ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

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

SISTEMA DE RECOMENDACIÓN DE PRODUCTOS PARA EMPRESAS DE RETAIL EN EL ECUADOR

AUTOR: AREVALO PELÁEZ, JOSE MATEO

DIRECTOR: MSC. DÍAZ ZUÑIGA, MAGI PAÚL

SANGOLQUÍ

2018

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"*SISTEMA DE RECOMENDACIÓN DE PRODUCTOS PARA EMPRESAS DE RETAIL EN EL ECUADOR*"** fue realizado por el señor ***Arévalo Peláez, José Mateo*** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 20 de diciembre del 2018

Firma:

.....

Ing. Díaz Zuñiga Magui Paúl MSc.

C.C.: 1707249072

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Yo, ***Arévalo Peláez, José Mateo***, con cédula de identidad n° 0103580072, declaro que el contenido, ideas y criterios del trabajo de titulación: **"*Sistema de recomendación de productos para empresas de retail en el Ecuador*"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas. Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 23 de octubre del 2018

Firma

.....

Ing. José Mateo Arévalo Peláez

C.C.: 0103580072

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

*Yo, **Arévalo Peláez, José Mateo**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Sistema de recomendación de productos para empresas de retail en el ecuador"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.*

Sangolquí, 23 de octubre del 2018

Firma

.....

Ing. José Mateo Arévalo Peláez

C.C.: 0103580072

DEDICATORIA

A mi hija Mayte el centro de mi universo,

A mis padres Mateo y Carmen por su infinito amor y apoyo,

A mi esposa por su paciencia y amor,

A mi hermano por enseñarme a ser mejor,

A mi hermana por su apoyo incondicional.

Mateo.

AGRADECIMIENTO

Al Ing. Paul Díaz director de mi trabajo de investigación, que gracias a su tiempo y dedicación contribuyo a que el presente trabajo de investigación pueda desarrollarse.

A mi amada esposa Dunia, por su paciencia y apoyo durante los años de estudio fueron largas horas de viaje juntos, gracias.

A mi hija Mayte, por comprender que papá no pudo estar en fechas importantes ya que tuvo que viajar para estudiar, gracias hija mía.

A mis padres y hermanos por el apoyo y motivación durante los años de estudio, gracias.

Mateo.

ÍNDICE DE CONTENIDOS

DEDICATORIA ..... **iv**
AGRADECIMIENTO ..... **v**
ÍNDICE DE CONTENIDOS ..... **vi**
ÍNDICE DE FIGURAS ..... **x**
RESUMEN ..... **xii**
ABSTRACT ..... **xiii**
1. CAPÍTULO I **1**
INTRODUCCIÓN **1**
1.1. Antecedentes 1
1.2. Planteamiento del problema 2
1.3. Justificación e Importancia 3
1.4. Objetivo general 4
1.5. Objetivos específicos 4
1.6. Formulación del problema 5
2. CAPÍTULO II **7**
FUNDAMENTACIÓN TEÓRICA **7**
2.1. Marco teórico **7**
2.1.1. Fundamentación de la variable Independiente 7
Fundamentación de la variable dependiente 9
2.2. Antecedentes del estado del arte **11**
2.2.1. Definición del objetivo 11
2.2.2. Definición de los criterios de inclusión y exclusión 11
2.2.3. Definición de la estrategia de búsqueda 12
2.2.4. Construcción de la cadena de búsqueda 14
3. CAPÍTULO III **22**
MEMORIA TÉCNICA METODOLÓGICA **22**

3.1. Metodología de Investigación 22
3.1.1. Definición de la tarea 22
3.1.2. Recolección y análisis de datos 22
3.1.3. Selección y configuración del modelo 23
3.1.4. Limpieza de datos 23
3.1.5. Análisis de resultados 23
3.1.6. Reporte a los tomadores de decisión 24
3.2. Ejecución del proceso de investigación 25
3.2.1. Definición de la tarea 25
3.2.2. Recolección y análisis de datos 25
3.2.3. Selección y configuración del modelo 25
3.2.4. Limpieza de datos 26
3.2.5. Análisis de resultados 26
3.2.6. Reporte a los tomadores de decisión 27
4. CAPÍTULO IV 29
RESULTADOS 29
4.1 Informe de Resultados 29
4.1.1. Definición de la tarea 29
4.1.2. Recolección y análisis de datos 30
4.1.3. Selección y configuración del modelo 39
4.1.4. Limpieza de datos 47
4.1.5. Análisis de resultados 52
4.1.6. Reporte a los tomadores de decisión 56
4.2 Metodología para ejecutar la propuesta 60
5. CAPÍTULO V 63
CONCLUSIONES Y RECOMENDACIONES 63

5.1. Conclusiones 63

5.2. Recomendaciones 64

BIBLIOGRAFÍA 66

ÍNDICE DE TABLAS

Tabla 1 *Estudios por grupo de control* 13
Tabla 2 *Construcción de la cadena de búsqueda* 14
Tabla 3 *Atributos para análisis de datos* 30
Tabla 4 *Criterios de selección de atributos* 42

ÍNDICE DE FIGURAS

Figura 1 Relación de variables 7
Figura 2 Artículos encontrados en los repositorios académicos.....15
Figura 3 Marco de Procesos para SME 24
Figura 4 Edad económica vs descripción del producto 33
Figura 5 Sexo cliente vs descripción del producto 34
Figura 6 Nivel de estudios Cliente vs descripción del producto 34
Figura 7 Profesión vs descripción del producto 35
Figura 8 Ingresos Cliente vs descripción del producto.....35
Figura 9 Tipo trabajo cliente vs descripción del producto 36
Figura 10 Número cargas cliente vs descripción del producto.....36
Figura 11 Tipo vivienda Cliente vs descripción del producto.....37
Figura 12 Cantón domicilio Cliente vs descripción del producto 37
Figura 13 Provincia domicilio cliente vs descripción del producto 38
Figura 14 Años antigüedad cliente vs descripción del producto 38
Figura 15 Tipo cliente vs descripción del producto 39
Figura 16 Proceso de carga de datos 40
Figura 17 Selección de tarea.....41
Figura 18 Clases del atributo a predecir 41
Figura 19 Selección de atributos de entrada del modelo 42
Figura 20 Selección de algoritmos de prueba.....43
Figura 21 Comparación de tipos de modelos según precisión 44
Figura 22 Comparación de tipos de modelos según margen de error 45
Figura 23 Simulación de modelo creado 46
Figura 24 Resultados del modelo creado.....46
Figura 25 Proceso del modelo Naive Bayes en RapidMiner parte 1 47
Figura 26 Proceso del modelo Naive Bayes en RapidMiner parte 2.....48
Figura 27 Proceso de Pre-Procesamiento en RapidMiner 49
Figura 28 Proceso de remplazo de valores faltantes en RapidMiner 49

Figura 29 Proceso de ordenado y filtrado en RapidMiner.....50
Figura 30 Proceso de separación de data y aplicación del modelo en RapidMiner 50
Figura 31 Proceso de validación del modelo en RapidMiner.....51
Figura 32 Resultado del modelo Naive Bayes.....51
Figura 33 Porcentajes de división de la data para validación del modelo.....52
Figura 34 Matriz de confusión obtenida del modelo.....52
Figura 35 Gráfico de dispersión de la predicción alcanzada con el modelo 53
Figura 36 Productos predichos por el modelo.....54
Figura 37 Clases con mayor porcentaje de predicción de acuerdo con el modelo.....55
Figura 38 Precisión del modelo.....55
Figura 39 Proceso de extracción de la data en RapidMiner 56
Figura 40 Carga de datos del modelo en Tableau 57
Figura 41 Reporte 1 de recomendación de productos en Tableau.....57
Figura 42 Reporte 2 de recomendación de productos en Tableau.....58
Figura 43 Reporte 3 de recomendación de productos en Tableau.....58
Figura 44 Reporte 4 de recomendación de productos en Tableau.....59
Figura 45 Reporte 5 de recomendación de productos en Tableau.....59

RESUMEN

Las empresas de ventas de productos al por mayor y menor gastan grandes sumas de dinero en campañas de publicidad masivas para vender sus productos. Dichas campañas consisten en correos, llamadas telefónicas, mensajes de texto, propaganda en radio y televisión. No obstante, se ha detectado que las campañas no tienen los resultados esperados, y que, por el contrario, llegan en algunos casos incluso a crear apatía en los clientes por los correos, llamadas y mensajes que reciben, con productos que no son de su interés. El presente proyecto mejora la efectividad de las campañas publicitarias de una empresa ecuatoriana dedicada a la venta de productos para el hogar, mediante el uso de metodologías de inteligencia de negocios. Se realizó un sistema de recomendación utilizando técnicas de minería de datos analizando la información histórica de la organización, a fin de encontrar patrones de compra entre sus clientes y productos. Para esto, se siguieron los lineamientos de la metodología de Dittert, Härting, Reichstein, & Bayer, 2018, la cual combina las mejores prácticas de las metodologías diseñadas para la analítica de datos de pequeñas y medianas empresas. Como resultado de esta implementación la organización cuenta con un sistema de apoyo para la mejora de la efectividad de su fuerza de ventas, ofreciendo un sistema que permite sugerir productos a sus clientes tomando en cuenta el segmento de mercado en el que se encuentran, sus ingresos, datos demográficos y preferencias.

PALABRAS CLAVES:

- **MODELO DE MINERÍA DE DATOS**
- **PATRONES DE COMPRA**
- **EMPRESAS DE PRODUCTOS**
- **SISTEMA DE APOYO**

ABSTRACT

Retail companies spend large sums of money on mass advertising campaigns to sell their products. These campaigns consist of emails, telephone calls, text messages, advertising on radio and television. However, it the company has detected that the campaigns do not have the expected results; on the contrary, in some cases they even create apathy in the clients due to the mails, calls and messages they receive, with products that are not of their interest. This project improves the effectiveness of the advertising campaigns of an Ecuadorian company dedicated to the sale of products for the home, by using business intelligence methodologies. A recommendation system was implemented, using data mining techniques, analyzing the historical information of the organization in order to find purchase patterns among its customers and products. For this, the guidelines of the methodology of Dittert, Härting, Reichstein, & Bayer, 2018 was followed, which combines the best practices of the methodologies designed for the data analytics of small and medium enterprises. As result of this implementation, the organization has a support system to improve the effectiveness of its sales force, offering a system that allows suggesting products to its customers taking into account the market segment in which they are located, their income, demographics and preferences.

KEYWORDS:

- **DATA MINING MODEL**
- **PURCHASE PATTERNS**
- **PRODUCT COMPANIES**
- **SUPPORT SYSTEM**

CAPÍTULO I

INTRODUCCIÓN

1.1. Antecedentes

Con la masificación del Internet y los comercios en línea a nivel mundial, muchas páginas web dedicadas a la comercialización de productos como Amazon, Wish, Ali Express, se han proliferado ofreciendo millones de productos a los usuarios, este gran catálogo de productos se ha convertido en un inconveniente para los usuarios ya que están sobrecargados de información lo que ha generado una reducción del interés y satisfacción en los clientes (Usmani, 2017) .

Históricamente han existido empresas exitosas como Blackberry, Palm y Nokia que en su momento fueron líderes y referentes del mercado a nivel mundial pero no innovaron a la velocidad que la tecnología avanzaba y se quedaron rezagadas en relación con su competencia llegando a ser absorbidas por otras empresas más grandes como el caso de Nokia que fue adquirida por Microsoft.

En el Ecuador las empresas dedicadas a las ventas al por mayor y menor están buscando la manera de innovar para evitar quebrar o ser absorbidas por empresas más grandes, para esto están utilizando las tendencias existentes en el mercado como venta cruzada, fidelización de clientes por medio de tarjetas de socio preferente, promociones masivas por medio de llamadas telefónicas y mensajes de texto. Se ha detectado que las llamadas masivas y los mensajes de texto no están teniendo los resultados esperados y más bien están creando apatía en los clientes hacia la empresa ya que se ven sobrecargados de información que no es de su interés (Figueroa, 2016).

Para contribuir con una solución efectiva al inconveniente de mantener el interés de los clientes, fidelizándolos con la empresa, al mejorar su satisfacción el presente proyecto usa técnicas

de inteligencia de negocios, para encontrar las relaciones entre los productos y los clientes, que permitan realizar recomendaciones de artículos que se ajusten a las necesidades y preferencias de los clientes.

1.2.Planteamiento del problema

Las empresas dedicadas a las ventas al por menor gastan grandes sumas de dinero anualmente en campañas publicitarias con el fin de atraer más clientes. Estas campañas generalmente van dirigidas al público en general como propaganda en radio y televisión, las cuales son costosas y no siempre tienen el retorno de inversión esperado. Por este motivo las organizaciones cuentan con personal especializado para realizar campañas publicitarias personalizadas por medio de llamadas telefónicas o correo electrónico a su base de datos de clientes para informarles de ofertas de productos. Los problemas con este tipo de campañas se dan ya que se ofertan productos a clientes que a lo mejor ya cuentan con el mismo o simplemente no les interesa; por ejemplo, un juego de comedor a un cliente con el perfil de estudiante o una computadora a un cliente de la tercera edad que por sus características no le interesa, esto provoca una reacción negativa del cliente hacia la empresa, porque lo saturan de ofertas que no son de su interés.

Por estos motivos, las campañas de mercadeo no obtienen los resultados esperados llevando a las empresas a buscar alternativas como el uso de la inteligencia de negocios para implementar sistemas que les permitan generar recomendaciones de productos de acuerdo con el perfil demográfico y económico del cliente para ofrecerle productos acordes a sus gustos evitando saturarlo de propaganda innecesaria.

1.3.Justificación e Importancia

Actualmente el Ecuador se encuentra en un proceso recuperación económica. Esto está obligando a las empresas a optimizar sus procesos y mejorar sus prácticas para ser más competitivas (Angulo, 2017). Entre estas prácticas está la optimización de la productividad de las empresas, teniendo como uno de sus parámetros de optimización el marketing, ya que en la actualidad se realizan grandes esfuerzos económicos y de recursos humanos para promocionar un producto o marca. Estas campañas pueden ser efectivas, pero sus costos son elevados, dejando de ser competitivas y sin poder fidelizar a sus clientes, al no poder propiciarles un servicio personalizado de acorde a sus preferencias.

Si las empresas no se ajustan a los cambios tecnológicos y aprovechan la información histórica que tienen corren el riesgo de desaparecer

El presente proyecto disminuye los costos de marketing y mejora los resultados obtenidos como resultado de las campañas publicitarias por medio de la utilización de la Inteligencia de Negocios e información histórica almacena en bases de datos de una empresa caso de estudio. Para lograr estos objetivos se utiliza metodologías para generar recomendaciones de productos a un segmento especifico del mercado que tenga mayor probabilidad de adquirir el mismo, lo cual generará un valor agregado para las empresas mediante la mejora de sus ventas y el aumento de la lealtad de sus clientes (Jannach, 2015).

En esta investigación se realizó la recolección y análisis de la información para definir las fuentes de datos a ser utilizadas, así como también los datos con valores atípicos para que sean excluidos, a fin de que no alteren los resultados, luego se procedió a la selección e implementación del modelo a ser adoptado acompañado del formato de los datos para que los mismos puedan ser

utilizados de acuerdo a los requerimientos del modelo, con esto se procedió a evaluar los resultados conjuntamente con los expertos del negocio quienes determinaron si los resultados obtenidos son óptimos, para esto se tomó en cuenta el 70% de los registros de la base de datos del sistema de ventas de la organización para las pruebas y aprendizaje del sistema, el 30% restante se utilizó para verificar la eficacia del sistema.

El presente proyecto no genera notificaciones automáticas a los potenciales clientes sobre los productos que les pueda interesar, el sistema genera un listado de clientes con los productos que puedan ser de interés para ellos, esto permitirá que la empresa proceda a buscar la mejor manera de contactar al cliente para ofertarle los productos.

1.4.Objetivo general

Implementar un sistema de recomendación de productos que, mediante el análisis de información histórica de los clientes, permita identificar clientes potenciales para productos determinados, con la finalidad de mejorar la efectividad de su fuerza de ventas.

1.5. Objetivos específicos

OE1: Realizar un análisis de la literatura mediante una revisión inicial, para determinar las técnicas de sistemas de recomendaciones existentes y seleccionar la que mejor se ajuste a las necesidades de la organización.

OE2: Recolectar y analizar datos de una organización caso de estudio mediante el uso de diagramas de dispersión, para identificar valores atípicos que puedan distorsionar los resultados.

OE3: Seleccionar y configurar el modelo de inteligencia de negocios a implementar, como producto del análisis de literatura, identificando el que mejor se ajuste a las necesidades de la empresa.

OE4: Formatear los datos para que puedan trabajar adecuadamente con el modelo seleccionado, eliminando datos nulos, discretizando los datos de ser necesarios para garantizar el correcto funcionamiento del modelo de inteligencia de negocios implementado.

OE5: Evaluar los resultados mediante el uso de técnicas de inteligencia de negocios, para determinar la probabilidad de que el modelo de inteligencia de negocios utilizado ayude a mejorar la efectividad de la fuerza de ventas.

1.6. Formulación del problema

Para la consecución del objetivo específico el proyecto de análisis, diseño e implementación de un sistema de recomendación de productos requiere se respondan las siguientes preguntas para cada objetivo específico:

OE1 -RQ1.1: ¿Cuáles son los estudios existentes en la actualidad sobre sistemas de recomendación para empresas de ventas al por mayor y menor?

OE1- RQ1.2: ¿Cuáles son las técnicas para la generación de sistemas de recomendaciones que mejor se ajustan a la realidad de la organización y por qué?

OE2 – RQ2.1: ¿Cuáles son las fuentes de datos con las que cuenta la organización?

OE2 – RQ2.2: ¿Existen valores atípicos que puedan afectar el resultado de la ejecución de los algoritmos de inteligencia de negocios?

OE3 - RQ 3.1: ¿Cuáles son los modelos de recomendaciones existentes en la actualidad?

OE3 – RQ3.2: ¿Qué modelo de recomendación se acopla mejor a las necesidades de la organización caso de estudio?

OE4 – RQ4.1: ¿Qué tipo de datos necesita el algoritmo de inteligencia de negocios para sistemas de recomendaciones para su correcto funcionamiento?

OE4 – RQ4.2: ¿Es posible acoplar los datos existentes para que se ajusten a las necesidades del algoritmo de inteligencia de negocios para sistemas de recomendaciones?

OE5 – RQ5.1: ¿Cuál es el nivel de confianza con el que se va a trabajar para el análisis de los resultados?

OE5 – RQ5.2: ¿Cuál es el margen de Error del modelo seleccionado?

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1. Marco teórico

La fundamentación teórica busca la congruencia con la hipótesis, para esto se realiza un análisis de la teoría usando las variables del problema, con la finalidad de investigar jerárquicamente cada categoría hasta llegar a la categoría que comprende y explica las variables dependientes e independientes del tema de estudio, para esto se propone la siguiente jerarquía de estudio:

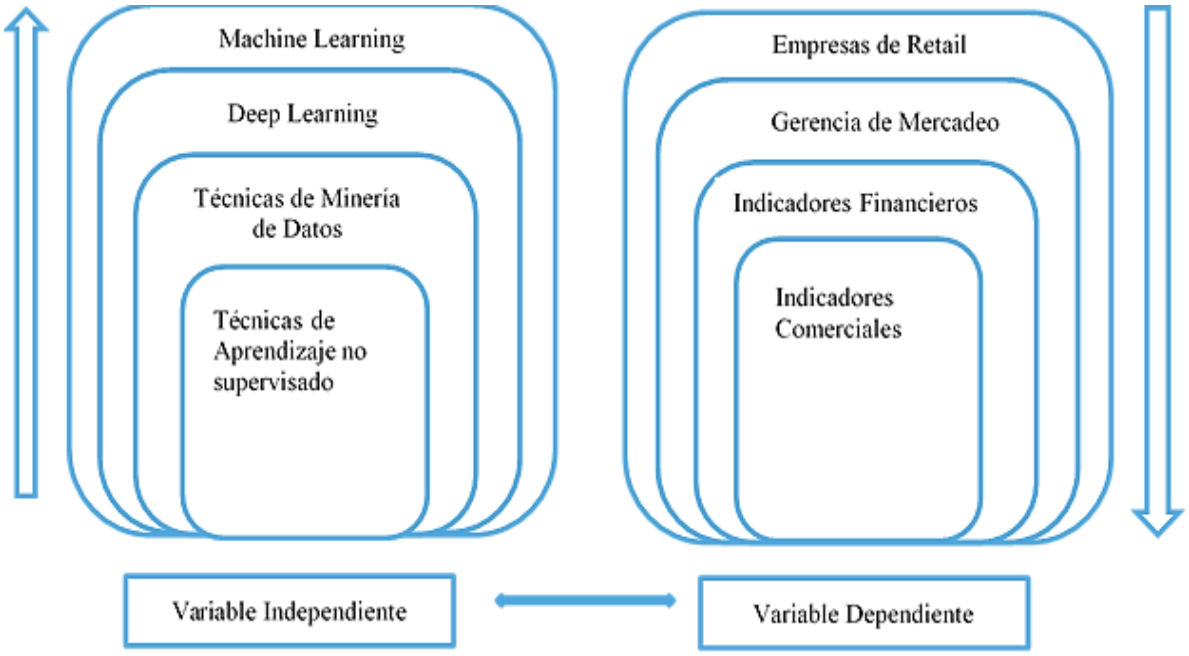


Figura 1 Relación de variables

2.1.1. Fundamentación de la variable Independiente

MACHINE LEARNING

Esta es una disciplina de la inteligencia de negocios cuyo término en español significa aprendizaje de máquina y se usa para referir a los algoritmos y técnicas que aprenden automáticamente a identificar patrones en los datos para predecir comportamientos futuros. Las

categorías principales de machine learning son: clasificación, descubrimiento de relaciones, regresión, clusterización, aprendizaje reforzado (Gallagher, Madden, & D'Arcy, 2015).

DEEP LEARNING

Es un subconjunto de machine learning que es usado para describir algoritmos y técnicas capaces de modelar conceptos abstractos de alto nivel utilizado principalmente en áreas de reconocimiento de imágenes y procesamiento de texto. Un ejemplo de estas técnicas son las redes neuronales las cuales son capaces de reconocer dígitos escritos a mano (Dean, 2014).

TÉCNICAS DE MINERÍA DE DATOS

La minería de datos hace referencia al proceso de extraer conocimiento nuevo y útil de los datos. El proceso de minería de datos permite encontrar anomalías, patrones y correlaciones dentro un conjunto de datos. Es utilizado para resolver varios problemas de negocios como perfilamiento de clientes, modelamiento del comportamiento de clientes, cálculo del historial crediticio, recomendación de productos entre otros (SAS, 2018).

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

Técnicas de aprendizaje supervisado. - En las técnicas de aprendizaje supervisado se tiene variables de entrada y salida. Lo que se hace es ajustar las variables de entrada para predecir las variables de salida.

Técnicas de aprendizaje no supervisado. - En las técnicas de aprendizaje no supervisado no tienen una variable de salida para predecir, solo se tiene variables de entrada. En vez de ajustar el modelo a las variables de entrada para predecir la variable de salida, estas técnicas buscan descubrir patrones dentro de los volúmenes de información. (Gorakala & Usuelli, 2015)

Fundamentación de la variable dependiente

EMPRESAS DE VENTAS AL POR MENOR

Según datos del INEC publicados en 2016, se registraron 843.745 empresas divididas en 19 actividades económicas como Comercio, Agricultura, Transporte, etc., según los últimos datos del Directorio de Empresas y Establecimientos 2016 publicado por el Instituto Nacional de Estadísticas y Censos (INEC).

De acuerdo con el tamaño, de las 843.745 empresas registradas en 2016, el 90,5% son microempresas, es decir con ventas anuales menores a 100 mil dólares y entre uno y nueve empleados; le sigue la pequeña empresa con el 7,5% y ventas anuales entre 100.001 a 1´000.000 entre 10 y 49 funcionarios.

La actividad económica con mayor número de empresas es el comercio al por mayor y por menor; reparación de vehículos automotores y motocicletas con 308.956, que representa el 36,6% del total de las empresas. En 2016, se registró ventas por 147.729 millones de dólares; 12.162 millones menos que el año anterior (INEC, 2017).

GERENCIA DE MERCADEO

El área de mercadeo de la empresa de ventas al por mayor y menor está a cargo de diseñar, coordinar, ejecutar e informar sobre las investigaciones de mercado, a la vez que están a cargo de mantener las relaciones con los proveedores y las jefaturas de mercadeo de mayoreo y minoreo, quienes a su vez son responsables de establecer las políticas y estrategias para las campañas de ventas ya sea en los locales o por medio de llamadas telefónicas, WhatsApp o página web.

INDICADORES FINANCIEROS

La empresa de ventas al por mayor y menor ha realizado un gran esfuerzo económico para consolidar sus indicadores financieros de tal manera que les permita tener información en tiempo real para la toma de decisiones sean estas para realizar inversiones, financiación en base a la evaluación de resultados que les permita realizar un análisis de su situación actual y proyectarse a futuro.

Los indicadores financieros utilizados son el retorno sobre la inversión, utilidad neta sobre la inversión, pasivo financiero, endeudamiento, gastos, ventas, flujo de caja, etc.

De estos indicadores para la investigación se usarán solo los indicadores pertenecientes a la parte comercial, de los cuales algunos sirven para determinar la efectividad de su fuerza de ventas.

INDICADORES COMERCIALES

Los indicadores comerciales permiten medir la rentabilidad de la empresa ya que provee de datos de ventas totales, ventas por región, ventas por agencia, margen total, margen comercial, plazos y gastos. Estos indicadores permiten realizar comparaciones entre los ingresos reales y los ingresos proyectados para determinar la situación actual de la empresa y tomar acciones correctivas a tiempo para la consecución de los presupuestos de ventas mensuales y anuales de la empresa. Otros indicadores son la comparación entre los gastos reales y el gasto presupuestado a fin de revisar si las áreas de la organización están llevando sus gastos según lo planificado y poder tomar medidas correctivas en caso de que el gasto supere lo presupuestado.

El nivel de granularidad de los indicadores de ventas de la empresa se lo puede reducir al punto que permite determinar la efectividad de una campaña específica de ventas, lo cual es útil ya que esto permite determinar si una determinada estrategia de ventas tuvo o no la efectividad esperada.

2.2.Antecedentes del estado del arte

Para el análisis del estado del arte se usaron las fases de criterios de inclusión y estrategia de búsqueda que son parte de un SMS[1], como fuentes de búsqueda de la información para la investigación se usaron los siguientes repositorios académicos Scopus, Springer, IEEExplore, ACM Digital Library.

2.2.1. Definición del objetivo

El objetivo del estudio del estado del arte está enfocado en resolver las preguntas de los objetivos específicos planteados en la sección anterior.

2.2.2. Definición de los criterios de inclusión y exclusión

Las búsquedas en las bases digitales retornan una gran cantidad de artículos dependiendo el tema, por lo cual es importante definir las características idóneas de los artículos a ser tomados en cuenta para el presente análisis, tomando en cuenta los siguientes criterios:

Criterios de inclusión

- Con el fin de analizar metodologías utilizadas en la actualidad, fueron tomados en cuenta artículos a partir del 2010.
- Se tomaron en cuenta solamente artículos científicos y documentos de conferencias publicados en el idioma inglés.

---

[1] Systematic Mapping Study (SMS): El SMS permite realizar un análisis de la literatura existente sobre un determinado tema a fin de identificar el estado del arte de un tema determinado.

- Que el artículo contenga información referente al uso de metodologías de inteligencia de negocios para sistemas de recomendaciones.
- En su mayoría se tomaron en cuenta artículos científicos y documentos de conferencias.

Criterios de exclusión

- Artículos que tengan temas de inteligencia de negocios no relacionados con sistemas de recomendaciones.
- Artículos que no estén en el idioma inglés.

2.2.3. Definición de la estrategia de búsqueda

Revisión Inicial: Se realiza una búsqueda inicial en los distintos repositorios académicos para buscar estudios relacionados con las preguntas de investigación.

Validación cruzada de estudios: En esta fase se procede a verificar que los estudios cumplan con los criterios de inclusión y exclusión, con lo cual finalmente se obtienen el listado inicial de documentos académicos con los cuales se va a trabajar en las siguientes fases del estudio.

Integración del Grupo de Control: El grupo de control está conformado por los estudios que cumplen con los criterios de inclusión y exclusión para lo cual se procede a realizar un análisis inicial del título de los estudios, introducción, conclusiones y palabras claves. Los estudios seleccionados para el grupo de control son los siguientes:

Tabla 1
Estudios por grupo de control

Grupo Control	Titulo	Palabras Clave
EC1	Adaptive KNN based Recommender System through Mining of User Preferences	Recommender system _ Personalization, Adaptive K- nearest neighbor, Ontology, Web mining, Clustering.
EC2	Recommendation system based on product purchase analysis	Dynamic networks, Viral marketing, Social Networks, Recommendation system, Amazon co-purchase Network, Market dynamics, Review trend analysis, Market-basket analysis.
EC3	A Predictive Approach for improving the sales of products in E-commerce	Data mining, e-commerce, online shopping, recommender systems, recommendation techniques, web mining.
EC4	Unsupervised Sparse Matrix Co-clustering for Marketing and Sales Intelligence	Automated Matrix Co-Clustering, Gaussian-based density estimator, product recommendations, targeted sales and marketing.
EC5	Collaborative Filtering and Deep Learning Based Recommendation System For Cold Start Items	Recommendation System, Data Mining, Deep Learning Neural Network, Collaborative Filtering, Cold Start Problem
EC6	Recommender systems using cluster-indexing collaborative filtering and social data analytics	Data mining; business analytics; social network; recommender system; cluster-indexing collaborative filtering

2.2.4. Construcción de la cadena de búsqueda

Para la construcción de la cadena de búsqueda se usan las palabras que más se repiten en cada contexto definido a partir de los estudios del grupo de control, para el presente estudio se definieron los siguientes contextos: Algoritmos de inteligencia de negocios, tipo de minería de datos, entorno de análisis.

Tabla 2
Construcción de la cadena de búsqueda

Contexto	Palabra Clave	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	Número de Repeticiones
Algoritmos de Inteligencia de Negocios	Neural Network	x				x		2
	Collaborative Filtering	x		x		x	x	4
	Clustering	x	x	x	x	x	x	6
	recommendation system	x	x	x	x	x	x	6
	K-nearest neighbor	x	x	x				3
Tipo de Minería de Datos	Web Mining	x	x	x		x	x	5
	Data Mining	x	x	x	x	x	x	6
	social Mining	x	x		x	x	x	5
Entorno de Análisis	Market Basket Analysis		x	x				2
	e-commerce	x	x	x				3
	online shopping			x		x		2

La cadena de búsqueda está formada por la unión de las palabras claves que más se repiten en cada contexto, los conectores usados son OR para las palabras que están dentro del mismo contexto y el conector AND para las palabras que están en contextos distintos, de esta manera se establece la siguiente cadena de búsqueda.

((recommendation system) OR (Clustering)) AND (data mining) AND (e-commerce)

Al aplicar la cadena de búsqueda en los repositorios académicos seleccionados para el estudio se obtienen los siguientes resultados, para esto se procede a filtrar solo los documentos que estén

en el idioma inglés, cuya fecha de publicación sea mayor al 2010 y que sean artículos científicos o documentos de conferencias.

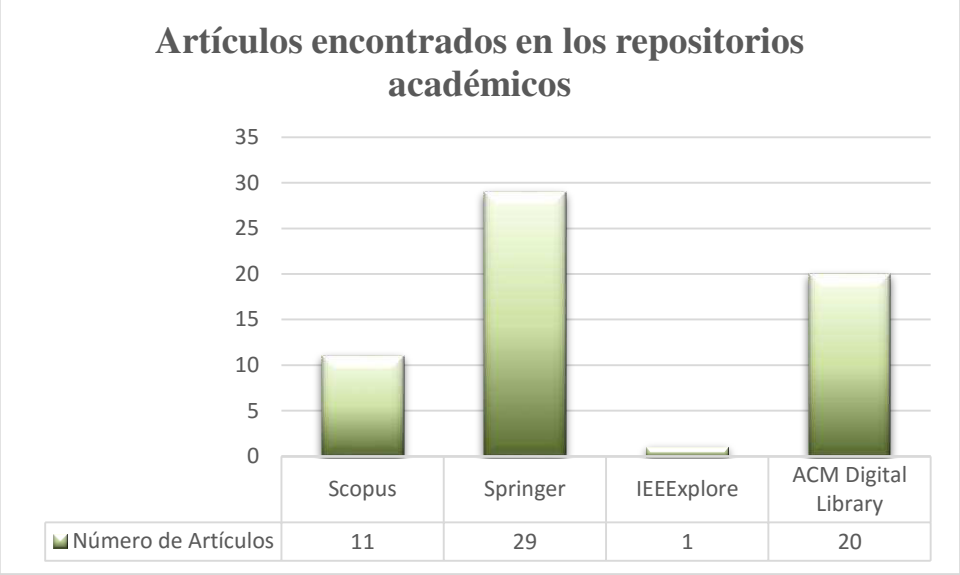


Figura 2 Artículos encontrados en los repositorios académicos.

Una vez obtenidos los resultados se realizó la revisión de los documentos encontrados los cuales se listan a continuación:

Analyzing cloud based reviews for product ranking using feature based clustering algorithm (Gobi, 2018)

En este documento los autores identifican como método común de recomendación el usar las evaluaciones realizadas por otros clientes sobre los productos identificando sus características principales, pero se dan cuenta que la mayoría de los sistemas de recomendaciones identifican solo las características mencionadas explícitamente sin tomar en cuenta las características implícitas ni las del contexto del revisor. Por este motivo, proponen una técnica que toma en cuenta las características implícitas, así como el contexto del revisor por medio de técnicas de clusterización. Esta técnica es puesta en práctica tomando como referencia información de la nube de Amazon sobre revisiones de celulares y cámaras.

A Predictive Approach for improving the sales of products in E-commerce (Usmani, 2017)

Los autores proponen un acercamiento predictivo para mejorar las ventas en el comercio electrónico, demuestran varios problemas que afectan a las técnicas de los sistemas de recomendaciones como la sobrecarga de información al usuario lo cual reduce la satisfacción del cliente con la tienda y su interés, proponen soluciones para resolver estos inconvenientes mediante el uso de técnicas como reglas de clasificación, Filtrado Colaborativo, minería por reglas de asociación and minería por reglas de secuencia.

Competent K-means for Smart and Effective E-Commerce (Akash Gujarathi, 2018)

Este estudio realiza una comparativa del algoritmo de cauterización K-means versus otros algoritmos tradicionales de clusterizacion. Los autores identifican las ventajas y desventajas de las implementaciones tradicionales de K-means y para las desventajas proponen mejoras para incrementar la efectividad y calidad del clúster, finalmente generan un pool colaborativo único basado en el algoritmo K-means modificando la técnica de similaridad de cosenos para determinar la similitud entre los objetos que pertenecen al clúster. Finalmente realizan un análisis matemático para probar que este algoritmo de cauterización mejorado tiene mejor desempeño que el algoritmo K-means tradicional.

Recommendation system based on product purchase analysis (Mitra, Ghosh, Basuchowdhuri, Manoj, & Sanjoy, 2016)

Proponen un sistema de recomendaciones basado en el análisis de la relación de los productos adquiridos con el fin de encontrar tendencias de compras, manifiestan que al analizar las compras de los clientes se revela tendencias de compra y su dependencia con futuras compras, lo cual puede permitir a las organizaciones definir una estrategia para incrementar las ventas.

Similar product clustering for long-tail cross-sell recommendations (Grozin & Alla, 2017)

En este documento los autores denotan que una de las razones principales para el rápido crecimiento del comercio electrónico es la habilidad que tienen las tiendas de proveer una gran variedad de productos ya que no están limitadas como las tiendas físicas al tamaño de la bodega. Sin embargo, las tiendas en línea tienen productos que generalmente tienen pocas compras comparados al resto de productos, a estos productos se los conoce como productos de cola larga "long-tail".

El uso de sistemas de recomendaciones para estos productos es complicado ya que no tienen tanta información como el resto de productos más vendidos, es así que los autores consideran el aplicar técnicas de clusterización para atacar este problema utilizando las visitas a los productos, datos del contenido de los productos como el árbol de categoría del producto y sus nombres para generar un sistema de recomendaciones que mejora la efectividad de la venta cruzada de productos permitiendo a las empresas recomendar con efectividad productos de cola larga.

Unsupervised Sparse Matrix Co-clustering for Marketing and Sales Intelligence (Zouzias, Vlachos, & Freris, 2012)

Los autores en su análisis proponen una metodología no supervisada mediante el uso de "Sparse Matrix Co-clustering" para identificar las relaciones entre los clientes y los productos a fin de realizar recomendaciones personalizadas, en este análisis los autores demuestran cómo al agrupar explícitamente el "co-clustering" con las recomendaciones de productos, utilizando datos de inteligencia de negocios del mundo real mejoran las recomendaciones. Al final obtienen un algoritmo robusto capaz de descubrir automáticamente estructuras del clúster superpuestas y disyuntivas incluso en medio de observaciones ruidosas.

A model for predicting outfit sales: Using data mining methods (Ullah, 2019)

En este estudio se desarrolla un modelo que utilizará técnicas de extracción de datos para predecir la suma de las ventas de equipos. Aquí se adoptan procedimientos de extracción de datos para luego entrenarlos en los diferentes clasificadores. Finalmente, se evaluó el rendimiento de cada clasificador mediante una validación cruzada de diez veces y se comparó los resultados, donde, el clasificador de percepción multicapa mostró el rendimiento más alto con 83.8% de precisión. Esta investigación también mostró el nivel estándar de mejora en el rendimiento (hasta un 9%), en comparación con los otros modelos previstos de ventas de equipos.

Product recommendation for E-commerce data using association rule and apriori algorithm (Bandyopadhyay, Thakur, & Mandal, 2019)

Este trabajo hace énfasis en el problema que existe en los sitios de comercio electrónico donde se proporcionan a los clientes una gran cantidad de información sobre los productos. Como resultado, estos clientes se confunden con la sobrecarga de información y les resulta difícil encontrar artículos satisfactorios. Estos sistemas de comercio electrónico no pueden ofrecer recomendaciones individuales como lo hace un vendedor y es por eso que los clientes no pueden elegir los productos requeridos según sus requisitos. Este trabajo analiza la recomendación basada en el contenido para el sitio de comercio electrónico, donde la minería de reglas de asociación y el algoritmo Apriori se utilizan para la predicción y recomendación de productos. El análisis de resultados muestra un mejor rendimiento del sistema en comparación con otras técnicas.

Research methodology for analysis of E-commerce user activity based on user interest using web usage mining (Diwandari, Permanasari, & Hidayah, 2018)

En este documento, los autores proponen un método para encontrar el interés de los usuarios en los productos ofrecidos en los sitios web de comercio electrónico basados en la minería del uso

de la web de los datos del flujo de clics. En este estudio, se investigó el interés del usuario utilizando técnicas de agrupación y clasificación. Los resultados experimentales mostraron que el método puede ayudar a analizar el comportamiento de los visitantes y el interés de los usuarios en los productos de comercio electrónico mediante la identificación de aquellos productos que despiertan el interés de los visitantes.

Evaluating discounts as a dimension of customer behavior analysis (Haghighatnia, Abdolvand, & Rajaee Harandi, 2018)

Este estudio está enfocado en buscar una mejor comprensión del comportamiento de los clientes mediante la identificación de clientes valiosos. Por lo tanto, este estudio utiliza un modelo basado en RFM denominado RdFdMd, en el que d es el nivel de descuento utilizado para analizar el comportamiento de compra del cliente y la importancia de los descuentos y la rentabilidad del negocio. Se utilizaron los algoritmos CRISP-DM y k-mean para la agrupación. Los resultados indican que el uso de RdFdMdmodel logra una mejor agrupación y valoración de clientes, y los descuentos se identificaron como un criterio importante para las compras de los clientes.

Analyzing customer behavior from shopping path data using operation edit distance (Syaekhoni, Lee, & Kwon, 2018)

Este estudio, analiza el comportamiento de los clientes a partir de sus datos de ruta de compra mediante el uso de un algoritmo de agrupación. Se propone una nueva medida de distancia para los datos de la ruta de compra, llamada la distancia de edición de la Operación y un método que permite que los datos de la ruta de compra de un cliente RFID se procesen de manera efectiva utilizando algoritmos de agrupamiento. Se recopilaron datos de una ruta de compras del mundo real de una tienda minorista y se aplicó el método al conjunto de datos. El método propuesto determinó efectivamente los patrones de compra de los clientes a partir de los datos.

A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization (Bhade, Gulalkari, Harwani, & Dhage, 2018)

Este estudio propone un enfoque sistemático para dirigirse a los clientes y proporcionar el máximo beneficio a las organizaciones. Un paso inicial importante fue analizar los datos de ventas adquiridos del historial de compras y determinar los parámetros que tienen la máxima correlación. Sobre la base de los grupos respectivos, los recursos adecuados pueden canalizarse hacia clientes rentables mediante algoritmos de aprendizaje automático. K-Means clustering se usa para la segmentación de clientes y Singular Value Decomposition se usa para proporcionar recomendaciones apropiadas a los clientes. Este documento también aborda los inconvenientes del sistema de recomendación, como el problema del arranque en frío, la escasez, etc. y la forma en que se pueden superar.

Mining consumer knowledge from shopping experience: TV shopping industry (Wen, 2018)

Este estudio utiliza el análisis de clústeres para identificar los perfiles de los consumidores de compras de TV. Al representar el mapa de conocimiento de marketing de los portavoces, se encuentra la mejor cartera de respaldo para hacer recomendaciones. Mediante el análisis de los voceros, el período, los perfiles de los clientes y los productos, se proponen cuatro modos de compra de televisión para los consumidores: nuevo producto, conocimiento, bajo precio y producto de lujo. Las recomendaciones relacionadas también se proporcionan para la referencia de la industria.

Conclusión

Al realizar la revisión de literatura se identifica que la manera como las empresas de comercio electrónico mejoran la efectividad de sus fuerzas de ventas es por medio de sistemas de

recomendaciones. Es común encontrar dentro de estos sistemas de recomendaciones algoritmos de clusterización que permiten segmentar el mercado tomando en cuenta las características explicitas e implícitas de los productos, así como el contexto de los clientes. Es por este motivo que para solventar el problema de las empresas de ventas al por mayor y menor en el Ecuador se propone la implementación de un algoritmo de inteligencia de negocios que analice la información histórica de los clientes para realizar recomendaciones de productos tomando en cuenta las características de los productos y usuarios con la finalidad de mejorar la efectividad de la fuerza de ventas de la empresa.

CAPÍTULO III

MEMORIA TÉCNICA METODOLÓGICA

3.1. Metodología de Investigación

La presente investigación está orientada por el marco teórico de procesos para SME propuesto por Dittert, Härting, Reichstein y Bayer en el 2018, esta es una metodología que combina lo mejor de metodologías como KDD-Process, SEMMA, CRISP-DM para la analítica de datos diseñada para pequeñas y medianas empresas (Dittert, Härting, Reichstein, & Bayer, 2017). Esta metodología compuesta por 6 etapas:

3.1.1. Definición de la tarea

La definición de la tarea debe ser realizada en conjunto con el área gerencial de la organización. Esto puede ser realizado internamente basado en grupos de trabajos o se puede contratar especialistas. Para esto se debe definir las necesidades de la organización para enlazarlas a las diferentes tareas de minería de datos como: clasificación, clusterización, análisis asociativo, aprendizaje estadístico, minería de datos. Para el presente estudio la tarea es la creación de un sistema de recomendaciones.

3.1.2. Recolección y análisis de datos

En esta etapa se analiza las fuentes de datos con las que cuenta la organización y que puedan ser útiles para la tarea a realizar, la mayoría de las pymes por lo general no tienen bien estructurada su información por lo que es necesario un análisis previo con gráficos de dispersión que permitan ver la distribución de los datos para identificar valores atípicos que vayan a afectar el resultado del estudio.

3.1.3. Selección y configuración del modelo

Dependiendo de la tarea de BI a ejecutar se va a contar con varias opciones de técnicas para su implementación, en este caso para sistemas de recomendaciones tenemos: clasificación,"Collaborative Filtering" (CF), reglas de asociación, clusterización, redes neuronales, arboles de decisión (Usmani et al., 2017). Para la selección del modelo se debe revisar que algoritmos se ajustan a las necesidades de la organización por medio de una serie de pruebas y errores.

3.1.4. Limpieza de datos

Esta tarea consiste en verificar que los tipos de datos a utilizar para la implementación del modelo sean adecuados ya que existen modelos como los arboles de decisión que pueden trabajar con casi cualquier tipo de dato, otros modelos como las redes neuronales solo aceptan datos numéricos. En caso de que la data no se ajuste automáticamente se debe buscar la manera de convertir los datos para que se ajusten al modelo a implementar.

3.1.5. Análisis de resultados

La evaluación de resultados va a depender del modelo a implementar para métodos como los de clasificación, redes neuronales se puede dividir la información en dos partes, la primera comprende la mayor parte de la información usualmente setenta por ciento para el aprendizaje y el treinta por ciento restantes se lo usa para la evaluación de resultados. Una vez obtenidos los resultados estos deben ser analizados juntamente con expertos de la organización quienes van a determinar si los resultados son óptimos.

3.1.6. Reporte a los tomadores de decisión

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

A continuación, se presenta el flujo de trabajo de la metodología a ser aplicada para el presente estudio basada en el marco de procesos para SME.

Figura 3 Marco de Procesos para SME

3.2. Ejecución del proceso de investigación

3.2.1. Definición de la tarea

En esta fase, se definieron los requerimientos del negocio, en la que mediante un análisis a personas que conforman y conocen el negocio, en este caso la experiencia propia, se definieron como objetivos del proyecto en primera instancia, obtener los datos de los clientes y los productos que han adquirido, en conjunto con todos los datos relacionados con los mismos. A continuación, se definió el objetivo final del proyecto que es proporcionar al negocio un sistema de recomendación de productos, tomando como caso de estudio una empresa de retail en el Ecuador

3.2.2. Recolección y análisis de datos

Con los datos recolectados en la etapa anterior, se procedió a analizar los mismos, con el propósito de encontrar aquella información mayormente relacionada con el objetivo final de este proyecto.

Es así, que mediante esta etapa se analizó las fuentes de datos con las que cuenta la organización, analizándola con gráficos de dispersión para observar la distribución de los datos e identificar valores atípicos que vayan a afectar el resultado del estudio. Dichos datos fueron cargados desde una base de datos Oracle a la herramienta RapidMiner para a continuación, proceder al análisis de valores atípicos y posterior creación del modelo de minería de datos.

3.2.3. Selección y configuración del modelo

En esta etapa de selección y configuración del modelo se utilizó Auto Model de RapidMiner, mediante el cual, se pudo realizar visualizar, las características de los datos obtenidos, el comportamiento de los atributos más importantes y sobre todo permitió seleccionar el modelo con

mejor confianza y adecuada a la data analizada. Además, para la selección del modelo se revisó que algoritmos se ajustan a las necesidades de la organización por medio de una serie de pruebas y errores.

3.2.4. Limpieza de datos

Luego de seleccionar el modelo de minería de datos a utilizar, se procedió a realizar la limpieza de la data seleccionada, esto utilizando las bondades de RapidMiner, que permite abrir un proceso del resultado obtenido al ejecutar Auto Model con el algoritmo seleccionado.

En esta fase se verificó que los tipos de datos a utilizar para la implementación del modelo sean adecuados, para que pueden trabajar de una manera confiable, ajustándose al modelo implementado.

3.2.5. Análisis de resultados

Ya obtenido el modelo y realizadas las respectivas transformaciones, para tener una data adecuada, se procedió a realizar la evaluación de resultados, dividiendo la información en dos partes, la primera comprendida por la mayor parte de la información en este caso el setenta por ciento, para el aprendizaje y el treinta por ciento restantes se lo usó para la evaluación de resultados. Dividida la data se realizó la evaluación utilizando una matriz de confusión que permitió relacionar el número de datos clasificados correctamente y el número de datos clasificados incorrectamente, determinando así, el nivel de confianza del modelo propuesto. Una vez obtenidos los resultados estos se analizaron juntamente con expertos de la organización quienes determinaron que los resultados son óptimos.

3.2.6. Reporte a los tomadores de decisión

Una vez obtenidos y validados los resultados estos se enlazaron con la herramienta Tableau, que permitió realizar un conjunto de reportes, disponibles, para ser presentados a los dirigentes de la organización, como resultado de aplicar minería de datos, y ofreciendo un sistema de recomendaciones de productos.

RESUMEN DEL CAPITULO III

En este capítulo se habla sobre la metodología utilizada y la manera como se ejecutó la misma. La presente investigación está orientada por la metodología propuesta por Dittert, Härting, Reichstein y Bayer en el 2018 compuesta por 6 etapas, en la primera se define la tarea de acuerdo a las necesidades de la organización para enlazarlas a las diferentes tareas de minería de datos. En la segunda etapa se recolecta y analiza los datos basándose en su utilidad en el proyecto a desarrollar. A continuación, se selecciona y configura el modelo, buscando algoritmos de minería de datos que se ajusten a las necesidades de la organización por medio de una serie de pruebas y errores. Luego se realiza la limpieza de datos con el propósito de buscar la manera de convertir los datos, para que se ajusten al modelo a implementar. Una vez obtenidos los resultados estos se analizan juntamente con expertos de la organización quienes determinan si los resultados son óptimos para finalmente pasar a la etapa de reportes para la toma de decisiones para compartir los resultados con las áreas gerenciales de la organización.

En la parte de ejecución de cada una de las etapas, se comenzó definiendo la tarea mediante un análisis a personas que conforman y conocen el negocio, para así determinar los requerimientos de éste. En la fase de recolección y análisis de datos, se analizó las fuentes de datos con las que cuenta la organización, mediante gráficos de dispersión para observar la distribución de los datos e identificar valores atípicos que vayan a afectar el resultado del estudio. Para la selección y configuración del modelo se pudo visualizar, las características de los datos obtenidos, el comportamiento de los atributos más importantes y seleccionar el modelo con mejor confianza y adecuada a la data analizada. Luego se procedió a realizar la limpieza de la data seleccionada, verificando que los tipos de datos a utilizar para la implementación del modelo sean adecuados y confiables. El análisis de resultados se realizó utilizando una matriz de confusión que permitió relacionar el número de datos clasificados correctamente y el número de datos clasificados incorrectamente, determinando así, el nivel de confianza del modelo propuesto. Una vez obtenidos y validados los resultados se realizaron un conjunto de reportes, para ser presentados a los dirigentes de la organización, como resultado de aplicar minería de datos, y ofreciendo un sistema de recomendaciones de productos.

CAPÍTULO IV

RESULTADOS

4.1 Informe de Resultados

4.1.1. Definición de la tarea

Las empresas de retail son aquellas que venden productos terminados a los consumidores a cambio de dinero. Pero con el acelerado mercado actual, es importante mantenerse relevante para el mercado y relacionarse con sus clientes. Los sistemas de recomendación ayudan a retener a los clientes, al sugerir productos en base a sus necesidades. Esto ayuda a la empresa en el aumento ventas y también pueden ayudarlo a crear lealtad de marca a través de la personalización relevante.

Este proyecto utiliza los registros del historial ventas de productos al cliente, perfil del cliente, dirección y datos relacionados.

Por cuanto, la tarea principal realizada fue que mediante la información recolectada se pudo predecir qué artículo es más probable que un cliente desee de acuerdo a sus características como edad, salario, dirección, entre otros. Esta predicción, es la recomendación que los vendedores de productos darán, basada en históricos de ventas de los productos a otros clientes con características similares.

Determinada la tarea principal, se establecieron como tareas secundarias:

- Análisis de los datos de la organización caso de estudio mediante el uso de diagramas de dispersión, identificando valores atípicos que distorsionaban los resultados.
- Selección y configuración del modelo de inteligencia de negocios que mejor se ajustó a las necesidades de la empresa.

- Formateo de los datos para que trabaje adecuadamente con el modelo seleccionado, eliminando datos nulos, discretizando los datos de ser necesarios, garantizando el correcto funcionamiento del modelo de inteligencia de negocios implementado.
- Evaluación de los resultados mediante el uso de técnicas de inteligencia de negocios, determinando la probabilidad de que el modelo de inteligencia de negocios utilizado ayude a mejorar la efectividad de la fuerza de ventas.

4.1.2. Recolección y análisis de datos

Se extrajo la información de las bases de los datos de la organización caso de estudio, almacenada en una base de datos Oracle, para facilidad de manejo de esta, se la almacenó en el repositorio de la herramienta RapidMiner para proceder con su análisis.

Desde la empresa se puedo obtener datos históricos de ventas realizadas anteriormente por la empresa, donde se tuvo una tabla con los siguientes atributos:

Tabla 3
Atributos para análisis de datos

	ATRIBUTO	DESCRIPCIÓN	VALORES
CLIENTE	EDAD ECONOMICA	Descripción de la capacidad económica del cliente.	ECOEST ECOINI JOVEN MAYOR
	DES_SEXO	Sexo del Cliente	MASCULINO FEMENINO
	DESC_NIV_EST	Nivel de estudios del Cliente.	BACHILLER BASICA SUPERIOR SECUNDARIA PRIMARIA NINGUNA INICIAL ELEMENTAL

Continúa

	DES_PROFESION	Profesión del Cliente	131 PROFESIONES
	SALARIO	Salario del Cliente	BAJO MEDIO_BAJO BASICO MEDIO ALTO MEDIO_ALTO
	TIPO_TRABAJO	Tipo de trabajo del Cliente	Relación de dependencia
	NRO_CARGAS	Número de cargas del Cliente	
	CLI_TIPO_VIVIENDA	Tipo de Vivienda del Cliente	FAM ARR PRO ANT HIP HER
	CANTON_DOM	Cantón del domicilio del Cliente	
	PROVINCIA_DOM	Provincia del domicilio del Cliente	59 cantones
	ANIOS_ANTIGUEDAD	Años de antigüedad del Cliente en el empleo actual	
	PER_NRO_IDENTIFICACION	Número de identificación del Cliente	
	TIPO_CLIENTE	Clasificación del Cliente de acuerdo a las compras realizadas.	RECURRENTE PREFERENTE RECURRENTE NORMAL BANCARIZADO A CLIENTE NUEVO BANCARIZADO B INCONSISTENCIA SIN DEFINIR
	TIPO_IDENTIFICACION	Tipo de identificación del Cliente	CED
VENTA PRODUCTO	VENTA_NETA	Datos de venta del producto	
	CANTIDAD		
	VENTA_F		
	TOTAL_FACTURA		

Continúa

PRODUCTO	PRECIO_DE_LISTA		
	DESCRIPCION_POLITICA	Políticas de venta de la empresa	POLITICA GENERAL MINOREO POLITICA EMPLEADOS MINOREO POLITICA GENERAL SALDOS BODEGA MIN CLIENTES ESTRELLA POL MIN CONTADO 30 DIAS POLITICA ACTIVACIONES DE AGENCIA
	FORMA_DE_VENTA	Forma de venta del producto	Crédito propio
	ID_PRODUCTO_1	Identificador del producto	
	CALIDAD_PRODUCTO	Calidad del producto	A B B RETIRADO
	DESCRIPCION_PRODUCTO	Descripción del producto	172 productos
	MODELO	Modelo del producto	
	LINEA_PRODUCTO	Línea del producto	LINEA AUDIO Y VIDEO LINEA ELECTRODOMESTICOS LINEA TECNOLOGIA SIN DEFINIR LINEA MUEBLES Y HOGAR PROMOCIONES ELECTRODOMESTICOS MENORES LINEA MOVILIDAD LINEA AGRICOLA Y FERRETERIA
	CATEGORIA	Categoría del producto	69 Categorías
	SUBCATEGORIA	Sub Categorías del Producto	140 Subcategorías
	MARCA	Marca del producto	46 Marcas

De acuerdo con los datos obtenidos y a los requerimientos de la empresa se definió como atributo a predecir para el sistema de recomendaciones al atributo DESCRIPCION_PRODUCTO, puesto que este permitirá definir cuál es el producto que el vendedor debe recomendar a los clientes.

Además, en base a los datos recolectados de la organización caso de estudio, se procedió a realizar un análisis de correlación con el atributo a predecir, para seleccionar los datos que mejores aportes darán al modelo de minería de datos.

Con la ayuda de la herramienta RapidMiner se procedió a relacionar los atributos concernientes a los clientes con el atributo seleccionado como label (DESCRIPCION_PRODUCTO).

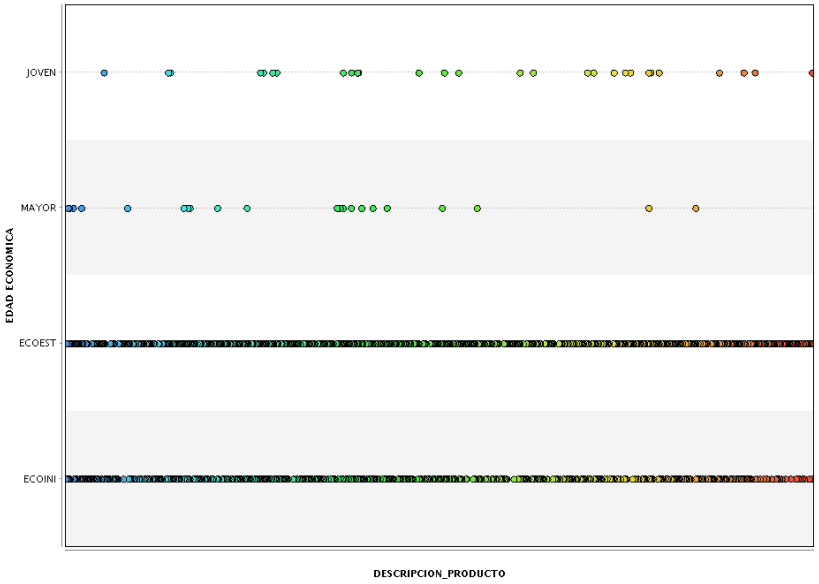


Figura 4 Edad económica vs descripción del producto

Resultado que indica que el grupo de productos ofrecidos en la empresa están más orientados a los clientes que tienen economía estable y que se están iniciando económicamente.

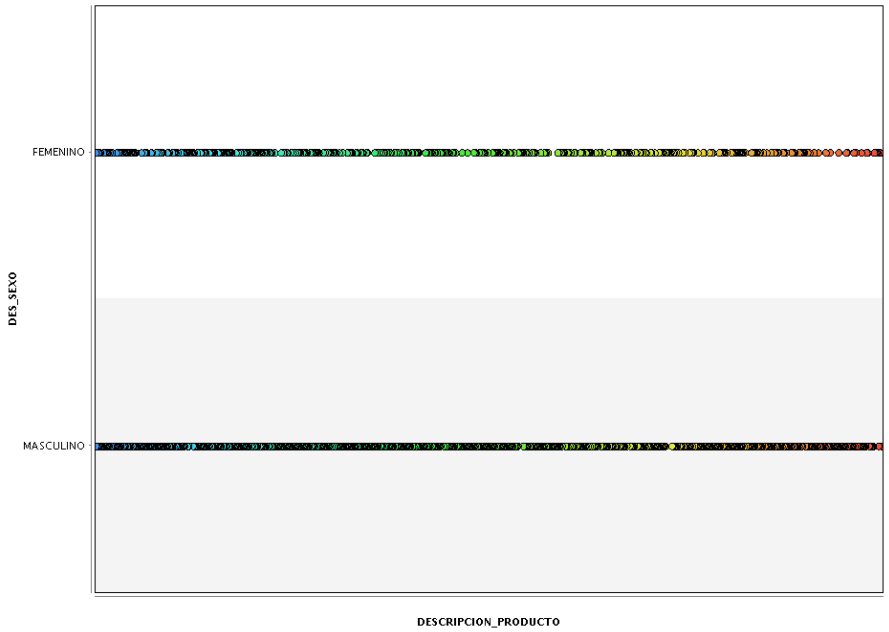


Figura 5 Sexo cliente vs descripción del producto

También, se puede observar que el género del cliente no afecta, en el nivel de compras de los productos.

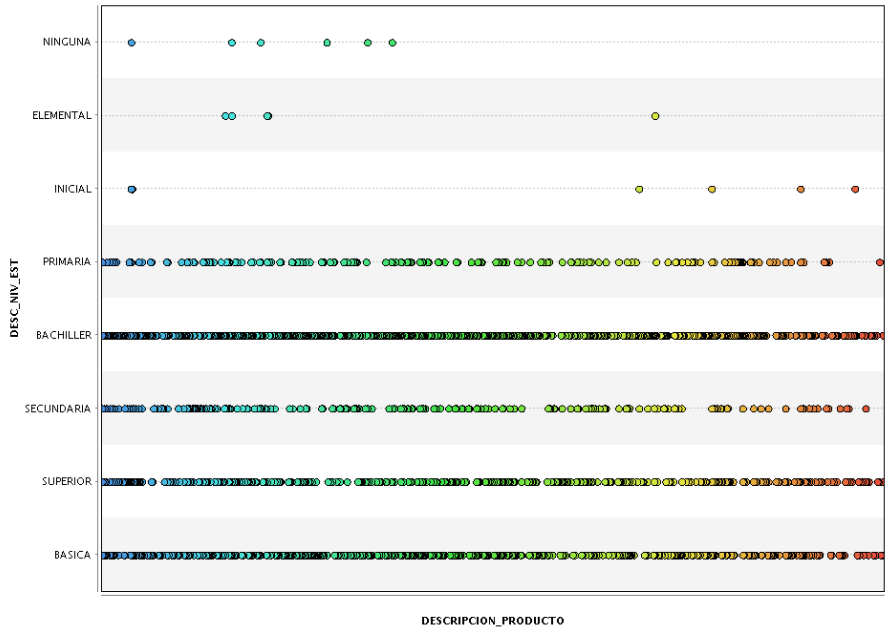


Figura 6 Nivel de estudios Cliente vs descripción del producto

Mayormente, el porcentaje de clientes que adquieren los productos de la empresa tienen un nivel de estudio mayor a primaria.

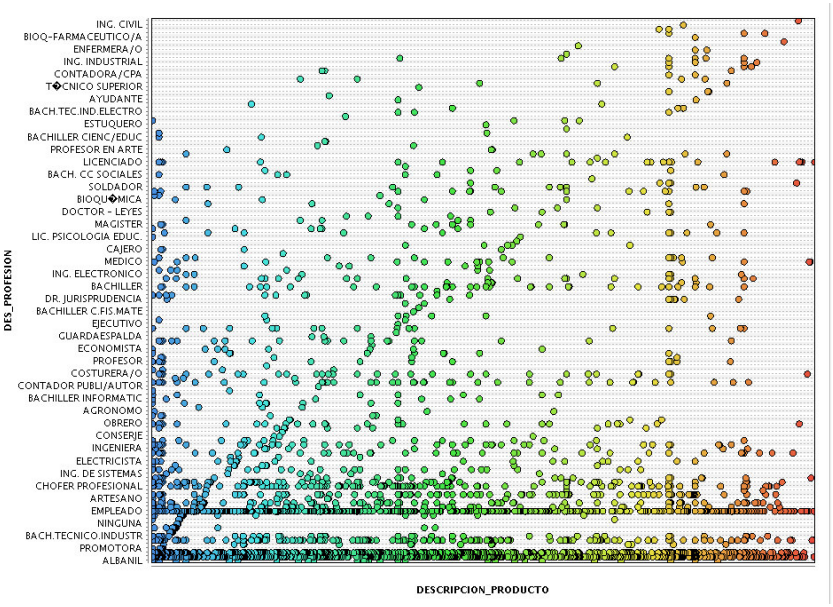


Figura 7 Profesión vs descripción del producto

Se requirió una normalización en los datos de profesión de los clientes para obtener mejores resultados.

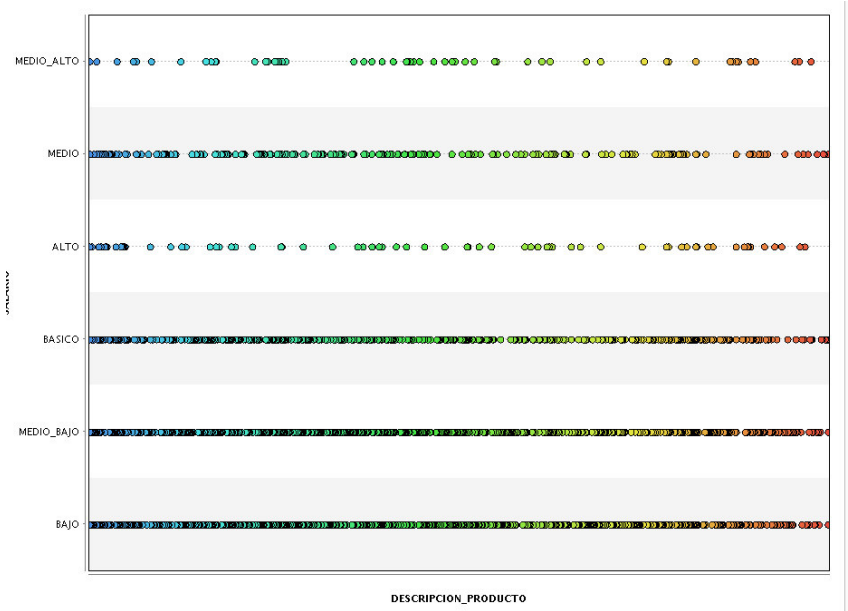


Figura 8 Ingresos Cliente vs descripción del producto

La venta de productos está mayormente concentrada en los clientes que tienen ingresos básicos, medio bajo y bajo.

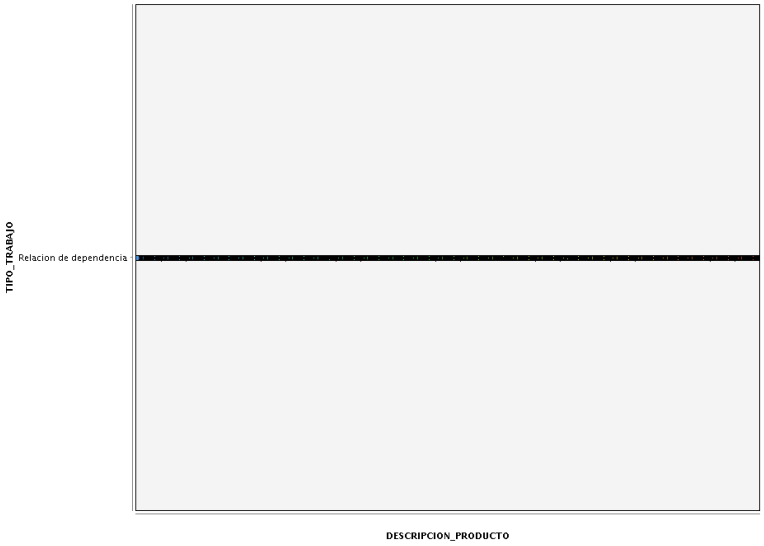


Figura 9 Tipo trabajo cliente vs descripción del producto

El tipo de trabajo del cliente no aporta al modelo a crear puesto que únicamente se tiene un solo valor.

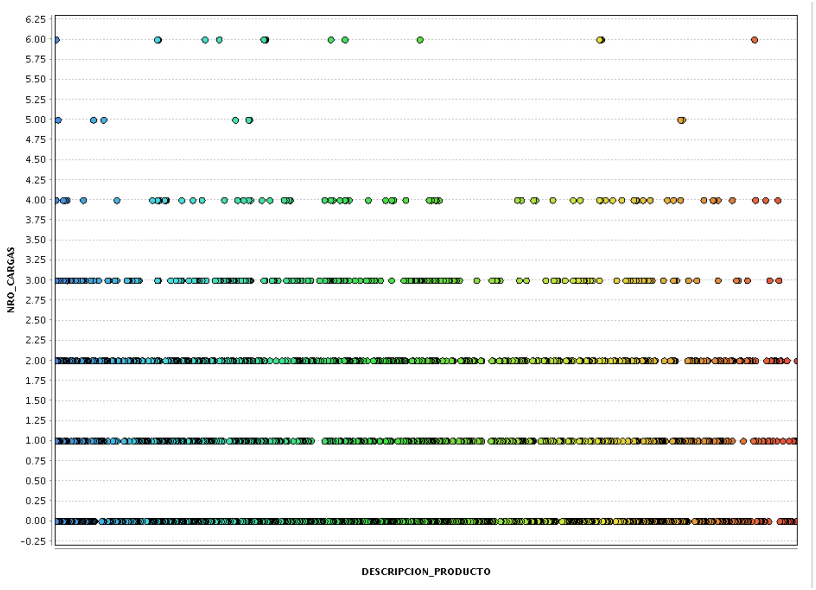


Figura 10 Número cargas cliente vs descripción del producto

Entre menos cargas tenga el cliente, mayor es su porcentaje de comprar un producto de la empresa caso de estudio.

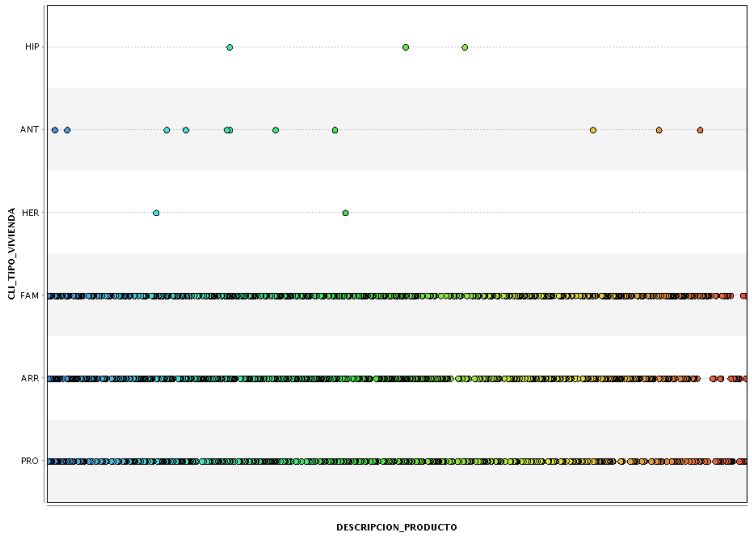


Figura 11 Tipo vivienda Cliente vs descripción del producto

El tipo de vivienda de los clientes que más compras han realizado en la empresa caso de estudio es de un familiar, arrendada y propia. Los demás tipos de vivienda se podrían omitir.

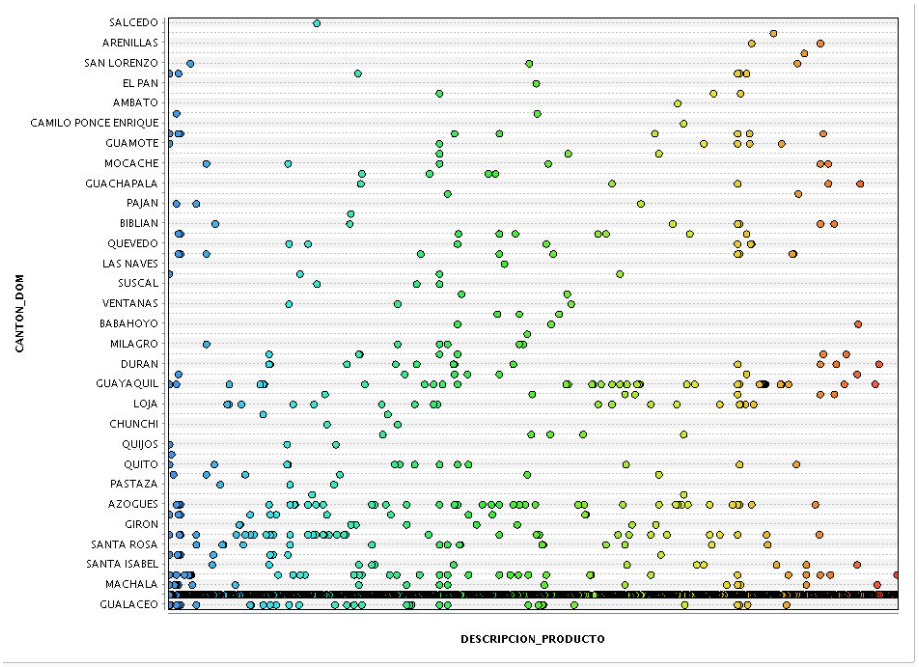


Figura 12 Cantón domicilio Cliente vs descripción del producto

Es necesario normalizar los datos del cantón de donde proviene el cliente, puesto la correlación que se tiene es pequeña por demasiados datos dispersos.

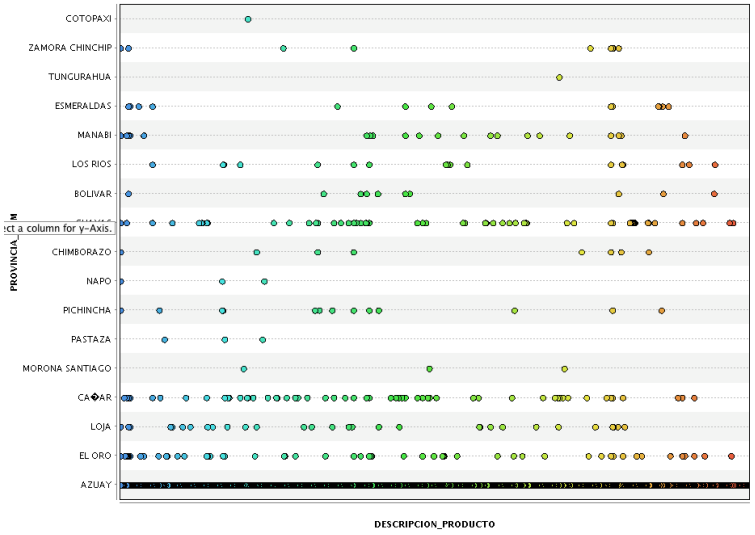


Figura 13 Provincia domicilio cliente vs descripción del producto

De igual manera que con los datos del cantón, es necesario normalizar los datos de provincia de donde proviene el cliente.

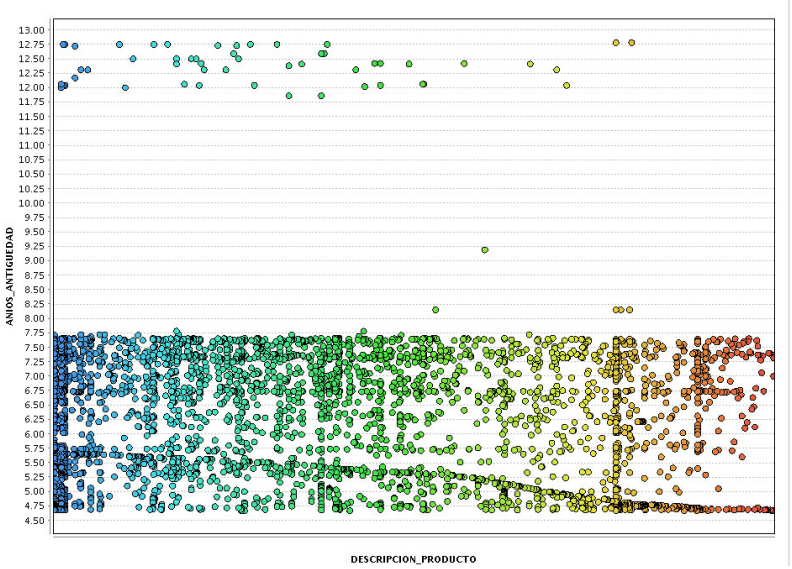


Figura 14 Años antigüedad cliente vs descripción del producto

Se registran datos atípicos de clientes que tienen más de 8 años de antigüedad en sus empleos, puesto que se registra poco interés en comprar.

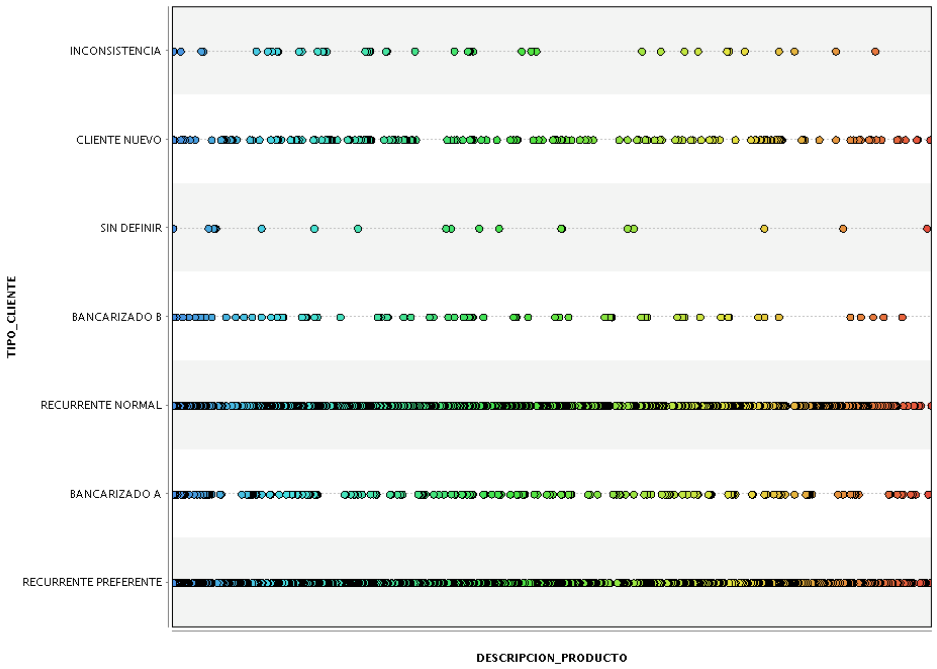


Figura 15 Tipo cliente vs descripción del producto

El atributo tipo de cliente, también se considera en la creación del modelo, ya que aporta características de los clientes.

4.1.3. Selección y configuración del modelo

En esta etapa de selección y configuración del modelo se utilizó Auto Model de RapidMiner, mediante el cual, se pudo realizar visualizar, las características de los datos obtenidos, el comportamiento de los atributos más importantes y sobre todo permitió seleccionar el modelo con mejor confianza y adecuada a la data analizada.

Se comenzó seleccionando los datos cargados previamente en el repositorio de RapidMiner:

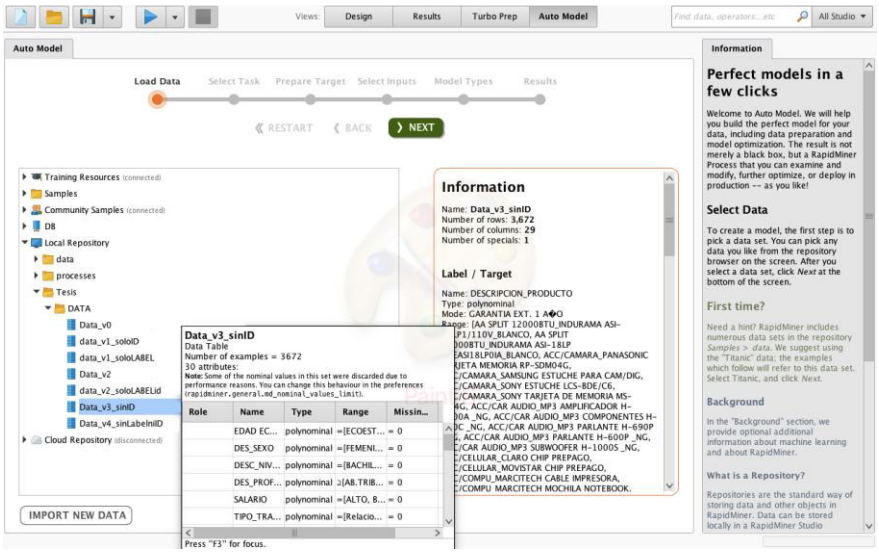


Figura 16 Proceso de carga de datos

En el siguiente paso se seleccionó el tipo de tarea a realizar, luego de haber seleccionado el conjunto de datos. Aquí se decide el tipo de problema que se quiere resolver. Se ofertan tres tareas: predecir, clústeres y valores atípicos.

Predecir, permite predecir los valores de una de las columnas de los datos e identifica la columna, mediante esta opción se construye un modelo de aprendizaje automático que predice los valores de esta columna en función de los valores de las otras columnas.

Clústeres, que agrupa los datos en clústeres. Esta opción no predice los valores de una sola columna, sino que encuentra conjuntos de puntos de datos que estén juntos.

Valores atípicos, permite encontrar puntos inusuales en sus datos. El objetivo aquí es encontrar puntos de datos individuales que estén lejos de todos los demás puntos de datos, posiblemente debido a errores en la recopilación de datos o debido a un comportamiento extraño o inesperado.

En esta ventana se escogió la opción “Predecir” y se seleccionó como atributo o columna a predecir la descripción del producto.

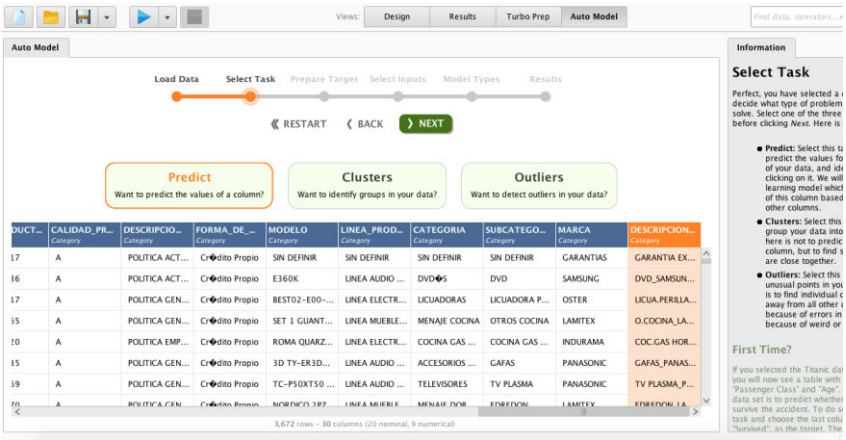


Figura 17 Selección de tarea

A continuación, la herramienta presenta las clases que se van a determinar mediante el atributo a predecir, es decir, en este paso se observan todos los productos ofertados por las empresas y que posteriormente serán los recomendados a los clientes de acuerdo con sus características.

Además, mediante el diagrama de barras se visualiza la cantidad de registros por cada una de las clases, estimando así la tendencia del modelo a crear. También, aquí se puede hacer una adecuación de los nombres de las clases en caso de requerirlo.

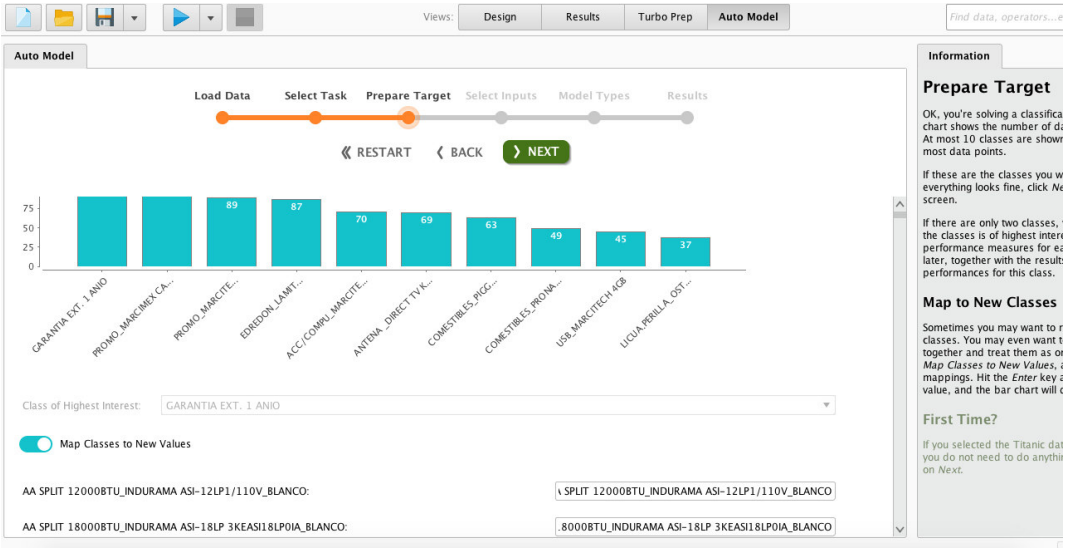


Figura 18 Clases del atributo a predecir

Una de las mayores ventajas presentadas por Auto Model de RapidMiner es que los datos a ser utilizados para crear el modelo no son los seleccionados por la herramienta, sino que permite

seleccionar los atributos que el usuario desea utilizar, y con la facilidad de mostrar la correlación, estabilidad, porcentaje de datos faltante y sobre todo la recomendación mediante colores de cuáles son los atributos que mejor se acoplan a los datos cargados.

En este paso se deseleccionó los atributos que no estaban relacionados a los clientes y los que no eran recomendados por la herramienta.

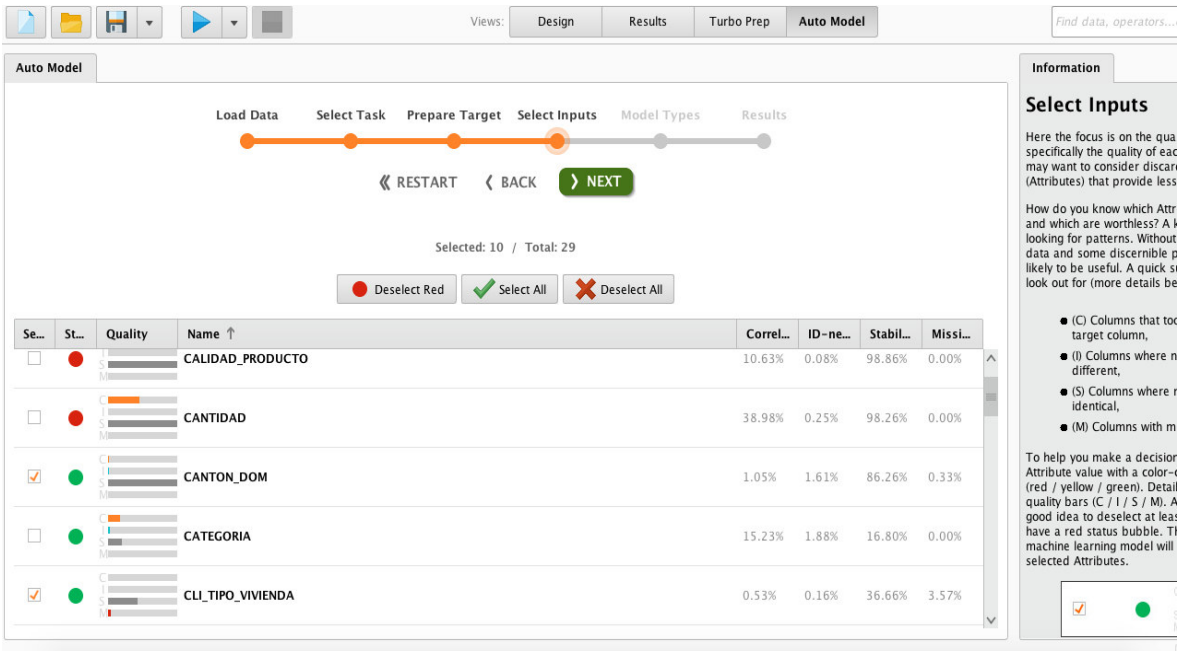


Figura 19 Selección de atributos de entrada del modelo

De acuerdo con esto se consideró los siguientes atributos:

Tabla 4
Criterios de selección de atributos

ATRIBUTO	SELECCIONADO	CRITERIO
EDAD ECONOMICA	SI	REFERENTE AL CLIENTE
DES_SEXO	SI	REFERENTE AL CLIENTE
DESC_NIV_EST	SI	REFERENTE AL CLIENTE
DES_PROFESION	SI	REFERENTE AL CLIENTE
SALARIO	SI	REFERENTE AL CLIENTE
TIPO_TRABAJO	NO	BAJA CORRELACIÓN
NRO_CARGAS	SI	REFERENTE AL CLIENTE
CLI_TIPO_VIVIENDA	SI	REFERENTE AL CLIENTE
CANTON_DOM	SI	REFERENTE AL CLIENTE

Continúa

VPROVINCIA_DOM	SI	REFERENTE AL CLIENTE
ANIOS_ANTIGUEDAD	SI	REFERENTE AL CLIENTE
PER_NRO_IDENTIFICACION	NO	BAJA CORRELACIÓN
TIPO_CLIENTE	SI	REFERENTE AL CLIENTE
TIPO_IDENTIFICACION	NO	BAJA CORRELACIÓN
VENTA_NETA	NO	REFERENTE AL PRODUCTO
CANTIDAD	NO	REFERENTE AL PRODUCTO
VENTA_F	NO	REFERENTE AL PRODUCTO
TOTAL_FACTURA	NO	REFERENTE AL PRODUCTO
PRECIO_DE_LISTA	NO	REFERENTE AL PRODUCTO
DESCRIPCION_POLITICA	NO	REFERENTE AL PRODUCTO
FORMA_DE_VENTA	NO	REFERENTE AL PRODUCTO
ID_PRODUCTO_1	NO	REFERENTE AL PRODUCTO
CALIDAD_PRODUCTO	NO	REFERENTE AL PRODUCTO
DESCRIPCION_PRODUCTO	LABEL	ATRIBUTO A PREDECIR
MODELO	NO	REFERENTE AL PRODUCTO
LINEA_PRODUCTO	NO	REFERENTE AL PRODUCTO
CATEGORIA	NO	REFERENTE AL PRODUCTO
SUBCATEGORIA	NO	REFERENTE AL PRODUCTO
MARCA	NO	REFERENTE AL PRODUCTO

Para seleccionar el algoritmo a ser utilizado para crear el modelo, mediante Auto Model de RapidMiner se seleccionó todos los modelos ofertados por la herramienta para verificar cuál de estos presenta mejor exactitud con los modelos creados.

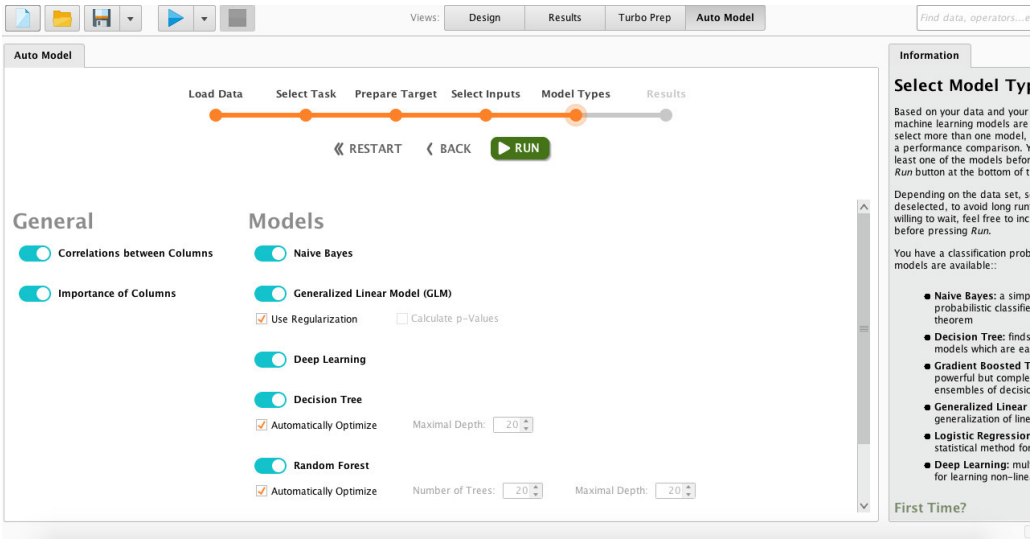


Figura 20 Selección de algoritmos de prueba

Luego de ejecutar el último paso, se obtuvo los siguientes resultados:

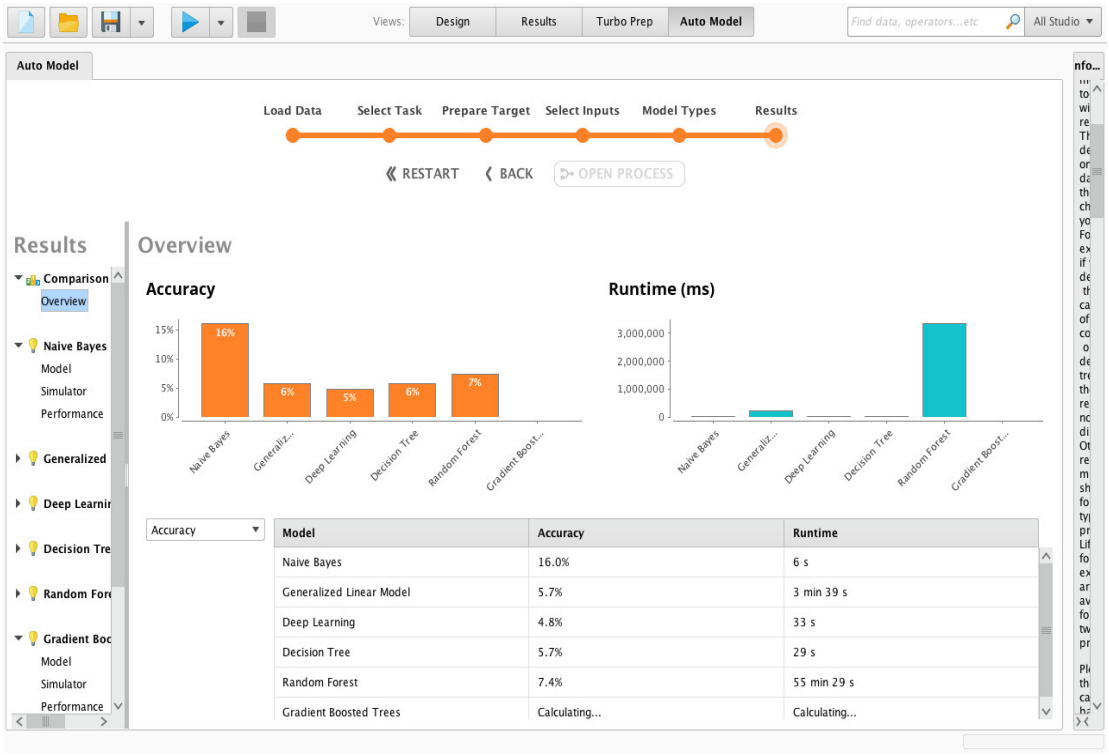


Figura 21 Comparación de tipos de modelos según precisión

De acuerdo con los resultados obtenidos, se escogerá como algoritmo a implementar en el modelo de minería de datos el de Naive Bayes por presentar mejor exactitud y menor margen de error con respecto a los otros algoritmos evaluados: modelo lineal generalizado, aprendizaje profundo, árbol de decisión, random forest.

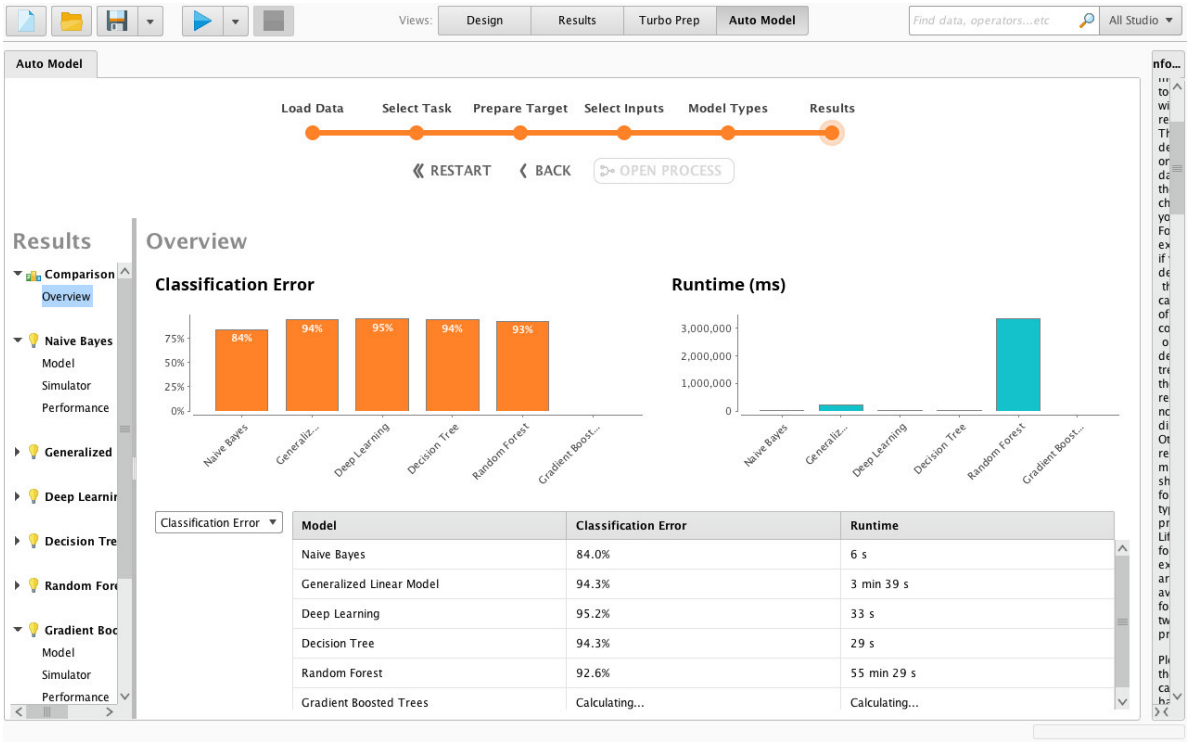


Figura 22 Comparación de tipos de modelos según margen de error

También, se sabe que el clasificador Naive Bayes calcula las probabilidades para predecir el producto a recomendar en función las entradas determinadas, en este caso los atributos que definen a un tipo de cliente. Luego selecciona el resultado con mayor probabilidad.

Este clasificador asume que las características son independientes. Se lo considera como un potente el algoritmo utilizado para:

- Predicción en tiempo real
- Clasificación de texto / filtrado de spam
- Sistema de recomendaciones

La herramienta Auto Model permite visualizar los resultados obtenidos del modelo seleccionado.

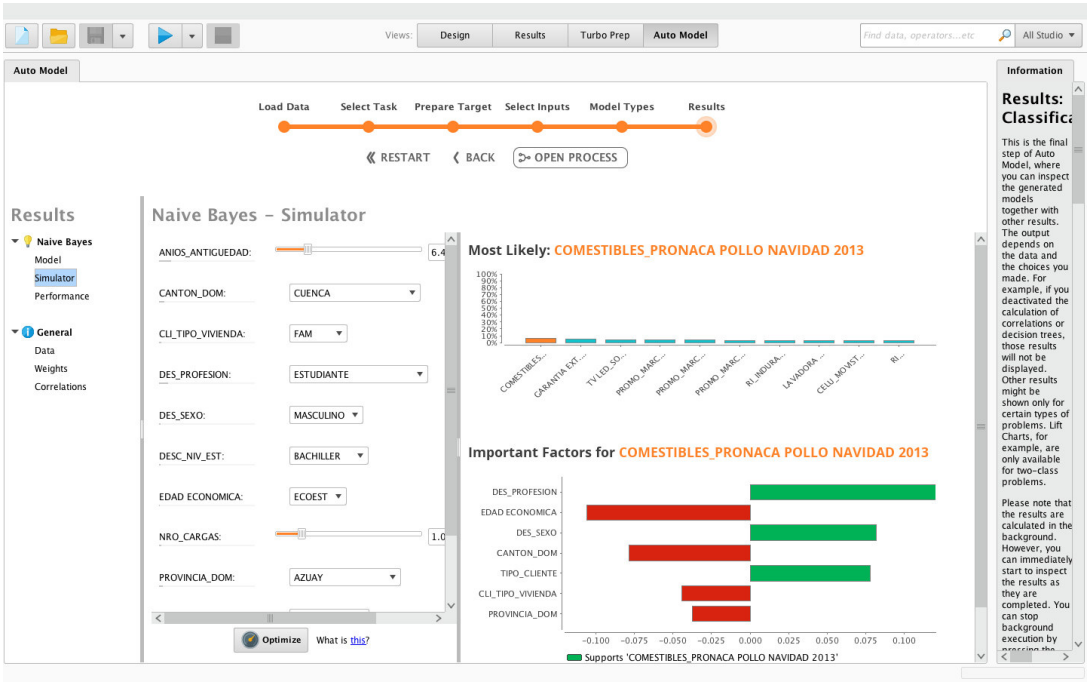


Figura 23 Simulación de modelo creado

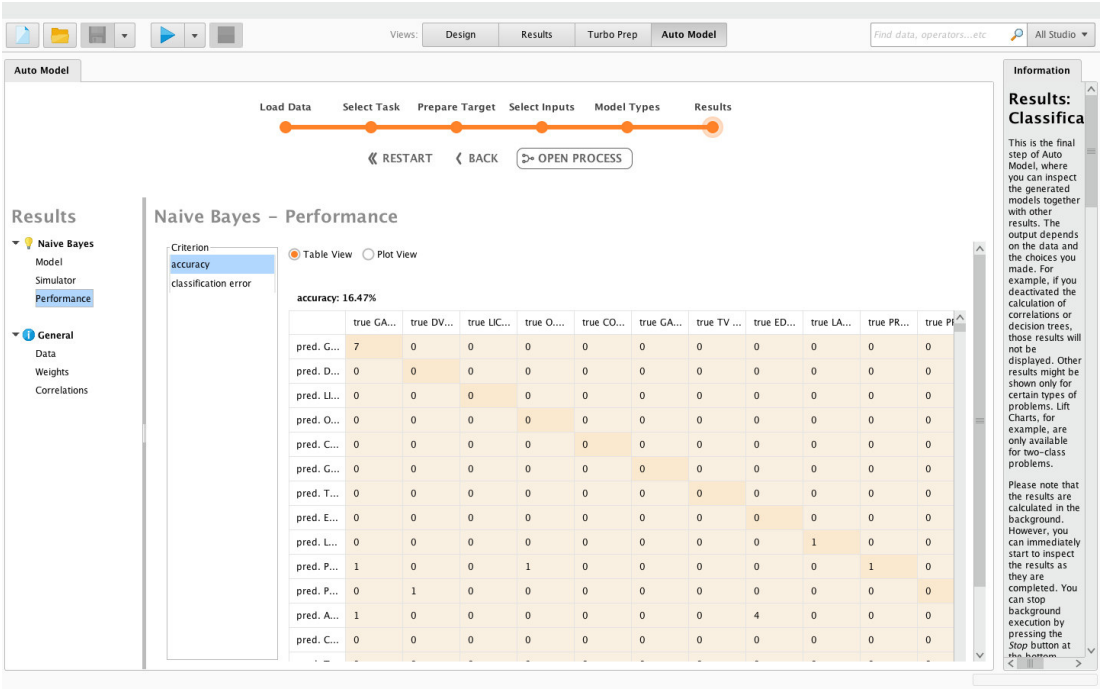


Figura 24 Resultados del modelo creado

4.1.4. Limpieza de datos

Luego de seleccionar el modelo de minería de datos a utilizar, se procedió a realizar la limpieza de la data seleccionada, esto utilizando las bondades de RapidMiner, que permite abrir un proceso del resultado obtenido al ejecutar Auto Model con el algoritmo seleccionado.

En esta fase se verificó que los tipos de datos a utilizar para la implementación del modelo sean adecuados, para que pueden trabajar de una manera confiable, ajustándose al modelo implementado.

Es así como se analizó cada uno de los procesos creados por la herramienta para verificar que los datos sean óptimos para presentar los resultados.

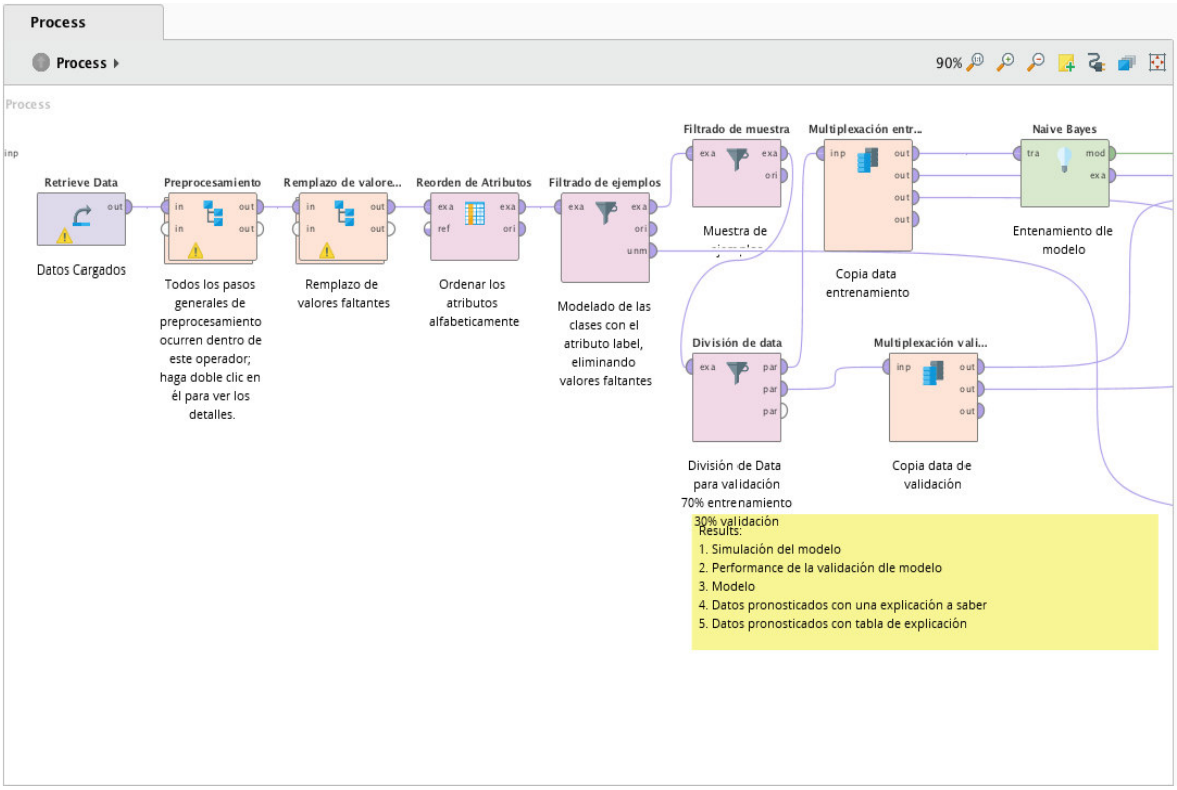


Figura 25 Proceso del modelo Naive Bayes en RapidMiner parte 1

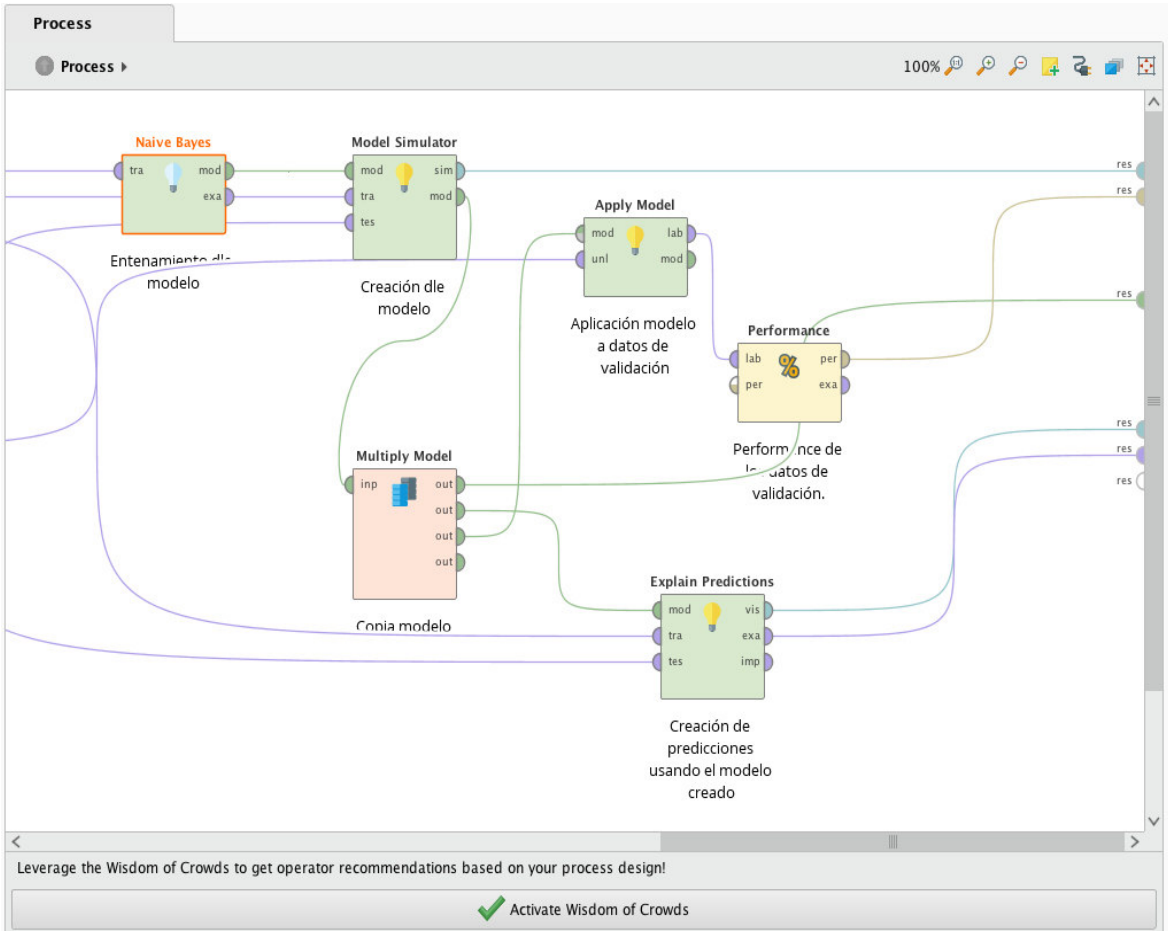


Figura 26 Proceso del modelo Naive Bayes en RapidMiner parte 2

En la primera parte del proceso se realizó la carga de los datos desde el repositorio de RapidMiner, para que seguidamente se haga el pre-procesamiento de los datos, en donde, se realiza la definición de la columna a predecir, se discretizan los valores, se mapea los datos de las clases, se filtran los atributos que aportan positivamente al modelo y por último se hace un remplazo de ñ por n.

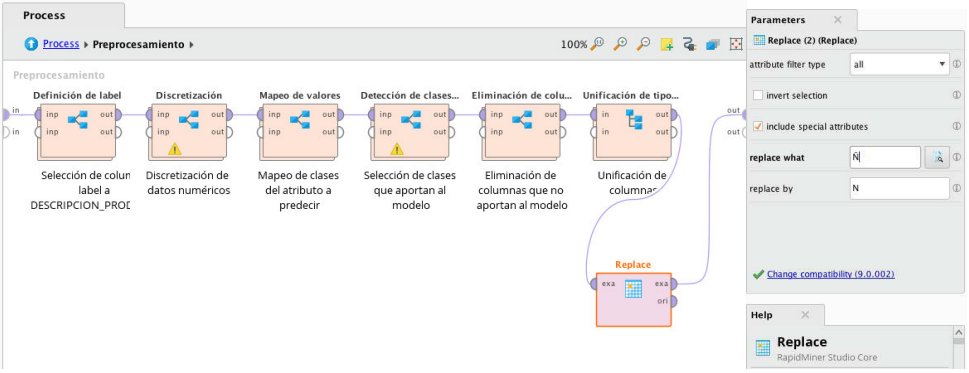


Figura 27 Proceso de Pre-Procesamiento en RapidMiner

A continuación, se realiza el cálculo y reemplazo de valores faltantes de todos los tipos de datos existentes.

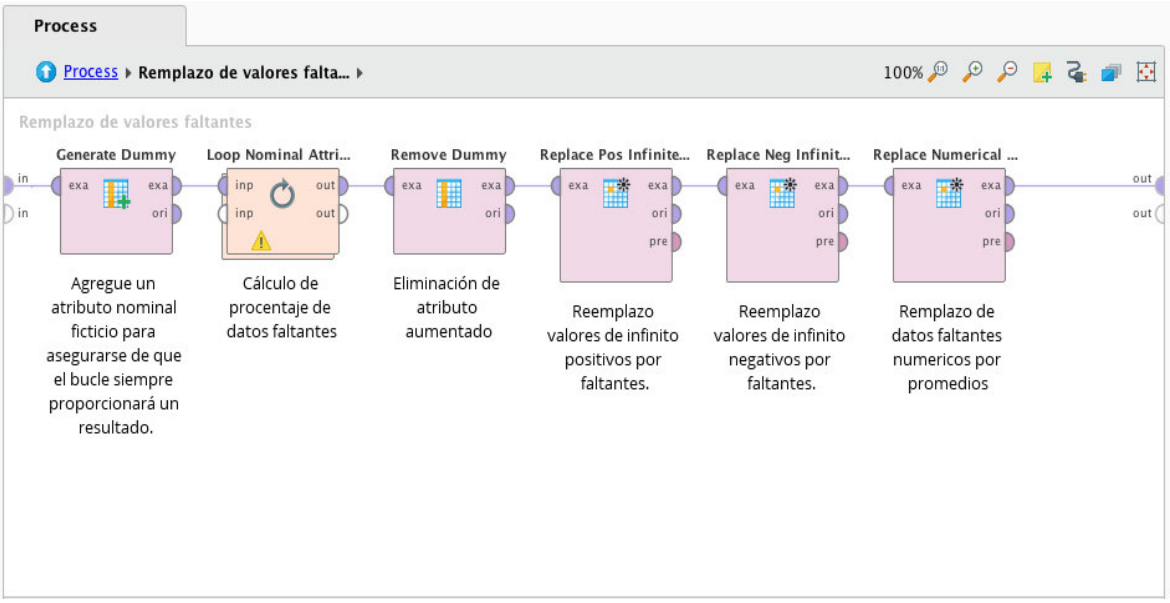


Figura 28 Proceso de remplazo de valores faltantes en RapidMiner

Seguidamente, en el proceso general se ordena los atributos alfabéticamente para su uso y se eliminan datos faltantes con filtros para su análisis.

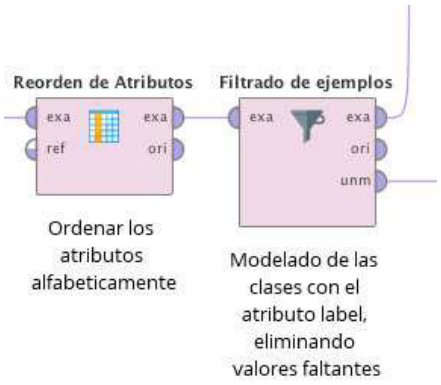


Figura 29 Proceso de ordenado y filtrado en RapidMiner

También, dentro del proceso se realiza la división de la data en 70% de datos para el modelado y 30% para el testing, que servirán para hacer la respectiva validación del modelo.

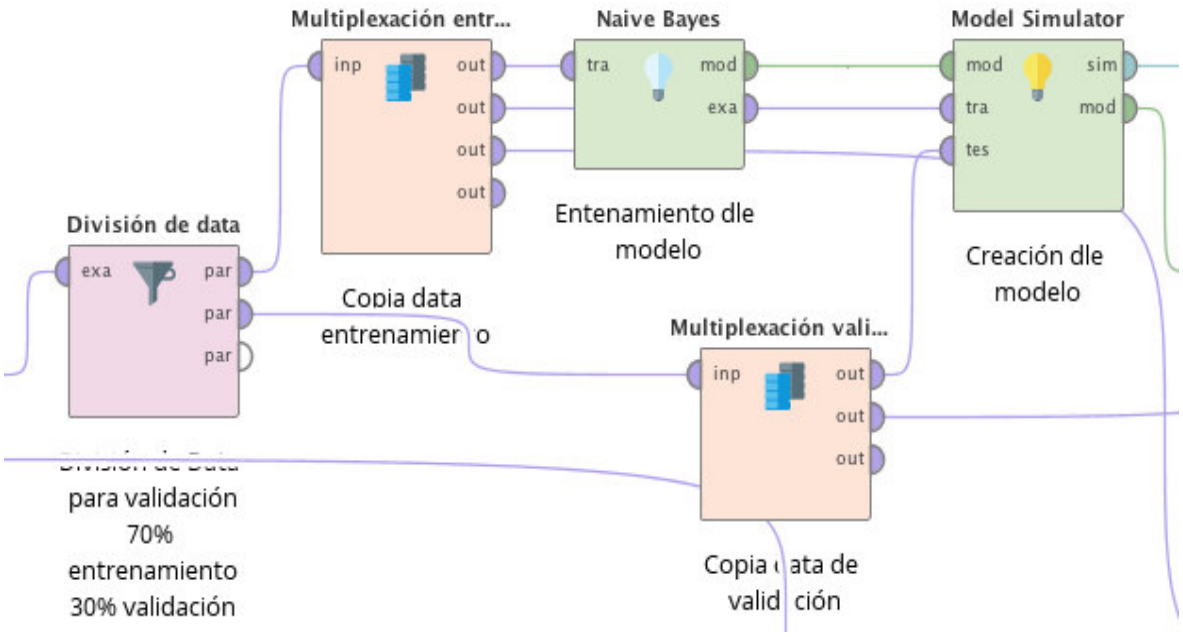


Figura 30 Proceso de separación de data y aplicación del modelo en RapidMiner

Por último, se realiza la validación del modelo luego de aplicarlo en los datos de testing, para obtener como resultado 3 salidas:

1. Simulación del modelo

2. Performance de la validación del modelo

3. Modelo

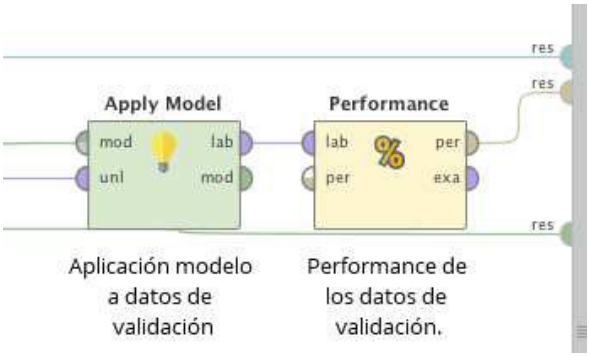
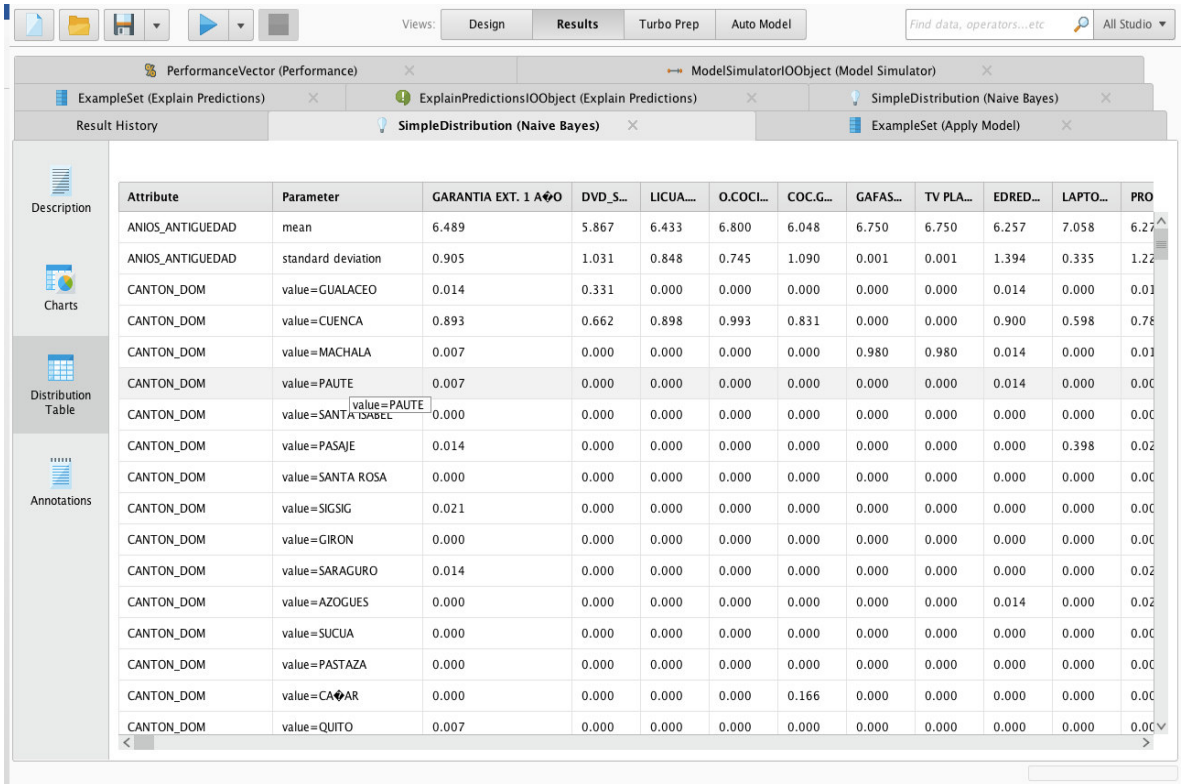


Figura 31 Proceso de validación del modelo en RapidMiner

Obteniendo como modelo final la relación de los atributos y sus clases sobresalientes con el atributo a predecir.



Attribute	Parameter	GARANTIA EXT. 1 A€O	DVD_S...	LICUA....	O.COCI...	COC.G...	GAFAS...	TV PLA...	EDRED...	LAPTO...	PRO
ANIOS_ANTIGUEDAD	mean	6.489	5.867	6.433	6.800	6.048	6.750	6.750	6.257	7.058	6.27
ANIOS_ANTIGUEDAD	standard deviation	0.905	1.031	0.848	0.745	1.090	0.001	0.001	1.394	0.335	1.22
CANTON_DOM	value=GUALACEO	0.014	0.331	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.01
CANTON_DOM	value=CUENCA	0.893	0.662	0.898	0.993	0.831	0.000	0.000	0.900	0.598	0.78
CANTON_DOM	value=MACHALA	0.007	0.000	0.000	0.000	0.000	0.980	0.980	0.014	0.000	0.01
CANTON_DOM	value=PAUTE	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.00
CANTON_DOM	value=SANTA ISABEL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=PASAJE	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.398	0.02
CANTON_DOM	value=SANTA ROSA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=SIGSIG	0.021	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=GIRON	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=SARAGURO	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.02
CANTON_DOM	value=AZOGUES	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.02
CANTON_DOM	value=SUCUA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=PASTAZA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=CA€AR	0.000	0.000	0.000	0.000	0.166	0.000	0.000	0.000	0.000	0.00
CANTON_DOM	value=QUITO	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00

Figura 32 Resultado del modelo Naive Bayes

4.1.5. Análisis de resultados

Ya obtenido el modelo y realizadas las respectivas transformaciones, para tener una data adecuada, se procedió a realizar la evaluación de resultados, utilizando el proceso que dividió la información en dos partes, la primera comprendida por la mayor parte de la información en este caso el setenta por ciento, para el aprendizaje y el treinta por ciento restantes se lo usó para la evaluación de resultados. Dividida la data se realizó la evaluación utilizando una matriz de confusión que permitió relacionar el número de datos clasificados correctamente y el número de datos clasificados incorrectamente, determinando así, el nivel de confianza del modelo propuesto.

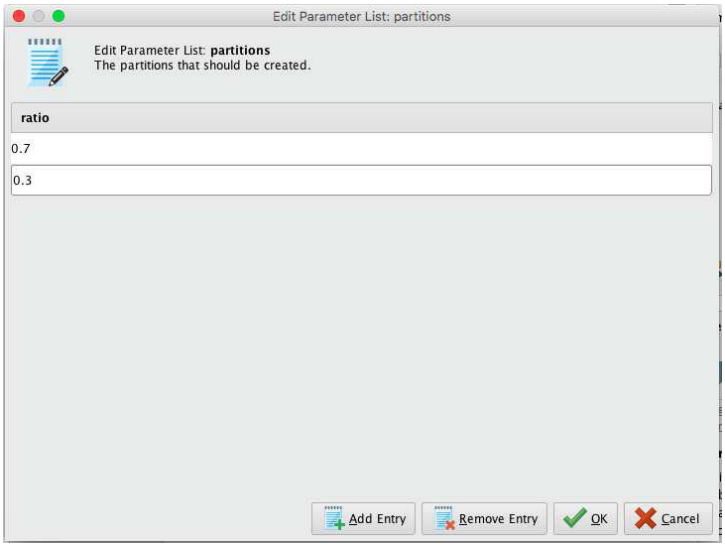


Figura 33 Porcentajes de división de la data para validación del modelo

Donde se obtuvo los siguientes resultados:

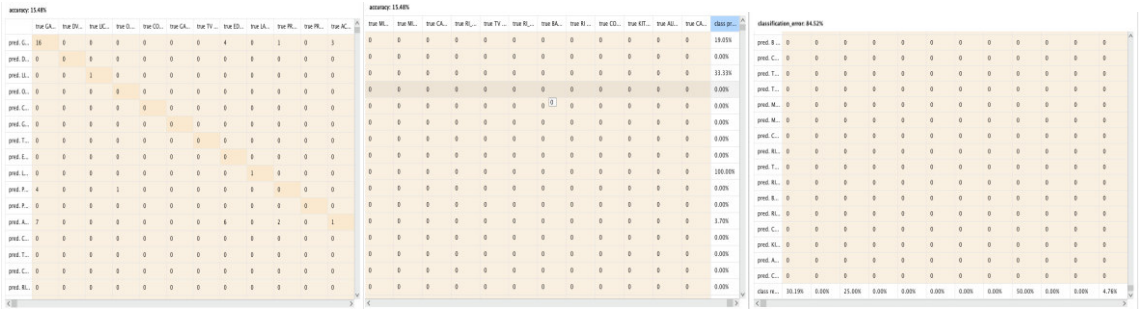


Figura 34 Matriz de confusión obtenida del modelo

En la matriz de confusión que se observa en la figura, se puede observar cómo se compara cada uno de los valores que se pueden obtener al predecir el producto a recomendar al cliente de acuerdo con las características de este.

Además, las facilidades de la herramienta permiten visualizar mediante gráficos el nivel de predicción del modelo creado.

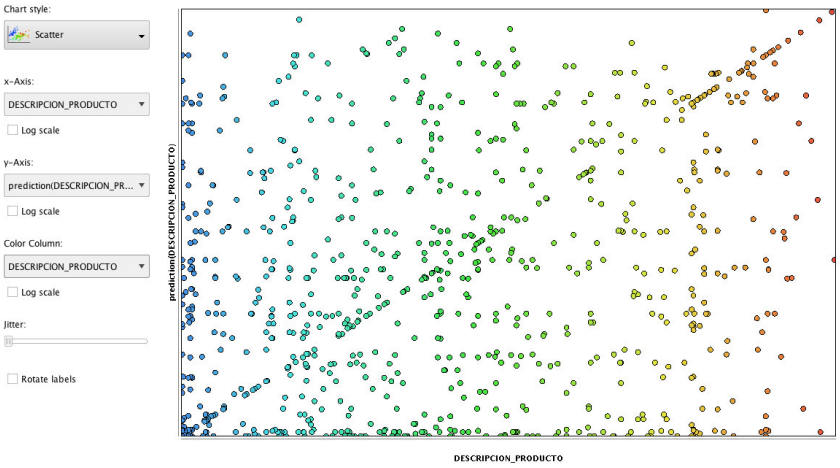


Figura 35 Gráfico de dispersión de la predicción alcanzada con el modelo

Esta figura, permite visualizar como la predicción de los diferentes productos tiene una distribución lineal, con pocos valores aislados, lo que indica que el porcentaje de predicción del modelo es alto.

También, se puede observar que el número de productos a recomendar es alto, por cuanto en la siguiente figura se puede observar las diferentes predicciones que se pueden dar en el modelo.

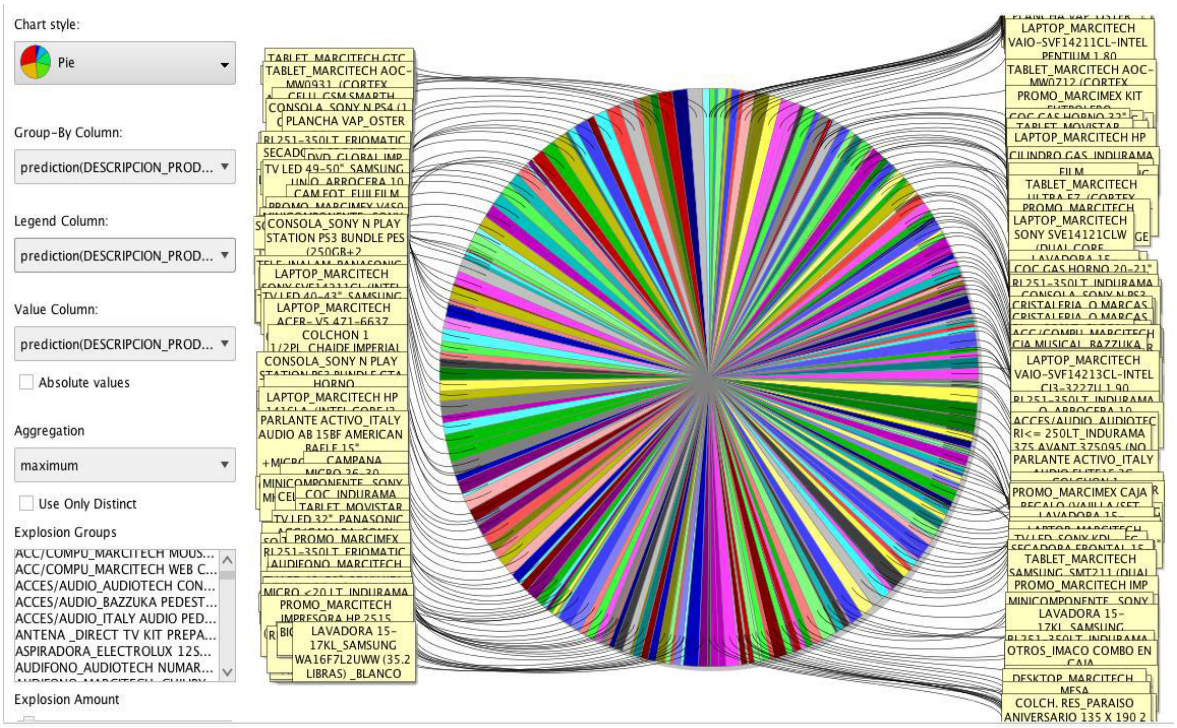


Figura 36 Productos predichos por el modelo

El producto con mayor predicción que se da en el modelo presentado es la garantía de 1 año, PROMO_MARCIMEX CAJA NAVIDAD 2013, entre otros como se puede ver en la figura.

Nominal values

Index	Nominal value	Absolute count	Fraction ↓
1	GARANTIA EXT. 1 ANIO	84	0.080
2	PROMO_MARCIMEX CAJA NAVIDAD 2013	44	0.042
3	EDREDON_LAMITEX NORDICO 2PZS. + 2 COJI	37	0.035
4	ANTENA _DIRECT TV KIT PREPAGO	29	0.028
5	ACC/COMPU_MARCITECH MOCHILA/MALETIN	27	0.026
6	COC.GAS HORNO 32"_INDURAMA PARMA QUARZO PAR019_CR	23	0.022
7	USB_MARCITECH 4GB	22	0.021
8	COMESTIBLES_PIGGIS PAVO NAVIDAD 2013	16	0.015
9	COMESTIBLES_PRONACA POLLO NAVIDAD 2013	16	0.015
10	ACC/CAR AUDIO_MP3 PARLANTE H-690P _NG	15	0.014
11	LAVADORA 15-17KL_L.G. WFS1634 _SILVER	14	0.013
12	PROMO_MARCITECH AUDIFONO GTC	14	0.013
13	PROMO_MARCITECH MOUSE	14	0.013
14	PROMO_MARCITECH REGULADOR GTC	14	0.013
15	PROMO_MP3 KIT CABLES (COMBO MP3) _-	11	0.010
16	CELU_CLARO LG L5-E612G	9	0.009
17	CONSOLA_SONY N PS3 LITTLEBIGPLANETKARTING 12GB_NG	9	0.009
18	PARLANTE ACTIVO_ITALY AUDIO ELITE15 3G (ACTIVA/10000 PMPO) _NG	9	0.009
19	PROMO_MARCITECH WEB CAM MARCITECH	9	0.009
20	CJA MUSICAL_BAZZUKA CMU-BAZZUKA BT01 + MICROF	8	0.008
21	LAPTOP_MARCITECH SONY SVF14221CL (INTEL PENTIUM 2117U 1.80GHZ/4...	8	0.008
22	LICUA.PERILLA_OSTER 4655-013	8	0.008
23	PROMO_MARCITECH IMPRESORA HP 1515 DESKJET INK ADVANTAGE	8	0.008
24	RI_INDURAMA 395 QUARZO 395054_CR	8	0.008
25	TV LCD_L.G. 32CS410 _NG	8	0.008

Figura 37 Clases con mayor porcentaje de predicción de acuerdo con el modelo

Por último, el porcentaje de precisión conseguido en el modelo es del 16.01%, aceptable en comparación con los otros modelos visualizados.

Table View Plot View

accuracy: 16.01%

	true GA...	true DV...	true LIC...	true O....	true CO...
pred. G...	8	0	0	0	0
pred. D...	0	0	0	0	0
pred. LI...	0	0	0	0	0
pred. O...	0	0	0	0	0

Figura 38 Precisión del modelo

El porcentaje de precisión alcanzado se debe a que los atributos con mayor correlación al atributo a predecir son los relacionados al producto que no se consideran en este trabajo debido a que el objetivo final del trabajo es determinar las características del cliente para con esto recomendar un producto determinado.

4.1.6. Reporte a los tomadores de decisión

Una vez obtenidos y validados los resultados, estos se enlazaron con la herramienta Tableau, que permitió realizar un conjunto de reportes, disponibles, para ser presentados a los dirigentes de la organización, como resultado de aplicar minería de datos, y ofreciendo un sistema de recomendaciones de productos.

Primero se exporto la data del modelo conseguido utilizando RapidMiner, para luego pasarla a la herramienta Tableau y visualizar los datos.

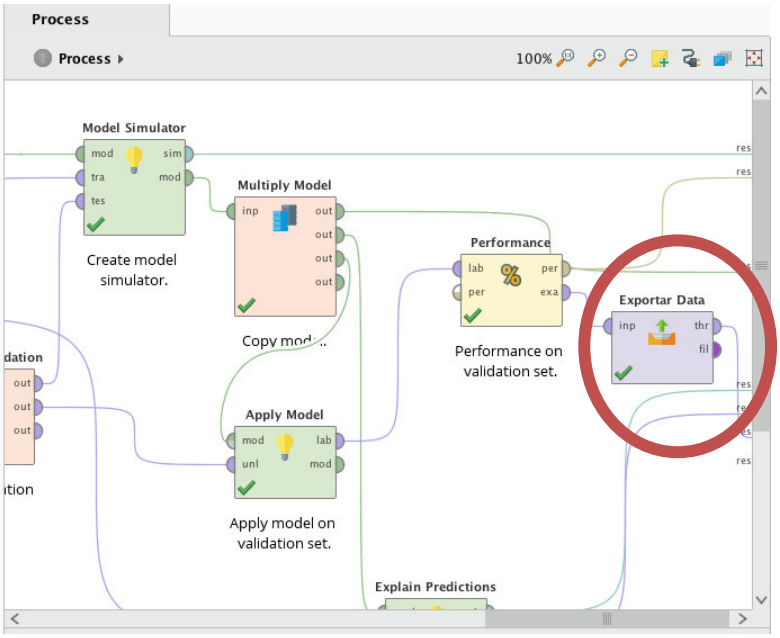


Figura 39 Proceso de extracción de la data en RapidMiner

Seguidamente, se procede a cargar la data en Tableau y crear los reportes que muestren a los dirigentes de la organización los beneficios de utilizar este modelo en un sistema de recomendaciones.

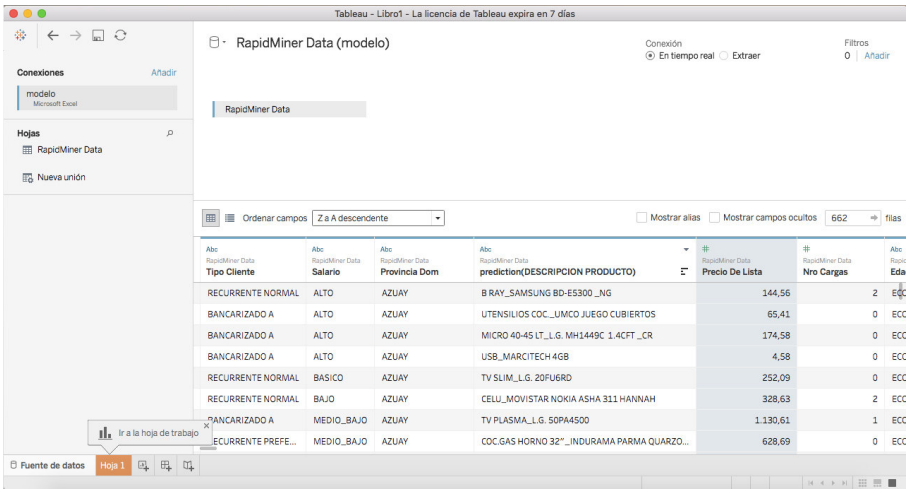


Figura 40 Carga de datos del modelo en Tableau

Ya con los datos en la herramienta Tableau se procedió a crear algunos reportes que muestren los resultados obtenidos con el modelo y como están las recomendaciones de productos dado por el modelo de acuerdo con las características de los clientes.

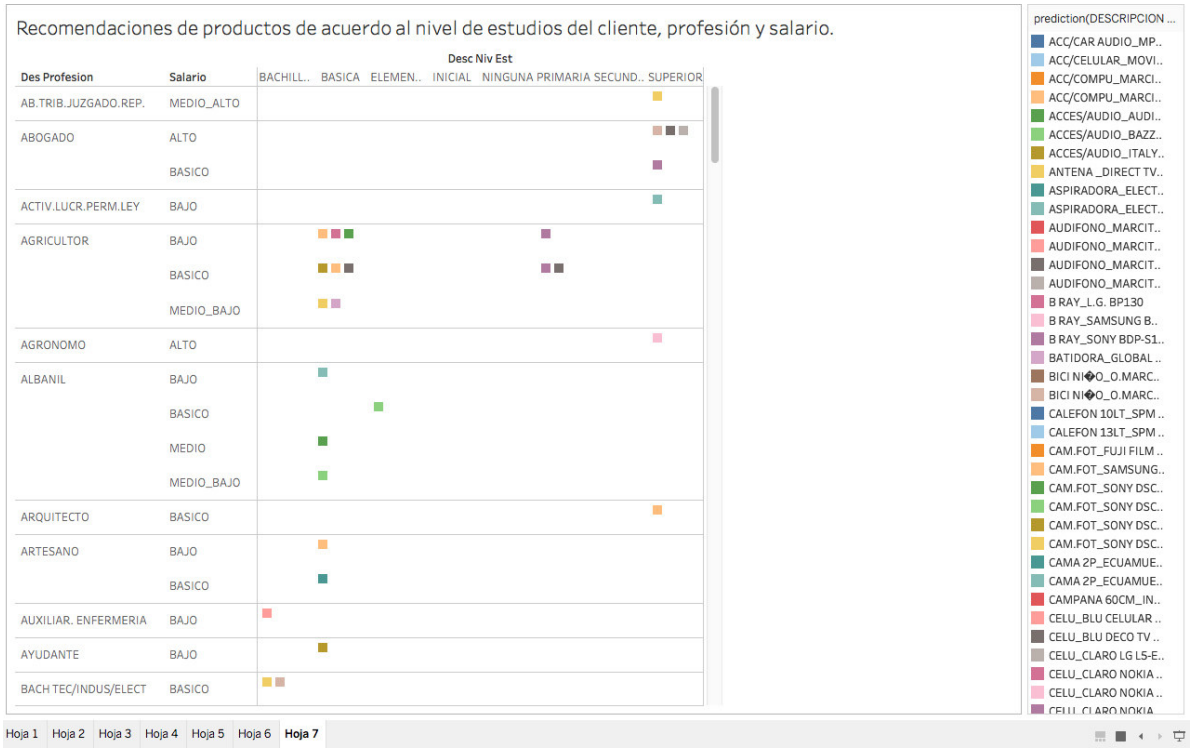


Figura 41 Reporte 1 de recomendación de productos en Tableau

Figura 42 Reporte 2 de recomendación de productos en Tableau

De la misma manera se puede visualizar reportes de cuáles son las características de clientes que se recomienda más productos.

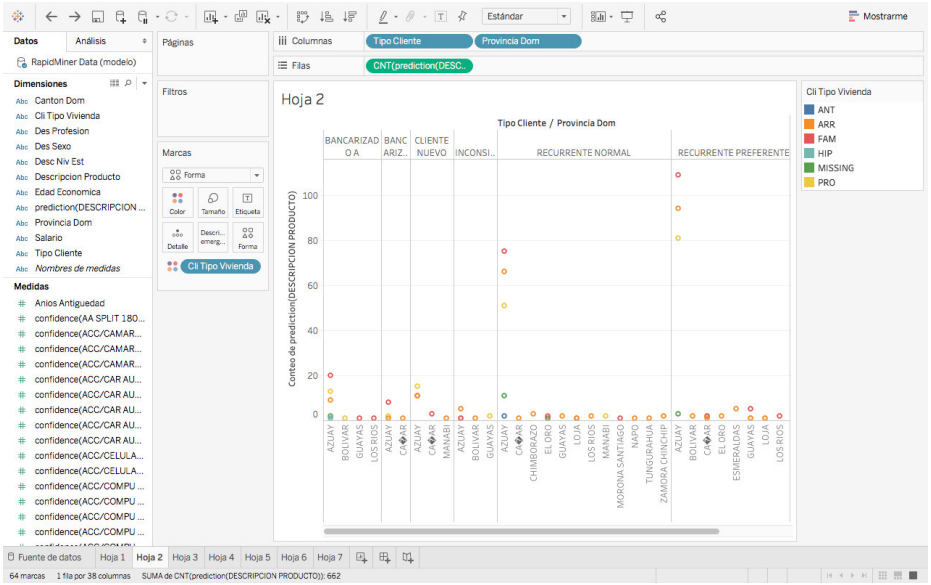


Figura 43 Reporte 3 de recomendación de productos en Tableau

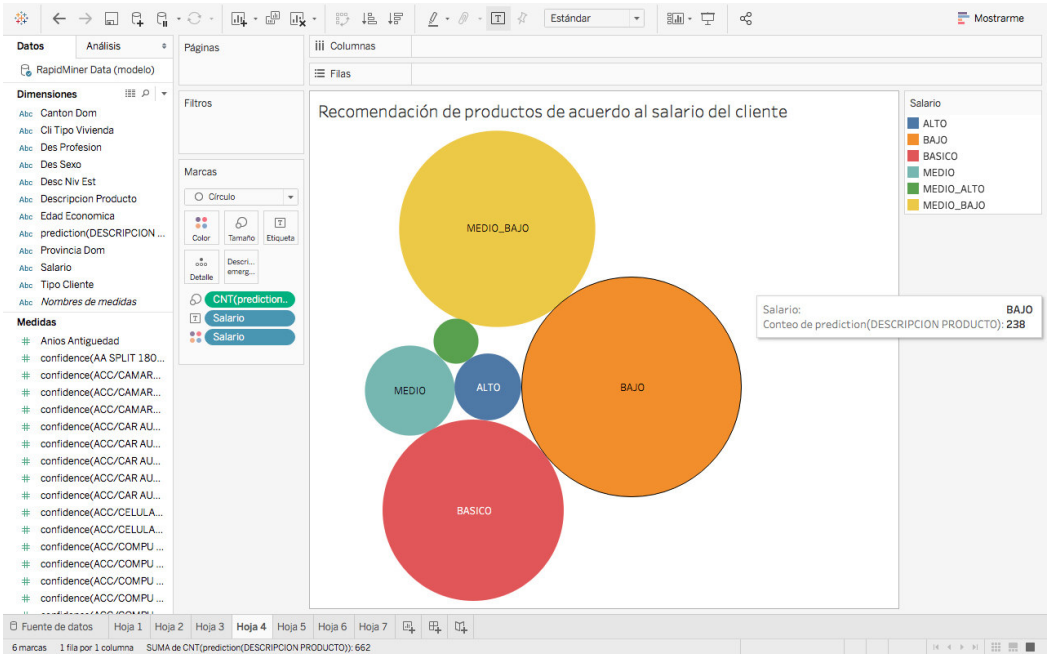


Figura 44 Reporte 4 de recomendación de productos en Tableau

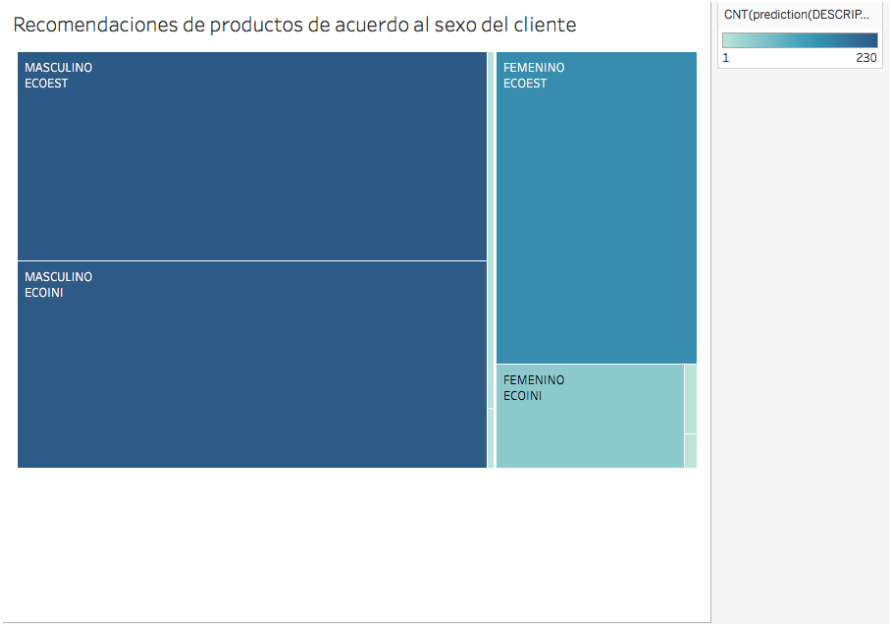


Figura 45 Reporte 5 de recomendación de productos en Tableau

4.2 Metodología para ejecutar la propuesta

De acuerdo con los resultados obtenidos se propone implementar la propuesta en una aplicación web que permita a los vendedores de productos de la organización ingresar los datos y características de los clientes para que, de esta manera, dicha aplicación recomiende que producto es el más optado a ser recomendado y así llegar de una mejor manera al cliente y ofrecerle el producto con mayor probabilidad a ser comprado.

Además, con los reportes creados, se sugiere a los dirigentes de la organización analizar qué productos tienen una menor probabilidad a ser recomendados, para así tomar decisiones de cómo mejorar la probabilidad de venta de este.

RESUMEN DEL CAPITULO IV

En el desarrollo de este capítulo se expusieron los resultados obtenidos luego de ejecutar este proyecto. Primero se presentó un informe de resultados donde se visualiza las tareas realizadas en cada una de las etapas de la metodología adoptada.

En la definición de la tarea se consideró que los registros a utilizar corresponden al historial ventas de productos al cliente. Por cuanto, la tarea principal realizada en el proyecto fue que mediante la información recolectada se prediga qué artículo es más probable que un cliente desee, de acuerdo a sus características. Esta predicción, es la recomendación que los vendedores de productos darán, basada en históricos de ventas de los productos a otros clientes con características similares.

Ese historial se extrajo de las bases de los datos Oracle de la organización, las cuales, para facilidad de utilización, se almacenaron en el repositorio de la herramienta RapidMiner. De acuerdo con los datos obtenidos y a los requerimientos de la empresa se definió como atributo a predecir

para el sistema de recomendaciones al atributo DESCRIPCION_PRODUCTO, puesto que este permitió definir cuál es el producto que el vendedor debe recomendar a los clientes. Además, en base a los datos recolectados de la organización, se procedió a realizar un análisis de correlación con el atributo a predecir mediante diagramas de dispersión utilizando la herramienta RapidMiner, para seleccionar los datos que mejores aportes darán al modelo de minería de datos.

En la etapa de selección y configuración del modelo se utilizó Auto Model de RapidMiner, donde se comenzó seleccionando los datos cargados previamente en el repositorio de RapidMiner, en seguida se escogió la opción “Predecir” y se seleccionó como atributo o columna a predecir la descripción del producto, a continuación, la herramienta presentó mediante un diagrama de barras la cantidad de registros por cada una de las clases, estimando así la tendencia del modelo a crear y también, se realizó adecuaciones a los nombres de algunas clases, luego se seleccionó los atributos que mejor se acoplan al modelo deseado basándose en la correlación, estabilidad, porcentaje de datos faltante, y finalmente se seleccionaron todos los algoritmos ofertados por la herramienta para verificar cuál de estos presenta mejor exactitud con los modelos creados. El algoritmo seleccionado fue Naive Bayes por presentar mejor exactitud y menor margen de error con respecto a los otros algoritmos evaluados: modelo lineal generalizado, aprendizaje profundo, árbol de decisión, random forest.

Luego de seleccionar el modelo de minería de datos a utilizar, se procedió a realizar la limpieza de la data seleccionada, esto utilizando las bondades de RapidMiner, que permitió abrir un proceso del resultado obtenido al ejecutar Auto Model con el algoritmo seleccionado, para verificar que los tipos de datos a utilizar para la implementación del modelo sean adecuados, y poder trabajar de una manera confiable, ajustándose al modelo implementado.

Ya obtenido el modelo y realizadas las respectivas transformaciones, para tener una data adecuada, se procedió a realizar la evaluación de resultados, dividiendo la data en dos partes, el 70% para el aprendizaje y 30% para la evaluación de resultados. Dividida la data se realizó la evaluación utilizando una matriz de confusión que permitió relacionar el número de datos clasificados correctamente y el número de datos clasificados incorrectamente, determinando así, el nivel de confianza del modelo propuesto.

Una vez obtenidos y validados los resultados estos se enlazaron con la herramienta Tableau, que permitió realizar un conjunto de reportes, disponibles, para ser presentados a los dirigentes de la organización, como resultado de aplicar minería de datos, y ofreciendo un sistema de recomendaciones de productos.

Como metodología para ejecutar la propuesta se propone implementar una aplicación web que permita a los vendedores de productos de la organización ingresar los datos y características de los clientes para que, de esta manera, dicha aplicación recomiende que producto es el más optado a ser recomendado y así llegar de una mejor manera al cliente y ofrecerle el producto con mayor probabilidad a ser comprado.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

El estudio del arte permite al investigador conocer artículos sobre el tema a desarrollar, en este caso determinó técnicas de sistemas de recomendaciones existentes, que fueron un apoyo en el momento de definir el algoritmo de minería de datos a utilizar, los atributos con mayor correlación y por último para definir la manera como se debe implementar el modelo diseñado. Mediante este análisis se confirmó que el modelo escogido es uno de los más usados y ha respondido a sistemas de minería de datos con respuestas confiables y con un grado de exactitud alta.

La fase de limpieza de datos es fundamental en todo proceso de inteligencia de negocios ya que en esta fase se identifica valores atípicos que pueden distorsionar los resultados y propiedades de los datos. Mediante estos diagramas se destacó características de los datos analizados como que los productos ofrecidos en la empresa están más orientados a los clientes que tienen economía estable y que se están iniciando económicamente; que el género del cliente no afecta en el nivel de compras de los productos; que el porcentaje de clientes que adquieren los productos de la empresa tienen un nivel de estudio mayor a primaria; que la venta de productos está mayormente concentrada en los clientes que tienen ingresos básicos, medio bajo que está acorde al segmento de mercado en el que se enfoca la empresa, más no en los que tienen ingresos altos; que entre menos cargas tenga el cliente, mayor es su porcentaje de comprar un producto de la empresa caso de estudio; que el tipo de vivienda de los clientes que más compras han realizado en la empresa caso de estudio es de un familiar, arrendada y propia.

Se obtuvo precisiones de 16% para Naive Bayes, 5.7% para el modelo Lineal Generalizado y Decision tree, 4.8% para modelo Deep Learning y 7.4% para Random Forest; donde se puede apreciar que en términos de precisión el algoritmo óptimo es el escogido para este trabajo. De igual manera el tiempo de respuesta del algoritmo es fundamental, es así que se obtuvo tiempos de respuesta de 6 segundos para Naive Bayes, 29 segundos para Decision Tree, 33 segundos para Deep Learning, 55 segundos para Ramdom Fores y 3 minutos 29 segundos para el modelo linear Generalizado. Lo que nos ayuda a certificar que el algoritmo seleccionado es el que mejor tiempo de respuesta tiene.

El proceso de formateo de los datos para que trabajen adecuadamente con el modelo seleccionado, eliminando datos nulos, discretizando los datos de ser necesarios y garantizando el correcto funcionamiento de este, permitió tener una data óptima para aplicar el modelo realizado. Tal es el caso que mejoro la precisión de un 10% a un 16%.

El sistema de recomendaciones implementado ayuda a la empresa no solo a predecir que producto es de interés del cliente si no también identificar los productos que tienen menos acogida entre los clientes como vitrina Indurama, vajilla andina entre otros, esto permite a la organización tomar medidas para mejorar la efectividad de la fuerza de ventas del negocio al lanzar promociones con los productos que menos acogida tienen entre los clientes.

5.2.Recomendaciones

Todo trabajo de Inteligencia de negocios debe estar basado en una metodología estándar con procesos utilizados por expertos en inteligencia de negocios para tener un resultado óptimo.

El modelo de inteligencia de negocios planteado puede mejorar si la empresa implementa un sistema en su página web que permita a los clientes valorar los productos lo que permitirá al

algoritmo tener un campo más para determinar de manera precisa las preferencias de los clientes por edad, situación económica, ubicación geográfica mejorando la precisión del algoritmo.

Los resultados obtenidos se recomienda analizarlos, para revisar la lista de productos y marcas ofertados, puestos algunos según el análisis no tendrán una predicción de compra grande por ejemplo una vitrina Indurama. De igual manera se podría analizar aquellos productos que tiene mayor probabilidad de ser vendidos como antena direct tv prepago.

El modelo de inteligencia de negocios se puede extender a otras áreas del negocio como por ejemplo cartera, para predecir la probabilidad de pago que tiene un cliente y así poder otorgarle un crédito, inventarios para mapear geográficamente donde deben estar ubicados los productos según la preferencia de los clientes y así evitar gastos por uso de bodegas, compras para identificar los productos que se deben adquirir con mayor frecuencia y los que ya se deben dejar de adquirir, para identificar lugares donde se debería incrementar locales de venta, para impulsar las ventas de las marcas que maneja el grupo, entre otros.

BIBLIOGRAFÍA

Akash Gujarathi, S. K. (2018). *Competent K-means for Smart and Effective E-commerce.* Obtenido de Springer: https://www.springerprofessional.de/competent-k-means-for-smart-and-effective-e-commerce/15547248

Angulo, S. (19 de 07 de 2017). Seis propuestas para reactivar la economía ecuatoriana. *El Comercio*.

Bandyopadhyay, S., Thakur, S., & Mandal, J. (2019). Product recommendation for E-commerce data using association rule and apriori algorithm. *International Conference on Modelling and Simulation, MS 2017*, (págs. 585-593). Kolkata; India.

Bhade, K., Gulalkari, V., Harwani, N., & Dhage, S. (2018). A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization. *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, (pág. 8494019). Bengaluru; India.

Dittert, Härting, Reichstein, & Bayer. (2017). *A Data Analytics Framework for Business in Small and Medium-Sized Organizations.* Obtenido de Springer: https://link.springer.com/chapter/10.1007/978-3-319-59424-8_16

Diwandari, S., Permanasari, A., & Hidayah, I. (2018). Research methodology for analysis of E-commerce user activity based on user interest using web usage mining. *Journal of ICT Research and Applications*, 54-69.

Figueroa, E. (22 de 04 de 2016). *Publicidad y medios masivos en Marketing.* Obtenido de Gestiopolis: https://www.gestiopolis.com/publicidad-medios-masivos-marketing/

Gobi, N. &. (19 de 02 de 2018). *Analyzing cloud based reviews for product ranking using feature based clustering algorithm.* Obtenido de Springer: https://link.springer.com/article/10.1007/s10586-018-1996-3#citeas

Grozin, V., & Alla, L. (2017). *Similar product clustering for long-tail cross-sell recommendations.* Obtenido de ResearchGate: https://www.researchgate.net/publication/319455031_Similar_product_clustering_for_long-tail_cross-sell_recommendations

Haghighatnia, S., Abdolvand, N., & Rajaee Harandi, S. (2018). Evaluating discounts as a dimension of customer behavior analysis. *Journal of Marketing Communications*, 321-336.

INEC. (31 de 10 de 2017). *Ecuador registró 843.745 empresas en 2016*. Obtenido de Ecuador en cifras: http://www.ecuadorencifras.gob.ec/ecuador-registro-843-745-empresas-en-2016/

Jannach, D. L. (2015). What recommenders recommend: An analysis of recommendation biases and possible countermeasures. . *User Modeling and User-Adapted Interaction*, 427–491.

Mitra, A., Ghosh, S., Basuchowdhuri, P., Manoj, K. S., & Sanjoy, K. S. (2016). *Recommendation system based on product purchase analysis.* Obtenido de Springer: https://link.springer.com/article/10.1007/s11334-016-0274-x

SAS. (2018). *Data Mining*. Obtenido de SAS: https://www.sas.com/en_us/insights/analytics/data-mining.html

Syaekhoni, M., Lee, C., & Kwon, Y. (2018). Analyzing customer behavior from shopping path data using operation edit distance. *Applied Intelligence*, 1912-1932.

Ullah, M. (2019). A model for predicting outfit sales: Using data mining methods. *International Conference on Emerging Technologies in Data Mining and Information Security, IEMIS 2018*, (págs. 711-720). Kolkata; India.

Usmani, Z. A. (2017). A predictive approach for improving the sales of products in e-commerce. *IEEE, Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference*, 188-192.

Wen, C.-H. L.-H.-F. (2018). Mining consumer knowledge from shopping experience: TV shopping industry. *International Arab Journal of Information Technology*, 1043-1051.

Zouzias, M., Vlachos, N., & Freris. (2012). *Unsupervised Sparse Matrix Co-clustering for Marketing and Sales Intelligence* . Obtenido de Springer: https://link.springer.com/chapter/10.1007/978-3-642-30217-6_49