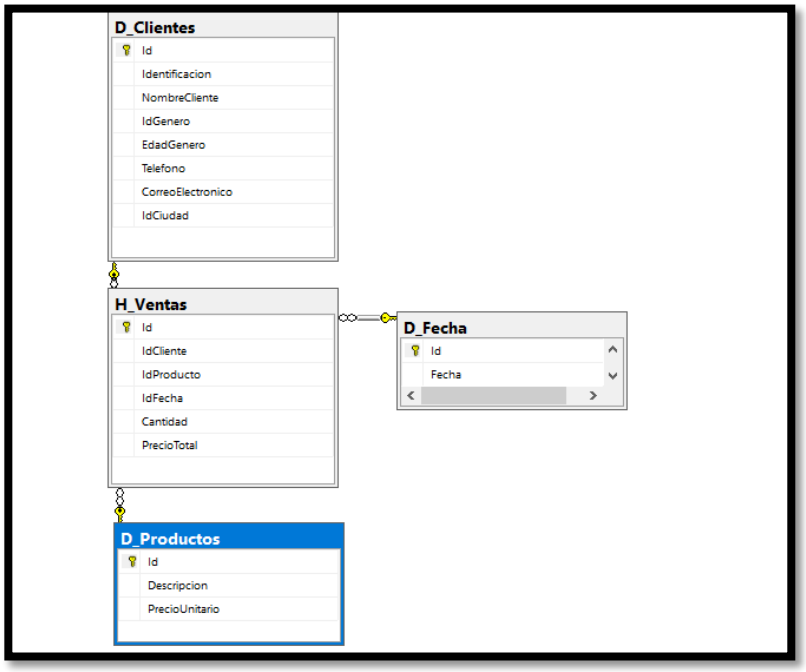


Fig. 12. Datamart ventas

Para realizar el proceso ETL (Extracción, Transformación y Carga) se utilizó la herramienta Microsoft Visual Studio 2019 con la extensión Integration Services, primero se crea un proyecto con el nombre de ETL-DW_Custom_Place.

Se agregarán 4 tareas de flujos de datos, una para cada tabla (D_Clientes, D_Productos, D_Fecha y H_Ventas), también se agregó una tarea de ejecución SQL para realizar el

proceso de depuración para que cuando se requiera agregar nuevos datos de la base de datos original y exportarlos a la data warehouse no se dupliquen los datos antes exportados.

D_Clientes: es la tabla de dimensiones que contiene los datos de los clientes como identificación, nombre, edad, etc.

D_Fecha: es la tabla de dimensiones que contienen los datos de las fechas como su identificador y la fecha en la que el cliente hizo la compra.

Producto: es la tabla de dimensiones que contiene los datos de los productos como precio unitario, nombre del producto y el identificador del producto.

H_Ventas: es la tabla de hechos que contiene los identificadores de las 3 tablas de dimensiones junto con la cantidad, la edad del cliente ya que será de suma importancia y el valor total de la compra.

La tarea de ejecutar SQL contiene un código que permite el limpiado de los datos de las tablas en caso se haya hecho antes un proceso ETL y no se esté conforme con lo transferido esta sentencia permite el borrado de todos los datos copiados y esto se hace consecutivamente se encienda el proceso.

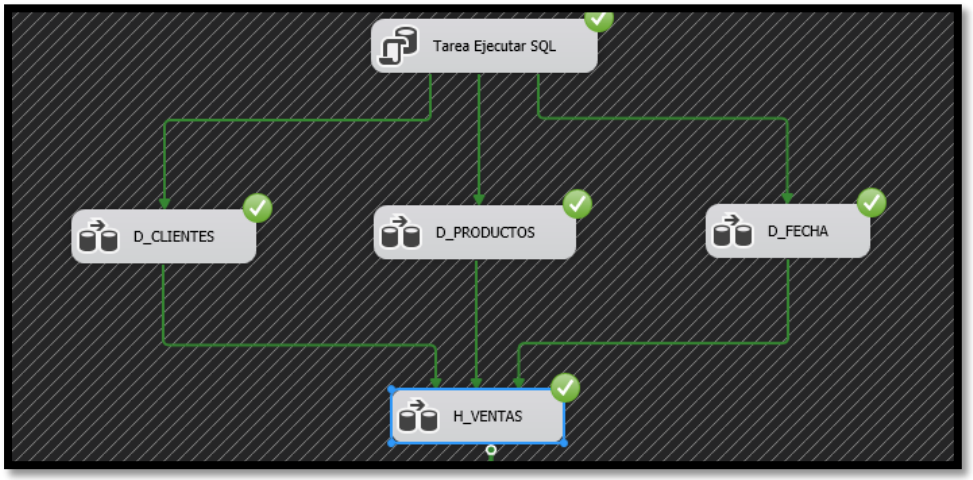


Fig. 13. Proceso general ETL

Para cada proceso se utilizó 2 origen de datos OLE DB en donde un origen será desde una base de datos en Microsoft SQL y el otro origen será de Microsoft Access, debido a que la información será extraída desde una base de datos, agregamos 1 conversor de datos para que los datos que serán extraídos sean compatibles con el tipo de dato que tendrá el dato en el data warehouse, y por último se agregó un destino de dato OLE DB, ya que estos datos se guardaran en otra base de datos. Y así el mismo proceso para las demás tablas.

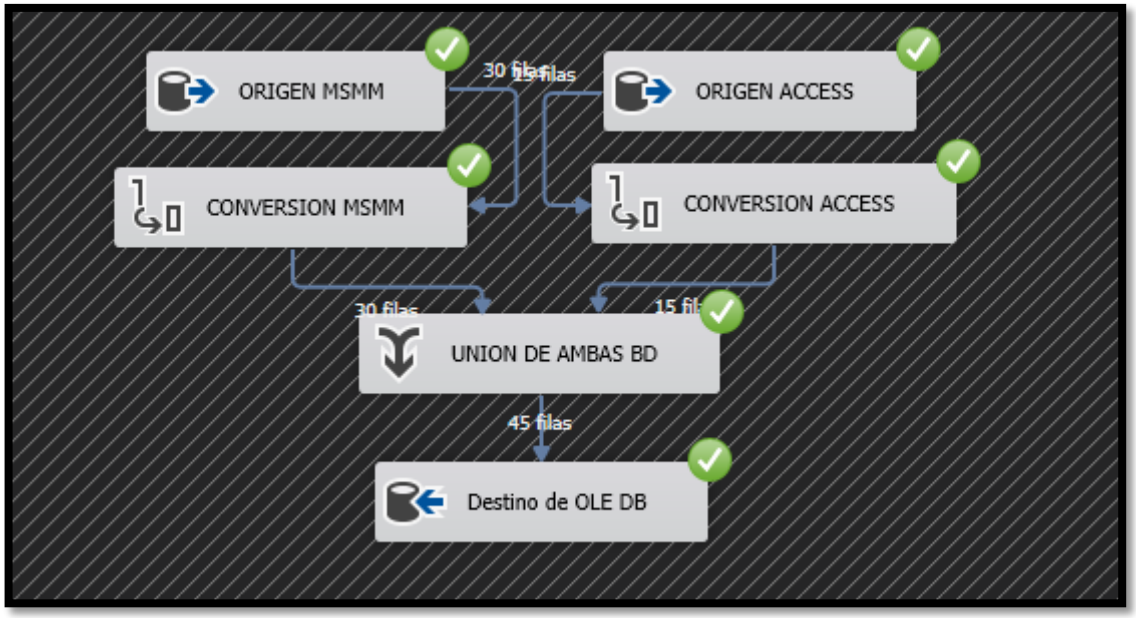


Fig. 14. Proceso ETL de clientes

Para seleccionar los campos que requeriremos dentro del datamart, se realiza una consulta dentro de origen de OLE DB. Primero se selecciona la conexión con el origen, luego se selección SQL Command, para posterior realizar la consulta y obtener los campos que se necesiten. Para registrar los datos y se coordinen con las columnas de la data warehouse, primero se hará la conexión con el destino que en este caso será la base de datos de la data warehouse, luego se procede a ubicar la tabla destino en este caso será la tabla Cliente de la data warehouse, y por último se hará un mapeo con los datos del transformado hecho en el proceso de conversión de datos y las columnas de la tabla cliente, y así de igual manera se realiza el proceso de las otras tablas.

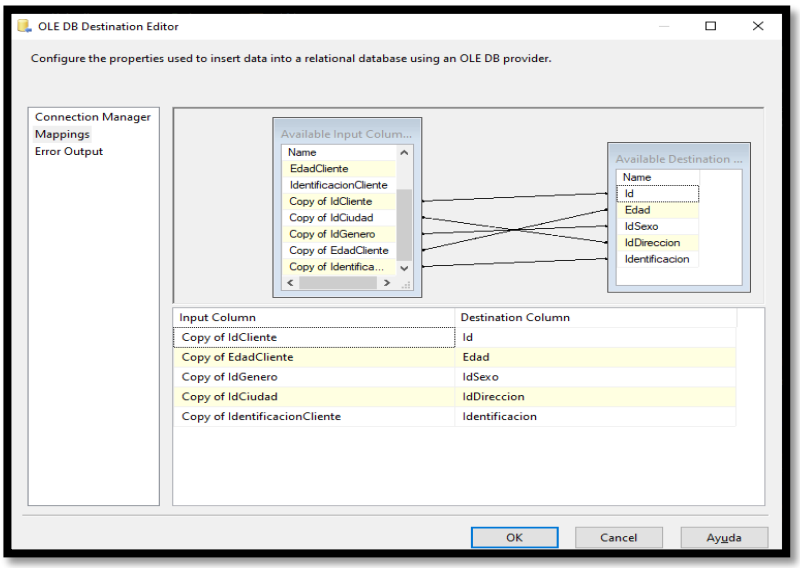


Fig. 15. Mapping origen-destino cliente

2.4.3. FASE TRES: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS.

2.4.3.1. EXPORTACIÓN DE LOS DATOS

Como primer paso para realizar el proceso de esta técnica se requiere que los datos o la información se encuentre en formato .csv para ello lo que se hará es exportar la data warehouse en el formato antes mencionado, siguiendo los siguientes pasos:

- Se selecciona el data warehouse, luego en tareas y exportar datos.
- Aparece una ventana el cual se pondrá el origen, es decir el data warehouse, se selecciona el tipo de origen que en este caso será de tipo OLE DB, se ubica la conexión donde se aloja el data warehouse y por último se elige la base de datos.

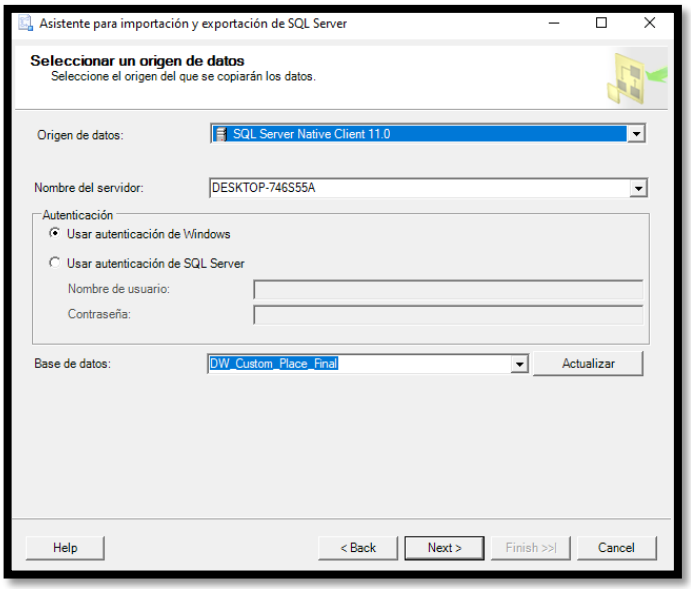


Fig. 16. Origen del tipo de modelo a exportar

- Luego aparecerá otra ventana en el cual ubicaremos el tipo de destino, como se requiere en formato .csv se lo exportara como un archivo de Excel, luego se escoge la ruta en donde se quiere que se guarde el archivo, se debe crear un archivo .xlsx (con extensión de Excel), el cual receptara los datos del archivo origen.

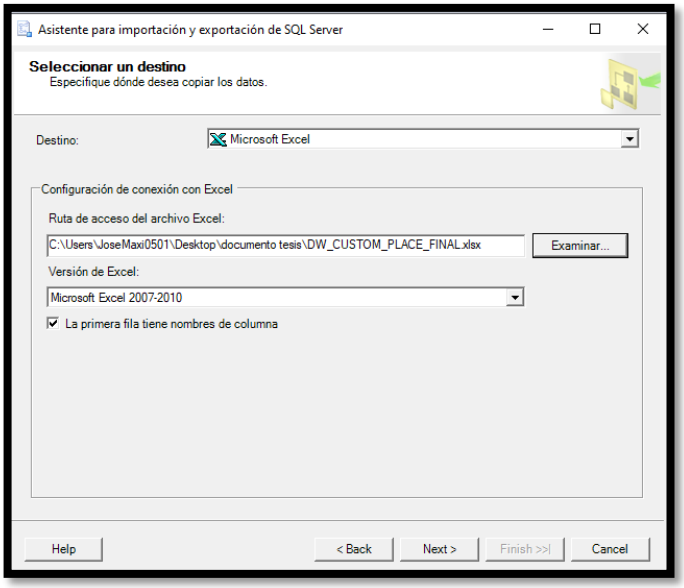


Fig. 18. Destino del tipo de modelo a exportar

- Posterior a lo realizados aparecerá otra ventana, se selecciona la primera opción que sería para copiar los datos que se encuentra en el origen. Luego se seleccionarán las tablas que queremos exportar y listo.

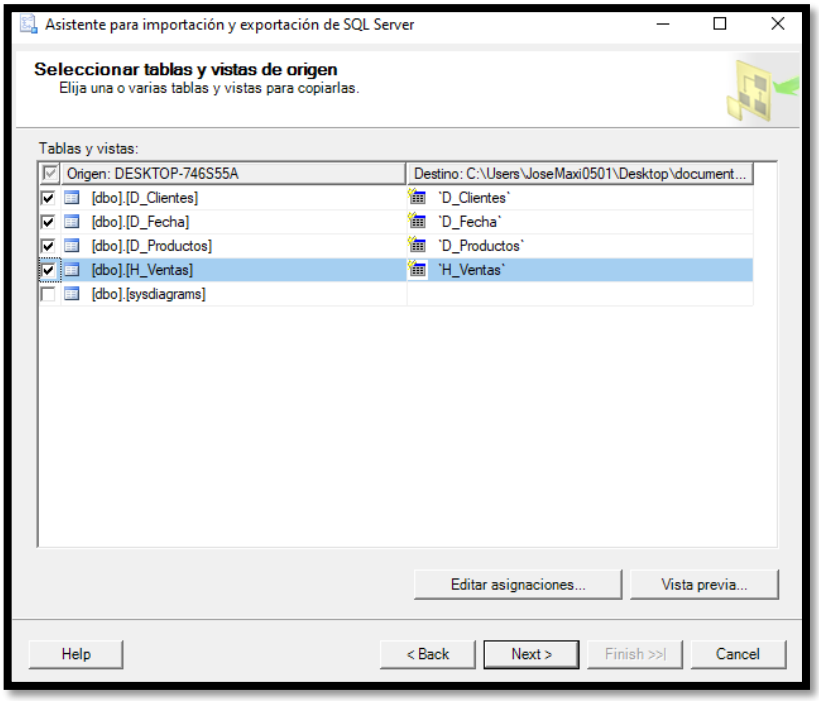


Fig. 17. Selección de tablas a exportar

Para realizar los 4 modelos que se presentaran en este trabajo se procedió a leer el archivo con extensión .csv realizado en el paso anterior, para la lectura correcta de este archivo se designaron ciertos parámetros para obtener una buena lectura de estos mismos, se utilizó la función read.csv que como su nombre lo indica permite la lectura de archivos .csv.

Se ubica la ruta en donde se encuentra alojado el archivo exportado, para luego seguir con la variable header que indica si el archivo contiene encabezado para sus columnas en este caso el archivo exportado cuenta con un header como primera columna por lo que se le ubicó un valor de TRUE, ya que por defecto viene con un valor de FALSO, y por último se le ubicó el tipo de separación que contiene el archivo exportado mediante la variable sep en este caso como el archivo es separado por comas se pondrá un valor de "," esto hará que se distribuyan de manera correcta los datos.

Como se mencionó en marco conceptual existen varios modelos o técnicas de minería de datos, de los cuales se puede definir que se dividen en dos grandes grupos como lo es por método de clasificación y regresión, las técnicas que se van a implementar son:

- Árbol de decisiones de regresión.
- Redes neuronales de regresión.
- Regresión lineal múltiple.
- Máquina de soporte de vectores.

Para que los datos puedan ser analizados de manera correcta, se convirtió las variables categóricas a numéricas utilizando el proceso one hot encoding lo que hace es crear un campo modificado de la variable que inicialmente se encontraba como categoría.

2.4.3.2. IMPLEMENTACIÓN DE LA TÉCNICA ÁRBOL DE DECISIONES.

Para realizar la técnica de árbol de decisiones se utilizar un programa llamado R Studio, lo primero que se hará en este programa es crear un nuevo archivo de tipo R Script, se utilizara el nombre "ArbolDecisionFinal", luego se importara las librerías que se utilizaran para esta técnica.

Tabla X

Librerías para la técnica árbol de decisiones

Librerías	
Rpart	para la creación del árbol
Rpart.plot	para graficar el árbol
Caret	para realizar las predicciones

Para la creación del árbol se tomó en cuenta ciertos aspectos que son los siguientes:

- Se escogerán aleatoriamente el 30 % de los datos para realizar las pruebas correspondientes.
- Variable a predecir.
- El set aleatorio fue seleccionado mediante la función set.seed al cual le dimos un valor de 123.

Tabla XI

Porcentaje de los nodos

Porcentaje de los nodos que no tenga hijos	
2	26%
4	21%
6	11%
9	10%
10	16%
11	16%

La profundidad máxima del modelo del árbol de decisiones de regresión es de 5, el nodo hijo con mejor rendimiento es el nodo número 2 con un porcentaje de observaciones de un 26% lo que quiere decir que en este segmento se aloja los datos con mejor probabilidad.

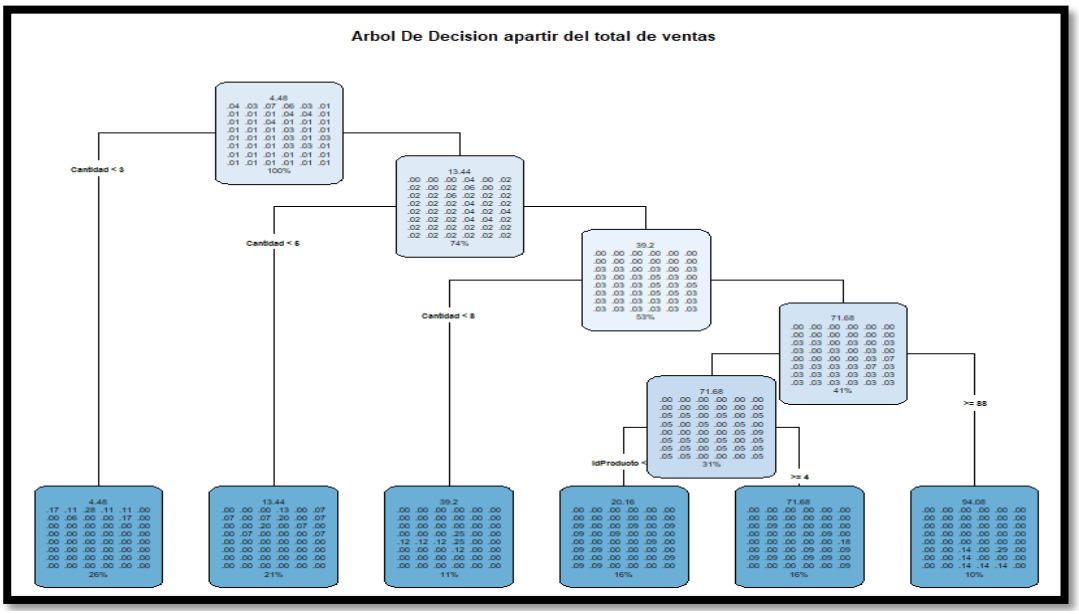


Fig. 19 Esquema del árbol de decisiones

2.4.3.3. IMPLEMENTACIÓN DE LA TÉCNICA REDES NEURONALES

Para realizar la técnica de redes neuronales se utilizará el mismo programa que se realizó el árbol de decisión llamado R Studio, lo primero que se hará en este programa es crear un nuevo archivo de tipo R Script, se utilizara el nombre "RedNeuronalFinal", luego se importara las librerías que se utilizaran para esta técnica.

Tabla XII

Librerías utilizadas para crear red neuronal

Librerías	
neuralnet	para la creación de la red neuronal
Ggplot2	para graficar la red neuronal
Caret	para realizar las predicciones

Para la creación de la red neuronal se tomó en cuenta ciertos aspectos que son los siguientes:

- Se escogerán aleatoriamente el 30 % de los datos para realizar las pruebas correspondientes.
- Variable a predecir.
- El set aleatorio de los datos de entrenamiento se los definió mediante una función llamada set.seed el cual le he dado un valor de 100.
- Cantidad de nodos por capas.

El valor con la ruta con las neuronas de mayor peso es: en el nodo 1-6 con un peso de 2.2091 en el nodo 2-5, con un peso de 5.2231, para el nodo 3-3 con un peso de 0.2423 y para el nodo 4-3 que vendría ser el último nodo que da como resultado la capa de salida tiene un peso de 3.0791.

Mientras que el valor de la ruta con el peor tamaño es decir que no tiene tanta relación con esas redes es: el nodo 1-4 con un peso de -9.6568, para el siguiente nodo con un peso no tan agradable es el nodo 2-2 con un peso de -5.2903, para el siguiente nodo 3-3 con un peso de -2.9558 y por último la capa de salida con peor peso se encuentra en el nodo 4-2 con un peso de -2.3169.

Para llegar a la capa de salida el modelo cuenta con 3 capas ocultas para obtener los datos con mejor precisión, en la primera capa oculta se cuenta con que cada nodo recibe un total de 6 entradas y para la segunda capa oculta se cuenta con que cada nodo recibe un total de 5 entradas, para el tercer nodo cuenta con 4 entradas, para dar como resultado en la capa de salida un total de 5 entradas con las cuales se destinan a un solo nodo que es el nodo resultante.

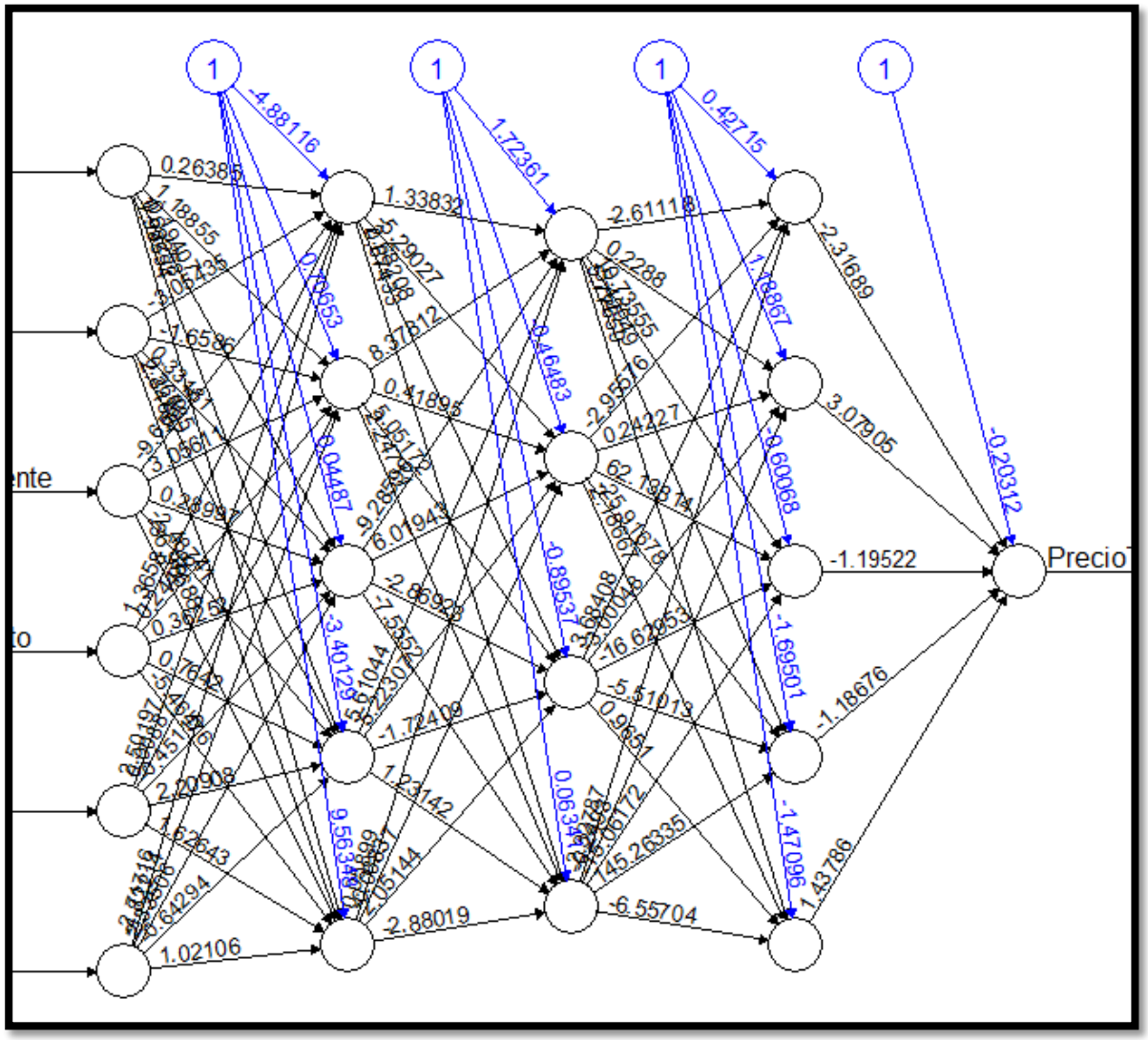


Fig. 20 Esquema de la red neuronal

2.4.3.4. APLICACIÓN DE LA TÉCNICA DE REGRESIÓN LINEAL MÚLTIPLE

Para realizar la técnica de regresión lineal múltiple, se utilizará el programa R Studio, primero se procedió a crear un Script con formato tipo R, con el nombre “RlmFinal”, una vez creado el script se procede a importar las librerías.

Tabla XIII

Librerías para Regresión lineal múltiple

Librerías	
Tidyverse	Necesario para la minería de datos.
Caret	Para realizar predicciones.

Para la creación del modelo de regresión lineal múltiple se tuvieron en cuenta ciertos aspectos:

- Se tomarán de forma aleatoria el 30% de los datos para las respectivas pruebas.
- Para obtener los datos aleatorios se utilizó la función set.seed al cual se le asigno un valor de 123.
- La variable a predecir.

Para interpretar el modelo de regresión lineal múltiple, se necesita de una variable que este correlacionada con la variable predictora a medida que los puntos del grafico estén más cercanos a la línea tiene mejor probabilidad de éxito en la predicción, el eje vertical está compuesto por una variable dependiente es decir los valores reales o los de prueba mientras que los puntos son los sets de datos proporcionados mediante la predicción y los valores reales.

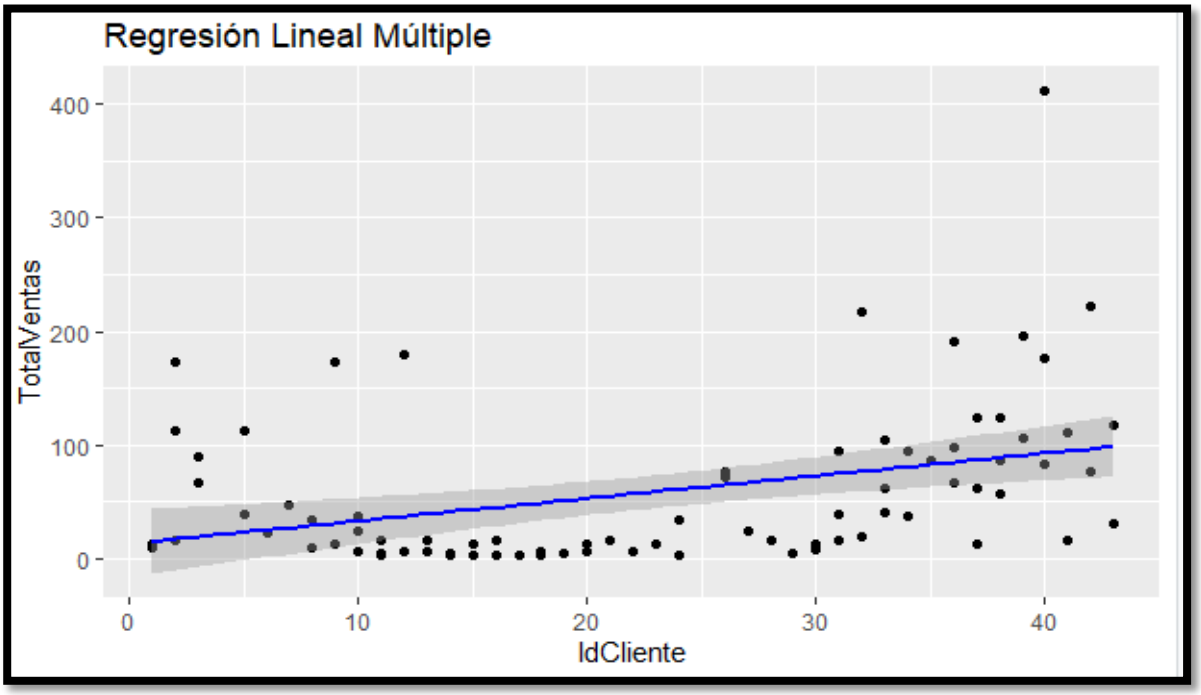


Fig. 21 Esquema de Regresión Lineal Múltiple

2.4.3.5. APLICACIÓN DE LA TÉCNICA DE MÁQUINA DE SOPORTE PARA VECTORES

Para realizar la técnica de máquina de soporte para vectores, se utilizará el programa R Studio, primero se procedió a crear un Script con formato tipo R, con el nombre "svmfinal", una vez creado el script se procede a importar las librerías.

Tabla XIV

Librerías para máquina de soporte para vectores.

Librerías	
E1071	para crear el modelo de máquina de soporte de vectores.
Caret	para realizar las predicciones

Para la creación del modelo SVM, se tomó en cuenta ciertos aspectos:

- Se escogerán aleatoriamente un 30% de datos para realizar.
- La semilla con la que se trabajó para seleccionar los valores aleatorios es de 123.
- Se le agregó un costo de 12.
- Un kernel de tipo lineal.
- Una variable gamma con un valor de 0.2.
- Un coeficiente0 con un valor de 2

Una vez parametrizado el modelo se procedió al respectivo diseño utilizando la variable que se seleccionó para predecir junto con el conjunto de variables destinadas a evaluar la predicción.

2.4.4. FASE 4: EVALUCIÓN DE RESULTADOS

2.4.4.1. EVALUACIÓN DE RESULTADOS DE LOS MODELOS.

Para poder evaluar los resultados de los modelos presentado en este trabajo, se utilizó la librería "Metrics", el cual como su nombre indica nos permite sacar los valores de los errores sin tener que hacer operaciones matemáticas complejas. Las siguientes métricas que se utilizaron son:

- Error Absoluto Medio (MAE).
- Error Cuadrático Medio (MSE).
- Raíz del Error Cuadrático Medio (RMSE).
- Coeficiente de Determinación (R^2).

La siguiente tabla muestra los resultados de las métricas evaluadas en el modelo de árbol de decisiones.

Tabla XV

Métricas del árbol de decisiones

Métricas del árbol de decisiones	
MAE	0,8811
MSE	0,8139
RMSE	0,9387
R^2	0,8992

La siguiente tabla muestra los resultados de las métricas evaluadas en el modelo de redes neuronales.

Tabla XVI

Métricas de la red neuronal

Métricas de Redes Neuronales	
MAE	0.3538
MSE	0,2249
RMSE	0,4742
R^2	0,7282

La siguiente tabla muestra los resultados de las métricas evaluadas en el modelo de regresión lineal múltiple.

Tabla XVII

Métricas de regresión lineal múltiple

Métricas del modelo de regresión lineal múltiple	
MAE	0,1850
MSE	0,3154
RMSE	0,4301
R^2	0,7972

La siguiente tabla muestra los resultados de las métricas evaluadas en el modelo de máquina de soporte de vectores.

Tabla XVIII

Métricas de máquina de soporte de vectores

Métricas del modelo de máquina de soporte de vectores	
MAE	0,1704
MSE	0,3025
RMSE	0,4128
R^2	0,8165

2.4.4.2. GRÁFICOS ESTADÍSTICOS DE LOS RESULTADOS

Estos gráficos que se presentara a continuación servirán para entender de mejor manera los resultados obtenidos de los procesos de las técnicas de “arboles de decisión”, de “redes neuronales”, de “regresión lineal múltiple” y de “máquina de soporte de vectores” con respecto a los datos obtenidos por la predicción y los datos reales planteados dentro de la base de datos.

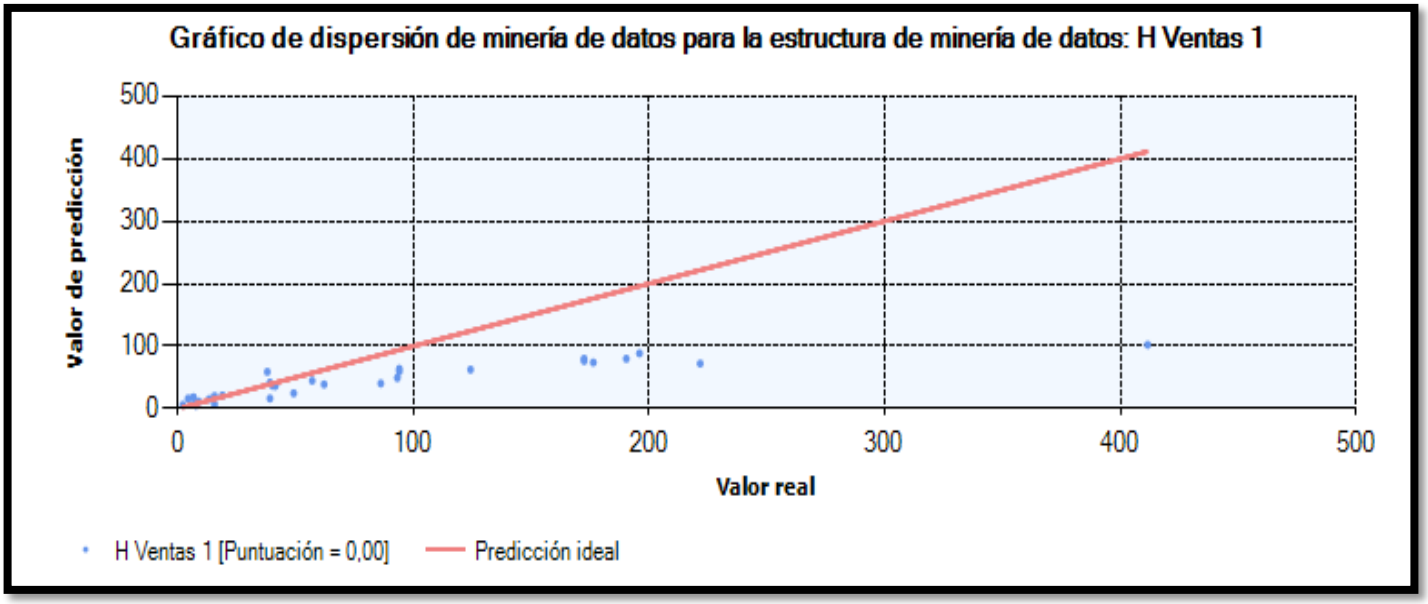


Fig. 23 Gráfico de dispersión del modelo de redes neuronales desde visual studio

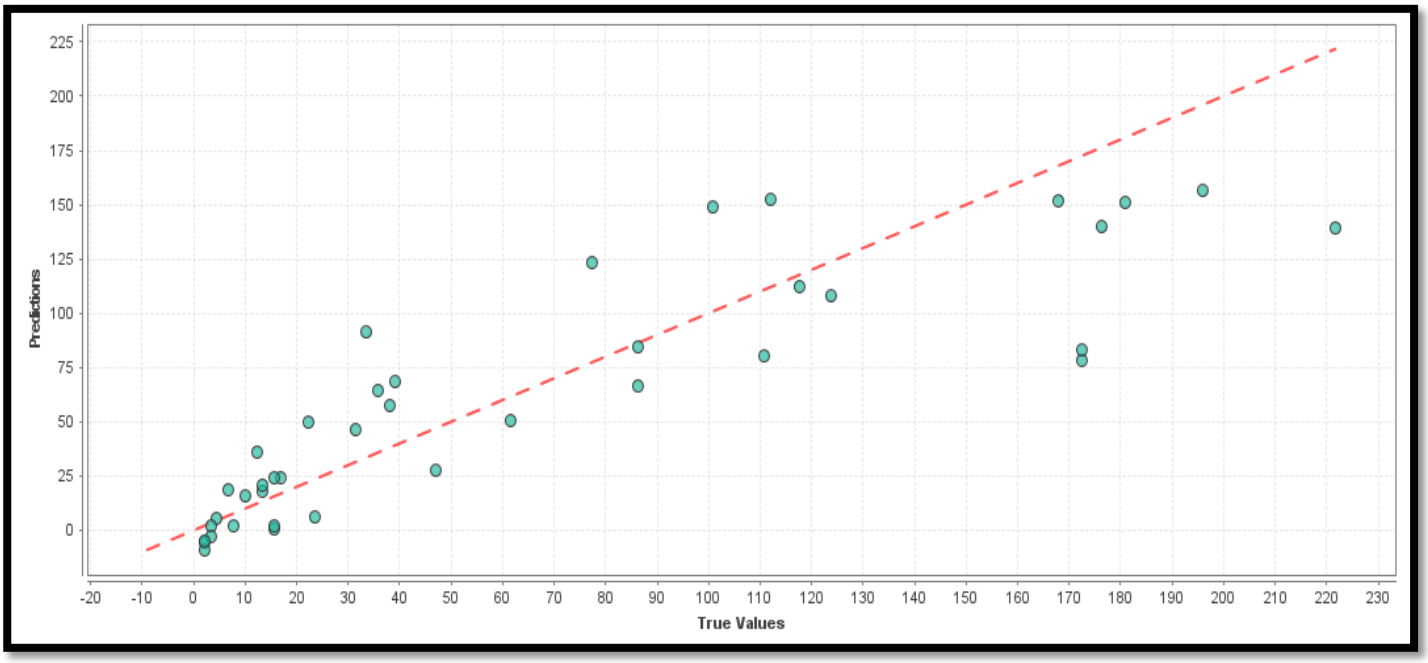


Fig. 22 Gráfico de dispersión del modelo de redes neuronales desde rapidminer

Como se puede observar en la Fig. 23 y la Fig. 22, es un gráfico de dispersión el cual cuenta con valores reales y los valores predictores, a menor dispersión quiere decir que el modelo es mucho óptimo, en la Fig. 23 que es extraída por el programa visual studio muestra menos puntos de distribución porque está tomando solo los valores de prueba mientras que la Fig. 22 que es extraída de rapidminer no hace particiones en set de entrenamiento y pruebas por lo que en el grafico se muestran todos los datos, también la relación de los datos influye

bastante a la hora de hacer el grafico de dispersión, en la Fig. 23 tiene un valor de 0. 672 mientras que la Fig. 22 tiene un valor de 0.865

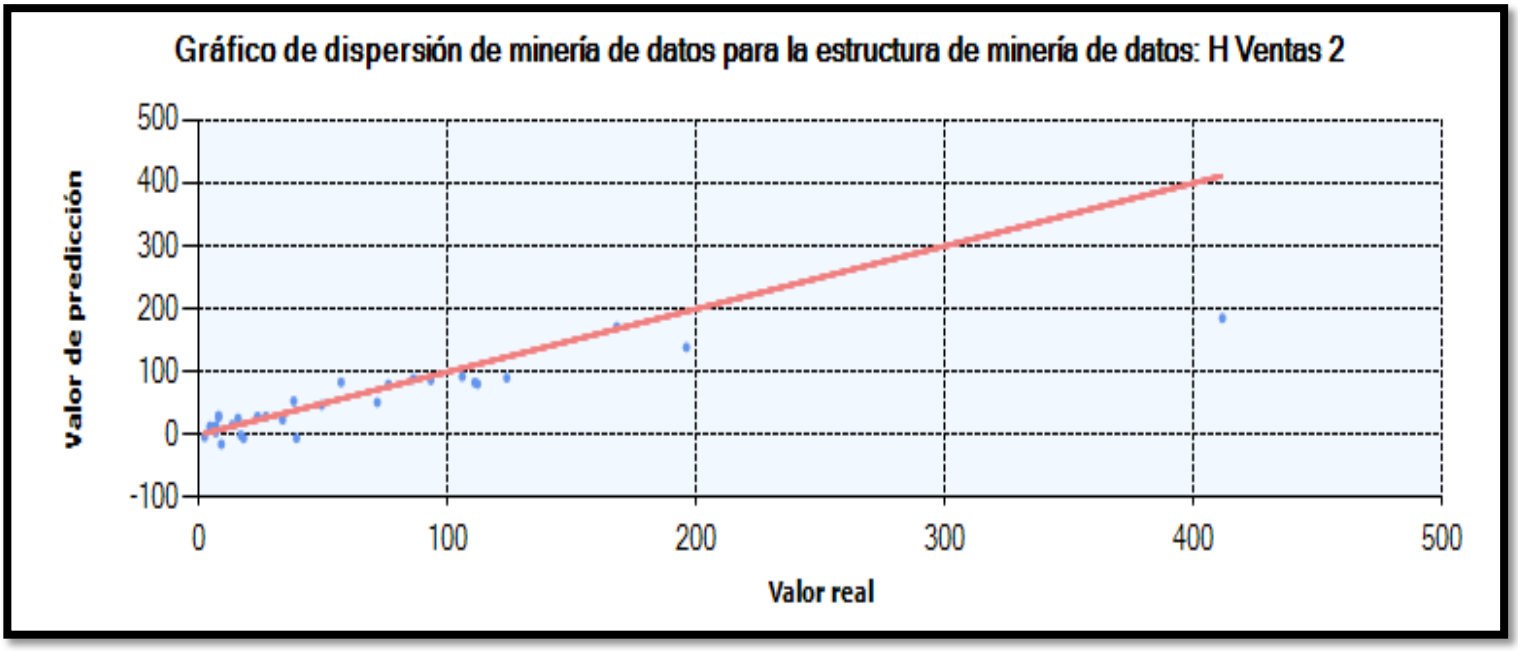


Fig. 25 Gráfico de dispersión del modelo de regresión lineal múltiple desde visual studio

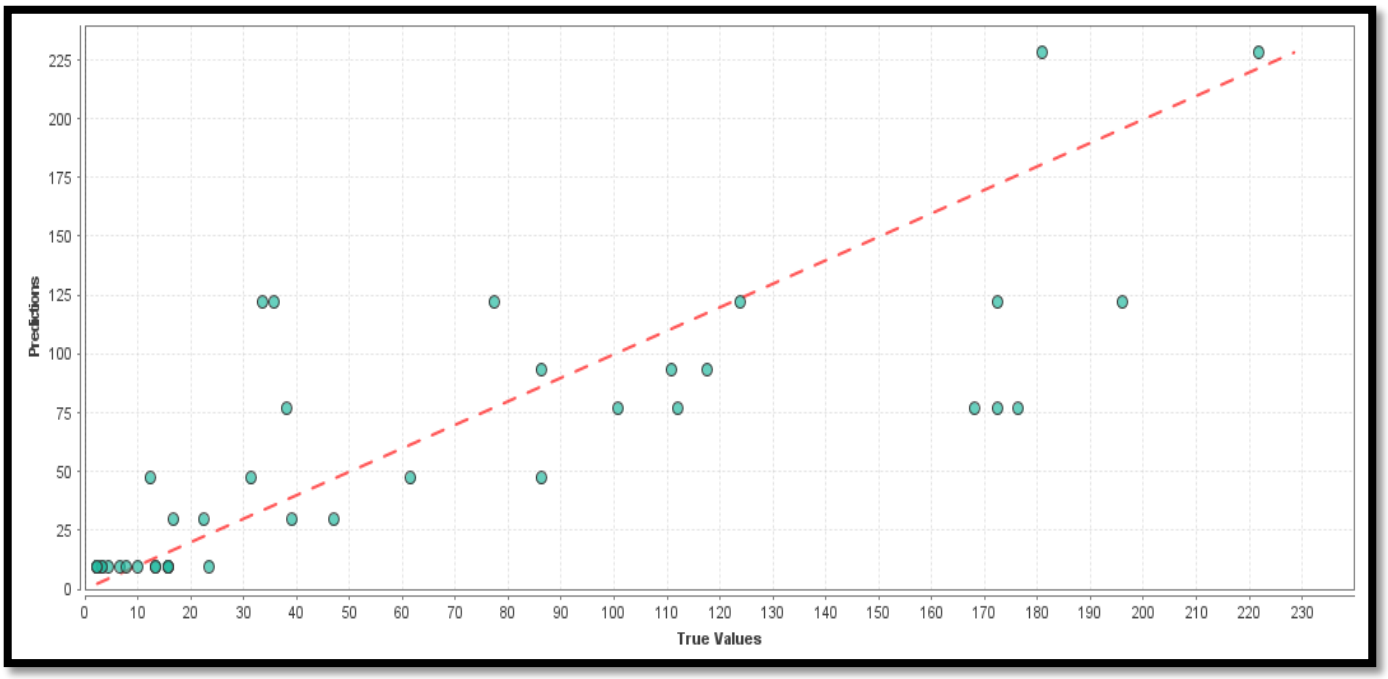


Fig. 24 Gráfico de dispersión del modelo de regresión lineal múltiple desde rapidminer

En la Fig. 25 el valor de correlación es de 0.873 lo que es un buen valor para que los valores de la variable de predicción no sean tan diferentes de los valores reales, en el Fig. 24 en este caso el valor es de 0.806 a medida que este valor de relación se acerque a 1 tendrá mejor

resultado de predicción, en este par de gráficos la mejor relación es la Fig. 25 puesto que a simple vista se ven que los valores están mucho más cerca de la línea diagonal mientras que el otro grafico se encuentra más dispersos.

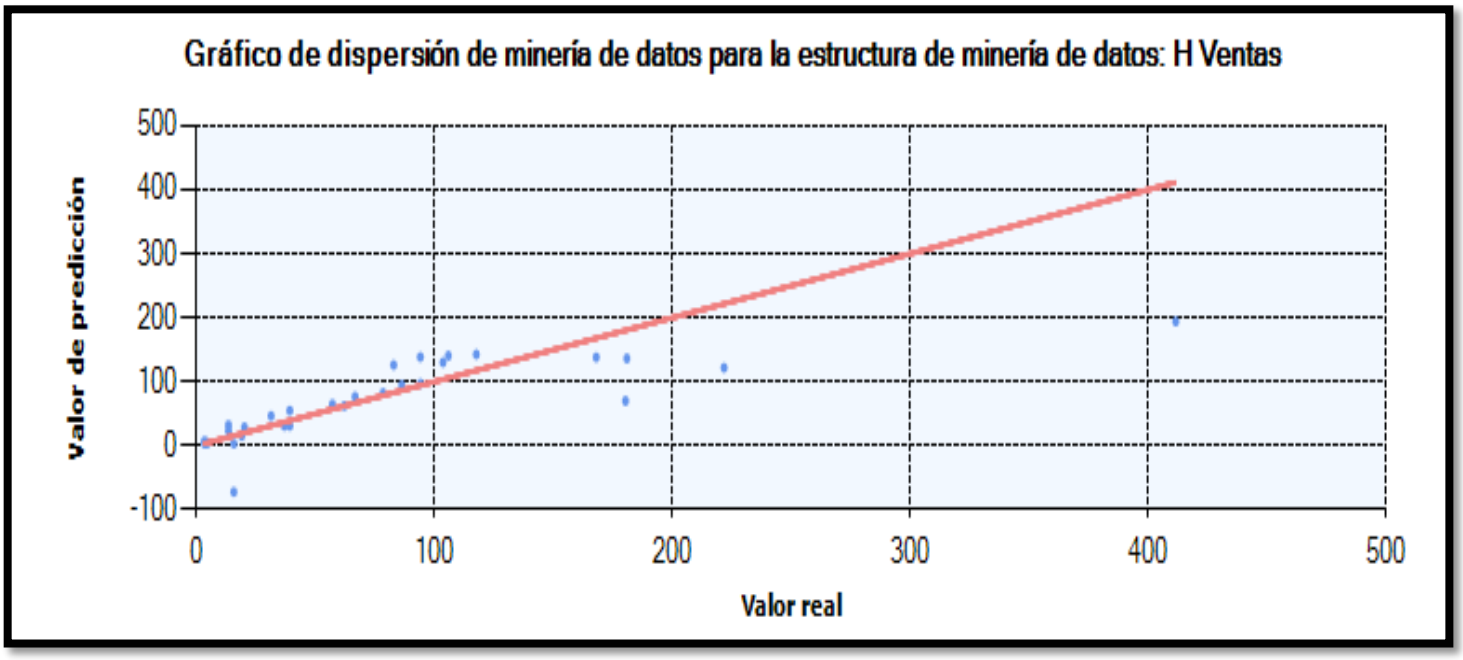


Fig. 26 Gráfico de dispersión del modelo de árboles de decisiones desde visual studio

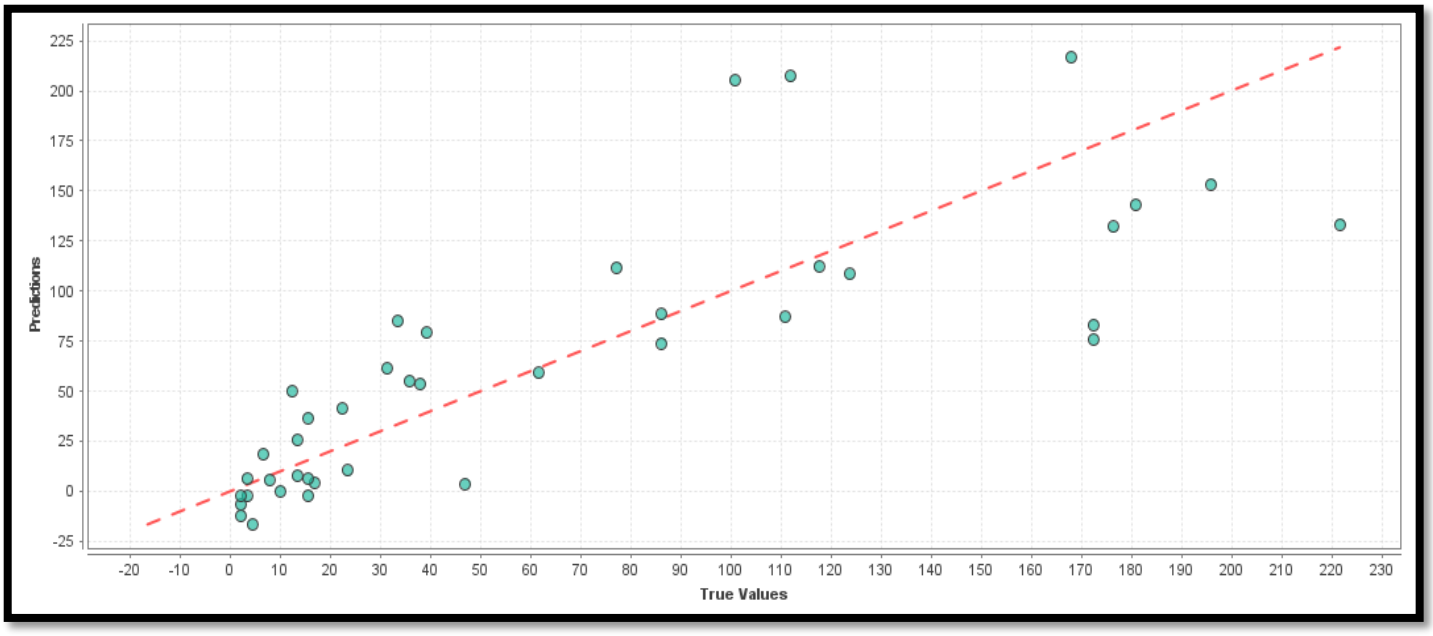


Fig. 27 Gráfico de dispersión del modelo de árboles de decisiones desde rapidminer

En la Fig. 26 y Fig. 27, son los gráficos de dispersión del modelo de árboles de decisión, uno es extraído de visual studio y el otro grafico es extraído de rapidminer, como se puede observar la Fig. 27 cuenta con valores mucho más dispersos teniendo un valor de correlación

de 0.817 mientras que la Fig. 26 cuenta con menos dispersión lo que da resultados óptimos con un valor de 0.856.

Attri...	Cantidad	EdadCli...	Id	IdCliente	IdFecha	IdProdu...	PrecioT...
Cantidad	1	-0.091	-0.195	-0.167	-0.195	0.036	0.674
EdadCli...	-0.091	1	-0.035	-0.026	-0.035	-0.135	-0.093
Id	-0.195	-0.035	1	0.951	1	0.738	0.296
IdCliente	-0.167	-0.026	0.951	1	0.951	0.746	0.323
IdFecha	-0.195	-0.035	1	0.951	1	0.738	0.296
IdProducto	0.036	-0.135	0.738	0.746	0.738	1	0.462
PrecioTo...	0.674	-0.093	0.296	0.323	0.296	0.462	1

Fig. 28 Correlación de variables

En la Fig. 28, se muestran los valores de correlación de las variables entre sí, por ejemplo, la variable idCliente tiene un valor de correlación de 0.951 con el idFecha lo que significa que estas dos variables son fuertemente relacionadas es decir que son directamente proporcionales, un valor cercano a -1 indicaría que estas variables son inversamente proporcionales.

2.5. RESULTADOS OBTENIDOS

2.5.1. RESULTADOS DE LOS MODELOS

Una vez realizados los modelos propuestos en la fase tres que habla sobre la implementación de las técnicas de minería de datos, se obtuvieron los siguientes resultados:

Tabla XIX

Resultados de las métricas de rendimiento

Métricas de rendimiento	Árbol de decisiones	Redes neuronales	Regresión lineal múltiple	Máquina de soporte de vectores
MAE	0,8811	0.3538	0,1850	0,1704
MSE	0,8139	0,2249	0,3154	0,3025

RMSE	0,9387	0,4742	0,4301	0,4128
R^2	0,8992	0,7282	0,7972	0,8165

Según los resultados obtenidos, mostrados en la Tabla XIX el modelo con mejor rendimiento es el de máquina de soporte de vectores, con un MSE de 0.3025, un MAE de 0.1704, un RMSE de 0.4128 y su coeficiente de determinación de 0.8165., mientras que el modelo con menos rendimiento es el de árbol de decisiones puesto que cuenta con un MSE de 0.8139, un MAE de 0.8811, un RMSE de 0.9387 y un coeficiente de determinación de 0.8992 debido a que sus métricas de rendimiento son superiores a las de máquina de soporte de vectores.

2.5.2. VARIABLE DEL PROYECTO

La variable que se evaluó durante el proyecto es el tiempo en el que la empresa toma decisiones hacia sus clientes, mediante la observación directa (ver anexo 2) se contrastó de que los clientes muchas veces no quedaban conformes con lo que realmente querían y esto consumía mucho tiempo por parte de los clientes por ello con esta propuesta se mejoró el tiempo de respuesta por parte de lo que realmente quieren los clientes.

Tabla XX

Tiempo de espera de los clientes

Variable del Proyecto	
Antes	**Después**
2 a 3 horas	1 hora

CONCLUSIONES

- La implementación de las técnicas de minería de datos sin duda es muy relevante a la hora de hacer un análisis de información o de datos, puesto que ayuda a optimizar de manera eficaz y a conocer la relación de variables de los datos almacenados, los tipos de datos que se almacenan, dando como resultado el descubrimiento de patrones que ayuden a la toma de decisiones de una tarea o de una responsabilidad.

- Trabajar con datos en cierta parte puede llegar a ser algo sensibles, puesto que, al trabajar con información de otras personas, ya sea como número de teléfono, la cedula o identificación de algún cargo, lugar de residencia e inclusive la edad puede llegar a ser una información delicada y sensible de tratar.

- Se implementó un almacén de datos el cual permitió que se desarrollaran las técnicas de manera correcta y eficaz, se realizó la integración de los datos a través del programa Microsoft visual studio, la creación del modelo o de la estructura del almacén de datos se desarrolló a través del programa Microsoft SQL Server Management Studio (SMSS).

- Existen un sin número de modelos de minería de datos, en este proyecto se abarcaron cuatro modelos de los cuales mediante métodos de regresión se establecieron métricas de rendimiento para comparar los resultados de los modelos y cual presentaba mejores resultados.

- El modelo con mejores resultados fue de el de máquina de soporte de vectores, con un MSE de 0.3025, un MAE de 0.1704 un RMSE de 0.4128 y un coeficiente de determinación de 0.8165, lo que convierte a este modelo en el que mejor rendimiento se obtuvo para el set de datos compartido de la empresa.

- La comprensión de los modelos obtenidos y de los resultados obtenidos, son importantes ya que con estos se puede definir patrones y conocer mejor las relaciones entre variables y datos, así como también tener conocimientos sobre bases de datos.

- Sin duda un papel muy importante a la hora de realizar minería de datos es la confidencialidad junto con la ética, puesto que trabajar con datos ajenos se debe asegurar que los datos con los que se trabaje de manera responsable, respetando los derechos de privacidad.

RECOMENDACIONES

- Para la aplicación de las técnicas de minería de datos, se pueden aplicar métodos de clasificación, otros modelos más profundizados los sets de datos con los que se requiera trabajar, ya que para cada set de datos existe un modelo con mejor rendimiento en este caso para este proyecto el modelo con mejor rendimiento fue el de máquina de soporte de vectores mientras que para otro set de datos quizás tenga mejor rendimiento un árbol de decisiones u otros modelos que no se aplicaron.

- En este trabajo se utilizó el entorno de trabajo centrado en R, para nuevas posibilidades el trabajo de minería de datos puede desarrollarse por otras plataformas como PostgreSQL, Python incluidas las librerías para el desarrollo y visualización de los modelos, Weka un software capaz de realizar modelos de minería de datos, Orange un software creado a partir de Python, y muchos otros programas.

- La minería no solo se centra en empresas de negocios de compra y venta, sino también entra en el ámbito de la medicina, la educación y muchos otros ámbitos más, por lo que para cada ámbito el proceso de la recolección de datos es diferente ya que se reúne o se recolecta los datos en base a lo que se necesita.

- Existen varias formas de graficar un modelo de minería de datos, en mi caso utilice la función plot de lenguaje r, pero también se podría elaborar los gráficos a partir de la librería Caret, de la librería ggplot2 y de otras más librerías para graficar.

REFERENCIAS

[1] P. Arcos Méndez, «Academia,» 2020. [En línea]. Available: https://www.academia.edu/44745426/Aplicaci%C3%B3n_de_miner%C3%ADa_de_datos_para_pron%C3%B3stico_de_ventas%20--use-spdy=off%20--disable-http2.

[2] Kyocera, «La minería de datos como herramienta estratégica,» [En línea]. Available: https://www.kyoceradocumentsolutions.es/es/smarter-workspaces/business-challenges/procesos/la-mineria-de-datos-como-herramienta-estrategica.html#:~:text=La%20miner%C3%ADa%20de%20datos%20tiene,ampliar%20el%20margen%20operativo%2C%20etc..

[3] Y. J. Marcano Aular y R. Talavera, «Scielo,» 01 2016. [En línea]. Available: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1012-15872007000100008.

[4] X. P. Silva Cama , «Análisis de la minería de datos aplicada en empresas del sector retail,» 1 2020. [En línea]. Available: https://repositorio.ucsp.edu.pe/bitstream/20.500.12590/16199/1/DONGO_POZO_ALD_MIN.pdf.

[5] A. J. Villao Balón, «Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela 'Lic. Angélica Villón L.',» 12 2021. [En línea]. Available: https://repositorio.upse.edu.ec/bitstream/46000/7330/1/UPSE-RCT-2022-Vol.8-No.2-008.pdf.

[6] A. J. Romero Fernández, «Minería de datos para la gestión de compras de medicamentos en el Hospital Básico El Puyo,» 08 2019. [En línea]. Available: https://dspace.uniandes.edu.ec/handle/123456789/10323.

[7] U. Shafique y H. Qaiser, «A comparative study of data mining process models (KDD, CRISP-DM and SEMMA),» *International Journal of Innovation and Scientific Research,* vol. 12, nº 1, pp. 217-222, 2014.

[8] Secretaria General UPSE, «Universidad Estatal Peninsula de Santa Elena,» 22 07 1998. [En línea]. Available: https://www.upse.edu.ec/secretariageneral/images/archivospdfsecretaria/4.REGLAMENTOS/1.%20NORMATIVAS%20ACAD%C3%89MICAS/REGLAMENTO_2019/RCS-SE-16-03-2019_REGLAMENTO_DEL_CENTRO_DE_INVESTIGACION_DE_SISTEMA_Y_TELECOMUNICACION.pdf.

[9] G. Huber, A Theory Of The Effects Of Advanced Information Technologies On Organizational Design., 1 ed., vol. 14, Academy Of Management Review, 1990, pp. 47-71.

[10] V. Alfonso Rodríguez y E. Chapis Cabrera, «Importancia de las tecnologías de la información y las comunicaciones, el internet y las redes sociales en el mejoramiento y desarrollo de las empresas,» *Revista Contribuciones a la Economía,* 2019.

[11] J. Amaya, Toma de decisiones gerenciales: Métodos cuantitativos para la administración, Segunda ed., Bogota D.C.: Ecoe ediciones., 2010.

[12] Observatorio Regional de Planificación para el Desarrollo de America Latina y el Caribe, «Plan de Creación de Oportunidades 2021-2025 de Ecuador,» [En línea]. Available: https://observatorioplanificacion.cepal.org/es/planes/plan-de-creacion-de-oportunidades-2021-2025-de-ecuador#:~:text=El%20Plan%20de%20Creaci%C3%B3n%20de,en%20el%20Plan%20de%20Gobierno.. [Último acceso: 20 06 2022].

[13] C. Rusu, «El alcance de la investigacion,» de *Metodologia de la Investigacion*, 2011.

[14] M. Sanchez, M. Fernandez y J. Diaz, «Técnicas e instrumentos de recolección de información: análisis y procesamiento realizado por el investigador cualitativo,» *Revista Científica UISRAEL,* vol. 8, nº 1, 4 2021.

[15] IBM, «Manual CRISP-DM de IBM SPSS Modeler - IBM Corporation,» 2020.

[16] H. E. Escobar Terán, M. Alcivar y A. Puris, «Aplicaciones de Minería de Datos en Marketing,» *Dialnet,* vol. 3, nº 8, pp. 503-512, 2016.

[17] V. Valcarcel Asencios, «Data Mining y el Descubrimiento del conocimiento,» *Revista de la Facultad de Ingeniería Industrial UNMSM,* p. 4, 2004.

[18] *LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES,* 2021.

[19] Asamblea Nacional Republica del Ecuador, «Registro oficial organo de la republica del ecuador,» 05 2021. [En línea].

[20] ASAMBLEA NACIONAL REPUBLICA DEL ECUADOR, LEY ORGÁNICA DE PROTECCIÓN DE DATOS, Quito, 2021.

[21] Microsoft, «Microsoft,» [En línea]. Available: https://powerbi.microsoft.com/es-es/. [Último acceso: 14 06 2022].

[22] G. Boccardo Bosoni y F. Ruiz Bruzzone, «Bookdown,» 03 07 2019. [En línea]. Available: https://bookdown.org/gboccardo/manual-ED-UCH/uso-basico-de-rstudio.html#que-es-rstudio-una-interfaz-para-usar-r.

[23] RapidMiner, «Estudio RapidMiner,» [En línea]. Available: https://docs.rapidminer.com/latest/studio/. [Último acceso: 14 06 2022].

[24] Microsoft, «Tareas básicas en Excel,» [En línea]. Available: https://support.microsoft.com/es-es/office/tareas-b%C3%A1sicas-en-excel-dc775dd1-fa52-430f-9c3c-d998d1735fca. [Último acceso: 14 06 2022].

[25] D. Strauss, Getting Started with Visual Studio 2019: Learning and Implementing New Features, Apress, 2019.

[26] Microsoft 2022, «Le damos la bienvenida al IDE de Visual Studio,» 28 10 2022. [En línea]. Available: https://learn.microsoft.com/es-es/visualstudio/get-started/visual-studio-ide?view=vs-2022.

[27] Microsoft 2022, «¿Qué es SQL Server Management Studio (SSMS)?,» 18 11 2022. [En línea]. Available: https://learn.microsoft.com/es-es/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver16.

[28] C. .. Aggarwal, Neural Networks and Deep Learning: A Textbook, Springer, 2018.

[29] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 2006.

[30] E. Porcel, G. Dapozo y M. Lopez, «Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios,» *XI Workshop de Investigadores en Ciencias de la Computación,* pp. 635-639, 2009.

[31] S. Hartshorn, Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners, 2016.

[32] C. Smith y M. Koning, Decision Trees and Random Forests: A Visual Introduction For Beginners, Independently, 2017.

[33] F. Perez Cruz, «Máquina de vectores soporte adaptativa y compacta,» *Dialnet,* 2000.

[34] N. Cristianini y J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[35] B. Etienne, Introduction to Machine Learning, Wolfram Media Inc, 2021.

[36] G. V. M. P. N. e. a. Chassagnon, «Aprendizaje profundo: definición y perspectivas para la imagen torácica.,» 2020.

[37] . N. Lavesson y P. Davidsson, «Evaluating learning algorithms and classifiers: : Cambridge University Press.,» 2011.

[38] B. Devlin, Data Warehouse: From Architecture to Implementation, Addison-Wesley, 1997.

[39] J. Reis y M. Housley, Fundamentals of Data Engineering: Plan and Build Robust Data Systems, O'Reilly Media, 2022.

[40] R. Kimball y M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Wiley, 2013.

[41] C. J. Date, Introduction to Database Systems, 8 ed., Pearson, 2003.

[42] C. Allen, S. Chatwin y C. Creary, Introduction to Relational Databases and SQL Programming, McGraw-Hill Osborne Media, 2003.

[43] M. V. Nevado Cabello, Introducción a las Bases de Datos relacionales, Vision libros, 2010.

[44] V. Valcarcel Asencios, «DATA MININGYEL DESCUBRIMIENTO DEL CONOCIMIENTO,» *Revista de la Facultad de Ingeniería Industrial,* vol. 7, nº 2, pp. 83-86, 2004.

[45] L. C. Molina Felix, «Data mining: torturando a los datos hasta que confiesen,» 2002. [En línea]. Available: https://www.uoc.edu/pdf/web/esp/art/uoc/molina1102/molina1102.pdf.

[46] H. Jiawei, . K. Micheline y P. Jian, Data Mining: Concepts an Techniques, 2012.

[47] C. M. Bishop, Pattern Recognition and Machine Learning, Springer New York, NY, 2006.

[48] T. Naeem, «Conceptos de Data Warehouse: enfoque de Kimball vs. Inmon,» 03 02 2020. [En línea]. Available: https://www.astera.com/es/type/blog/data-warehouse-concepts/.

[49] Google, «Google Maps,» 18 01 2023. [En línea]. Available: https://www.google.com/maps/@-2.229812,-80.937022,17z?hl=es.

[50] C. Manzano Munizaga, «Minería de datos para el diagnóstico de deterioro neuropsicológico a individuos expuestos a pesticidas organofosforados,» *Thesis for: Master of Information Technology,* 2015.

[51] Redaccion KeepCoding , «KeepCoding,» 2022. [En línea]. Available: https://keepcoding.io/blog/red-neuronal-en-deep-learning/.

[52] Z. Mohammed J. y W. Meira Jr, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press, 2020.

[53] R. Kimball y M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Wiley, 2013.

ANEXOS

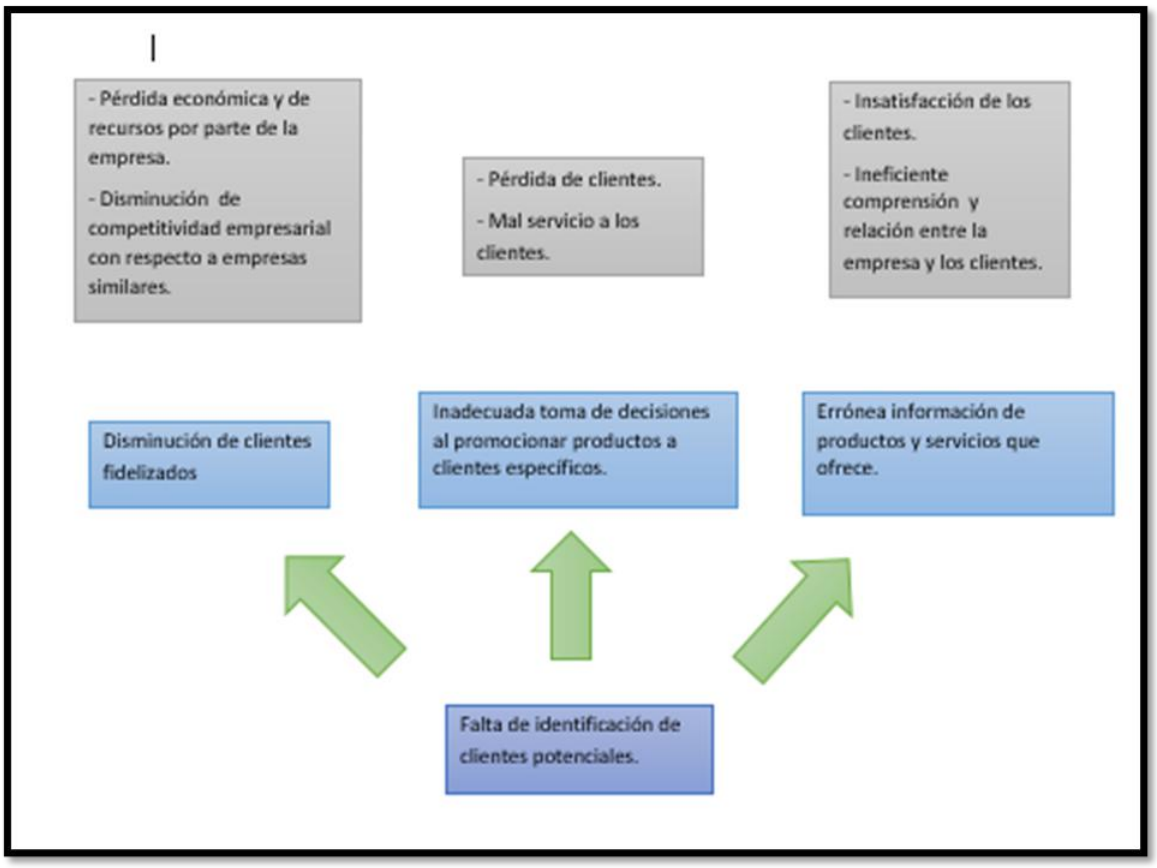
Universidad Estatal Península de Santa Elena
Facultad de Sistemas y Telecomunicaciones
Carrera de Tecnología de la Información

Entrevista dirigida al gerente de la empresa "Custom Place"	
Objetivo: Conocer información importante sobre la empresa "Custom Place"	
1.	¿En qué año inició la empresa con sus funciones?
2.	¿Cuáles han sido los inconvenientes que ha tenido la empresa en los 2 años de pandemia?
3.	¿Qué tecnologías utiliza la empresa para dar información?
4.	¿Cuántos clientes se han integrado a la empresa en estos años?
5.	¿Cuántas personas están delate de la empresa?
6.	¿Cuánta es la cantidad vendida y cuanto es la cantidad de dinero invertido?
7.	¿Qué es lo que desea conseguir la empresa?
Responsable:	José Suarez Alvarado

Anexo 1.- Preguntas de la entrevista

Universidad Estatal Península de Santa Elena **Facultad de Sistemas y Telecomunicaciones** **Carrera de Tecnología de la Información**	
Prueba de Observación	
Lugar: Empresa Custom Place, Ciudadela "Puerta del Sol" entre Las Milinas (Salinas) y Barrio 9 de octubre (José Luis Tamayo). **Tipo de Observación:** Observación Directa	
Duración de la prueba de observación: 1 hora	
• La empresa toma gran cantidad de tiempo en solventar las necesidades de los clientes. • Algunos clientes simplemente no ordenan porque no encuentran lo que desean llevar. • La empresa muchas veces se queda sin materiales porque no tienen idea del comportamiento de los clientes. • Sus registros de clientes son muy amplios por lo que les resulta complicado el análisis de los mismos.	
Responsable:	José Suarez Alvarado

Anexo 2. Prueba de observación



Anexo 3. Árbol de problemas. Técnica de identificación de problema en empresa "Custom Place”