UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

TEMA:

Elaboración de una propuesta tecnológica basada en algoritmos de aprendizaje automático para fidelización de clientes

AUTOR:

Ingrid Gabriela Barrezueta Flores

Trabajo de titulación previo a la obtención del título de
INGENIERO EN SITEMAS COMPUTACIONALES

TUTOR:

Ing. Colon Mario Celleri Mujica, Mgs.

2022

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación **ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA FIDELIZACIÓN DE CLIENTES** fue realizado en su totalidad por **Ingrid Gabriela Barrezueta Flores**, como requerimiento para la obtención del título de **Ingeniero en Sistemas Computacionales**.

TUTOR

f. _____
Ing. Colon Mario Celleri Mujica, Mgs.

Guayaquil, a los 15 días del mes de septiembre del año 2022

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Ingrid Gabriela Barrezueta Flores**

DECLARO QUE:

El Trabajo de Titulación, **ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIAJE AUTOMÁTICO PARA FIDELIZACIÓN DE CLIENTES** previo a la obtención del título de **Ingeniero en Sistemas Computacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

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

Guayaquil, a los 15 días del mes de septiembre del año 2022

EL AUTOR

f. _____

Ingrid Gabriela Barrezueta Flores

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

AUTORIZACIÓN

Yo, **Ingrid Gabriela Barrezueta Flores**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA FIDELIZACIÓN DE CLIENTES** cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 15 días del mes de septiembre del año 2022

EL AUTOR:

f. _____
Ingrid Gabriela Barrezueta Flores

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

UNIVERSIDAD CATÓLICA DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA INGENIERÍA EN SISTEMAS COMPUTACIONALES

f. _____

ING. ANA CAMACHO CORONEL, MGS

DIRECTORA DE CARRERA

f. _____

ING. EDISON TOALA QUIMI, MGS

DOCENTE DE LA CARRERA

f. _____

ING. FERNANDO CASTRO AGUILAR, PhD

OPONENTE

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERIA

CARRERA DE INGENIERIA EN SISTEMAS COMPUTACIONALES

REPORTE URKUND

URKUND

Documento	ProyectoTitulacion2022-Barrezueta.docx (D143267969)
Presentado	2022-08-25 16:42 (-05:00)
Presentado por	ingrid.barrezueta@cu.ucsg.edu.ec
Recibido	colon.celleri.ucsg@analysis.urkund.com
Mensaje	Tesis Mostrar el mensaje completo
	3% de estas 23 páginas, se componen de texto presente en 8 fuentes.

TUTOR

f. _____

Ing. Celleri Mujica Mario Colon, Mgs.

AGRADECIMIENTO

Quiero agradecer a todas las personas que me ayudaron a alcanzar este objetivo en mi vida profesional, especialmente quiero agradecer a José, Daniel, y Mercedes por todo su apoyo a lo largo de este nuevo logro en vida. Agradezco a la Universidad Católica de Santiago de Guayaquil y a mis docentes por la formación académica brindada. Agradezco a la empresa Lotería Nacional por permitirme llevar a cabo este proyecto. Finalmente quiero agradecer a mi familia por toda la fe y apoyo que pusieron en mí, sin ellos no habría sido posible lograr este objetivo en mi vida.

DEDICATORIA

Dedico el presente trabajo de titulación a mis padres Jenny y Franklin quienes con su esfuerzo y apoyo incondicional me permitieron cumplir cada uno de mis objetivos; gracias por su esfuerzo y por inculcar en mí valores que me permitieron hacer frente a los obstáculos y adversidades a lo largo de esta etapa en mi vida profesional y personal. A mis hermanos, John, Steven, Lissette, por ser parte de este proceso.

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE ANEXO

RESUMEN

El aprendizaje de maquina o aprendizaje automático se encuentra presente en diferentes industrias modernas, pero, es popularmente aplicado en el sector comercial para el análisis de clientes, su aplicación permite entre otros aspectos descubrir patrones en el comportamiento de clientes que las empresas pueden utilizar para aplicar estrategias comerciales, como retener o fidelizar clientes. El agrupamiento o clustering es una técnica muy utilizada en el aprendizaje automático para este tipo de análisis, se basa en la partición de un conjunto de datos en varios grupos en donde cada grupo contiene elementos similares entre sí y mantiene una diferencia respecto a los otros grupos. El presente trabajo de titulación tiene como objetivo obtener la segmentación de clientes de la empresa Lotería Nacional mediante la aplicación de algoritmos de aprendizaje automático, para ello se crearon variables que permitieron identificar el nivel de lealtad de los clientes de la empresa Lotería Nacional. Para el desarrollo del presente trabajo de titulación, se aplicó la metodología CRISP-DM que sirvió para el proceso de minería de datos. El análisis de los datos se lo realizó en base al modelo RFM (Recencia, Frecuencia, Monto) y sobre este modelo se aplicaron los algoritmos de agrupamiento k means, k nearest neigbor y árbol de decisión. Para validar el resultado de los algoritmos se separaron los datos para entrenamiento y pruebas que permitieron evaluar la precisión de los algoritmos, finalmente se utilizó la herramienta Power BI para presentar los resultados de una forma amigable y sencilla.

Palabras Clave: *Aprendizaje Automático, Clustering, Fidelización de Clientes, Modelo RFM.*

ABSTRACT

Machine learning is present in different modern industries, but it is popularly applied in the commercial sector for customer analysis, its application allows to discover patterns in customer behavior that companies can use to apply business strategies, such as retaining or building customer loyalty. Grouping or clustering is a machine learning technique for this type of analysis, it is based on the partition of a data set into several groups where each group contains elements like each other and maintains a difference with respect to the other groups. The objective of this titling work is to obtain the segmentation of the clients of the National Lottery company through the application of automatic learning algorithms, for which variables were created that allowed identifying the level of loyalty of the clients of the National Lottery company. For the development of this degree work, the CRISP-DM methodology was applied, which served for the data mining process. Data analysis was performed based on the RFM model (Recency, Frequency, Amount) and the k-means, k-nearest neighbor, and decision tree clustering algorithms were applied to this model. To validate the result of the algorithms, the data was separated for training and tests that allowed evaluating the accuracy of the algorithms, finally the Power BI tool was used to present the results in a friendly and simple way.

Key words: *Machine Learning, Clustering, Customer Loyalty, RFM model.*

INTRODUCCIÓN

Las empresas inteligentes utilizan su información de forma constante para generar conocimiento y cumplir sus objetivos estratégicos, esta información puede abarcar grandes volúmenes de datos, dentro de los cuales podemos descubrir información valiosa mediante técnicas que permiten explorar y explotar estos datos de manera automática. El aprendizaje automático se basa en una serie de técnicas usadas para el aprendizaje autónomo y no autónomo a partir de información suministrada, los algoritmos aprenden de los datos para poder realizar predicciones y descubrir patrones dentro de la información suministrada.

“Un algoritmo en minería de datos (o aprendizaje automático) es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias” (Microsoft, 2022).

La implementación de algoritmos de aprendizaje automático es ampliamente utilizado en diferentes industrias, pero la aplicación más común es en el análisis del comportamiento de clientes, especialmente del sector comercial. Las empresas recopilan gran cantidad de información de sus clientes, la inserción y ejecución de esta información a través de algoritmos de aprendizaje automático permite a las empresas conocer mejor al cliente y predecir hábitos de compras, tendencias del mercado, los productos populares, etc; lo que permite armar estrategias y tomar decisiones comerciales basadas en la información obtenida.

Basado en este contexto se plantea elaborar una propuesta tecnológica basada en la aplicación de algoritmos de aprendizaje automático para la clasificación de clientes de un tipo de negocio en particular dedicado a la venta de loterías y apuestas. Esta propuesta tiene como objetivo analizar y descubrir patrones e información de interés para armar estrategias comerciales acorde a las preferencias del cliente con el fin de aplicar un programa de fidelización.

En el presente trabajo de titulación se estructura los capítulos de la siguiente manera: Capitulo 1.- Se describe le problema, su ubicación, sus causas y consecuencias, se definen el objetivo general y objetivos

específicos, el alcance, la justificación y circunstancias; Capitulo 2.- Contiene el marco teórico, donde se argumenta los conceptos, normas, estándares, leyes y reglamentos que soportan la presente investigación; Capitulo 3.- Se describe la metodología de la investigación, se dimensiona la población y muestra y se especifica los instrumentos de recolección de datos; Capitulo 4.- se presenta la propuesta tecnológica, resultado del presente trabajo de titulación donde se detalla las herramientas utilizadas, las técnicas de procesamiento de datos y otros aspectos utilizados en el desarrollo; Conclusiones y recomendaciones. - Donde se muestran el resultado del trabajo de titulación y se da respuesta a los objetivos y propósitos planteados.

1 CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

EL PROBLEMA

1.1 UBICACIÓN DEL PROBLEMA EN UN CONTEXTO

La transformación digital conlleva hoy en día que las empresas adopten técnicas basadas en tecnología para mejorar su competitividad y productividad, la cantidad información que generan a lo largo del tiempo puede llegar a enormes lagos de datos, por lo que puede resultar complejo realizar un análisis y obtener conclusiones que puedan aplicarse de manera ágil a estrategias de negocio.

Las compañías suelen fijar un 10% como objetivo para mejorar, pero esto supone una innovación sostenida, es decir, se mejora los productos o servicios existentes, pero no se crea valor si no simplemente evoluciona los existentes. Con el tiempo esto impide aprovechar las oportunidades que tienen los negocios si se implementan mecanismos para el tratamiento de información para lograr verdaderos avances en beneficio del negocio (Cukier, 2015)

Un artículo publicado por WSI, una de las agencias digitales más grandes del mundo, menciona. “Mejorar la experiencia del cliente es una tendencia al alza, una empresa que no tiene clientes no puede persistir en el tiempo. Por ello, las empresas tienen el reto de conocer a sus clientes, los cuales se encuentran sometidos a un entorno cambiante” (WSI, 2021)

1.2 CAUSAS Y CONSECUENCIAS DEL PROBLEMA

De acuerdo con lo mencionado por un representante del área de marketing de la empresa Lotería Nacional, resulta complejo realizar un análisis de la información disponible de clientes ya que esta se encuentra dispersa en diferentes fuentes de datos y realizar cruces de información resulta una tarea compleja. La mayor parte de los reportes son realizados en archivos de Excel que son alimentados por otros archivos que se descargan desde un portal de reportaría interno de la empresa, por lo que mantener actualizada esta información resulta compleja, además la manipulación diaria de los archivos puede incurrir en la corrupción del mismo. Por otro lado, la

empresa lleva activa varios años y ha amasado una gran cantidad de información de diferentes características, al manejar la mayor parte de los reportes en archivos impide acceder a información de gran volumen y solo se limitan a analizar a un grupo de datos muy reducido o datos fraccionados por lotes, lo que impide tener una visión global de la información y se excluyen grupos que podrían resultar valiosos para el negocio. Como consecuencia, al depender de archivos e incurrir a la manualidad en los datos se puede caer en sesgos y perder agilidad en la toma de decisiones.

1.3 DELIMITACIÓN DEL PROBLEMA

El presente proyecto pretende, observar o medir el comportamiento de compras realizadas por clientes de Lotería Nacional en sus puntos de venta, mediante la aplicación de algoritmos de Aprendizaje Automático, los datos a considerar para el desarrollo del proyecto corresponden al periodo de enero 2018 a diciembre 2021.

1.4 FORMULACIÓN DEL PROBLEMA

Durante las últimas décadas, las grandes empresas han presentado necesidades complejas dentro de su estructura organizacional debido a la constante innovación y mejora que deben aplicar a sus procesos para sostener los diferentes niveles de su estructura organizativa a efectos de mantenerse competitivos dentro del mercado para brindar servicios o productos de calidad a sus clientes. Las nuevas formas de hacer negocios están relacionadas estrechamente con el continuo desarrollo e innovación de tecnologías emergentes, permitiendo que las empresas puedan aplicar modelos de toma de decisiones más precisas y orientadas a convertirse en organizaciones inteligentes.

La empresa mencionada anteriormente, cuenta varios puntos de venta distribuidos en varias ciudades, lleva activa en el mercado varios años, su actividad económica está centrada en la venta de loterías y apuestas, la empresa durante sus años de actividad ha amasado gran volumen de información de sus clientes y ha presentado problemas al momento de clasificar y analizar el comportamiento de compra de sus clientes para aplicar estrategias de fidelización.

Dada esta problemática se formula el problema de la siguiente manera:

¿La implementación de algoritmos de Aprendizaje automático ayudaría a la empresa a tomar mejores decisiones comerciales para fidelizar clientes?

1.5 EVALUACIÓN DEL PROBLEMA

Actualmente fidelizar clientes es tan importante como captar nuevos, incluso los costos e inversión para retenerlos pueden ser mucho menor a la inversión destinada para captar o generar nuevos. Al respecto (Sanchez, 2017) comenta, la fidelización de clientes ahorra gastos en marketing ya que un cliente que ha efectuado una compra inmediatamente conoce la marca, por lo que es probable que vuelva a comprar a diferencia de un cliente nuevo, el cliente habitual requiere de menos operaciones en el proceso de compra. Fidelizar clientes es una forma de garantizar ventas a largo plazo, ya que resulta más sencillo y barato conseguir que un cliente vuelva a comprar a que un cliente nuevo compre (Pierrend, 2020). Lograr retener clientes se ha convertido en un objetivo primordial para las empresas que desean mantenerse competitivas, para ello recurren a herramientas de analítica especializadas mediante las cuales se pueden conocer o descubrir información de sus clientes, información que no puede ser descubierta con herramientas tradicionales de analítica, especialmente si la información a analizar es de gran volumen.

El cliente de hoy no es el mismo de hace 10 años, las exigencias se dan al entorno actual altamente competitivo, por lo que esperan que conozcan a profundidad quienes son, sus preferencias y que cada experiencia sea personalizada. Esto lleva a las empresas a un escenario en el que necesitan comprender y planificar estrategias de fidelización a fin de mantener sus clientes. La influencia de la tecnología en el comportamiento de clientes contribuye a obtener información de manera rápida, pues al estar disponible de forma virtual es más accesible. Para que se genere la fidelización de clientes se deben aplicar tres acciones importantes sobre el uso de la tecnología, una de ellas es la recopilación y organización de los datos, que es la base del análisis para tomar acciones correctivas, como segunda acción es la implementación del programa de fidelización una vez que se hayan

identificado los clientes, sus necesidades e intereses se desarrollan estrategias de fidelización para lograr la lealtad de estos clientes. Y como tercera acción, se realiza seguimiento de los clientes, sus preferencias y comportamiento con el fin de lograr relaciones a largo plazo (Pierrend, 2020)

1.6 OBJETIVOS

1.6.1 OBJETIVO GENERAL

Elaborar una propuesta tecnológica para fidelización de clientes aplicando algoritmos de aprendizaje automático para la empresa Lotería Nacional.

1.6.2 OBJETIVOS ESPECÍFICOS

- Analizar y evaluar los datos y las variables a considerar de la información generada por los clientes de Lotería Nacional
- Desarrollar un prototipo algorítmico basado en técnicas de aprendizaje automático que identifique el nivel de lealtad de los clientes de la empresa Lotería Nacional
- Evaluar y comparar la precisión de los algoritmos desarrollados
- Desarrollar paneles de visualización para presentar los hallazgos obtenidos

1.7 ALCANCES DEL PROYECTO

El alcance del proyecto comprende el desarrollo de un modelo de aprendizaje automático prototipo que clasifica e identifica grupos de clientes, los resultados se presentarán en una interfaz web amigable mediante la cual se podrá visualizar los resultados obtenidos del algoritmo. No es parte del alcance de este proyecto la implementación del prototipo en ambiente producción.

1.8 JUSTIFICACION E IMPORTANCIA

La aplicación de algoritmos de aprendizaje automático tiene un impacto positivo dentro de las empresas que deciden incorporarlos a su núcleo para motivos estratégicos, su implementación ofrece beneficios, tales como, mejor posicionamiento del negocio en el mercado, capacidad de descubrir patrones y correlaciones, clasificar y personalizar actividades de clientes, aumentar la

participación del cliente, disminuir costos que se pueden traducir en aumento de ingresos para la empresa.

El enfoque de los negocios inteligentes se basa en la toma de decisiones en base a sus datos, el análisis de datos mediante algoritmos automatizados genera conocimiento que impulsa al crecimiento del negocio a partir de los resultados obtenidos, utilizando estratégicamente estos resultados las empresas pueden obtener una ventaja competitiva sobre sus pares en el mercado.

Por ello el desarrollo de una propuesta basada en algoritmos de aprendizaje automático se considera que podría resultar de gran ayuda para la toma de decisiones de la empresa.

1.9 HIPÓTESIS O PREGUNTA DE INVESTIGACIÓN

Para el proyecto en cuestión, se plantea la siguiente pregunta de investigación.

¿La implementación de algoritmos de Aprendizaje automático ayudaría a la empresa a tomar mejores decisiones comerciales para fidelizar clientes?

2 CAPÍTULO II

MARCO TEÓRICO

2.1 FIDELIZACIÓN DE CLIENTES

La palabra fidelizar puede contener varios significados según el contexto en el que se la aplique. La Real Academia Española (RAE), define fidelizar como “Conseguir de diferentes modos, que los empleados y clientes de una empresa permanezcan fieles a ella" (RAE, 2021), basándonos en este contexto. La fidelización es una estrategia o método que aplican las empresas para mantener a clientes fieles al negocio.

2.2 TRANSFORMACIÓN DIGITAL EN LAS EMPRESAS

El termino transformación digital tiene sus inicios en el siglo XVIII, con el inicio de la mecanización del trabajo denominada revolución industrial 1.0, este fue el primer paso a lo que hoy en día conocemos como transformación digital, e esta época aparecieron las primeras máquinas de vapor que lograban multiplicar la productividad en las fábricas y aumentar la velocidad en la distribución de productos. Luego en la industria 2.0 marcó varios aspectos, quizá el más importante es el que marcó Henry Ford con las líneas de montaje a gran escala en las fábricas, lo que supuso un gran avance ya que permitió trabajar en cadena a bajo costo. Además, otros factores importantes como el uso industrial de la electricidad. Tiempo después surge la industria 3.0 allá por los años 70, fue entonces en esta época donde las grandes corporaciones informáticas aparecieron en el mercado con la comercialización y diversificación de los computadores y herramientas de software. Lo cual día paso a la Industria 4.0, la transformación digital de la economía y con ello la transformación de los negocios. La transformación digital es un conjunto de tecnologías que se integran a todas las áreas de la empresa, cambiando profundamente la forma en la que opera y brinda valor al cliente. “Es una transformación a nivel corporativo mediante nuevas operaciones y modelos digitales que agregan valor a negocios corporativos, mejorando la productividad y rentabilidad de las operaciones corporativas” (Schallmo, 2017)

“Proceso evolutivo que aumentan las capacidades digitales para generar nuevos modelos de negocios o experiencias de clientes que agregan mayor valor” (Morakanyane, 2017)

2.3 APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es una rama de la inteligencia artificial, que como su nombre lo indica, permite que las maquinas aprendan de manera autónoma, en este aspecto aplicar aprendizaje automático es una habilidad indispensable para desarrollar sistemas capaces de predecir e identificar patrones en los datos, esta tecnología se encuentra en varias aplicaciones modernas, tales como el correo, plataforma de streaming, redes sociales. “Un algoritmo de aprendizaje automático o de minería de datos, es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias” (Microsoft, 2022)

2.4 TIPOS DE APRENDIZAJE AUTOMÁTICO

En la literatura podemos encontrar que existen diferentes tipos de aprendizaje automático, los más comunes se nombran a continuación.

2.4.1 Aprendizaje Automático Supervisado

Este tipo de aprendizaje se caracteriza por entrenar el algoritmo dándole atributos (características) y las respuestas (etiquetas) para que así en un futuro el algoritmo pueda hacer una predicción conociendo las reglas ya antes ingresadas (Sandoval, 2018). En este tipo de aprendizaje existen dos tipos de algoritmos.

TABLA I. ALGORITMOS DE APRENDIZAJE SUPERVISADO

Aprendizaje Supervisado	Algoritmo de Clasificación
	Algoritmo de Regresión

Tabla 1.- *Algoritmos de Aprendizaje Supervisado*

2.4.2 Aprendizaje Automático no Supervisado

En este tipo de aprendizaje no existe ninguna categorización o etiquetado de los datos, sólo se ingresan las características, se aplica este tipo de aprendizaje cuando queremos que se agrupen los datos según sus características, el algoritmo solo sabe que los datos comparten ciertas características por lo que busca similitudes y asume que puedan pertenecer a un mismo grupo. (Russell, 2018)

TABLA II. ALGORITMOS MAS COMUNES DEL APRENDIZAJE NO SUPERVISADO

Aprendizaje Automático No Supervisado	Agrupamiento,Clustering: K – Means, KNN
	Arboles de decisión, Cluster jerárquico
	Density Based Scan Clustering (DBSCAN)
	Modelo de Agrupamiento Gaussiano

Tabla 2.- *Algoritmos más comunes del aprendizaje no supervisado*

Supongamos que tenemos muchos datos dispersos de visitantes de una página web dedicada a la venta de artículos de tecnología, si estos datos los ingresamos en un algoritmo de agrupamiento con el objetivo de detectar grupos de visitantes similares, se podría identificar, por ejemplo, que el 65% de los visitantes son hombres mayores de 18 años les interesa adquirir consolas de video juegos y el 30% les interesa los juegos de acción. En este caso, el algoritmo de agrupamiento dividirá a cada grupo en subgrupos de características similares. Para abordar este caso se necesita de la combinación de varias características relacionadas a una sola característica, por ejemplo, la combinación de un computador con su modelo. A esta técnica se le denomina extracción de características.

2.4.3 Aprendizaje automático por refuerzo

En este tipo de aprendizaje, se implementa un agente que sondea un espacio desconocido y determinará las acciones que deberían llevarse a cabo, mediante técnicas de prueba y error, aprenderá de manera autónoma cada vez que se le recompense o penalice sus acciones. El agente deberá actuar o crear estrategias de la mejor forma posible para obtener la mayor recompensa en tiempo y forma. (Torres, 2021)

Es importante acotar que tanto el aprendizaje no supervisado como el aprendizaje por refuerzo requiere que el entrenamiento sea desarrollado por un humano, en el caso del aprendizaje no supervisado se definen los objetivos que se quieren lograr cómo, por ejemplo, obtener la agrupación correcta de imágenes, por otro lado, en el aprendizaje por refuerzo se definen recompensas o penalizaciones dependiendo del comportamiento como, por ejemplo, obtener una puntuación en un juego según las acciones tomadas. Si bien es cierto que la persona que desarrolla el entrenamiento no interviene en el proceso de aprendizaje, sí definen las reglas y límites del aprendizaje en ambos casos.

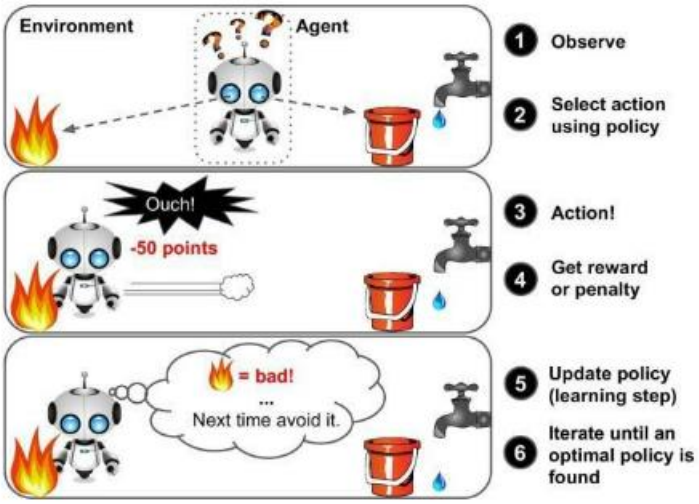


Ilustración 1 .- *Aprendizaje por refuerzo*

2.4.4 Aprendizaje automático por lote

El aprendizaje por lote es un sistema que no puede aprender de manera paulatina, al sistema se deben ingresar todos los datos necesarios, por lo que la forma en la que trabaja este tipo de aprendizaje es de manera offline ya que necesita una gran cantidad de recursos y tiempo para su ejecución. Para trabajar en este tipo de aprendizaje, primero hay que capacitar al sistema y luego ejecutarlo (Russell, 2018)

2.4.5 Aprendizaje automático en línea

Este tipo de aprendizaje es lo opuesto al aprendizaje por lote, puede aprender de manera paulatina al ingresar todos los datos como instancias ya sean como grupo o conjunto de datos o datos individuales, es decir, el sistema puede ir aprendiendo conforme se vayan suministrando datos, como su

nombre lo indica “en línea”, no necesita trabajar de manera offline (Russell, 2018). Se puede usar este tipo de aprendizaje automático para casos en los que se requiera de flujo de datos continuos los cuales necesiten adaptarse rápidamente a nuevos cambios. Además, este tipo de aprendizaje soporta grandes cantidades de datos. Para implementar este tipo de aprendizaje se debe saber que tan rápido el sistema pueda adaptarse a cualquier cambio.

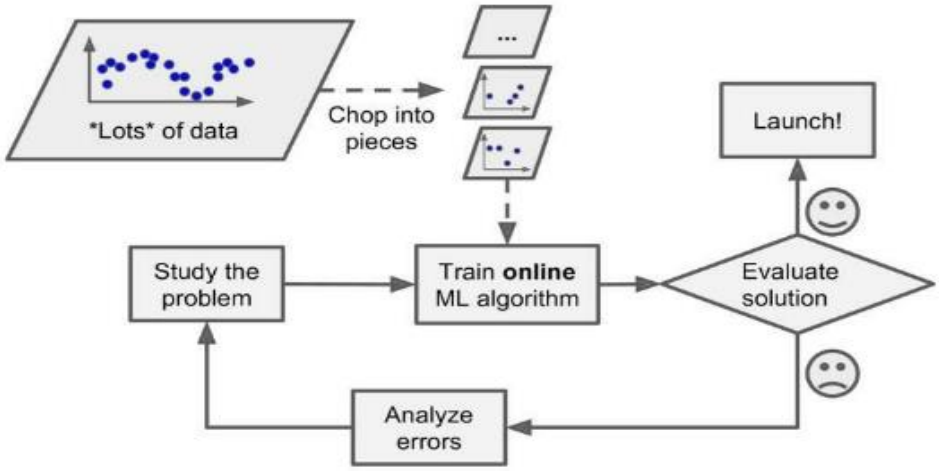


Ilustración 2 .- *Aprendizaje en Línea*

2.5 CLUSTERING O AGRUPAMIENTO

Permite identificar segmentos o grupos donde sus elementos comparten características similares. Es un proceso que se utiliza para moldear o revelar estructuras y patrones ocultos en un conjunto de datos, es como una búsqueda de semejanzas y diferencias en los datos, el resultado es la división de diferentes conjuntos de datos en diferentes conglomerados o clases. (Kovera, 2017) Un ejemplo común que se suele abordar para explicar de manera sencilla es la clasificación de frutas, estas se pueden agrupar en frutas maduras, algo maduras y verdes según el color y la dureza que observamos.

2.6 ALGORITMOS JERÁRQUICOS

Los algoritmos jerárquicos se utilizan cuando se requiere agrupar datos de los cuales se desconoce su estructura interna, en otras palabras, no existe conocimiento acerca de las etiquetas de la clase a la que pertenecen. Trabajan uniendo o separando en cada iteración el par de grupos mas semejantes. Estos algoritmos producen una secuencia enlazada a las particiones del conjunto de datos, es decir la estructura de los grupos se

organizan de manera jerárquica y cada clúster (grupo) puede verse como la unión de otros clústeres (grupo), obteniendo diferentes niveles de jerarquía de grupos. Tradicionalmente se muestra esta organización en forma de un dendograma, el cual proporciona una clasificación de manera jerárquica de la información procesada. (EcuRED, n.d.)

Aglomerativos: Este tipo de algoritmo se basa en distancias empiezan considerando a cada elemento como un grupo individual y cada iteración se une a los grupos más cercanos hasta que se obtenga un único grupo o se cumpla el criterio de parada. Los algoritmos más comunes de este tipo son Single Link y Cure.

Divisos: Este tipo de algoritmo comienza considerando el conjunto de elementos como un único grupo y en cada iteración se particionan en dos hasta que queden varios grupos como objetos individuales o hasta que se cumpla el criterio de parada.

Criterio de parada: Los criterios de parada más comunes en la aplicación de algoritmos jerárquicos son: cuando se obtengan “n” grupos o cuando la distancia entre el par de grupos más próximos sea mayor a un umbral definido

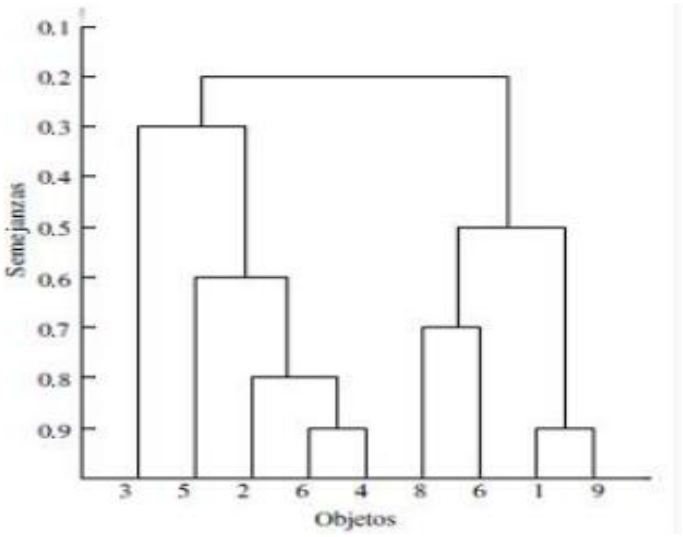


Ilustración 3 .- *Ejemplo Dendograma; Algoritmo Jerárquico*

2.7 ALGORITMOS DE PARTICIONAMIENTO

Son utilizados para la clasificación de individuos (no de variables) en K grupos. Se selecciona una partición de los individuos en K grupos e intercambiar los elementos entres clústeres para obtener una partición mejor (Chamba Jimenez , 2015). Los algoritmos más comunes son los siguientes k-means, k-medoids, Self-Organizing Maps SOM.

2.8 ALGORITMOS DE CLASIFICACIÓN

Como su nombre lo indica, los algoritmos de clasificación clasifican los datos en diferentes grupos, el objetivo de la implementación de este algoritmo es saber a qué grupo pertenece el elemento de estudio. El algoritmo encuentra patrones en los datos suministrados y los clasifica en grupos, cuando se ingresan nuevos datos los compara y los ubica en uno de los grupos, es así como el algoritmo puede predecir de los datos y separar por grupos. Las variables pueden ser tipo categórico o discreto. Pueden ser.

TABLA II. TIPOS DE VARIABLES

Binaria	{Si,No}, {Verde,Rojo} .. etc
Múltiple	Compara {Producto1..Producto2}
Ordenada	Riesgo. {Alto,Medio,Bajo}

Tabla 3 .- *Tipos de variables - Algoritmos de clasificación*

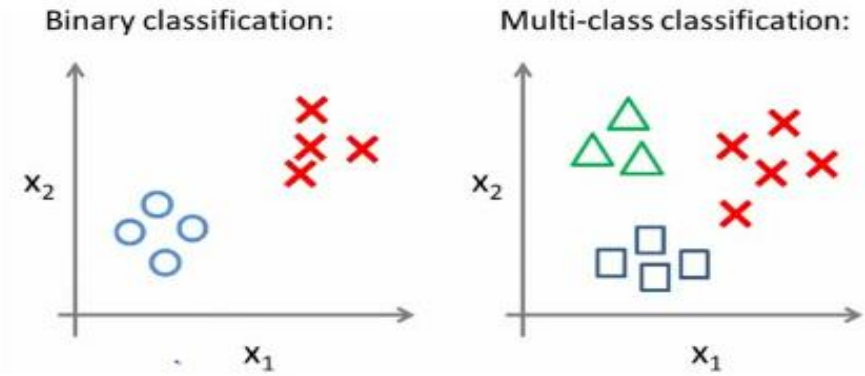


Ilustración 4 .- *Ej. Algoritmo de clasificación*

2.9 ALGORITMOS DE REGRESIÓN

En este método se espera como resultado un número, no lo ubica en grupos, sino que devuelve un valor especifico. Se trata de un algoritmo que se utiliza en técnicas aprendizaje automático y estadística, esta técnica establece una recta para proporcionar la tendencia en un conjunto de datos.

Cuando se realiza regresión, esperamos como resultado un número, es decir el resultado de la técnica de aprendizaje será un valor numérico dentro de un conjunto infinito de posibles resultados (Heras, 2020). Algunos ejemplos de casos de uso son los siguientes:

- Predecir el valor de una bien inmueble
- Predecir posible deserción de clientes o empleados
- Estimar el tiempo en el que un cliente vuelva a comprar algún articulo

Para llegar a tales resultados, al algoritmo se le suministra información referente al caso de estudio que se requiera analizar, y por medio de un gráfico de dispersión puede predecir el elemento de estudio.

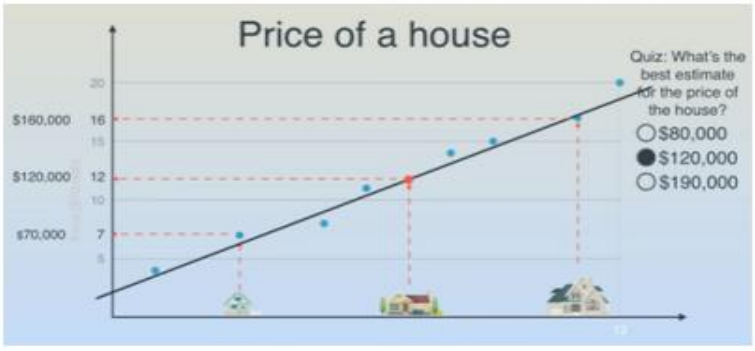


Ilustración 5 .- *Ej. de gráfico de dispersión en algoritmo de regresión*

2.10 ALGORITMO K-MEANS

El algoritmo K-Means es uno de los algoritmos de clustering más usados, este algoritmo pertenece a los algoritmos de partición, en el nombre K se refiere al número de grupos a diferencia de los algoritmos jerárquicos en este se especifica el número de K grupos. (Kovera, 2017). En el libro guía publicado por Artem Kovera “Machine learning with clustering”, menciona los siguientes pasos a seguir para la implementación de K-means.

1. Colocar K centroides (grupos) en ubicaciones aleatorias, el número de conglomerados que se tendrá como resultado final es igual al número de centroides.
2. Para cada centroide seleccione el más cercano, para esto se calcula la distancia euclidiana y se asigna al grupo cuyo centro sea el más cercano o próximo.
3. Recalcular las posiciones de los centroides.
4. Reasignar los elementos a los centroides más cercanos.

5. Repetir los pasos 3 y 4 hasta que los centroides ya no se muevan.

2.11 ÁRBOLES DE DECISIÓN

Los algoritmos basados en árboles son los más utilizados en el aprendizaje automático, son algoritmos de aprendizaje no supervisado. Los usamos para resolver problemas de clasificación o regresión. En los modelos de clasificación se predice un valor de una variable a partir de otras variables. Por ejemplo, cuando queremos pronosticar si un cliente comprará un producto de una determinada marca, agrupándolo mediante la clasificación entre marcas. Por otro lado, en los modelos de regresión se predice los valores de las variables en función de variables que son independientes entre sí, por ejemplo, cuando queremos predecir el valor de un terreno en función de variables como localización, superficie, distancia a la playa. (Kovera, 2017)

Los árboles de decisión son estructuras formadas por ramas y nodos de distintos tipos:

- **Nodos internos**: Son las características que se consideran para tomar las decisiones.
- **Ramas:** Representan la decisión tomada dependiendo de cada condición.
- **Nodos finales:** Representa el resultado de la decisión tomada.

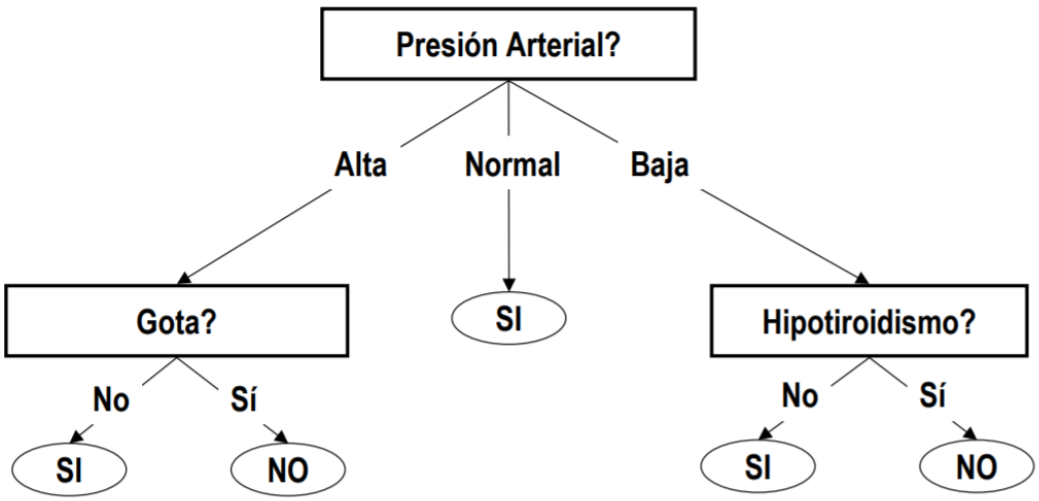


Ilustración 6 .- *Ej. Árbol de decisión*

2.12 ALGORITMO K NEAREST NEIGBOR (K VECINOS MÁS CERCANOS)

K nearest neigbor (KNN) es un algoritmo de aprendizaje no supervisado, donde “K” significa la cantidad de punto vecinos más cercanos para clasificar los n grupos que ya se conocen previamente (Reed, 2020). Se divide en las siguientes etapas:

- Selecciona el número de K vecinos más cercanos.
- Calcula la distancia entre elementos, la más usada es la distancia euclidiana.
- Toma laos K vecinos más cercanos según las distancias calculadas.
- Con los K vecinos calculados, contamos el número de puntos en cada categoría.
- Integrar el punto a la categoría más representativa entre los K vecinos.
- El modelo está listo.

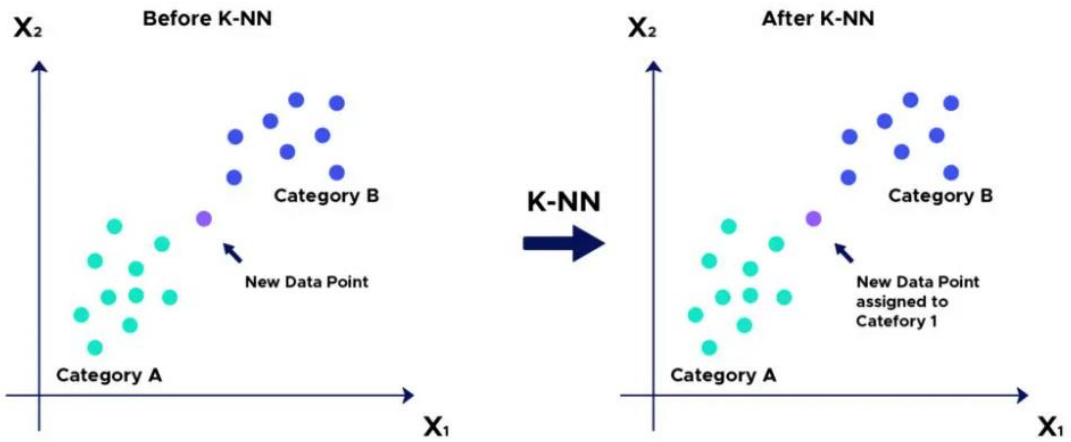


Ilustración 7 .- *Ej. Algoritmo K nearest neigbor*

2.13 ANÁLISIS RFM

RFM (Recencia, Frecuencia, Monto) es una técnica de mercadeo que es usada para estudiar el comportamiento de clientes de forma que se examina lo que el cliente ha comprado, utilizando 3 valores (R) recencia de compra (F) frecuencia de compra (M) Monto total de compras como valor monetario. Esta técnica se basa en la ley de Pareto o también conocida como la regla del 80/20, la ley de Pareto es usada en infinidad de situaciones, en el caso de

RFM se podría decir que el 80% de las ventas las generan el 20% de clientes, aunque parezca exagerado es perfectamente comprobable en la mayoría de los negocios. (Chamba Jimenez , 2015) .

Para aplicar la técnica RFM, a cada cliente se le asigna un puntaje del 1 al 4, calificándolos bajo los indicadores antes mencionados (Recencia, Frecuencia y Monto), los clientes que obtengan puntajes de 4-4-4 en un periodo determinado, serán los clientes ideales para el negocio.

2.14 METODOLOGÍAS DE MINERIA DE DATOS

2.14.1 Metodología CRISP-DM

La metodología CRISP-DM es ampliamente usada para el desarrollo de proyectos de minería de datos o proyectos que involucre análisis de datos, actualmente IBM es el principal promotor de esta metodología, aunque también es muy usada en otras empresas y medios académicos, se considera como el principal promotor a IBM ya que incorporó la metodología CRISP-DM a uno de sus productos SPSS. (Espinosa, 2020).

La metodología CRISP-DM contiene las siguientes fases:

Comprensión del negocio: En esta fase se realiza un levantamiento de información de los objetivos de la empresa y sus necesidades y se enfoca en el planteamiento de los objetivos y exigencias del proyecto que posteriormente se diseña un plan preliminar para lograr los esos objetivos.

Comprensión de los datos: En esta fase se recolecta los datos necesarios para el proyecto, se realiza una exploración de los datos para comprender su naturaleza, calidad y validez para el análisis.

Preparación de los datos: Se realizan las actividades necesarias para construir el conjunto de datos final que será analizado o entrenado a través de herramientas de Aprendizaje de máquina, este proceso es el que toma más tiempo, incluso puede representar un 80% del tiempo invertido en el proyecto ya que incluye la selección, limpieza y creación de nuevas variables e integración de datos y formateo de los datos.

Modelado: En esta fase de obtiene el modelo de minería de datos, selección de la técnica de minería de datos a usar, selección de los datos de prueba y obtención del modelo.

Evaluación del Modelo: En esta fase se determina la calidad del modelo y se compara con resultados previos o se analiza con expertos que dominan el problema, si los resultados de esta fase son satisfactorios se da paso a la siguiente fase.

Implementación del modelo: En esta fase se implementa el modelo una vez ya validado, se explota y se muestra los resultados de manera clara obtenido de las fases anteriores.

3 CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1 METODOLOGÍA

El tipo de metodología de investigación que se siguió es de enfoque cuantitativo ya que las características del presente trabajo de investigación se desarrollaron siguiendo varios pasos de manera secuencial, con el objetivo de medir fenómenos en base a un análisis y tratamiento previo a los datos proporcionados por la empresa Lotería Nacional, qué mediante fundamentos estadísticos se interpretaron resultados con cierto grado de precisión.

En el libro Metodología de la Investigación 6ta edición de Sampieri, menciona que el enfoque cuantitativo de una investigación representa un conjunto de procesos secuenciales y probatorios en el que en cada etapa precede a la siguiente y no podemos eludir pasos ya que cada paso depende del que le antecede de estos se extraen variables en un determinado contexto y se analizan la mediciones obtenidas mediante métodos estadísticos y se sustrae una serie de conclusiones (Hérnandez Sampieri , 2014).

3.2 Métodos y Técnicas

Técnica de Exploración: Para conocer e identificar las variables a considerar para la aplicación del proyecto, exploración de los datos proporcionados por la empresa.

Técnica de observación y entrevista: Para conocer la situación de la empresa se usaron técnicas de observación.

Método inductivo: Permitió comprender los objetivos del negocio mediante los cuales se pudo identificar la necesidad de crear estrategias de fidelización de clientes.

Método deductivo: Se utilizó para detectar problemas o circunstancias que no permiten a la empresa aplicar estrategias de fidelización.

Método científico: Se utilizó el método científico para recopilar información en base a casos de éxito de minería de datos y aprendizaje automático aplicados en empresas.

3.3 Población y Muestra

Lotería Nacional es una empresa muy reconocida en en el mercado ecuatoriano, su actividad económica corresponde a la venta de loterías y apuestas, su expansión geográfica le ha permitido llegar a varios cantones y provincias del Ecuador por lo que su población abarca diferentes lugares del país. Durante sus años de actividad ha logrado amasar una base de datos de clientes de alrededor 5536 registros de clientes que frecuentemente compran productos, a estos clientes se les denomina Loteros. De los cuales se tomará como muestra los clientes que realizaron compras en los puntos de venta distribuidos en diferentes provincias del país, correspondientes al periodo de enero 2018 a Diciembre 2021.

TABLA IV. NUMERO TRANSACCIONES 2018 - 2021

Año	**#Trans**
2018	3.060.037
2019	4.185.961
2020	3.140.345
2021	5.272.282
Total	**15.658.625**

Tabla 4 .- *Número transacciones 2018 - 2021*

4 CAPÍTULO IV

PROPUESTA TECNOLOGICA

4.1 METODOLOGÍA

La metodología que se usó en el presente proyecto es la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), esta metodología se compone de cinco fases: Muestreo, Exploración, Modificación, Modelado y Valoración.

4.2 HERRAMIENTAS DE DESARROLLO

Sql Server: Gestor de base datos donde se almacenará la información proporcionada por la empresa.

Lenguaje Python: Es un lenguaje de programación de código abierto que es ampliamente utilizado para análisis de datos y aprendizaje automático.

Power BI: Es una herramienta mediante la cual se pueden generar tableros de visualización para reportes en la nube.

4.3 ANÁLISIS DE DATOS

4.3.1 TÉCNICAS PARA EL PROCESAMIENTO Y ANÁLISIS DE DATOS

4.3.1.1 Examinar la información existente.

Esta tarea corresponde a una de las fases de la metodología CRISP-DM y abarca las tareas de comprensión de los datos, preparación de los datos y modelado de los datos.

4.3.1.2 Comprensión de los datos

La fase de comprensión de los datos abarca diferentes actividades, tales como recopilación, exploración y verificación de la calidad de los datos.

4.3.1.3 Recopilación de datos Iniciales

El primer paso de este proceso fue la adquisición de los datos necesarios para llevar a cabo el presente proyecto, los datos proporcionados corresponden al registro de cliente y sus transacciones realizadas desde el

año 2018 hasta el año 2021. Los datos recopilados para el proyecto se los han categorizado de la siguiente manera.

Clientes: Contiene información personal de los clientes, como Fecha de Nacimiento, Genero, estado civil, tipo de cliente, medios de contacto y código único del cliente.

Transacciones: Abarca los registros de facturación en ventas mensuales realizadas por los clientes, nos proporciona información detallada de las compras realizadas durante el periodo 2018 – 2021.

Productos: Contiene información de los grupos, subgrupos y precios de los productos comercializados por la empresa.

Puntos de Venta: Contiene información geográfica de los puntos de venta distribuidos en diferentes ciudades.

Esta información fue generada en archivos de Excel por lo que se creó un proceso en sql para cargar la información en una base de datos que en este caso se ha elegido como gestor de base de datos Sql Server.

4.3.1.4 Descripción de los datos

El gestor de base de datos que maneja la empresa es Sql Server. Por lo que se usó la herramienta Sql Magnament Studio para la exploración de los datos. A continuación, se describen las tablas.

Tabla de cliente: Contiene información personal de los clientes de la empresa, la tabla cuenta con 197.415 registros, a continuación, se describen los atributos a los cuales se tuvo acceso.

TABLA V. REGISTRO DE CLIENTES QUE COMPRARON DURANTE 2018 - 2021

Atributos	Descripción	Tipo
Id Cliente	Código de identificación del cliente	bigint
Fecha Nacimiento	Fecha de nacimiento del cliente	Date
Tipo Cliente	Cliente Final o Lotero	Varchar(10)

Tabla 5 .- *Tabla de Clientes*

Tabla Transacciones: Contiene toda la información de la facturación en venta realizada por la empresa en periodo Enero 2018 – Diciembre 2021.

Cuenta con un total de 15.658.625 registros de facturas. A continuación, se describen los atributos a los cuales se tuvo acceso.

TABLA VI. TABLA DE REGISTROS TRANSACCIONALES

Atributos	Descripción	Tipo
Codigo Factura	Código de identificación único de la factura	Varchar(200)
Fecha	Fecha de emisión de la factura	Datetime
Id Punto Operacion	Identificador único del punto de venta	Bigint
Id Cliente	Identificador único del cliente que realizó la compra	Bigint
Id Producto	Identificador único del producto vendido	Int
Id Sorteo	Identificador único del sorteo vendido	bigint
Subtotal	Valor a pargar antes de impuestos	Money
Impuesto	El impuesto a pagar en la compra	Money
Total	Valor total a pagar	Money

Tabla 6 .- *Tabla de registros transaccionales*

Tabla Puntos de Venta: Contiene toda la información referente a los puntos de ventas de la empresa, cuenta con 1237 registros. Se detalla a continuación los campos a los cuales se tuvo acceso.

TABLA VII. TABLA DE PUNTOS DE VENTA

Atributos	Descripción	Tipo
Id Punto Operacion	Código de identificación único del punto de venta	bigint
Punto Operacion	Nombre del punto de venta	Varchar(200)
Ciudad	Ciudad de donde se encuentra ubicado el punto de venta	Varchar(50)
Region	Region donde ha sido categorizado el punto de venta	Varchar(20)
Provincia	Provincia de donde pertenece el punto de venta	Varchar(50)

Tabla 7 .- *Tabla de Puntos de venta*

Tabla de Productos: Contiene información de los productos y sorteos que se han vendido durante el periodo 2018 - 2021, cuenta con 68 registros. Se detalla a continuación los campos a los cuales se tuvo acceso.

TABLA VIII. TABLA DE PRODUCTOS

Atributos	Descripción	Tipo
Id Producto	Código de identificación único del Producto	int
Producto	Nombre del Producto	Varchar(50)
Id Tipo Sorteo	Código de identificación único del tipo de sorteo	int
Tipo Sorteo	El tipo de sorteo al que pertenece el producto	Varchar(50)
Precio	Precio del producto por tipo de sorteo	float

Tabla 8 .- *Tabla de Productos*

4.3.1.5 Modelo Entidad Relación

Luego de realizar un proceso de normalización de los datos, el modelo entidad relación queda de la siguiente manera.

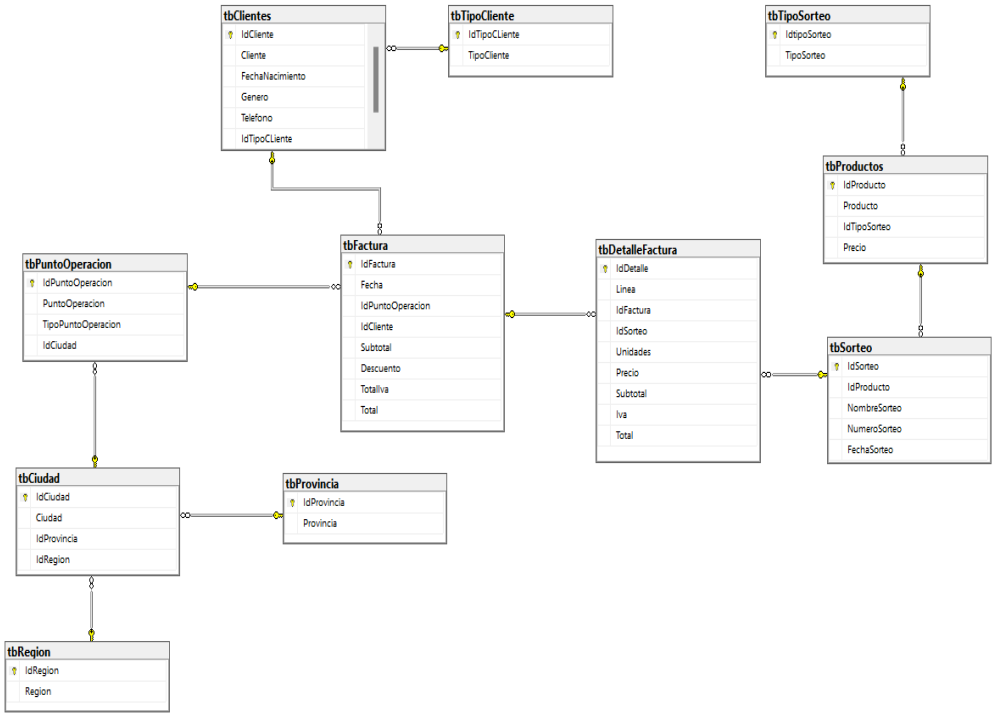


Ilustración 8 .- *Esquema de base de datos*

4.4 EXPLORACIÓN DE LOS DATOS

Considerando que la empresa desea descubrir los niveles de lealtad de sus clientes para aplicar estrategias de fidelización, se ha realizado un análisis de los datos transaccionales para tener una idea de cómo se encuentra la actividad de compra de los clientes de la empresa.

- **Ventas realizadas en los últimos 4 años**

El diagrama de barras de la siguiente figura (Ver figura 3) indica que en el año 2019 se incrementó el número de ventas respecto al número de ventas del año 2018, el año 2020 se ve un leve incremento respecto al año 2018, pero un decremente importante respecto a 2019 ya que debido a la pandemia mundial COVID-19 no se pudo lograr los objetivos comerciales para ese año. Sin embargo, para 2021 se observa un incremento considerable respecto a los años anteriores.

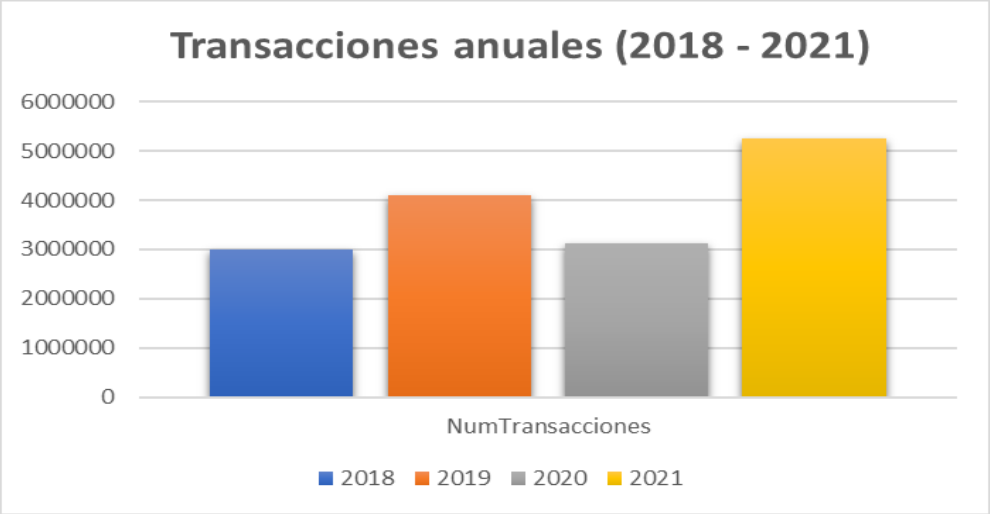


Ilustración 9 .- *Transacciones anuales*

- **Monto en dólares de ventas de los últimos 4 años**

El diagrama de barras de la siguiente figura (Ver figura 4) indica que a pesar de que el año 2018 (Ver figura 3) tuvo un menor número de transacciones respecto a los otros años, el monto en dólares del año 2018 supera al resto de años. Como se puede observar existe un decrecimiento en las ventas en 2021 a pesar de que este año tuvo un mayor número de transacciones.

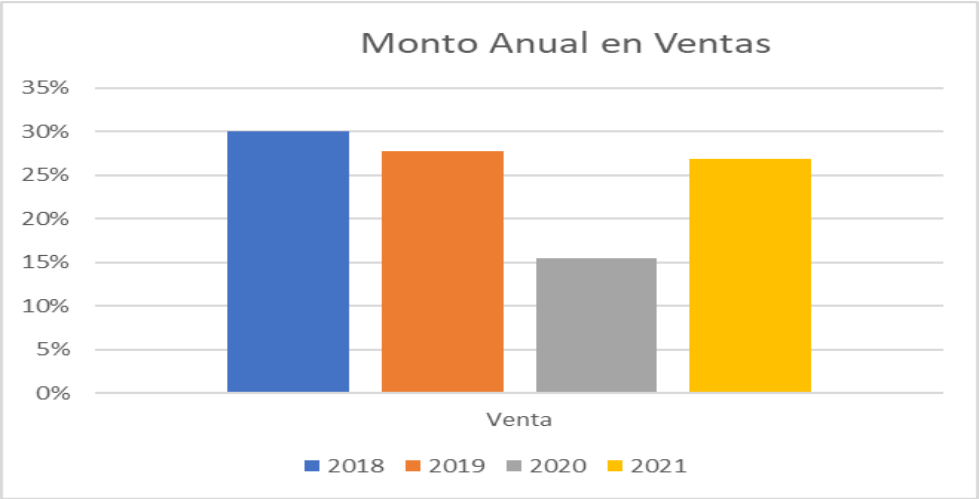


Ilustración 10 .- *Monto Anual en Ventas*

Durante el año 2021 la empresa dio más impulso a la venta de boletos de $0.50 ctvs. de dólar y boletos de $1 dólar, puesto que debido a la recesión derivada de los efectos de la pandemia COVID-19, la empresa se vio en la necesidad de realizar ajustes en sus precios.

TABLA IX. TABLA COMPARATIVA DE PRECIOS 2018 - 2021

Precio	2018	2019	2020	2021	Total
0.25	0.97%	0.90%	0.55%	0.18%	2.59%
0.50	0.16%	0.89%	0.90%	1.41%	3.36%
1.00	21.31%	18.76%	9.89%	22.18%	72.14%
1.75		0.04%	0.04%	0.09%	0.17%
2.00	5.50%	5.12%	3.94%	3.48%	18.04%
3.00	1.85%	1.86%			3.70%
Total	**29.78%**	**27.56%**	**15.32%**	**27.33%**	**100.00%**

Tabla 9 .- *Tabla comparativa de precios 2018 - 2021*

- **Número de Clientes por lugar geográfico**

La mayor concentración de clientes de Lotería Nacional se da en las provincias de Guayas, Pichincha y Manabí. El 36% del total de clientes, se concentra en la provincia del Guayas, seguido del 28% en Pichincha, el 17% en la provincia de Manabí y el 19% restante repartido en otras provincias.

TABLA X. PORCENTAJE DE CLIENTES POR PROVINCIA

Provincia	Porcentaje
Guayas	42%
Pichincha	17%
Manabí	11%

Otras Provincias	30%

Tabla 10 .- *Porcentaje de clientes por provincia*

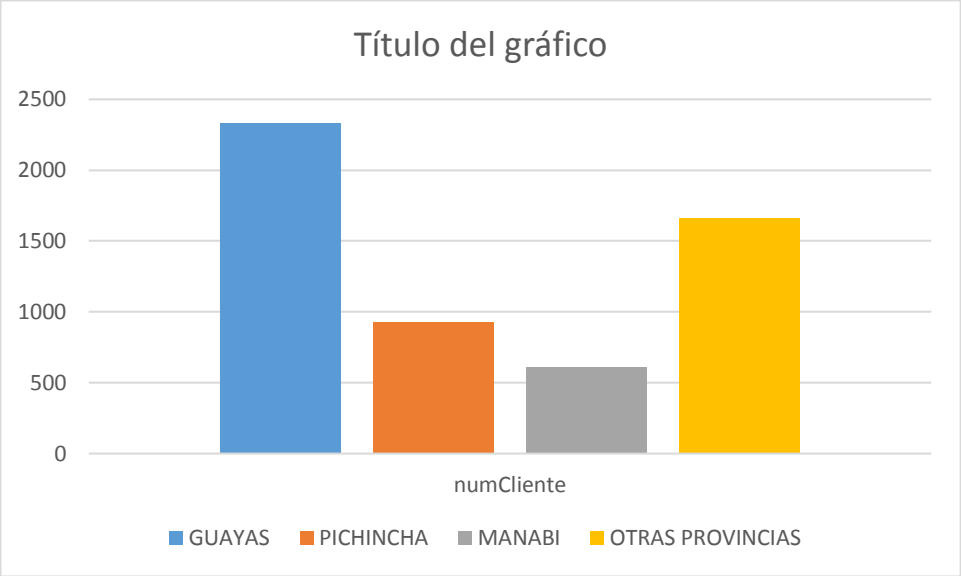


Ilustración 11 .- *Histograma número de clientes*

4.5 PREPARACIÓN DE LOS DATOS

La preparación de los datos abarca procesos de limpieza, selección, construcción de nuevos datos, integración y dar formato a los datos.

4.5.1 Selección de los datos

Los datos se seleccionaron en base a los objetivos del proyecto en cuestión, en el que se plantea elaborar una propuesta tecnológica para fidelización de clientes de Lotería Nacional, con el objetivo de poder realizar estrategias de fidelización de clientes.

Se ha seleccionado los siguientes datos de la tabla Facturas, CodigoTransaccion, IdCliente, Fecha, Total. De los 15.658.625 registros se seleccionaron las transacciones correspondientes a facturas emitidas y se excluyen las transacciones registradas como notas de crédito, quedando un total de 15.463.155

Se han seleccionado los siguientes atributos de la tabla de Clientes. IdCliente, Fecha de Nacimiento, Tipo de Cliente. De estos clientes se han seleccionado aquellos que poseen transacciones entre Enero 2018 a Diciembre 2021, quedando un total de 5.536 registros.

TABLA XI. TAMAÑO INICIAL DEL CONJUNTO DE DATOS SELECCIONADO

	Datos Lotería Nacional
Años a analizar	4 años (2018 a 2021)
Número de clientes	5.536
Número de Transacciones	15.463.155

Tabla 11 .- *Tamaño inicial del conjunto de datos seleccionado*

4.5.2 Limpieza de datos

Algunas tareas de limpieza de datos se realizaron mediante el lenguaje SQL, en la herramienta SQL Server Management Studio, se crearon procesos almacenados para importar información desde archivos csv a la base de datos en la cual se desarrolló el proyecto.

Se excluyeron las transacciones realizadas como cliente final, esto redujo drásticamente el número de transacciones a 3.296.124 registros.

Se realizaron correcciones en el atributo ciudad, ya que presentaba valores atípicos en los nombres, se omitieron caracteres especiales y se estableció un estándar para no afectar el análisis.

TABLA XII. TAMAÑO FINAL DEL CONJUNTO DE DATOS SELECCIONADO

	Datos Lotería Nacional
Años a analizar	4 años (2018 a 2021)
Número de clientes	5.536
Número de Transacciones	3.296.124

Tabla 12 .- *Tamaño final del conjunto de datos seleccionado*

TABLA XIII RESUMEN DE LOS DATOS SELECCIONADOS

Categoría de atributo	Atributo	Tipo	Descripción
Identificador del Cliente	Id Cliente	Continuo	Código único del cliente
Comportamiento de compra	Recencia	Discreto	Número de días transcurridos desde la última compra efectuada
	Frecuencia	Discreto	Número de veces que el cliente realiza compras en el año
	Monto	Continuo	Suma monetaria del valor total en compras
Demográfico	Tipo Cliente	Categórico	Identifica si el cliente es cliente final o cliente Lotero

Demográfico	Edad	Discreto	Edad del cliente
Geográfico	Ciudad	Categórico	Ciudad en la que el cliente efectuó la compra
Descriptivo	Precio	Continuo	Precio del producto
Descriptivo	Producto	Categórico	Nombre del producto vendido

Tabla 13 .- *Resumen de datos seleccionado*

4.5.3 Generación de las variables RFM

Para la creación de las variables RFM (Recencia, Frecuencia, Monto), se dividió a los clientes por tipo ya que en la empresa existen dos tipos de clientes, el cliente final que es aquel cliente que compra boletos sin ser asociado a la empresa y el Lotero, este tipo de cliente es el que se asocia a la empresa y la empresa a cambio le da un descuento por distribuir el producto. Para el cálculo de la Recencia se tomó la última fecha de compra del cliente y se calculó la diferencia entre la fecha final que en este caso es el 31 de Dic del 2021 y la fecha de la última compra realizada por el cliente; La frecuencia se calculó contando el número transacciones realizadas por el cliente en el periodo de enero 2018 a diciembre 2021; El monto, es la suma total de las compras realizadas durante el periodo ya antes mencionado.

4.5.4 Definición de las escalas RFM

Para la definición de las escalas RFM, se separaron los datos en cuartiles, divididos en 0.25, 0.50 y 0.75 para los cuartiles 1, 2 y 3 correspondientemente, se aplicó esta técnica a las variables RFM creadas, (ver sección 2.13 del Capítulo 2).

TABLA XIV. TABLA DE ESCALAS RFM

Escala	Nombre Escala	Recencia	Frecuencia	Monto $
1 Punto	Alto	7 - 0 días	[706+]	[5428+]
2 Puntos	Medio	8 - 31 días	[200 - 705]	[1633 – 5427]
3 Puntos	Bajo	32 - 895 días	[23 - 199]	[317 – 1632]
4 Puntos	Muy Bajo	895+ días	[1 - 22]	[0.25 – 316]

Tabla 14 .- *Escalas de variables RFM*

Puntuación Recencia: Se otorga la puntuación más alta (4) a clientes con transacciones más recientes.

Puntuación Frecuencia: Se otorga puntuación más alta (4) a clientes que su número de transacciones sea mayor a 705 para clientes Loteros y 351 para clientes finales.

Puntuación Monto: Se otorga la puntuación más alta (4) a clientes con mayor volumen en compras.

4.5.5 Normalización de las variables RFM

Una fase muy importante en el tratamiento de los datos es la normalización de variables ya que el conjunto de datos objeto a estudio puede contener información atípica (Outliers) que podría afectar el resultado final. En este caso se aplicó normalización a las variables RFM (Recencia, Frecuencia y Monto) ya que son las variables de las que obtendremos información de las compras realizadas por los clientes y que nos ayudará a comprender su nivel de lealtad. Para normalizar estas variables se extrajo valor mínimo de cada variable RFM y se calculó la diferencia para cada una de ellas y luego se las dividió por la diferencia entre el mínimo y máximo de cada variable, a continuación, se muestra la fórmula empleada.

$$Norm\ Rec = \frac{\text{Rec} - \text{Rec Min}}{\text{Rec Max} - \text{Rec Min}}$$

Ecuación 1 .- *Fórmula de normalización de Recencia*

$$Norm\ Fre = \frac{\text{Fre} - \text{Fre Min}}{\text{Fre Max} - \text{Fre Min}}$$

Ecuación 2 .- *Fórmula de normalización de Frecuencia*

$$Norm\ Mon = \frac{\text{Mon} - \text{Mon Min}}{\text{Mon Max} - \text{Mon Min}}$$

Ecuación 3 .- *Fórmula de normalización de Monto*

Antes de aplicar la normalización de las variables RFM, graficamos en un histograma la distribución de las mismas, para darnos una idea de cómo se encuentran las variables antes de estandarizarlas.

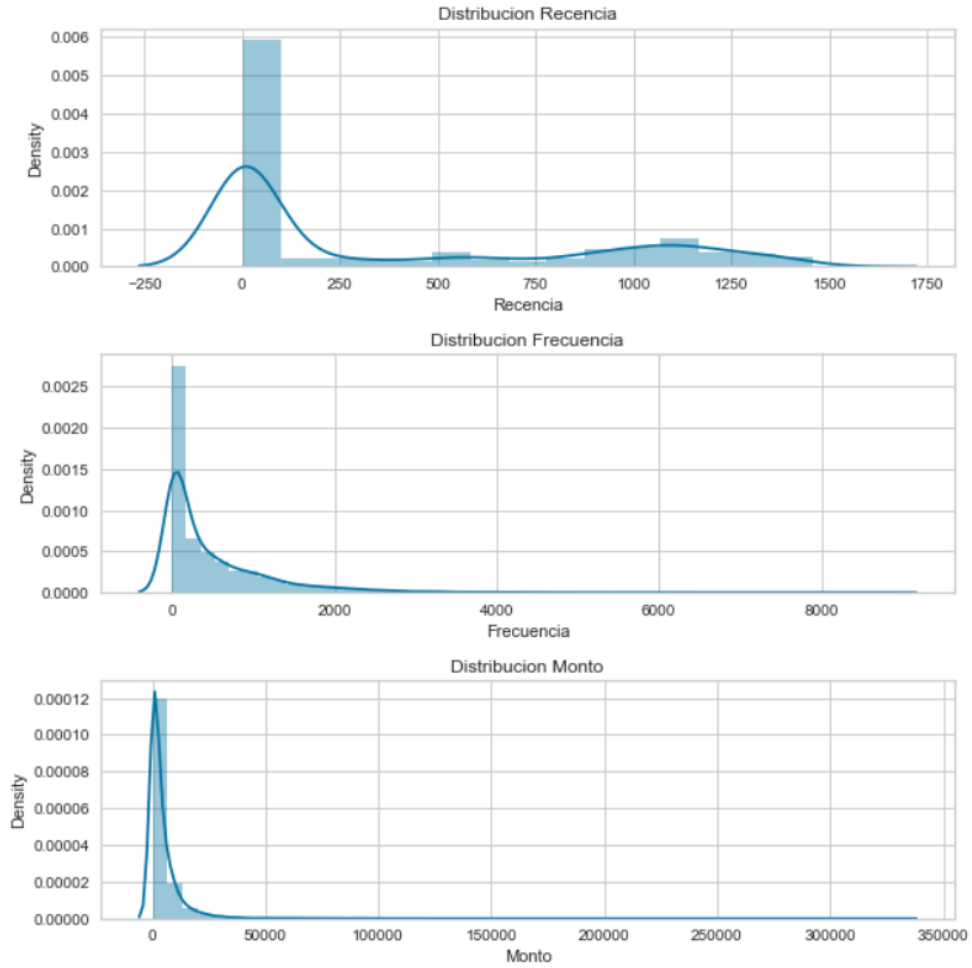


Ilustración 12 .- *Distribución de variables RFM antes de normalizar*

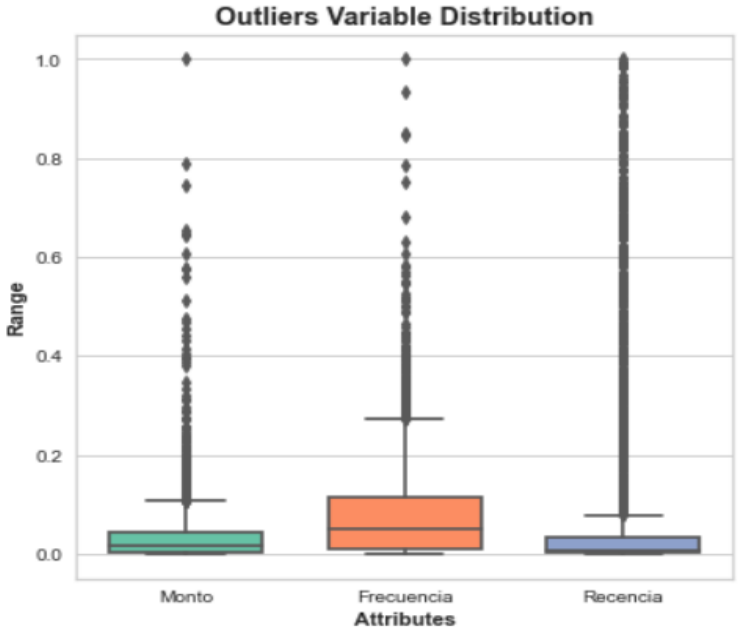


Ilustración 13 .- *Gráfico de cajas de Outliers de las variables RFM*

Como se puede apreciar en la figura anterior, las variables RFM se encuentran muy sesgadas, por lo que realizaremos transformaciones para escalar las variables y que estas se encuentren un rango de 0 a 1, para que no afecte el entrenamiento del modelo.

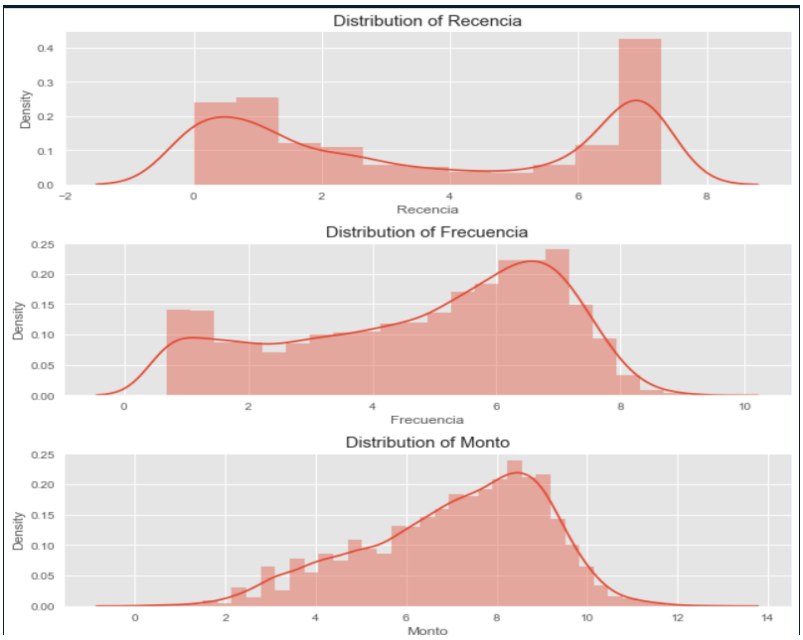


Ilustración 14 .- *Histograma de variables RFM normalizadas*

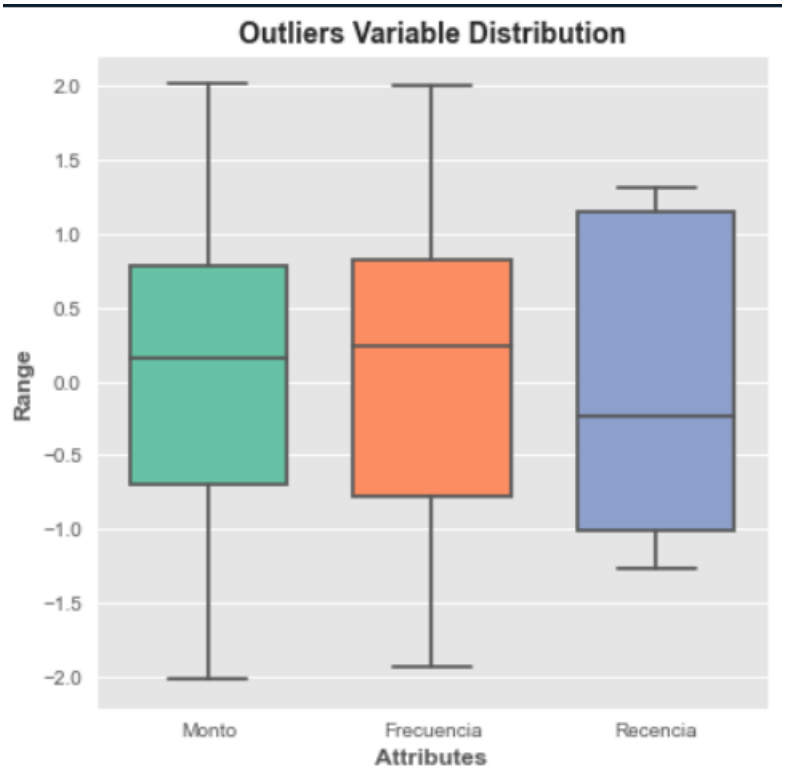


Ilustración 15 .- *Gráfico de cajas de variables RFM normalizadas*

Una vez aplicada las fórmulas, procedemos a observar la distribución de las variables rfm, para darnos una idea de como quedaron las variables después de aplicar la normalización.

4.5.6 División de datos de para pruebas y entrenamiento

Para la división de los datos se empleó el módulo train_test_split del paquete sklearn de python, se consideraron el 75% del total del conjunto de datos para entrenamiento y el restante 25% para pruebas. A continuación, se muestra el código que se empleó.

```
##DIVIDIR DATOS EN TABLAS DE TESTING Y TRAINING

from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(x,y,train_size=0.75,random_state=0)
[11]  ✓ 0.7s
```

Ilustración 16 .- *Código en python, división de datos de entrenamiento y pruebas*

TABLA XV. REGISTROS PARA DATOS DE ENTRENAMIENTO Y PRUEBAS

Entrenamiento (75%)	Pruebas (25%)
4.152 registros	1.384 registros

Tabla 15 .- *Numero de registros para datos de entrenamiento y pruebas*

4.6 MODELADO DE DATOS

4.6.1 Aplicación de algoritmos de aprendizaje automático

Ya preparados y normalizados los datos, estos están listos para poder aplicarlos a los algoritmos detallados a continuación.

4.6.2 Algoritmo K means

El primer paso para aplicar K means, es identificar el número de grupos óptimos, se utilizó el método de la curva de distorsión o método del codo (Ver sección 2.10 del capítulo 2), como en nuestro conjunto de datos tenemos dos tipos de cliente, Lotero y Cliente final, se realizó el método mencionado para ambos tipos de clientes, para clientes identificados como loteros, la curva de distorsión arrojó 3 grupos óptimos mientras que para clientes final se encontraron 4 grupos óptimos.

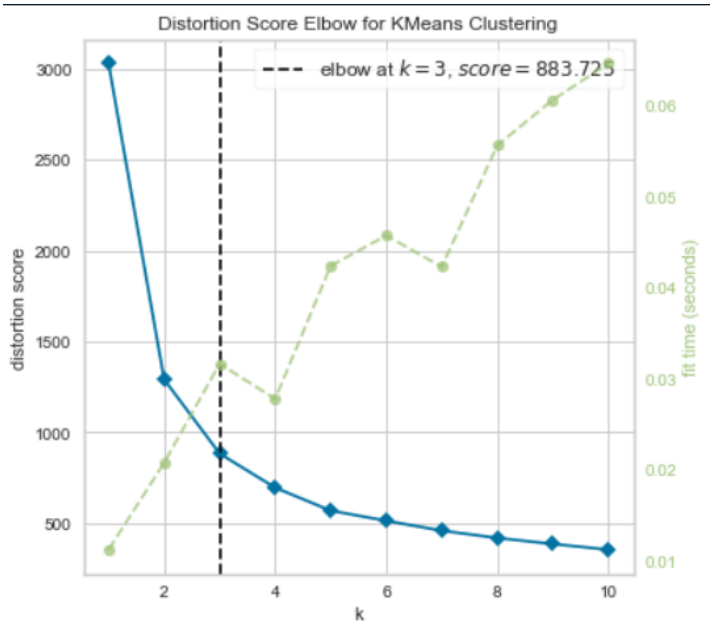


Ilustración 17 .- *Gráfico curva de distorsión - Número de grupos óptimos*

4.6.3 Entrenamiento del Modelo K means

Ya definido el número de grupos óptimos, se procede con el entrenamiento del modelo. a continuación, se muestra los clústeres formados para Loteros, el número de iteraciones en las que el algoritmo converge fueron 1000.

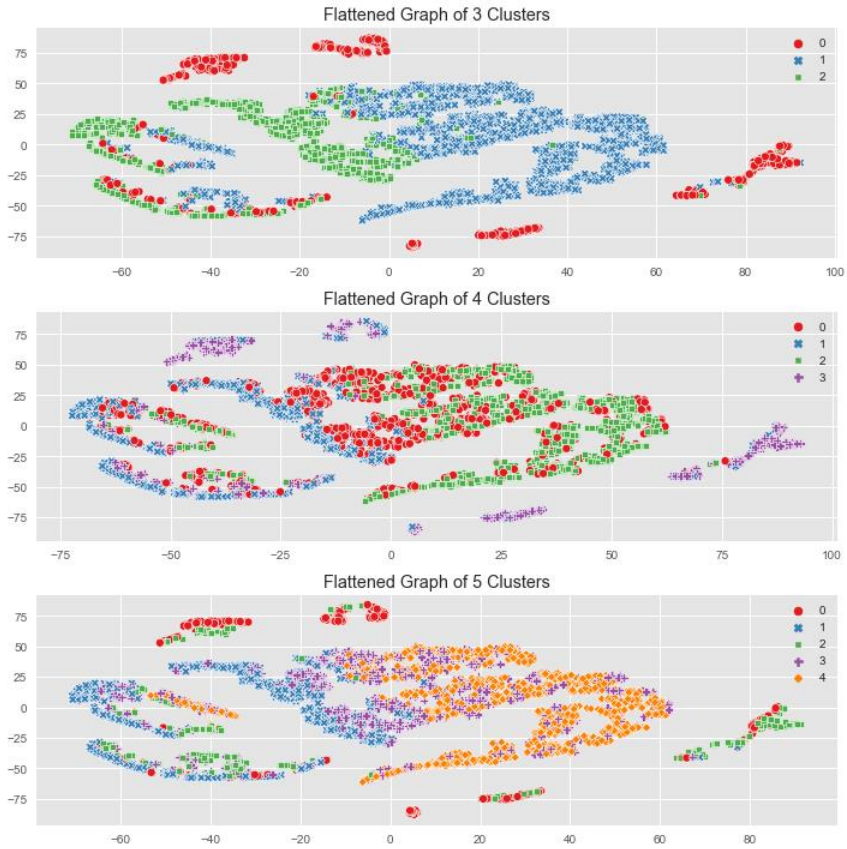


Ilustración 18 .- *Resultado de clusteres formados*

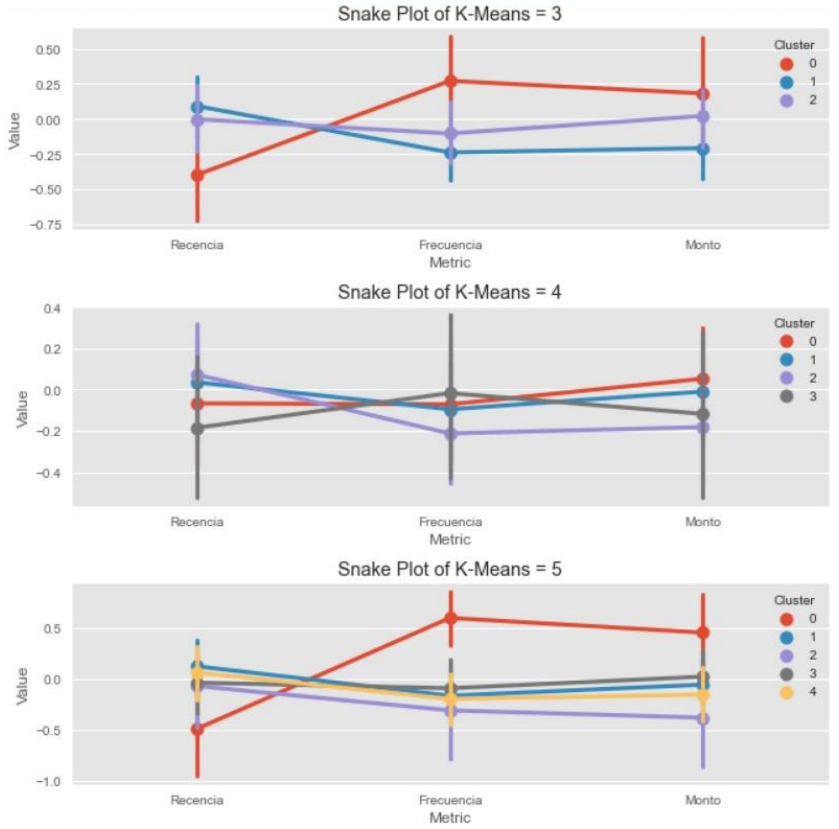


Ilustración 19 .- *Gráfico snake plot clientes*

A partir de los gráficos anteriores, es evidente que, si dividimos a nuestro conjunto de datos en 3 grupos, se segmenta mejor que si segmentamos con más de 3 grupos, aunque también se podría optar por segmentar con un mayor número de clústeres, pero esto ya dependería completamente de cómo la empresa quiera segmentar a sus clientes.

4.6.3.1 Algoritmo de Hunt Árbol de Decisión

La aplicación de este algoritmo se usó para clasificar el nivel de lealtad de los clientes según las escalas creadas para las variables RFM, a continuación, se muestra los 5 primeros registros de la tabla creada a partir de las combinaciones de las escalas de las variables RFM.

	Recencia	Frecuencia	Monto	R_clasificacion	F_clasificacion	M_clasificacion	Score	NivelLealtad
0	1.0	1.0	1.0	Alto	Alto	Alto	111	Alto
1	2.0	1.0	1.0	Medio	Alto	Alto	211	Alto
2	3.0	1.0	1.0	Bajo	Alto	Alto	311	Medio
3	4.0	1.0	1.0	Muy bajo	Alto	Alto	411	Muy bajo
4	1.0	1.0	2.0	Alto	Alto	Medio	112	Alto

Ilustración 20 .- *Tabla de clasificación de variables RFM*

Para clasificar las variables RFM, primero debemos separar las variables predictoras de la variable a predecir, en este caso la variable a predecir es el nivel de lealtad del cliente, por lo que la vamos a separar de nuestro conjunto de datos.

	Recencia	Frecuencia	Monto	R_quartil	F_quartil	M_quartil	RFM_Segmento	RFM_Total
0	0	1176.0	8462.25	1	1	1	111	3
1	0	1835.0	28474.75	1	1	1	111	3
2	0	2339.0	5624.00	1	1	1	111	3
3	0	1047.0	12515.25	1	1	1	111	3
4	1	3645.0	37503.50	1	1	1	111	3

Ilustración 21 .- *Conjunto de datos de variables predictoras*

Luego de separar las variables predictoras de la variable a predecir, creamos el modelo, en este caso se usó el módulo DecisionTreeClassifier del paquete sklearn.tree. En la siguiente figura se muestra el código empleado para crear el modelo en el que se llama al constructor de la clase DecisionTreeClassifier y le enviamos como parámetro el número de niveles que queremos en el árbol dentro de la variable max_depth=4, en este caso solo queremos que nos grafique hasta 4 niveles.

```
from sklearn.tree import DecisionTreeClassifier

#Llamamos al constructor del arbol de decision
arbol = DecisionTreeClassifier(max_depth=4)
```

Ilustración 22 .- *Código en Python, creación de modelo árbol de decisión*

Luego del paso anterior, entrenamos el modelo con los datos de entrenamiento seleccionados previamente. (ver sección # Capitulo 4)

```
#Entrenamos el modelo
loyal_tree = arbol.fit(X_train,Y_train)
```

Ilustración 23 .- *Código en Python, Entrenamiento del modelo, árbol de decisión*

Como último paso, se grafica el resultado obtenido. Como podemos observar, el árbol contiene 4 niveles, donde el nodo de Gini nos indica la pureza de cada nodo, a mayor índice de Gini, menor pureza

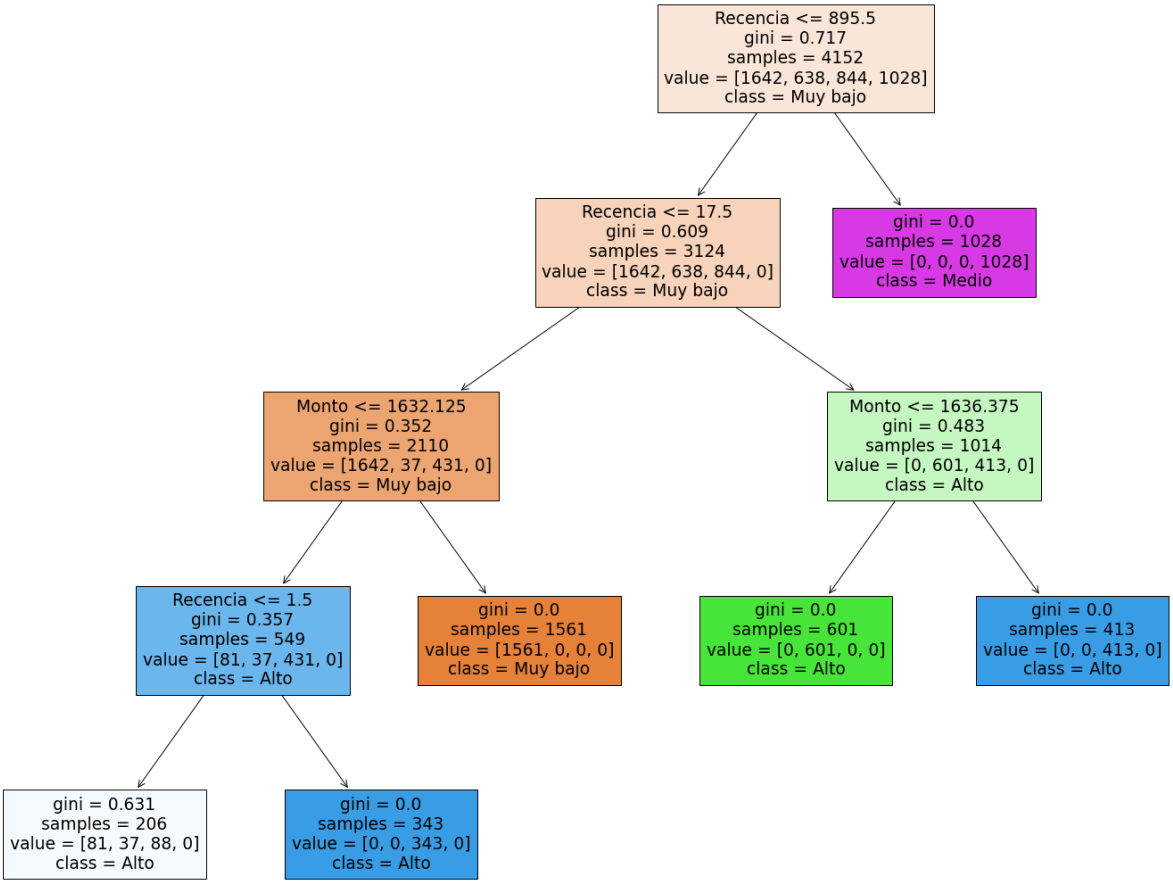


Ilustración 24 .- *Resultado, árbol de decisión*

4.6.4 Algoritmo KNN (K Nearest Neigbors)

Para aplicación del algoritmo K vecinos mas cercanos se usaron las variables RFM como variables predictoras, a efectos de la aplicación del algoritmo la variable NivelLealtad se transformo a valores numéricos, siendo 1 el nivel de lealtad Alto y 4 el nivel de lealtad más Bajo. A continuación, se muestra la línea de código usada para entrenar el modelo.

```
1  from sklearn.neighbors import KNeighborsClassifier
2  clf = KNeighborsClassifier(n_neighbors=4)
3  clf.fit(X_train, Y_train)
```

[57] ✓ 0.2s

```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=4)
```

Ilustración 25 .- *Código en Python, Entrenamiento modelo K nearest neigbor (KNN)*

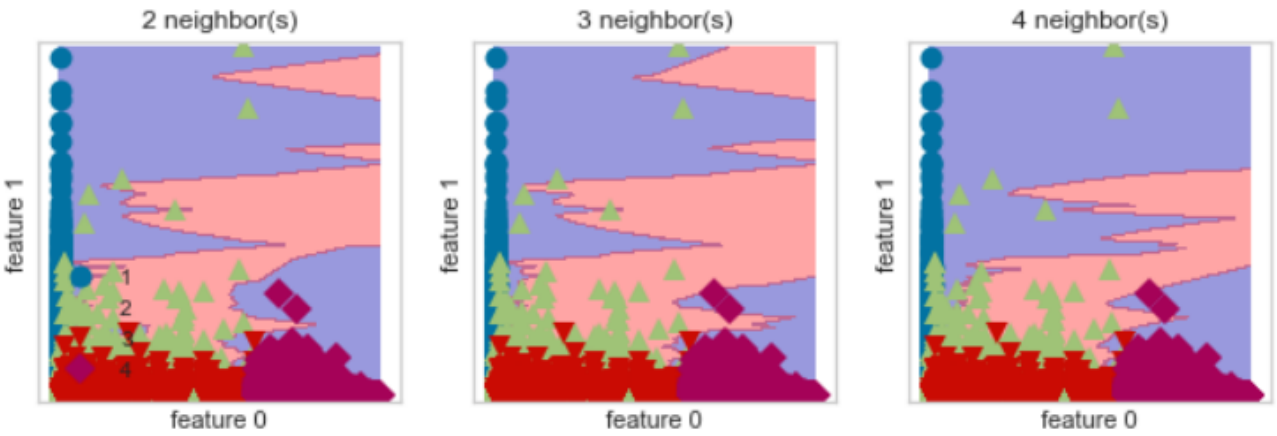


Ilustración 26 .- *Resultado, gráfico KNN*

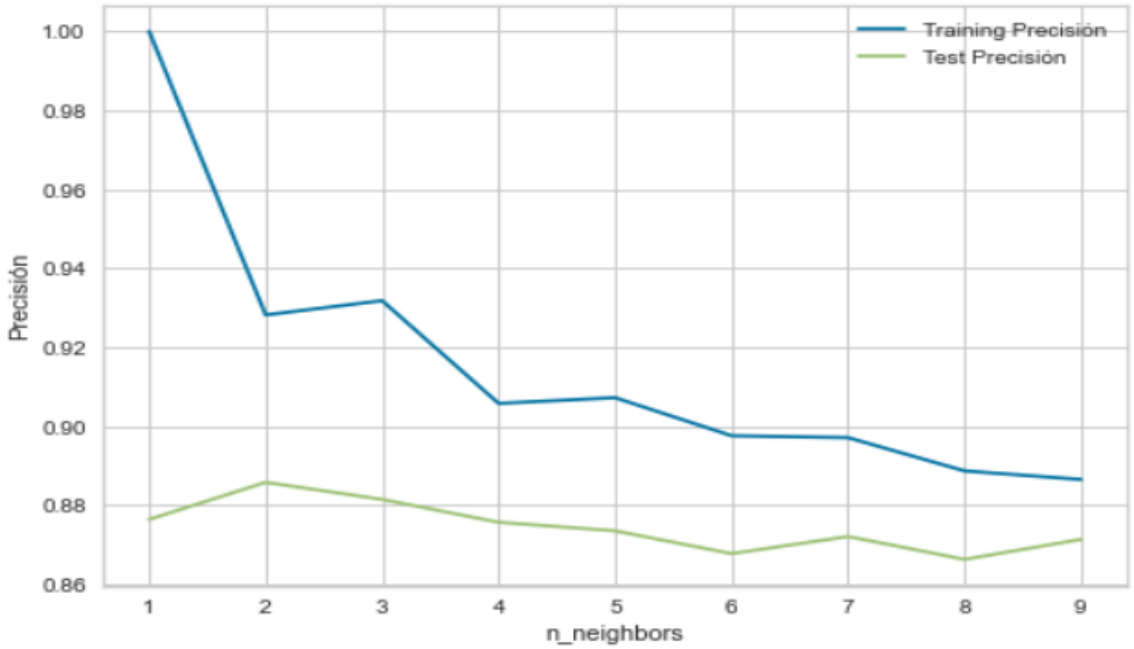


Ilustración 27 .- Gráfico de precisión KNN

4.7 EVALUACIÓN DE LOS ALGORITMOS

Se desarrolló código en Python para obtener la precisión de cada algoritmo y poder evaluar que algoritmo se ajusta mejor a nuestro modelo, a continuación, se muestra los resultados obtenidos.

4.7.1 Árbol de decisión

En el algoritmo de árbol decisión se obtuvo una precisión global de 98%, para las clases Alta, Media, Baja y Muy baja que corresponden a los niveles de lealtad, se obtuvo 0.96% para la clase alta, 0.95% para clase media, 1.0% para la clase baja y 1.0% para la clase muy baja.

TABLA XVI. TABLA DE PRECISIÓN ÁRBOL DE DECISIÓN

Descripción	Precisión
Precisión Global	0.98%
Clase Alta	0.96%
Clase Media	0.95%
Clase Baja	1.0%
Clase Muy Baja	1.0%

Tabla 16 .- *Precisión del Árbol de decisión*

4.7.2 K Nearest Neigbor (K vecinos más cercanos)

Para el algoritmo K Nearest Neigbor se obtuvo una precisión de 0.87%, siendo uno de los arrojó un porcentaje de precisión significativamente menor al del árbol de decisión.

```
1  print("Precision del algoritmo: {:2f}".format(clf.score(X_test, Y_test)))
```

```
Precision del algoritmo: 0.875723
```

Ilustración 28 .- *Precisión Algoritmo K nearest neigbor*

4.8 IMPLEMENTACIÓN

Esta sección corresponde a una de las fases de la Metodología CRISPDM, en la cual se detalla la implementación del modelo una vez validado en fase anterior.

Para la implementación y visualización de los resultados obtenidos, se moldeó los datos en la herramienta Power BI, para visualizar de una manera más amigable los resultados obtenidos después de la aplicación de los algoritmos.

A continuación, se muestra los gráficos y métricas desarrolladas en la herramienta Power BI.

Los resultados obtenidos, se los dividieron en 4 secciones, catalogados como hallazgos, en la sección hallazgo 1; se muestra un resumen de los resultados obtenidos; en la sección hallazgo 2; se muestra el comparativo por año según los resultados obtenidos; en la sección hallazgos 3; se muestra un resumen comparativo por producto y precio; en la sección hallazgo 4; se muestra un comparativo por provincias y rango de edades

Ilustración 29 .- *Portada Dashboard de hallazgos*

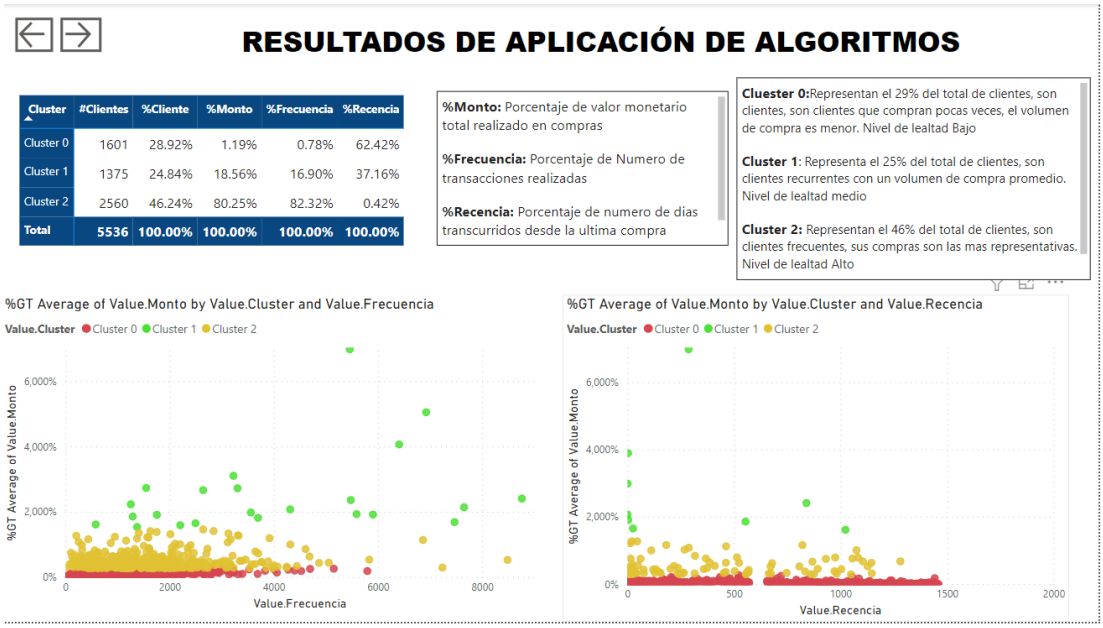


Ilustración 30 .- *Dashboard visualización hallazgos 1*

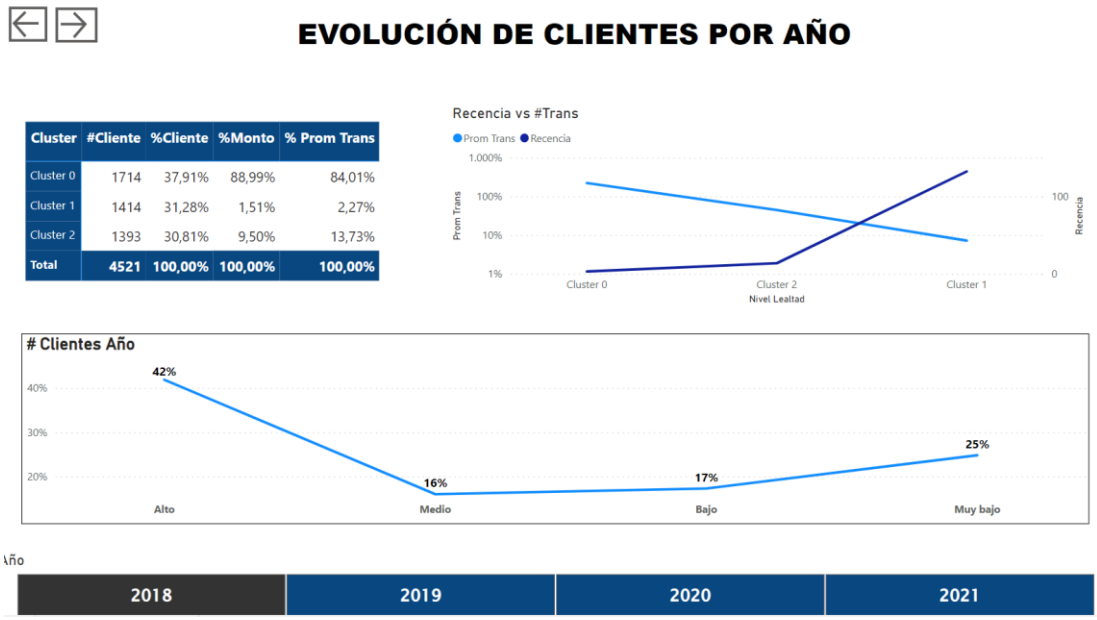


Ilustración 31 .- Dashboard visualización hallazgo 2

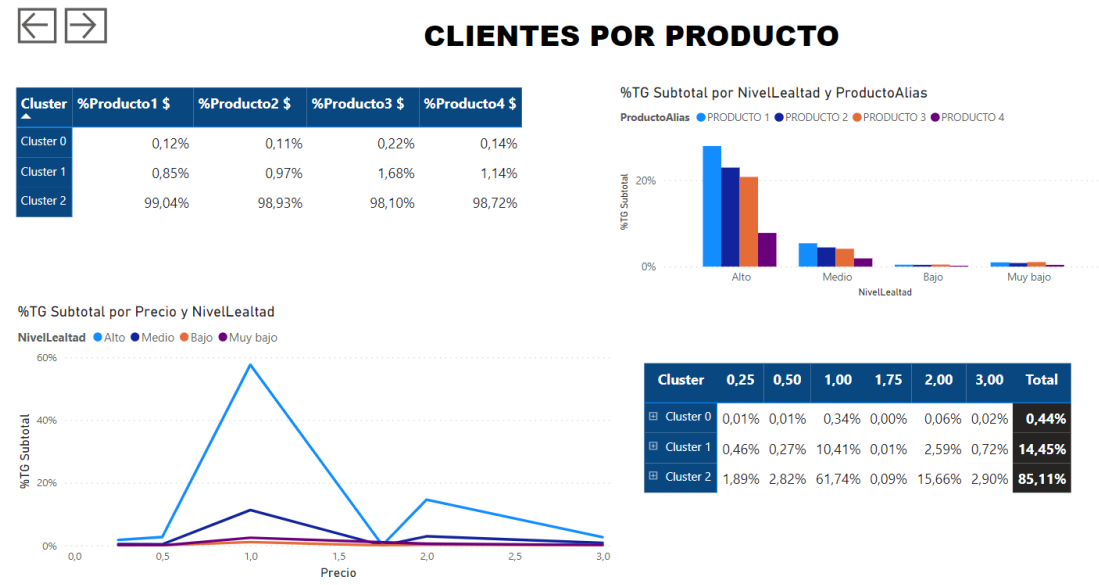


Ilustración 32 .- *Dashboard visualización hallazgo 3*

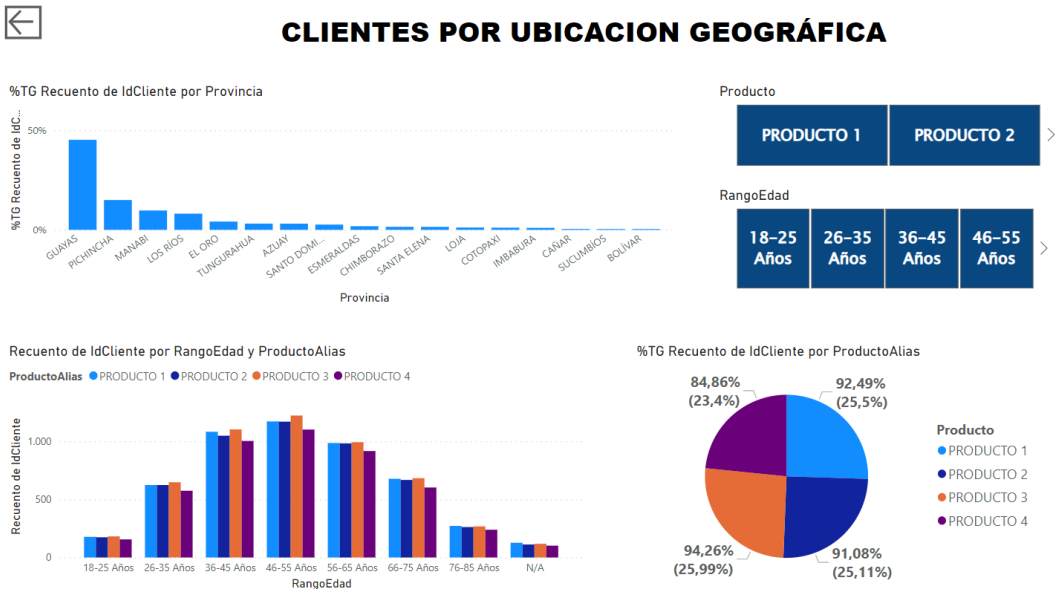


Ilustración 33 .- *Dashboard visualización hallazgo 4*

5 CONCLUSIONES

A partir de los datos transaccionales de compras, proporcionados por la empresa Lotería Nacional, se logró aplicar técnicas para el desarrollo de algoritmos de aprendizaje automático a partir de las variables RFM, que permitieron conocer el nivel de lealtad de los clientes y tomar acciones respecto a la información obtenida.

La elección de algoritmos se hizo en base a una búsqueda de los algoritmos popularmente aplicados para el análisis del comportamiento de clientes.

Para la validación y comprobación de los algoritmos se separaron en un 25% en datos para pruebas y el 75% en datos de entrenamiento, a partir de esto se pudo determinar la precisión de los algoritmos desarrollados, que se detallan a continuación. Árbol de decisión 0.96% y K nearest neigbor 0.87%, para obtener el número de clústeres óptimos en la aplicación del algoritmo K means, se utilizó la técnica de la curva de distorsión o método del codo dando como resultado que el numero óptimos para aplicación de k means es de 3 grupos, también se hicieron pruebas para 4 y 5 grupos para observar si los datos pueden agruparse en más de 3 grupos, como resultado se obtuvo que para 4 y 5 grupos el algoritmo no se compacta bien, por lo que el número de grupos en los que se compacta mejor el algoritmo es 3.

Con los resultados obtenidos se desarrolló en la herramienta Power BI el modelado de datos y paneles gráficos mediante los cuales se pudo interpretar los resultados arrojados de manera más amigable para el usuario.

6 RECOMENDACIONES

Que la persona encargada haga uso de los grupos de clientes obtenidos en la presente investigación de la forma que estime conveniente, por ejemplo, se podría aplicar estrategias para premiar a los mejores clientes y así mantener su lealtad o crear promociones para atraer a clientes que tienen un nivel de lealtad bajo o también se podría ofrecer descuentos para animar a compradores regulares a aumentar su valor monetario.

Que se realice un mejor control de la información que se registra en la base de datos, para evitar duplicidad de datos, campos importantes vacíos o información sesgada, como por ejemplo validar fechas de nacimiento o ciudades, variables que pueden ser importantes al momento de analizar la información.

Se recomienda a la empresa Lotería Nacional, considerando el gran volumen de información que almacenan y la importancia de esta, consideren alimentar una base de datos donde no se omita ningún dato acerca de sus clientes, como ejemplo, edad, genero, correo electrónico, estado civil, etc. ya que estas son variables muy importantes al momento de aplicar un análisis de clientes o armar una estrategia de clientes.

7 REFERENCIAS BIBLIOGRÁFICAS

Arana, C. (Febrero de 2021). *Modelos de aprendizaje automático mediante árboles de decisión.* Obtenido de https://www.econstor.eu/bitstream/10419/238403/1/778.pdf

Arora, P., Varshney, S., & Deepali, D. (Diciembre de 2015). *Analysis of K-Means and K-Medoids Algorithm For Big Data*. Obtenido de https://www.researchgate.net/publication/301234359_Analysis_of_K-Means_and_K-Medoids_Algorithm_For_Big_Data

BBVA. (2015). *Reinventar la Empresa en la Era Digital.* Obtenido de https://www.bbvaopenmind.com/wp-content/uploads/2015/01/BBVA-OpenMind-libro-Reinventar-la-Empresa-en-la-Era-Digital-empresa-innovacion1-1.pdf

Chamba Jimenez , S. F. (Octubre de 2015). *Minería de Datos para segmentación de datos para segmentación de clientes en la empresa tecnológica Master PC*. Obtenido de https://dspace.unl.edu.ec/jspui/bitstream/123456789/10462/1/Chamba%20Jim%C3%A9nez%2C%20Sairy%20Fernanda.pdf

Chashif Syadzali, S. S. (2020). *Business Intelligence using the K-Nearest Neigbor Algorithm to Analyze Customer Behavior in Online Crowdfunding Systems.* Obtenido de https://www.e3s-conferences.org/articles/e3sconf/pdf/2020/62/e3sconf_icenis2020_16005.pdf

Cuadros, J., Gozales, C., & Jiménez , P. (26 de Febrero de 2017). *Análisis multivariado para segmentación de clientes basada en RFM*. Obtenido de http://www.scielo.org.co/pdf/tecn/v21n54/0123-921X-tecn-21-54-00041.pdf

Cukier, K. (2015). *Los Big Data y el futuro de los negocios .* Obtenido de Reinventar la empresa en la era digital: https://www.bbvaopenmind.com/wp-content/uploads/2015/01/BBVA-OpenMind-libro-Reinventar-la-Empresa-en-la-Era-Digital-empresa-innovacion1-1.pdf

EcuRED. (s.f.). *Algoritmos Jerárquicos*. Obtenido de https://www.ecured.cu/Algoritmos_jer%C3%A1rquicos

Espinosa, J. (Marzo de 2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública*. Obtenido de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000100008

Gutierrez, R., Carrasco, R., & Sanchez, C. (30 de Junio de 2021). *An RFM Model Customizable to Product Catalogues and*. Obtenido de https://www.mdpi.com/2227-7390/9/16/1836/htm

Heras, J. M. (29 de Septiembre de 2020). *¿Clasificación o Regresión?* Obtenido de IArtificial.net: https://www.iartificial.net/clasificacion-o-regresion/#Clasificacion

Hérnandez Sampieri , R. (2014). *Metodología de la Investigación 6ta Edición.* Obtenido de https://www.uca.ac.cr/wp-content/uploads/2017/10/Investigacion.pdf

Jaramillo, L., Galindo, M., & Real, J. (2020). *Análisis clúster para big data: una aplicación con variables*. Obtenido de http://www.estadisticas.med.ec/Publicaciones/JOHAMSC-61-45-50-2020-WEB-03-04.pdf

Kovera, A. (2017). *Machine Learning with Clustering.*

Lopez, C., Tucker, S., Salameh, T., & Tucker, C. (2018). *An unsupervised machine learning method for discovering patient clusters.* Obtenido de Journal of Biomedical Informatics: https://reader.elsevier.com/reader/sd/pii/S1532046418301308?token=85C8C2275522333072CFDE6EB670401926C4FC8C4BD6E9E93AD83D17DF40B0DC648FA3EA1DA7ED589CFB53AC74331645&originRegion=us-east-1&originCreation=20220628004413

M. Ambigavathi, D. S. (Julio de 2020). *Analysis of Clustering Algorithms in Machine.* Obtenido de https://www.researchgate.net/publication/343031883_Analysis_of_Clustering_Algorithms_in_Machine_Learning_for_Healthcare_Data

Microsoft. (18 de 04 de 2022). *Microsoft*. Obtenido de https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions#:~:text=Un%20algoritmo%20en%20miner%C3%ADa%20de,espec%C3%ADficos%20de%20patrones%20o%20tendencias.

Morakanyane, R. G. (Junio de 2017). *Conceptualizing Digital Transformation in Business Organizations: A systematic review of literature .* Obtenido de https://www.researchgate.net/publication/321805933_Conceptualizing_Digital_Transformation_in_Business_Organizations_A_Systematic_Review_of_Literature/link/5a32a9a40f7e9b2a288d7ee9/download

Moroke, N. D. (2015). *A TWOSTEP CLUSTERING ALGORITHM AS APPLIED TO CRIME*. Obtenido de https://virtusinterpress.org/IMG/pdf/cocv12i2c4p8.pdf

Pierrend, S. D. (20 de 10 de 2020). *Customer Loyalty and Customer Retention.* Obtenido de https://revistasinvestigacion.unmsm.edu.pe/index.php/administrativas/article/view/18935/15876

RAE. (2021). *Real Academia Española*. Obtenido de https://dle.rae.es/fidelizar

Reed, M. (2020). *Python Machine Learning The ultimate beginner's Guide to Learn Python Machine Learning Step by Step.*

Russell, R. (2018). *Step-by-Step Guide To Implement Machine Learning Algorithms with Python.*

Sanchez. (2017). *Negocios y Empresas*. Obtenido de https://www.puromarketing.com/14/28784/fidelizacion-clientes

Sandoval, L. J. (19 de Julio de 2018). *Machine Learning algorithms for data analysis and prediction.* Obtenido de REVISTA TECNOLÓGICA N° 11. ENERO:

http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Schallmo, D. R. (30 de Noviembre de 2017). *DIGITAL TRANSFORMATION OF BUSINESS MODELS BEST PRACTICE, ENABLERS,.* Obtenido de https://www.researchgate.net/publication/322467178_History_of_Digital_Transformation/link/5be2942892851c6b27ac7133/download

Torres, J. (16 de Abril de 2021). *Introducción al aprendizaje por refuerzo*. Obtenido de https://medium.com/aprendizaje-por-refuerzo/1-introducci%C3%B3n-al-aprendizaje-por-refuerzo-92c9239aed90

WSI. (2 de Julio de 2021). *La importancia de conocer al cliente*. Obtenido de https://wsiconecta.com/importancia-de-conocer-al-cliente/

8 ANEXOS

8.1 ANEXO 1: CARGA DE INFORMACIÓN A BASE DE DATOS LOCAL

Script realizado en sql para cargar a una base de datos, los archivos entregados en formato csv.

```
INSERT INTO tbSales(Año,Mes,IdpuntoOperacion,CodigoTransaccion,Producto,Subtotal)
SELECT [Año],[Mes],[IdPuntoOperacion],[CodigoTransaccion],[Producto],[Subtotal]
FROM OPENROWSET('Microsoft.ACE.OLEDB.12.0',
   'Excel 12.0 Xml;Database=C:..\\Data\2022\Transacciones.xlsx;', [Hoja1$])
```

```
INSERT INTO tbCliente
SELECT [IdCliente],[FechaNacimiento],[Telefono],[Correo],[TipoCliente]
FROM OPENROWSET('Microsoft.ACE.OLEDB.12.0',
   'Excel 12.0 Xml;Database=C:..\\Data\2022\Clientes.xlsx;', [Hoja1$])
```

```
INSERT INTO tbProducto
SELECT [IdProducto],[Producto],[Precio]
FROM OPENROWSET('Microsoft.ACE.OLEDB.12.0',
   'Excel 12.0 Xml;Database=C:..\\Data\2022\Producto.xlsx;', [Hoja1$])
```

8.2 ANEXO 2: LECTURA DE DATOS DESDE LA BASE DE DATOS

```
import pyodbc
import numpy as np
import pandas as pd
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import plotly.express as px
from yellowbrick.cluster import KElbowVisualizer
from sklearn.preprocessing import StandardScaler
from feature_engine.outliers import Winsorizer
```

```
server = 'localhost'
bd = 'proyecto'
usuario = 'gbarrezueta'
pwd = ''

try:
    cnxn = pyodbc.connect(driver='{ODBC Driver 17 for SQL Server}', host=server, database=bd,
                          user=usuario, password=pwd)
except Exception as e:
    print('Ocurrio un error de conexion ',e)

cursor = cnxn.cursor()

query = "select * from tbAnalisisRfm"
df = pd.read_sql_query(query,cnxn)
df
```

8.3 ANEXO 3: CREACIÓN DE VARIABLES RFM

A continuación, se muestra el código utilizado para generar las variables RFM, primero se consulta la última fecha que contiene df_lotero, que representa el dataset de los datos leídos previamente, luego se agrupa los datos por el identificador del cliente y mediante la función “agg” se realiza las operaciones para la creación de las variables. En el caso de recencia, se calcula la diferencia en dias entre la última fecha contenida en el dataset y la ultima fecha en la que el cliente realizó la última transacción, para la frecuencia, se cuenta el numero de facturas generadas por el cliente y el Monto es la suma de lo que ha comprado el cliente.

```
from datetime import timedelta

ultima_fecha = df_lotero['Fecha'].max() + timedelta(days=1)

rfm = df.groupby(['IdCliente']).agg({
        'Recencia': lambda x: (ultima_fecha - x.max()).days,
        'Frecuencia': 'count',
        'Monto': 'sum'})
```

8.4 ANEXO 4: ASIGNACIÓN DE ESCALAS A LAS VARIABLES RFM

Se dividió los datos en cuartiles, para poder asignar las escalas RFM a cada cliente, y mediante una función se le asignó el puntaje correspondiente. La función RScore asigna el puntaje a la variable recencia, mientras la función FMScore asigna el puntaje a frecuencia y monto.

```
quantiles = rfm_lotero.quantile(q=[0.25,0.50,0.75])
quantiles = quantiles.to_dict()

```

```
def RScore(x,p,d):
  if x <= d[p][0.25]:
    return 1
  elif x <= d[p][0.50]:
    return 2
  elif x <= d[p][0.75]:
    return 3
  else:
    return 4

def FMScore(x,p,d):
  if x <= d[p][0.25]:
    return 4
  elif x <= d[p][0.50]:
    return 3
  elif x <= d[p][0.75]:
    return 2
  else:
    return 1
```

A continuación, se le asigna el puntaje a cada cliente usando las funciones anteriormente desarrolladas

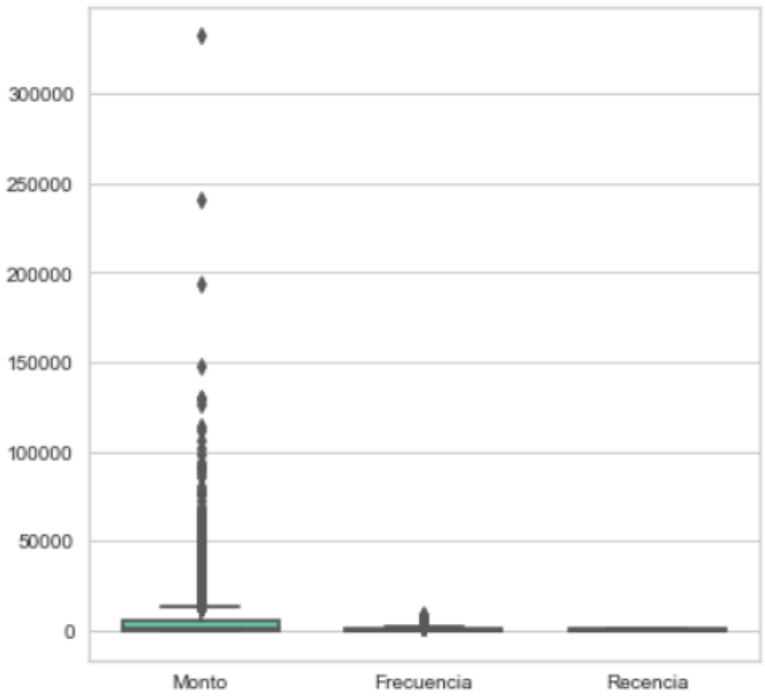
```
rfm_lotero['R_quartil'] = rfm_lotero['Recencia'].apply(RScore,args=('Recencia',quantiles))
rfm_lotero['F_quartil'] = rfm_lotero['Frecuencia'].apply(FMScore,args=('Frecuencia',quantiles))
rfm_lotero['M_quartil'] = rfm_lotero['Monto'].apply(FMScore,args=('Monto',quantiles))
rfm_lotero['RFM_Segmento'] = rfm_lotero.R_quartil.map(str)+rfm_lotero.F_quartil.map(str)+rfm_lotero.M_quartil.map(str)
rfm_lotero['RFM_Total'] = rfm_lotero[['R_quartil','F_quartil','M_quartil']].sum(axis=1)
rfm_lotero
```

8.5 ANEXO 5: DETECCIÓN DE DATOS ATIPICOS

Se muestra el código implementado para observar si las variables creadas presentan datos atípicos

```

attributes = ['Monto','Frecuencia','Recencia']
plt.rcParams['figure.figsize'] = [6,6]
sns.boxplot(data = rfm_lotero[attributes], orient="v", palette="Set2" ,whis=1.5,saturation=1, width=0.7)
plt.ylabel("Range", fontweight = 'bold')
plt.xlabel("Attributes", fontweight = 'bold')
```



8.6 ANEXO 6: DETECCIÓN DE DATOS ATIPICOS

Se muestra el código implementado para normalizar las variables contenidas en el dataset

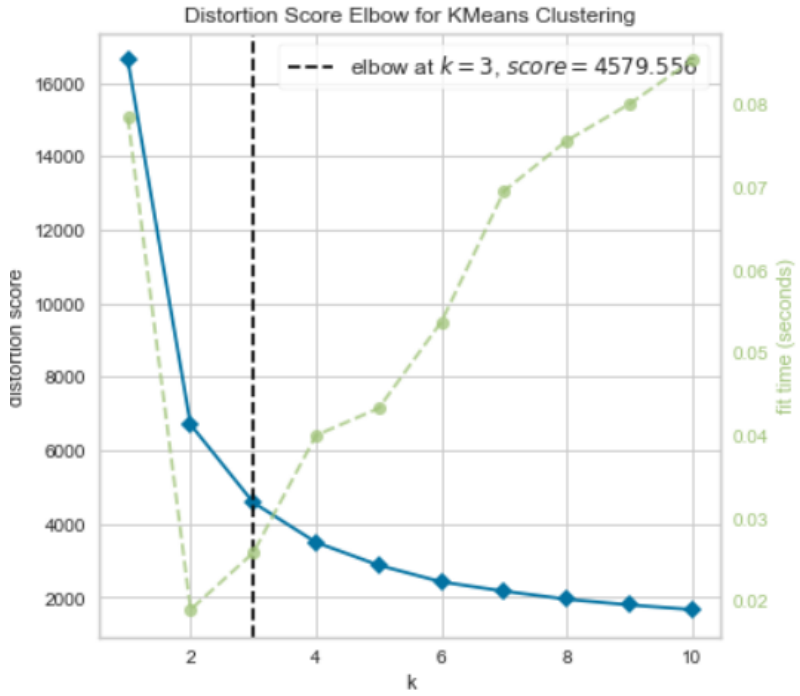
```
rfm_normalized = (rfm_df-rfm_df.min())/(rfm_df.max()-rfm_df.min())
rfm_normalized.shape
```

```
(5536, 3)
```

8.7 ANEXO 7: NÚMERO DE GRUPOS OPTIMOS

Se muestra el código implementado para obtener el número de grupos óptimos para aplicarlo en el modelo K means, las librerías que se usaron.

```
model = KMeans()
visualizer = KElbowVisualizer(model, k=(1,11))
visualizer.fit(RFM_Table_scaled)       # Entrenamos con los datos
numero_cluster = visualizer.elbow_value_
visualizer.show()        # Renderizamos la imagen
```



8.8 ANEXO 8: APLICACIÓN DE ALGORTIMO K MEANS

A continuación, se muestra el código utilizado para la aplicación del algoritmo k means

```
clustering = KMeans(n_clusters=3,max_iter=1000) #Crea el modelo
clustering.fit(RFM_Table_scaled) #Aplica el modelo a los datos
```

```
KMeans
KMeans(max_iter=1000, n_clusters=3)
```

```
KMeans(n_clusters = 3 ,init='k-means++', n_init = 10 ,max_iter=1000,
       tol=0.0001,  random_state= 111  , algorithm='elkan')
```

```
KMeans
KMeans(algorithm='elkan', max_iter=1000, n_clusters=3, random_state=111)
```

8.9 ANEXO 9: INSERCION A LA BASE DE DATOS

A continuación, se muestra el código utilizado para la inserción a la base de datos de los resultados obtenidos por k means

```
for index, row in rfm_lotero.iterrows():
    cursor.execute("INSERT INTO tbClusterKmeans (IdCliente,Recencia,Frecuencia,Monto,R_quartil,F_quartil,M_quartil,Tot_rfm,Cluster,Puntaje
                   "VALUES(?,?,?,?,?,?,?,?,?,?)",
                   row.IdCliente,row.Recencia,row.Frecuencia,row.Monto,row.R_quartil,row.F_quartil,row.M_quartil,row.RFM_Total,row.Cluster
cnxn.commit()
cursor.close()
```

8.10 CARTA DE CONFORMIDAD

LOTERIA NACIONAL
JUNTA DE BENEFICENCIA DE GUAYAQUIL

Guayaquil, 18 de Agosto 2022

Carta de Conformidad

Ingeniero
Edison Toala Quimí, Mgs.
Coordinador de la unidad de titulación

De mis consideraciones. -

Por medio de la presente certifico que la estudiante **Ingrid Gabriela Barrezueta Flores**, con C.I. **0940442007**, estudiante de titulación de la Facultad de Ingeniería carrera Sistemas Computacionales de la Universidad Católica de Santiago de Guayaquil, realizó a nuestra entera satisfacción el proyecto **"ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMATICO PARA FIDELIZACIÓN DE CLIENTES"**, motivo por el cual cumplo comunicar que el proyecto mencionado en cuestión cumplen con los requerimientos solicitados necesarios y será de gran utilidad para la empresa.

Atentamente.

Ing. Jorge Enrique Medina
Gerente General
H. Junta de Beneficencia de Guayaquil
Lotería Nacional

8.11 CARTA DE ACEPTACIÓN

Guayaquil, 30 de mayo 2022

Sr. Ingeniero
Jorge Enrique Medina
GERENTE GENERAL DE LOTERIA NACIONAL

Presente,

De mi consideración.

Yo, Ingrid Gabriela Barrezueta Flores con C.I 0940442007, estudiante de la unidad de titulación de la carrera de Ingeniería en Sistemas Computacionales de la Universidad Católica de Santiago de Guayaquil, por medio de la presente me dirijo a usted para realizar la apertura de mi proyecto de titulación dentro de su institución, el mismo trata de "ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMATICO PARA FIDELIZACIÓN DE CLIENTES", el cual pretende aportar información valiosa sobre problemas con retención de clientes y sería muy importante realizarlo en tan distinguida institución.

Por la atención a la presente, de antemano mi más sincero agradecimiento.

Atentamente.

Ingrid Gabriela Barrezueta Flores
C.I 0940442007

Aceptado
02-06-2022

DECLARACIÓN Y AUTORIZACIÓN

Yo, **Ingrid Gabriela Barrezueta Flores**, con C.C: # **0940442007** autor/a del trabajo de titulación: "**ELABORACIÓN DE UNA PROPUESTA TECNOLÓGICA BASADA EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA FIDELIZACIÓN DE CLIENTES."** previo a la obtención del título de **Ingeniero en Sistemas Computacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 15 de septiembre de 2022

_____

Nombre: **Ingrid Gabriela Barrezueta Flores**

C.C: **0940442007**

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	**Elaboración de una propuesta tecnológica basada en algoritmos de aprendizaje automático para fidelización de clientes**		
AUTOR(ES)	**Ingrid Gabriela Barrezueta Flores**		
REVISOR(ES)/TUTOR(ES)	**Colon Mario Celleri Mujica**		
INSTITUCIÓN:	**Universidad Católica de Santiago de Guayaquil**		
FACULTAD:	**Ingeniería**		
CARRERA:	**Ingeniería en Sistemas Computacionales**		
TITULO OBTENIDO:	**Ingeniero en Sistemas Computacionales**		
FECHA DE PUBLICACIÓN:	**15** de **septiembre** de **2022**	**No. DE PÁGINAS:**	**76**
ÁREAS TEMÁTICAS:	**Inteligencia artificial, Análisis de datos, Análisis de clientes**		
PALABRAS CLAVES/ KEYWORDS:	**Aprendizaje Automático, Clustering, Fidelización de clientes, Modelo RFM**		

RESUMEN/ABSTRACT:

El aprendizaje de maquina o aprendizaje automático se encuentra presente en diferentes industrias modernas, pero, es popularmente aplicado en el sector comercial para el análisis de clientes, su aplicación permite entre otros aspectos descubrir patrones en el comportamiento de clientes que las empresas pueden utilizar para aplicar estrategias comerciales, como retener o fidelizar clientes. El agrupamiento o clustering es una técnica muy utilizada en el aprendizaje automático para este tipo de análisis, se basa en la partición de un conjunto de datos en varios grupos en donde cada grupo contiene elementos similares entre sí y mantiene una diferencia respecto a los otros grupos. El presente trabajo de titulación tiene como objetivo obtener la segmentación de clientes de la empresa Lotería Nacional mediante la aplicación de algoritmos de aprendizaje automático, para ello se crearon variables que permitieron identificar el nivel de lealtad de los clientes de la empresa Lotería Nacional. Para el desarrollo del presente trabajo de titulación, se aplicó la metodología CRISP-DM que sirvió para el proceso de minería de datos. El análisis de los datos se lo realizó en base al modelo RFM (Recencia, Frecuencia, Monto) y sobre este modelo se aplicaron los algoritmos de agrupamiento k means, k nearest neigbor y árbol de decisión. Para validar el resultado de los algoritmos se separaron los datos para entrenamiento y pruebas que permitieron evaluar la precisión de los algoritmos, finalmente se utilizó la herramienta Power BI para presentar los resultados de una forma amigable y sencilla.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:**+593-995-205406	E-mail: **gabrielabarrezueta94@gmail.com**
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Toala Quimí, Edison José**	
	Teléfono: +593-990-976776	
	E-mail: edison.toala@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA

Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	