UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

TESIS PARA OBTAR EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

"MODELO DE ANALISIS PREDICTIVO PARA LA GESTION DE ABASTECIMIENTO DE LA EMPRESA TOP LLANTAS UTILIZANDO LENGUAJE R"

Área de investigación:
Gestión de Datos y de Información

Autor(es):
Br. Jose Armando Principe Arteaga
Br. Jhon Cristian Saavedra Campos

Jurado Evaluador:

Presidente: Abanto Cabrera, Heber Gerson
Secretario: Castillo Robles, Edward Fernando
Vocal: Meléndez Revilla, Karla Vanessa

Asesor:
Ms. Agustín Eduardo Ullón Ramirez
Código Orcid: https://orcid.org/0000-0003-1198-1855

TRUJILLO – PERÚ
2021

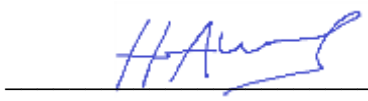
Fecha de sustentación: 2021/08/13

“MODELO DE ANALISIS PREDICTIVO PARA LA GESTION DE ABASTECIMIENTO DE LA EMPRESA TOP LLANTAS UTILIZANDO LENGUAJE R”

Elaborado por:

Br. Jose Armando Principe Arteaga

Br. Jhon Cristian Saavedra Campos



Ing. Heber Gerson Abanto Cabrera
Presidente
CIP: N° 106421

Ing. Edward Fernando Castillo Robles
Secretario
CIP: N° 192352

Ing. Karla Vanessa Meléndez Revilla
Vocal
CIP: N° 120097

Ing. Agustín Eduardo Ullón Ramírez
Asesor
CIP: 137602

PRESENTACIÓN

Señores Miembros del Jurado:

Acorde a los requerimientos dispuestos en el reglamento de grados y Títulos de la universidad y el reglamento interno de la escuela profesional de ingeniería de computación y sistemas ponemos a disposición el Trabajo de investigación titulado:

“MODELO DE ANALISIS PREDICTIVO PARA LA GESTION DE ABASTECIMIENTO DE LA EMPRESA TOP LLANTAS UTILIZANDO LENGUAJE R”

La presente investigación fue desarrollada bajo el marco de referencia de los lineamientos establecidos por la Facultad de Ingeniería, la Escuela Profesional de Ingeniería de Computación y Sistemas, así como lo aprendido durante el periodo de formación profesional en la universidad.

Los autores.

DEDICATORIA

En primer lugar, un agradecimiento a Dios por guiarme y haberme permitido cumplir una de mis metas: culminar satisfactoriamente mi carrera profesional.

A mis padres y a mi hermana por aconsejarme, guiarme y ayudarme durante mis años de estudio, siendo siempre mi gran motivación. Es meritorio brindar unas palabras de eterna gratitud a todos mis amigos, quienes aportaron a mi formación profesional logrando siempre nuestro objetivo.

Br. Jose Armando Principe Arteaga

Esta tesis se la dedico a Dios y toda mi familia por darme el apoyo y las fuerzas para continuar con mis objetivos; y por guiarme en el camino del bien y los buenos valores.

A mis padres por su comprensión, consejos y las enseñanzas que formaron la persona que soy. A mi abuela por ser mi gran motivación, mis tíos y hermano por ser mi ejemplo de lucha y superación ante los obstáculos que la vida me presento.

A mis amigos con quienes compartí muchas experiencias y conocimientos que nos hicieron lograr nuestro objetivo y por superar los problemas que el grupo y cada uno tenía.

Br. Jhon Cristian Saavedra Campos

AGRADECIMIENTO

Un especial agradecimiento a la empresa Top Llantas de Trujillo, quienes nos permitieron acceder a toda la información necesaria para el desarrollo de la tesis.

También un agradecimiento a nuestro asesor Ing. Agustín Ullón, por la orientación durante el desarrollo de inicio a fin de a tesis.

Agradecemos también a todos los docentes, compañeros y personas que nos acompañaron durante los años de aprendizaje en la universidad.

Los autores.

RESUMEN

“MODELO DE ANALISIS PREDICTIVO PARA LA GESTION DE ABASTECIMIENTO DE LA EMPRESA TOP LLANTAS UTILIZANDO LENGUAJE R”

Por:

Br. Jose Armando Principe Arteaga

Br. Jhon Cristian Saavedra Campos

Actualmente ha incrementado la cantidad de vehículos que circulan por las calles, debido a esto existe un aumento en la demanda de neumáticos para los diferentes tipos de vehículos. Esto genera que los distribuidores tengan dificultad para medir el abastecimiento e inversión dentro del mercado, ya que no poseen las herramientas que faciliten esta gestión.

Es por eso que la presente investigación propone desarrollar un modelo de análisis predictivo para la gestión de abastecimiento de la empresa top llantas utilizando lenguaje R; basado en la evaluación de cuatro modelos de aprendizaje supervisado, como son Árbol de decisiones, Random Forest, Naive Bayes, SVM.

Para el desarrollo de la solución del proyecto se utilizó la Herramienta Rstudio junto al leguaje R; la biblioteca de paquetes que proporciona nos da la facilidad del manejo y desarrollo de los diferentes algoritmos de los modelos, permitiéndonos realizar el proceso de todas las fases del análisis.

ABSTRACT

"PREDICTIVE ANALYSIS MODEL FOR SUPPLY MANAGEMENT OF THE COMPANY TOP LLANTAS USING R LANGUAGE"

By:

Br. Jose Armando Principe Arteaga

Br. Jhon Cristian Saavedra Campos

Currently, the number of vehicles that circulate on the streets has increased, due to this there is an increase in the demand for tires for different types of vehicles. This makes it difficult for distributors to measure supply and investment within the market, since they do not have the tools to facilitate this management.

That is why this research proposes to develop a predictive analysis model for the supply management of the company Top Llantas of Trujillo city using R language; based on the evaluation of four supervised learning models, such as Decision Tree, Random Forest, Naive Bayes, SVM.

For the development of the project solution, the Rstudio Tool was used together with the R language; the package library that it provides gives us the ease of handling and developing the different algorithms of the models, allowing us to carry out the process of all the phases of the analysis.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1. Planteamiento del problema

Actualmente en el Perú, encontramos más de 160 marcas de neumáticos que abastecen a diferentes empresas de llantas. El crecimiento de consumo de llantas anual en el Perú asciende en un promedio de 100 mil unidades de neumáticos, y cabe recalcar que la reposición de desgaste de llanta promedio es de 3 años, por cual el rubro de neumáticos viene teniendo un gran crecimiento en el mercado. Debido al crecimiento y expansión de mercado de estas empresas, implican un manejo de data mayor, la cual se perjudica ya que no se tienen las herramientas para poder realizar un procesamiento y análisis de la data.

Según el APP, informó que, en enero de 2020, aumentó la cifra de demanda de autos en comparación al año anterior, pero según la empresa indico que no se tiene una exploración de su data la cual pueda proporcionarle un modelo de tendencia y le ayude a tomar mejores decisiones al momento de realizar la inversión en la adquisición de los neumáticos y garantice una mayor venta.

En la empresa Top Llantas, no se tiene identificado las cantidades necesarias para satisfacer las necesidades de los clientes, ni los productos que mayores ganancias le dan a la empresa. No se tiene un conocimiento previo de la cadena de suministros para atender las demandas del mercado y mejorar la gestión del inventario. No hay una evaluación de la información, por lo que la empresa desconoce la cantidad de abastecimiento de los suministritos necesarios para los requerimientos de sus clientes, siendo su principal problema la escases de productos generando un pedido improvisado de la cantidad exacta que no se cuenta en ese momento, esto genera un gasto adicional ya que al realizar un pedido de mayor cantidad, los costos y gastos serían menores generando un pequeño margen de ganancia; por otra parte, se tiene algunos repuestos sobrantes que han quedado obsoleto en la empresa.

Este proyecto analizará la data de las ventas de llantas de años anteriores con el fin de estimar la cantidad necesaria y clasificándolas a través de algoritmos de aprendizaje supervisado poder tomar mejores decisiones en la gestión de abastecimiento.

Características problemáticas

La empresa presenta los siguientes principales problemas:

- No se cuenta con una evaluación de la información, no hay un correcto análisis, ni mucho menos una optimo conocimiento.
- No se cuenta con un control de los suministros, por ejemplo, el control de cantidades para el abastecimiento.

Análisis de características problemáticas

- ✓ La empresa no cuenta con una herramienta tecnológica que le permita procesar su data y pueda dar resultados a nuevos conocimientos los cuales le den soporte a tomar buenas decisiones en las inversiones para el abastecimiento de neumáticos y aros.
- ✓ No hay conocimiento de métodos de análisis predictivos para el control y gestión de abastecimiento de suministros.

1.2. Delimitación del problema

El siguiente proyecto se delimitará en el Análisis de un modelo predictivo utilizando la herramienta R para mejorar la toma de decisiones en el abastecimiento de mercancía para la empresa Top Llantas.

1.3. Formulación del problema

¿En qué medida un modelo de análisis predictivo sobre la información del abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo influye en la toma de decisiones sobre la gestión de suministros?

1.4. Hipótesis

1.4.1 General

El desarrollo de un modelo de análisis predictivo sobre la información de abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo, permitirá tener una mejor toma de decisiones sobre la gestión de los suministros.

1.5. Objetivos

1.5.1 General

Desarrollar un modelo de análisis predictivo para la gestión de abastecimiento de la empresa top llantas utilizando lenguaje R.

1.5.2 Específicos

- Recabar todos los datos necesarios relacionados a las ventas e inventarios de los diferentes tipos de modelos y marcas de llantas.
- Desarrollar la limpieza y procesamiento de los datos obtenidos de la empresa top llantas.
- Crear el modelo predictivo y realizar la validación “hold out” y “k-fold” a través del lenguaje R, para elegir el modelo más eficaz.
- Evaluar e interpretar los modelos de análisis predictivo de abastecimiento de llantas.

1.6. Justificación de la investigación

Académica

Utilizaremos los conocimientos aprendidos en los diferentes cursos tales como machine learning, sistemas inteligentes y la metodología KDD, también utilizaremos conceptos de minería de datos, donde aplicaremos diferentes algoritmos en este caso aplicaremos los que pertenecen al aprendizaje supervisado, los cuales nos ayudaran a afrontar esta situación problemática.

Organización

El avance tecnológico ha desarrollado muchas herramientas para el análisis de grandes datos, por consecuencia el modelo de análisis predictivo permitirá a la empresa mejorar la toma de decisiones respecto a la gestión del abastecimiento, para así luego hacer una mejor inversión de tal forma que pueda generar una mayor ganancia.

Tesista

Nos permitirá mejorar y obtener conocimientos acerca del lenguaje R y sus amplias librerías para la exploración y manipulación de datos, también vamos adquirir conocimientos en la minería de datos y los patrones que nos permitirán desarrollar proyectos sobre el análisis de datos a futuro.

2. MARCO TEÓRICO

2.1. Antecedentes de la Investigación

(Espino Timón, 2017) "Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo – herramientas Open Source que permiten su uso", el proyecto presenta un estudio acerca de las herramientas tecnológicas para el análisis predictivo, ya que hay una enorme importancia en el tema de análisis de grandes volúmenes de data en la empresa y se requiere sacar ventaja de esto y aprovecharla para poder competir. Este análisis requiere de herramientas las cuales te permitan detectar ciertos patrones para generar un conocimiento nuevo. Se realizaron las pruebas en dos herramientas R-Studio y Weka, en ambos se aplicó el algoritmo de árbol de decisiones y otros modelos de agrupamiento, siendo R-Studio la herramienta más flexible y con mayor potencia.

(Apolaya Torres & Espinosa Diaz, 2018) "Técnicas de Inferencias, Predicción y Minería de Datos", el objetivo del proyecto es implementar un modelo basado en árbol de decisiones que le permita una mejor toma de decisiones en la Escuela de Ing. de Sistemas y Computación de la UPC. Se utilizó la metodología KDD. La investigación inicio con la selección de datos en

este caso se utilizó la data de alumnos, depurando los campos innecesarios, haciendo una limpieza y normalización de la data, luego realizo la transformación y se aplicó la minería de datos, en este caso se seleccionó el algoritmo árbol de decisiones para realizar las pruebas, obteniendo un porcentaje de error de un 9,13% en la ejecución del algoritmo.

(Grández Márquez, 2017) "Aplicación de Minería de Datos para Determinar Patrones de Consumo Futuro en Clientes de una Distribuidora de Suplementos Nutricionales", el proyecto busca determinar los patrones de consumo de una distribuidora de suplementos nutricionales, se utilizó la metodología CRISP-DM y los algoritmos que se utilizaron fueron un modelo de asociación, modelo de Clúster y modelo de red neuronal, siendo el más efectivo según los datos el modelo de asociación logrando cumplir con los objetivos planteados.

(Morales Tabares, 2016) "Modelo Multivariado de predicción del stock de repuestos para equipos médicos", el proyecto presenta un modelo MPREDSTOCK, el cual realiza la predicción del stock de repuestos mediante algoritmos matemáticos. Primero se realizó la recolección de datos, luego se hizo la exploración identificando algunos indicadores como por ejemplo frecuencia de las piezas por fallas, por rotura. Luego se verifico a través del coeficiente correlacional lineal de Pearson, se aplicaron y luego evaluar la precisión de la predicción. En la evaluación del modelo de regresión múltiple resulto ser el más adecuado para realizar el pronóstico arrojando un 92% de precisión. Los patrones permitieron identificar diferentes tipos de indicadores, permitiendo tener conocimiento de las frecuencias de las fallas, disponibilidad de stock.

(Angeles Gonzales, 2017) "Analítica de negocios en la gestión de ventas de la empresa Inversión Generales Fabrizio", este proyecto propone el diseño de un dashboard para mejorar la gestión de ventas mediante un análisis empleando la minería de datos para la toma de decisión a futuro. El principal motivo del desarrollo de este modelo es hacer una exploración

de los datos obteniendo una información futura detallada. Los resultados muestran un dashboard con un control total en la gestión de las ventas, demostrando que la analítica de negocios a través de la minería de datos y la presentación de un dashboard son una propuesta tecnológica para la toma de decisiones.

(Jiménez Chura, 2017) "Análisis Predictivo para los procesos de admisión de la universidad de la universidad nacional del altiplano", el objetivo de este proyecto es predecir la tendencia de postulantes de acuerdo a la data de las escuelas públicas y privadas. Se utilizó la metodología CRISP-DM para el desarrollo del análisis predictivo, y se desarrollará en el software R utilizando los paquetes RMySQL, ggplot, plynom, entre otros, los cuales nos permitirán explorar y manipular la data y también poder graficarla para la interpretación del conocimiento obtenido. El resultado final nos predijo y confirmo el crecimiento en las escuelas como ing. Civil, Contabilidad, también nos muestra el nivel de formación de las escuelas que postularon a la universidad.

2.2. Fundamentación teórica de la investigación

2.2.1 Análisis de Predicción

En el análisis de predicción se utiliza la estadística, aprendizaje automático y diferentes técnicas de base de datos, así a partir de un conjunto de datos históricos y actuales poder hacer un pronóstico a futuro; también agrupa técnicas de modelamiento de procesos, tecnologías de la información. (Vaibhav & M. L., 2018)

2.2.1.1 Proceso

Para realizar un análisis de predicción en necesario llevar a cabo una serie de pasos, los cuales facilita el trabajo del analista a la hora de predecir a futuro. (Vaibhav & M. L., 2018)

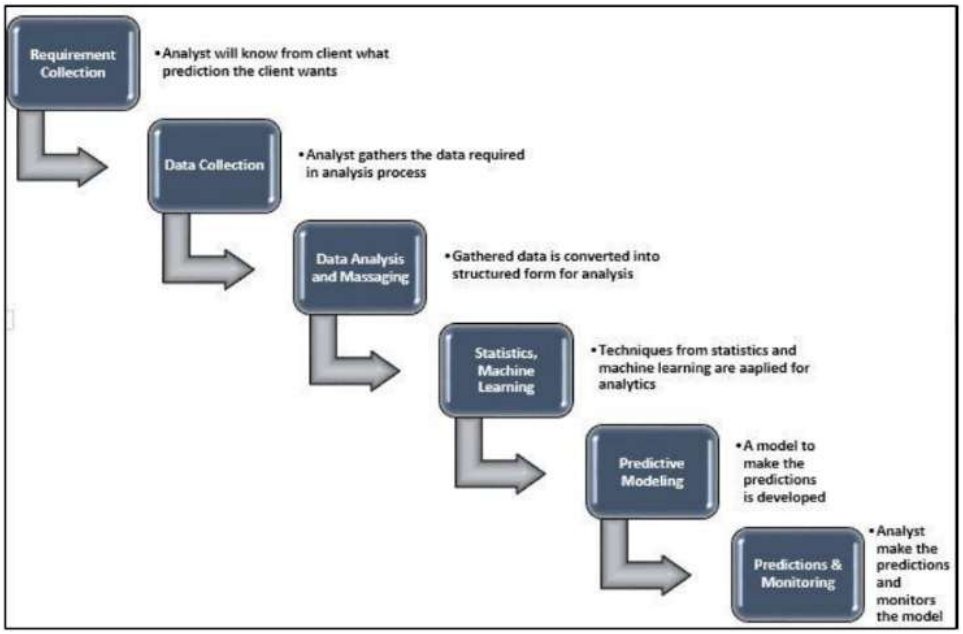


Figura 1 Proceso del Análisis Predictivo (Vaibhav & M. L., 2018)

2.2.1.1.1 Recolección de Requerimientos

Antes de realizar un modelo de predicción, se debe conocer cuál es la finalidad de éste. Se tiene que precisar cuál será la información que se quiere obtener. (Vaibhav & M. L., 2018)

2.2.1.1.2 Recolección de Datos

Una vez definido los requerimientos, se realiza la recolección de datos. Estos datos pueden proceder de distintas fuentes y pueden ser tanto estructuradas como no estructuradas. (Vaibhav & M. L., 2018)

2.2.1.1.3 Análisis de los Datos

Los datos recolectados deben ser analizados y preparados antes de incluirlos en el modelo. Se hace la conversión de los datos no estructurados a estructurados; luego de ello, se debe de verificar la calidad de los datos, ya que la precisión del modelo predictivo depende de la calidad. (Vaibhav & M. L., 2018)

2.2.1.1.4 Estadística, aprendizaje automático

Para el desarrollo de un modelo de predicción es necesario aplicar distintos conocimientos estadísticos y aprendizajes automáticos. Todos los modelos relacionados con predicción se fundamentan en procedimientos estadísticos. (Vaibhav & M. L., 2018)

2.2.1.1.5 Modelos de Predicción

En este paso se procede con el desarrollo del modelo de predicción. Una vez desarrollado, se realiza un test con los datos de prueba, que son una parte del total de datos recolectados. Verificada la efectividad y es considerado idóneo, se arregla para aplicar el modelo y hacer la predicción a través de los datos restantes. (Vaibhav & M. L., 2018)

2.2.1.1.6 Predicción y monitoreo

Luego de la efectividad del modelo de predicción, este se desarrolla en el sistema para poder realizar las predicciones y mejorar la toma de decisiones de manera diaria. (Vaibhav & M. L., 2018)

2.2.1.2 Ventajas del modelo predictivo

Según (Vaibhav & M. L., 2018) referencia que en la actualidad existe bastante demanda por parte de empresas que buscan realizar predicciones, ya que buscan efectividad en el mercado y mejorar sus beneficios. Dentro de todas las ventajas que se obtiene, algunas de los más comunes son:

- ✓ Detección de Fraudes
- ✓ Reducción de Riesgos
- ✓ Optimización de la Campaña de Marketing
- ✓ Sistema de apoyo en las decisiones

2.2.2 Machine Learning

El aprendizaje automático es una tecnología que abarcando gran parte del sistema web por ejemplo la red de internet nos recomienda lugares de acuerdo a nuestro historial de búsqueda, en las redes sociales nos sugieren usuarios amigos, sitios web como Amazon, recomendación de productos, etc., estamos expuestos todos los días sin saber al aprendizaje automático. (Bhatia, 2019)

Es considerado como un método científico que consiste en la enseñanza a computadoras, sin la necesidad de ser programadas, a la extracción de patrones y similitudes que pueda haber entre nuestros datos; para después poder pronosticar comportamientos y mejorar la toma de decisiones. (Valdez Alvarado, 2018)

El aprendizaje automático, facilita el entrenamiento de los modelos con los grupos de datos. Existen modelos de aprendizaje que por ellos mismos son capaces de ajustarse constantemente, dependiendo de la continuidad de ingreso de nuevos datos. También, encontramos modelos que se originan de algoritmos de aprendizajes de máquinas, los cuales no varían al desarrollarse. (Hurwitz & Kirsch, 2018)

2.2.2.1 Tipo de Aprendizajes

Para lograr una mayor exactitud de los modelos de predicción, se requieren de diferentes métodos de aprendizaje. Existen distintas perspectivas dependiendo el grado del problema que se esté planteando. (Hurwitz & Kirsch, 2018)

2.2.2.1.1 Aprendizaje Supervisado

Este tipo de aprendizaje tiene como misión hallar patrones dentro del grupo de datos que sirvan para ser analizados, lo cuales poseen cualidades en común que identifican a los datos. (Hurwitz & Kirsch, 2018)

Árboles de Decisiones

Este es un modelo de clasificación predictivo utilizando condiciones, el árbol está compuesto por una raíz que es el nodo principal por el cual se empezara la condición, seguidos de los nodos contienen los atributos que se ingresan y por ultimo las ramas que vienen hacer el resultado de las condiciones. Los algoritmos que se aplican en este modelo se llaman ID3 y C4.5. (Bhatia, 2019)

El algoritmo de árbol de decisiones realiza la clasificación introduciendo la poda y crecimiento del árbol, y este crecimiento depende de las características que se asignan, para luego realizar la poda y el árbol pueda optimizar el rendimiento. La precisión del algoritmo depende de la recolección y el pre procesamiento de los datos generados. (Berry, Yap, & Mohamed, 2020)

Redes Bayesianas

Es un modelo de clasificador y funciona muy diferente al árbol de decisiones, este algoritmo se basa en calcular la probabilidad de que una hipótesis sea cierta, clasifican las características de una clase particular. La precisión de los algoritmos se mide a través de una matriz de confusión las cuales nos da la precisión a través de los verdaderos positivos y los falsos positivos. (Bhatia, 2019)

El modelo de Naive Bayes es considerado un algoritmo de aprendizaje supervisado y no supervisado porque es usado tanto en los modelos de agrupación como los de clasificación, es conocido a su gran teorema de probabilidades, dado que el algoritmo trabaja con redes bayesianas y crea grafos los cuales generan probabilidades los cuales lleguen al resultado. El modelo como clasificador tiene que proporcionar las etiquetas correspondientes las cuales ayudan a determinar el objetivo probabilístico. (Berry, Yap, & Mohamed, 2020)

Support Vector Machines (SVM)

Es un modelo clasificador el cual utiliza los hiperplanos para realizar la clasificación de características, estos hiperplanos dividen las clases y para encontrar el hiperplano resultante más óptimo, el margen que los divide debe ser el máximo posible. Su objetivo es descubrir el más alto margen de separación y así aumentar la precisión del algoritmo. (Swamynathan, 2017)

El algoritmo SVM utiliza márgenes los cuales hacen posible que la clasificación tenga un error mínimo, estos márgenes describen las distancia que se genera entre las clases a través de un hiperplano. La precisión del modelo depende en las infracciones del margen y la clasificación errónea de las características en ambos lados. Estas infracciones se cometen cuando encontramos algunas clases quedan fuera de los vectores correspondientes creando un margen de error, entonces mientras mayor sea la distancia del margen los puntos serán clasificados con mayor precisión o efectividad. (Berry, Yap, & Mohamed, 2020)

2.2.2.1.2 Aprendizaje No Supervisado

Suele utilizarse en datos muy extensos y que no poseen características entre ellos. Para poder deducir su significado es necesario aplicar algoritmos que analicen y empiecen a buscar similitudes para ser clasificados fundamentándose en patrones. (Hurwitz & Kirsch, 2018)

Este tema nos permite abordar ciertas situaciones en donde no se sabrá los resultados u objetivos que se quiere lograr, es decir, no se tiene características de salida y los algoritmos funcionan a través de la agrupación de características similares, ya que los datos no usan etiquetas. No existe una retroalimentación por lo que no hay alguien para que te corrija. (Bhatia, 2019)

El objetivo del aprendizaje no supervisado es conocer e investigar eventos desconocidos y descubrir los patrones semejantes que pueden agruparse en clases. (Swamynathan, 2017)

Support Vector Machines (SVM)

En este caso, el algoritmo SVM trabaja como clustering en este caso se obtienen los datos en la información previa aprendida para después reconocer las clases similares dentro dataset. Este modelo es el indicado para tratar casos con grandes conjuntos de datos, los cuales reducen ciertas características y así tener una mayor precisión. (Berry, Yap, & Mohamed, 2020)

K-Means

El modelo se utiliza para descubrir nuevas clases en un conjunto de datos que no son etiquetas, el algoritmo se desarrolla a través de las distancias medias que hay en las clases. El modelo lo que realiza es asignar valores a las clases de acuerdo a la proximidad media más cerca al menor error. Este los agrupa en k grupos las clases para luego descubrir las características similares de las clases a través de las distancias minimizadas entre clases. (Berry, Yap, & Mohamed, 2020)

2.2.3 Minería de Datos

Se define como minería de datos al proceso de exploración y análisis de información importante de una gran base de datos, el cual es realizado por medio de algoritmos el cual se encarga de identificar patrones, con la finalidad de que se pueda tener conocimiento del negocio, tener una mejor toma de decisiones y poder realizar pronósticos. (Hurwitz & Kirsch, 2018)

2.2.3.1 Fases de la Minería de Datos

Este proceso implica siete fases:

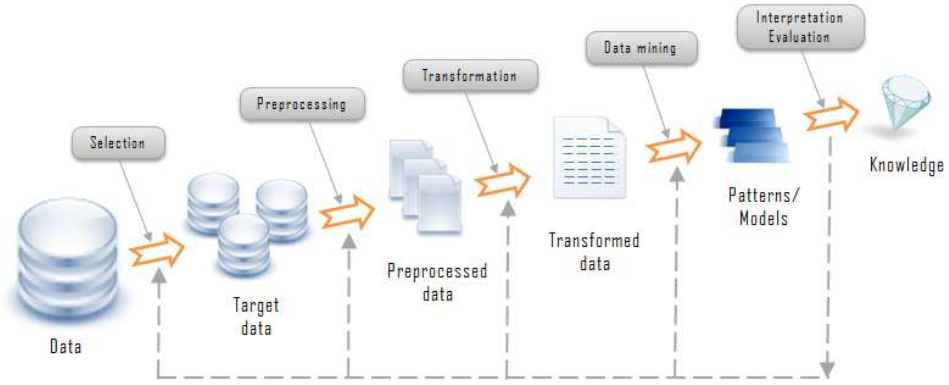


Figura 2 Fases de la Minería de Datos (Sumiran, 2018)

- **Fase 1 – Integración de los datos**

  Implica realizar la recolección de los datos

- **Fase 2 – Selección de datos**

  Consiste en seleccionar solo los datos que sean necesarios.

- **Fase 3 – Limpieza de los datos**

  En esta fase se realiza la depuración de datos.

- **Fase 4 – Transformación de los datos**

  Normalización de los datos

- **Fase 5 – Data Mining**

  Se aplicarán los diferentes métodos de minería

- **Fase 6 – Evaluación**

  Identificación y descarte de patrones repetidos

- **Fase 7 – Presentación del conocimiento**

La extracción de los datos se fracciona por lo general en dos, la minería de datos descriptiva, que consiste en examinar patrones para identificar los datos; y la minería de datos predictiva,

la cual se encarga de predecir la conducta del modelo apoyándose en el grupo de datos. (Sumiran, 2018)

2.2.3.2 Técnicas

2.2.3.2.1 Asociación

Es la más popular, consiste en encontrar patrones enfocándose en el parentesco que existe entre un elemento en concreto con los demás de idéntica transacción. (Sumiran, 2018)

2.2.3.2.2 Patrones secuenciales

Se basa en el hallazgo de patrones en la base de datos de secuencias, estos se aplican en la examinación siguiente con la finalidad de identificar la relación con los datos. (Sumiran, 2018)

2.2.4 R para la Ciencia de Datos

El objetivo de R es el análisis de los datos sin procesar y convertirlos en nuevo conocimiento, R nos brinda las mejores librerías y es una herramienta muy útil para la ciencia de los datos. (Grolemund & Wickham, 2017)

2.2.4.1 R

Se compone de servidores espejos los cuales distribuyen todos los paquetes de R que se utilizan para la ciencia de datos. (Grolemund & Wickham, 2017)

R es un lenguaje de programación que deriva otro lenguaje llamado S, cuenta con más de 10.000 paquetes cubriendo así una gran variedad de campos desde estudios financieros, bayesianas, análisis de datos, etc. Dentro de sus características podemos destacar su variedad de herramientas estadísticas que posee para el análisis de datos, es multiplataforma, permite a los usuarios definir sus propias funciones, posee capacidades graficas muy destacables y es libre. (Jiménez, 2019)

Python podría ser clasificado, dentro de los lenguajes de programación, como un lenguaje de alto nivel, pudiendo ser interpretado por diferentes sistemas operativos y aceptando distintas técnicas de programación; también posee con unas reglas de estilo en su estructura de código, con la finalidad de estandarizar la forma de programar. (Bahit, 2018)

Ambos lenguajes de programación son de código abierto y gratuitos, sin embargo, R se orientas más hacia el lado estadístico y reportes en general, ya que cuenta con una gran variedad de paquetes dentro de los cuales la mayoría van orientados al análisis de datos, por ende, es utilizado con más frecuencia en lo que respecta a la ciencia de los datos. Por otro lado, Python tiene una orientación más hacia el desarrollo, y posee mayor practicidad a la hora de manipular grupos de daros masivos y de diferentes plataformas.

2.2.4.2 RStudio

Es el entorno para la programación de R, tiene un panel de consola muy bien distribuida con sus salidas gráficas. (Grolemund & Wickham, 2017)

2.2.4.2.1 Importación de Datos

Para este paso se necesitará instalar el paquete “tidyverse” para leer archivos rectangulares de texto plano (videos, audios, Word, pdf, Excel, etc.)

Este paquete contiene las siguientes funciones:

- Read_csv (): Lee archivos separados por comas
- Read_csv2 (): Lee archivos separados por punto y comas.
- Read_tsv (): lee archivos delimitados por tabulaciones
- Read_delim (): lee archivos con cualquier formato delimitador.

Cuando se ejecuta cualquiera de estas funciones se imprime todas las especificaciones que contiene el documento.

2.2.4.2.2 Datos Ordenados

Para organizar nuestros datos necesitaremos instalar el paquete “Tidyr”, el cual nos permite dar variables a cada columna, su fila y celdas. (Grolemund & Wickham, 2017)

2.2.4.2.3 Separar y Unir

Aquí se emplean las funciones de “separate()” la cual se utiliza para separar una columna en varias columnas y la función “unite()” para volver a unir las columnas a una sola, ejemplos separar un formato de fecha y hora que vienen juntos en una columna. (Grolemund & Wickham, 2017)

2.2.4.2.4 Visualización

Aquí se instalará el paquete ggplot2 el cual se utiliza para la visualización de datos procesados para luego ser interpretados. La función ggplot() crea un sistemas de coordenadas x e y donde se pueden especificar ciertos argumentos como el tipo de gráfico, el color, etc.). (Grolemund & Wickham, 2017)

2.3. Metodología del Proyecto

Knowledge Discovery in Databases es modelo que se utiliza para el descubrimiento de nuevos conocimientos a través de la integración tecnológica. El conocimiento se descubre a partir de la gestión y tratamiento de los datos, se emplean algoritmos estadísticos los cuales van hacer entrenados de acuerdo al enorme conjunto de datos, manipulación y transformación de data, y por último los resultados son visualizados e interpretados. (Swamynathan, 2017)

2.3.1 Etapas del Proceso KDD

EL proceso es iterativo, lo que permite que puede regresar a las fases anteriores para reajustarlos:

2.3.1.1 Selección

En esta fase se determinan los datos para el descubrimiento del conocimiento, se integran los atributos necesarios que apoyaran con el descubrimiento de nuestro objetivo. Se seleccionan los datos correctos y relevantes para el análisis de los datos. (Swamynathan, 2017)

2.3.1.2 Procesamiento

A menudo cuando vemos un conjunto de datos real sabemos que están incompletos, esto se refiere a que faltan datos o hay espacio en blancos, etc., encontramos valores erróneos o atípicos, valores inconsistentes, lo que nos lleva a decir que los datos recogidos no incongruentes o se encuentran desordenados. Estos datos no fiables hacen que el procedimiento de minería de datos llegue hacer inválidos y confusos. (Swamynathan, 2017)

El procesamiento y limpieza sirven para mejorar los resultados que se realizaran en el proceso de la minería y se deben tomar las medias correspondientes, por ejemplo; eliminación de datos duplicados e innecesarios, reemplazar datos en blanco, aplicar técnicas para el tratamiento de datos faltantes, etc. (Swamynathan, 2017)

2.3.1.3 Transformación

En esta etapa, los datos se consolidan de tal forma se pueda dar una mejor forma a la minería de datos y se puedan encontrar los atributos necesarios para dicho proceso. Se reducen el número de las características y se cambian algunos de los formatos. Existen varias formas de transformación como, por ejemplo, la normalización de valores, reducción o división de la data, agregación de datos faltantes, suavización de datos, etc. (Swamynathan, 2017)

2.3.1.4 Minería de Datos

Una vez ya limpiada y transformada la data se aplican los algoritmos de machine learning, estos pueden ser supervisados y no supervisados los cuales nos ayudaran a descubrir los patrones para poder llegar a descubrir el conocimiento. Tenemos modelos predictivos los

cuales pueden ser de clasificación y regresión. Estos modelos predicen valores a futuro a través de patrones que ayudan a descubrir las características similares que tienen los datos. El objetivo es encontrar el modelo con más precisión y menos margen de error para la predicción. (Swamynathan, 2017)

2.3.1.5 Interpretación y Evaluación

En este paso final se muestran los resultados descubiertos o encontrados por los patrones de minería y son mostrados al usuario de una forma amigable, es decir; a través de una visualización de gráficos. (Swamynathan, 2017)

Presentamos los modelos utilizados en la minería de datos y se dan los resultados según la precisión que se han identificado. (Swamynathan, 2017)

3. MATERIAL Y MÉTODOS

3.1. Población

Empresas dedicadas a la venta de autopartes o accesorios automovilísticos en el departamento La Libertad.

3.2. Muestra

Registro de inventario y ventas del 2018 al 2019 de la empresa Top Llantas que se dedica a la compra venta de Llantas al público en general.

3.3. Unidad de Análisis

Abastecimiento de los productos

3.4. Metodología

3.4.1 Nivel de Investigación

Aplicada

3.4.2 Diseño de Investigación

Diseño Pre-experimental con pre-prueba y post-prueba

3.4.3 Variables de estudio y Operacionalización

VI: Modelo de análisis predictivo en la empresa Top Llantas.

VD: Mejor toma de decisiones en la gestión de abastecimiento de llantas.

Tabla 1 Operacionalización de las Variables

Variable	**Dimensión**	**Indicador**	**Unidad de medida**	**Instrumento de Investigación**
VI	Precisión	% de precisión de cada modelo	% precisión	Hoja de datos
VD	Satisfacción	Grado de satisfacción	Rango de satisfacción	Tabla de satisfacción

3.5. Técnicas e instrumentos de recolección de datos

- Observación
- Encuesta

3.6. Técnicas de procesamiento y análisis de datos

Para analizar la información se aplicará los instrumentos mencionados en el punto anterior para recopilar la información necesaria.

Para la evaluación e interpretación de los resultados se utilizará la prueba estadística T-student.

4. RESULTADOS: APLICACIÓN DE LA METODOLOGÍA

4.1. Integración de los datos

El objetivo institucional es reducir la problemática en el "abastecimiento de neumáticos" por lo que el objetivo esencial para este trabajo se presenta a través de un modelo predictivo utilizando el lenguaje R.

- Mejorar la toma de decisiones en la gestión de abastecimiento
- Conocer la predicción de la futura demanda según el tipo de auto

La data fue obtenida de un Excel, que es donde se almacenaba las ventas realizadas, inversiones y abastecimiento desde enero del 2018 hasta diciembre del 2019. En ella podemos observar la variedad de marcas, modelos y medidas que se venden en el establecimiento a diario, también su resistencia de cada uno de ellos junto con su precio tanto de compra para el abastecimiento, como de venta al público y su costo de traslado hacia el local. Estos neumáticos están clasificados según su índice de carga en autos, SUV, furgones y camionetas.

Tabla 2 Descripción de los datos

N°	CAMPO	DESCRIPCIÓN	VALORES
1	Fecha	El día, mes y año en los que se vendieron las llantas de la empresa TOP Llantas.	Desde enero del 2018 hasta diciembre del 2019
2	Marca	Clases de neumáticos	GoodYear, Hilo, Lima Caucho, etc.
3	Medida	Modelos de los neumáticos según el aro	Alfanuméricos
4	PR	Indica la resistencia del neumático según las capas que contenga.	4-6-8-10-12-14-16-18-20
5	Tipo_auto	Son las clases de automóviles	Auto, SUV, Furgón, Camión
6	Procedencia	El país de donde son adquiridos.	América, Brasil, China, Consignación, Ecuador, India, Japón, Korea, México, Perú, Polonia, Reino Unido, Taiwán, Tailandia, Vietnam
7	Cost_flete	El flete es el costo a pagar por el desplazamiento de una carga en un medio de transporte.	3-5-7
8	Cant_ingreso	Cantidad de la adquisición de neumáticos.	----------
9	Pre_dolar	Precio dólar de la compra del neumático al proveedor	----------
10	Pre_soles	Conversión de precio de compra del neumático al tipo de cambio a la fecha.	----------
11	Pre_venta	Precio de venta al publico	----------
12	Cant_venta	Cantidad de neumáticos vendidos.	----------

4.2. Selección de los datos

Exploración

La finalidad de la exploración es poder saber las características que poseen los datos y así descubrir su patrones y relaciones que existen entre ellos.

- **Total de ganancia por tipo de auto**

	tipo_auto <fctr>	total_ganancia <dbl>
1	Auto	6532482
2	Camion	112095
3	Furgon	747525
4	SUV	2524504

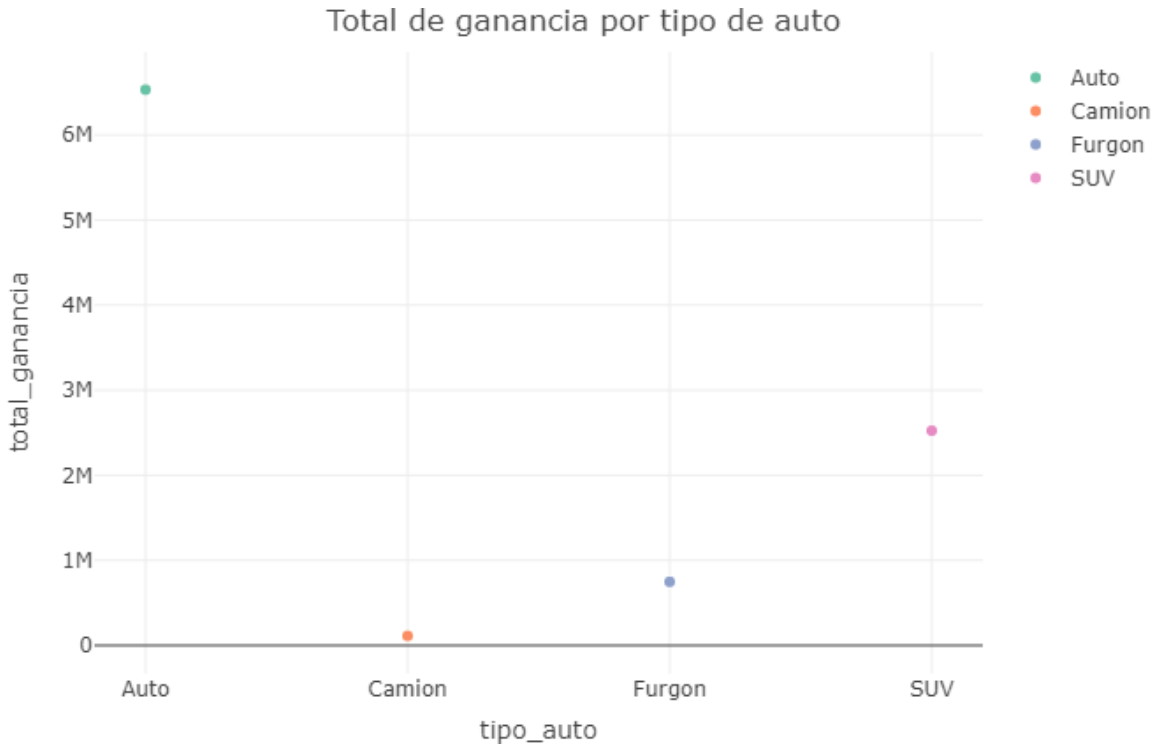


Figura 3 Total de ganancia por tipo de auto

- **Cantidad de ventas por tipo de auto**

	tipo_auto <fctr>	recuento_ventas <dbl>
1	Auto	192479
2	Camion	2361
3	Furgon	18116
4	SUV	50993

Figura 4 Cantidad de ventas por tipo de auto

- **Total de ventas en cada mes del año 2019 según tipo de auto**

mes <dbl>	tipo_auto <fctr>	total_vendidos <dbl>
1	Auto	1432376.10
1	Camion	66290.72
1	Furgon	343376.11
1	SUV	832300.40
2	Auto	1294254.29
2	Camion	26797.80
2	Furgon	286027.30
2	SUV	681369.18
3	Auto	1380983.38
3	Camion	58102.33
3	Furgon	318904.60
3	SUV	728364.19
4	Auto	1382037.83
4	Camion	59708.25
4	Furgon	327844.58
4	SUV	760661.92
5	Auto	1517737.14
5	Camion	65711.12
5	Furgon	324212.92
5	SUV	800591.08
6	Auto	1355276.78

1-21 of 48 rows Previous 1 2 3 Next

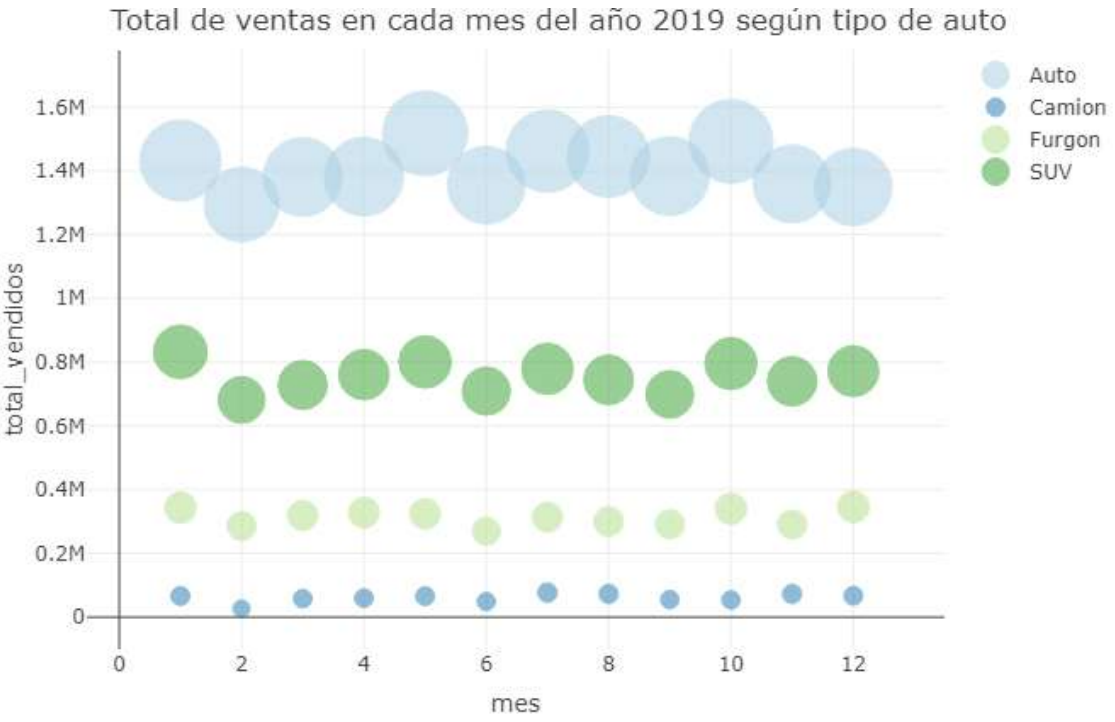


Figura 5 Total de ventas en cada mes del año 2019 según tipo de auto

- **Total ganancia en cada mes del año 2018**

mes <dbl>	total_ganancia <dbl>
1	413814
2	384308
3	423350
4	399600
5	436316
6	406550
7	386765
8	429717
9	406005
10	426124
11	426779
12	417289

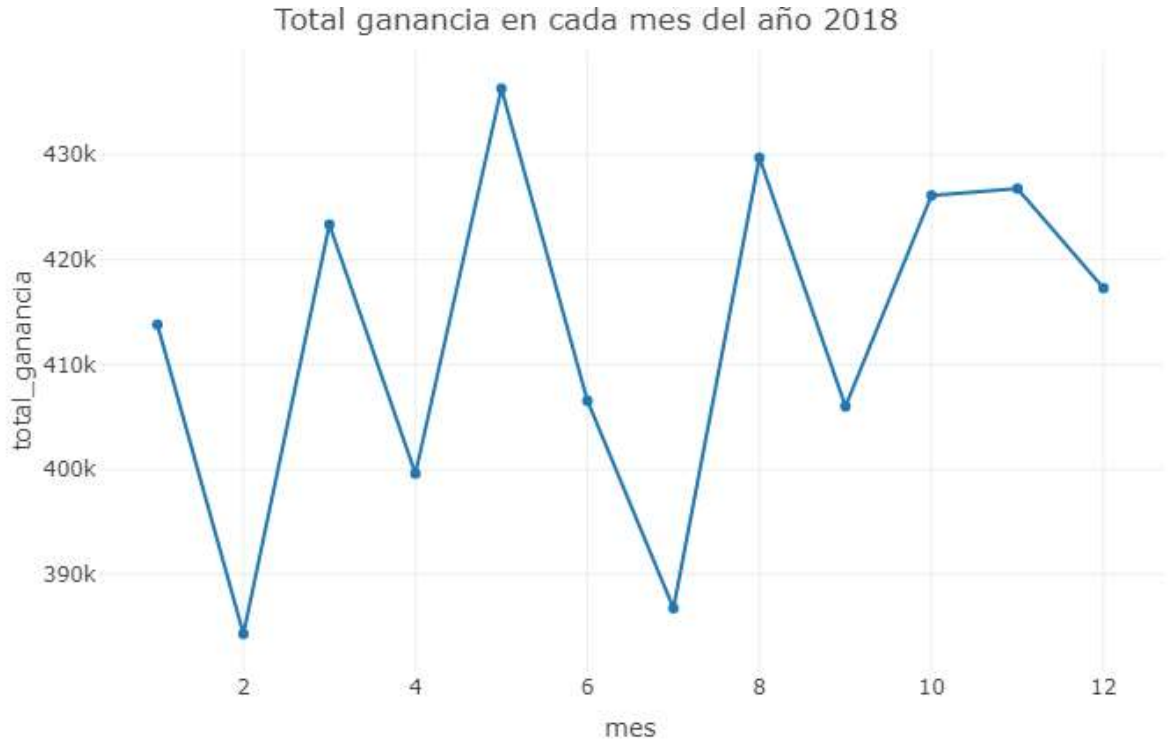


Figura 6 Total ganancia en cada mes del año 2018

Preparación y Limpieza

Para la preparación y limpieza de la data empezaremos reemplazando los valores vacíos de cada columna.

```
apply(is.na(tpll),2,sum)
```

```
      fecha        marca       medida           pr    tipo_auto
          0            0            0            0            0
procedencia   cost_flete cant_ingreso    pre_dolar    pre_soles
       3784            0            0            0            0
  pre_venta   cant_venta
          0            0
```

Figura 7 Columnas con valores NA

En la figura 7 podemos observar que la única columna que contiene valores NA es "procedencia", para una mejor manipulación de datos serán reemplazados.

```
tpll$procedencia[is.na(tpll$procedencia)] <- "OTROS"
sum(is.na(tpll$procedencia))
```

```
sum(is.na(tpll$procedencia))
```

```
[1] 0
```

Figura 8 Comprobación de inexistencia de valores NA

Estos valores NA fueron reemplazados por "OTROS", después se corrobora que el cambio haya sido efectuado.

Una vez realizada la limpieza de los datos, se seleccionarán las columnas necesarias para las fases posteriores, así se obtuvieron: Tipo de auto, precio soles, precio venta, cantidad de venta, total de ventas, ganancia e inversión flete.

Como resultado, se obtuvo una estructura de datos adecuada para su posterior transformación.

4.3. Transformación de los datos

Procederemos a realizar la transformación de los datos generando 3 nuevas variables "total_venta", "ganancia" y "cost_flete", estas fueron creadas a partir de las ya existentes con una estructura de datos apropiada.

```
#Creamos la columna Total Venta
tpll$total_venta = tpll[,11]*tpll[,12] #pre_venta*cant_venta

#Creamos la columna Ganancia
tpll$ganancia = tpll[,13]-(tpll[,12]*tpll[,10]) #total_venta -
(cant_venta * pre_soles

#Creamos cost_flete
tpll$inv_flete =tpll[,7]*tpll[,12]

head(tpll)
```

TopLlantas - RStudio Source Editor

tpll

Filter

procedencia	cost_flete	cant_ingreso	pre_dolar	pre_soles	pre_venta	cant_venta	total_venta	ganancia	inv_flete
CHINA	5	0	86.00	286.3800	326.3800	8	2611.040	320	40
INDIA	5	0	89.56	298.2348	348.2348	8	2785.878	400	40
CHINA	5	0	100.00	333.0000	373.0000	8	2984.000	320	40
CHINA	5	0	99.00	329.6700	399.6700	8	3197.360	560	40
CHINA	5	0	84.60	281.7180	321.7180	8	2573.744	320	40
PERU	5	0	98.50	328.0050	348.0050	8	2784.040	160	40
CHINA	5	0	85.00	283.0500	323.0500	8	2584.400	320	40
INDIA	5	0	89.90	299.3670	339.3670	8	2714.936	320	40
INDIA	5	0	107.00	356.3100	396.3100	8	3170.480	320	40
CHINA	5	0	99.00	329.6700	409.6700	8	3277.360	640	40
CHINA	5	0	86.00	286.3800	326.3800	8	2611.040	320	40
CHINA	5	0	83.30	277.3890	317.3890	8	2539.112	320	40

Figura 9 Generación de nuevas variables

También realizaremos la coerción de los datos, es decir la conversión del tipo de dato.

```
tpll$pre_dolar = as.numeric(tpll$pre_dolar)
tpll$pre_soles = as.numeric(tpll$pre_soles)
tpll$pre_venta = as.numeric(tpll$pre_venta)
tpll$cant_ingreso = as.numeric(tpll$cant_ingreso)
tpll$cost_flete = as.numeric(tpll$cost_flete)
tpll$cant_venta = as.numeric(tpll$cant_venta)
tpll$marca = as.factor(tpll$marca)
tpll$medida = as.factor(tpll$medida)
tpll$tipo_auto = as.factor(tpll$tipo_auto)
tpll$procedencia = as.factor(tpll$procedencia)
tpll$fecha = as.Date(tpll$fecha)
```

Así también se realizaron operaciones de agregación o normalización, consolidando los datos de una forma idónea para la fase siguiente.

4.4. Data Mining

En esta fase seleccionaremos los modelos para solucionar el problema, los criterios de evaluación son los siguientes:

- Los modelos deben ser del tipo de aprendizaje supervisado, ya que se utilizará una etiqueta de salida para cumplir con el objetivo.
- Se utilizarán algoritmos de clasificación del tipo multi-clase, debido a que nuestra etiqueta tiene más de dos salidas.
- Los modelos fueron elegidos también según el tipo de la métrica que vamos a evaluar, en nuestro caso la métrica es "precisión o exactitud".

Aplicaremos 4 métodos de aprendizaje supervisado, principalmente de clasificación. Los modelos seleccionados son los siguientes:

- ✓ Random Forest
- ✓ SVM
- ✓ Naive Bayes
- ✓ Arboles de decisiones

Antes de la aplicación de los modelos, realizaremos una serie de pasos para poderlos desarrollar con facilidad.

1. Selección del Dataset: como primero paso, asignaremos una variable con el nombre "dataset" en el cual almacenaremos los datos necesarios que se aplicaran en los modelos anteriormente mencionados.

```
dataset = tpll[,c(5,10,11,12,13,14,15)]
```

tipo_auto <fctr>	pre_soles <dbl>	pre_venta <dbl>	cant_venta <dbl>	total_venta <dbl>	ganancia <dbl>	inv_flete <dbl>
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0
Auto	0.0000	0.0000	0	0.0000	0	0

1-10 of 225,148 rows

Figura 10 Datos seleccionados

En los modelos de clasificación se requiere una variable objetivo, en nuestro caso será "tipo_auto". Las demás variables se le denominan predictores, los cuales nos ayudaran a separar y encontrar los patrones.

2. Creación de las particiones: una vez definido las variables que utilizaremos procederemos dividir toda la data en 2 grupos; el primero lo denominaremos Entrenamiento que nos permitirá crear y preparar el predictivo. El segundo será establecido con la finalidad de comprobar la fiabilidad de la predicción al cual llamaremos Prueba.

   Utilizaremos la función "createDataPartition()" para crear nuestras particiones, donde el 70% de los datos serán asignados para el grupo entrenamiento y el 30% serán empleados en la prueba.

   Al momento de generar las particiones, es necesario realizar la separación de manera aleatoria, por ello agregaremos una semilla antes a través de la función "set.seed()", la cual nos servirá más adelante para que el experimento sea replicable.

```
#Semilla
set.seed(123456)

#Creando las particiones
datos=createDataPartition(dataset$tipo_auto, p = 0.7, list = FALSE)

#Particion Entrenamiento
entrenamiento = dataset[datos,]

#Particion Prueba
Prueba = dataset[-datos,]
```

TopLlantas - RStudio Source Editor

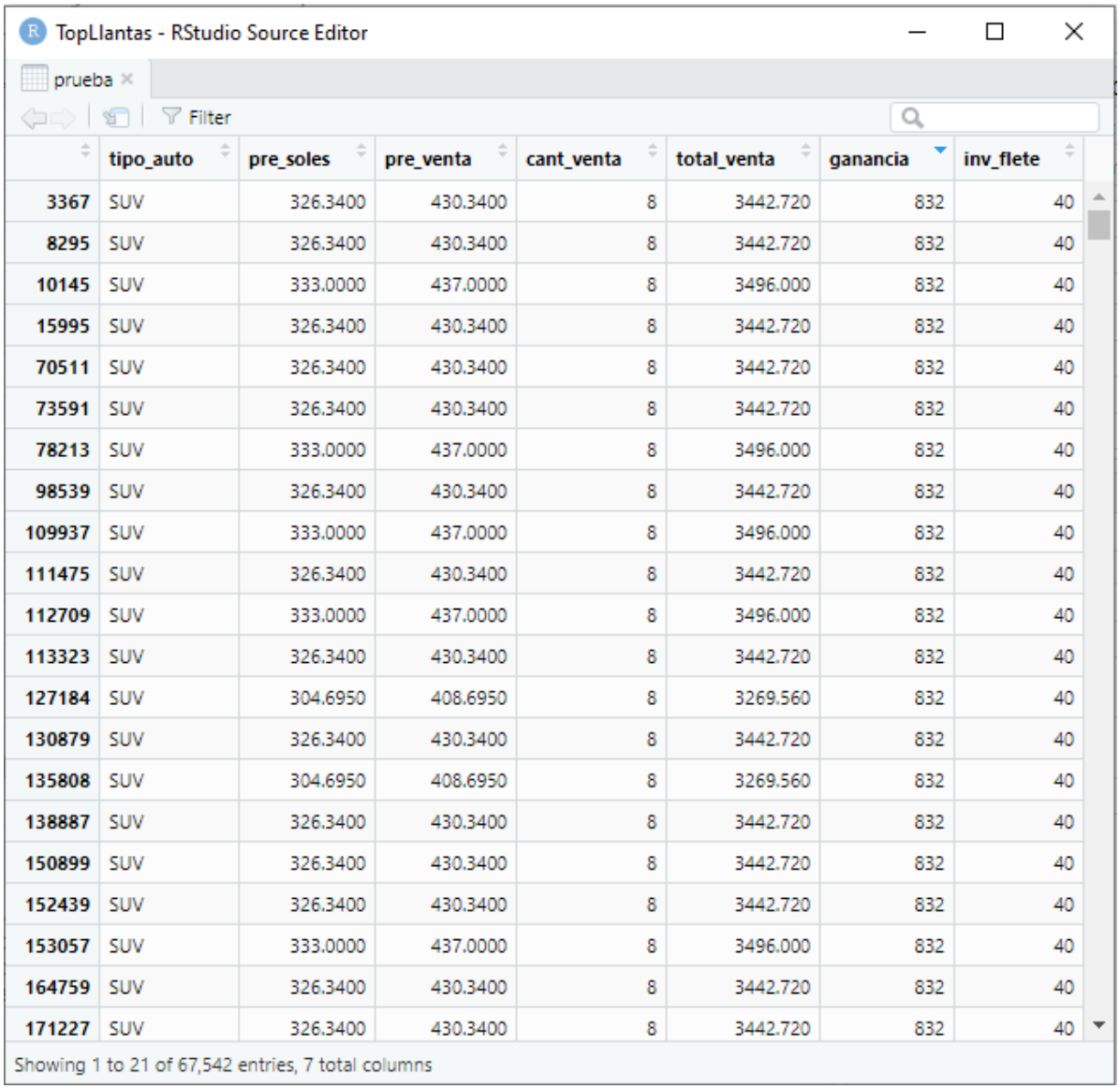
entrenamiento

Filter

	tipo_auto	pre_soles	pre_venta	cant_venta	total_venta	ganancia	inv_flete
1211	SUV	326.340	430.340	8	3442.72	832	40
1521	SUV	333.000	437.000	8	3496.00	832	40
1829	SUV	333.000	437.000	8	3496.00	832	40
2443	SUV	326.340	430.340	8	3442.72	832	40
3059	SUV	326.340	430.340	8	3442.72	832	40
4601	SUV	333.000	437.000	8	3496.00	832	40
6140	SUV	304.695	408.695	8	3269.56	832	40
10452	SUV	304.695	408.695	8	3269.56	832	40
13532	SUV	304.695	408.695	8	3269.56	832	40
13533	SUV	333.000	437.000	8	3496.00	832	40
13839	SUV	326.340	430.340	8	3442.72	832	40
16921	SUV	333.000	437.000	8	3496.00	832	40
18152	SUV	304.695	408.695	8	3269.56	832	40
18153	SUV	333.000	437.000	8	3496.00	832	40
24005	SUV	333.000	437.000	8	3496.00	832	40
25235	SUV	326.340	430.340	8	3442.72	832	40
26777	SUV	333.000	437.000	8	3496.00	832	40
28316	SUV	304.695	408.695	8	3269.56	832	40
31705	SUV	333.000	437.000	8	3496.00	832	40

Showing 1 to 20 of 157,606 entries, 7 total columns

Figura 11 Dataset entrenamiento con 157,606 registros (70% de datos aleatorios)



	tipo_auto	pre_soles	pre_venta	cant_venta	total_venta	ganancia	inv_flete
3367	SUV	326.3400	430.3400	8	3442.720	832	40
8295	SUV	326.3400	430.3400	8	3442.720	832	40
10145	SUV	333.0000	437.0000	8	3496.000	832	40
15995	SUV	326.3400	430.3400	8	3442.720	832	40
70511	SUV	326.3400	430.3400	8	3442.720	832	40
73591	SUV	326.3400	430.3400	8	3442.720	832	40
78213	SUV	333.0000	437.0000	8	3496.000	832	40
98539	SUV	326.3400	430.3400	8	3442.720	832	40
109937	SUV	333.0000	437.0000	8	3496.000	832	40
111475	SUV	326.3400	430.3400	8	3442.720	832	40
112709	SUV	333.0000	437.0000	8	3496.000	832	40
113323	SUV	326.3400	430.3400	8	3442.720	832	40
127184	SUV	304.6950	408.6950	8	3269.560	832	40
130879	SUV	326.3400	430.3400	8	3442.720	832	40
135808	SUV	304.6950	408.6950	8	3269.560	832	40
138887	SUV	326.3400	430.3400	8	3442.720	832	40
150899	SUV	326.3400	430.3400	8	3442.720	832	40
152439	SUV	326.3400	430.3400	8	3442.720	832	40
153057	SUV	333.0000	437.0000	8	3496.000	832	40
164759	SUV	326.3400	430.3400	8	3442.720	832	40
171227	SUV	326.3400	430.3400	8	3442.720	832	40

Showing 1 to 21 of 67,542 entries, 7 total columns

Figura 12 Dataset prueba con 67,542 registros (30% de datos aleatorios)

4.4.1 Validación “Hold Out”

4.4.1.1 Modelo Árbol de decisiones

Creando el Modelo

Comenzaremos con el entrenamiento del modelo usando la función del paquete C50 que nos permitirá crear el árbol de decisiones. La función requiere de una formula donde irá especificado la variable objetivo de la clasificación, en este caso será expresada como “tipo_auto~.”.

La fórmula se encargará de hacer la clasificación de los datos del dataset “Entrenamiento” basándose en el tipo de auto respeto a las demás variables que serán utilizadas como predictores.

```
arbol = c5.0(tipo_auto~., data = entrenamiento)
```

Evaluando el modelo

Una vez realizado el entrenamiento, procederemos al análisis de los resultados

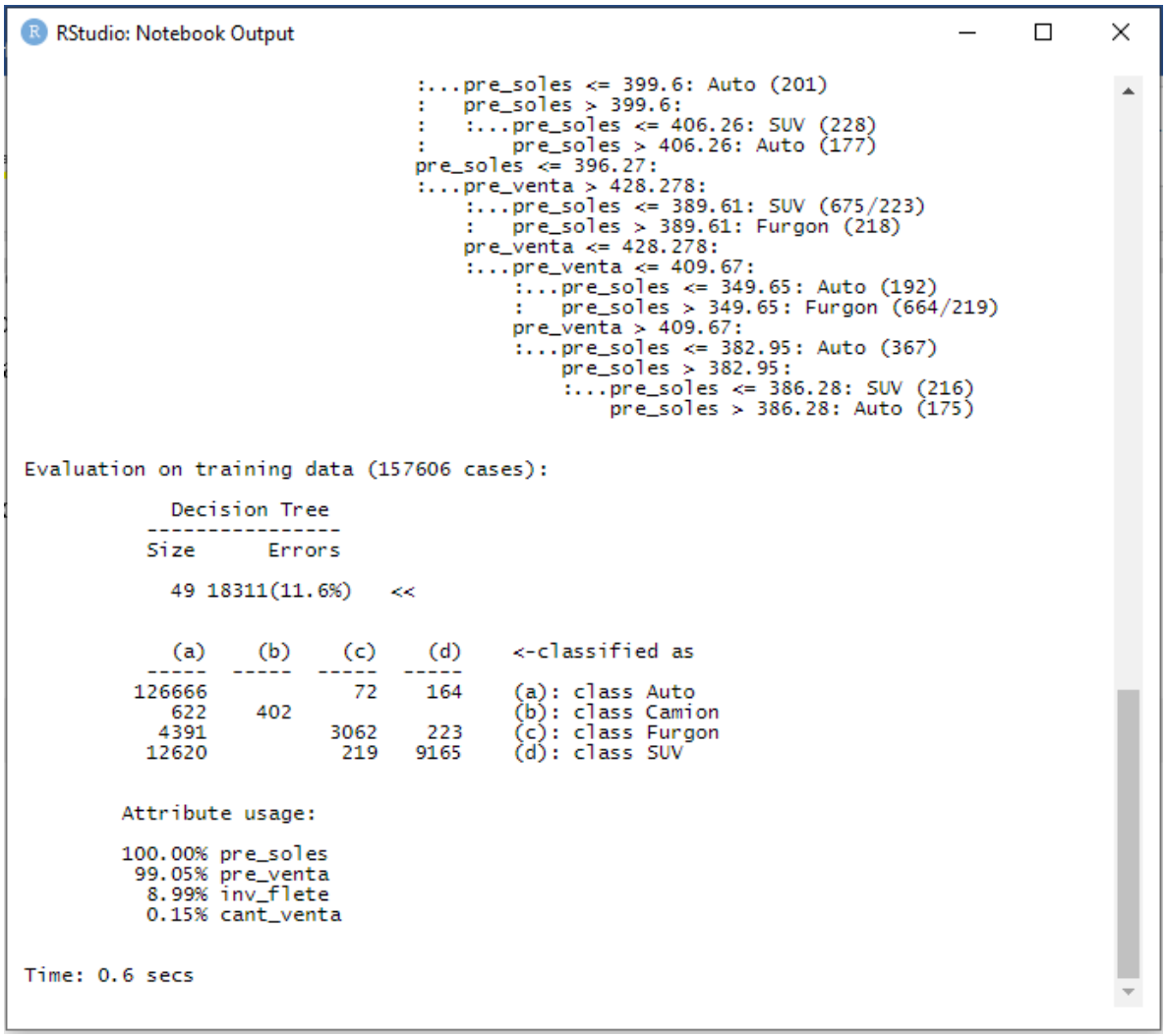


Figura 13 Resultados del modelo Árbol de decisiones

Observamos que se crearon un total de 49 hojas o nodos (size) y hubo 18311 clasificaciones erróneas (errors) de los 157606 casos de entrenamiento, obteniendo un margen de error del 11.6% en la predicción.

Según la importancia de los atributos se observa que se priorizo la variable pre_soles obteniendo el 100% de casos clasificados, seguido de pre_venta con un 99.05%, el valor de inv_flete con 8.99% y cant_venta con un 0.15% de casos que clasificaron.

Comprobación del modelo predictivo

Una vez se ha realizado el entrenamiento del modelo, procederemos a realizar la predicción y a evaluar su eficacia.

Primero aplicaremos la función "predict()" sobre el dataset denominado "prueba" el cual ya fue creado anteriormente y contiene valores similares a la data con la que entrenamos.

```
arbolprediccion = predict(arbol, prueba, type = "class")
```

Luego crearemos nuestra matriz de confusión a través de la función "confusionMatrix()".

```
mc = confusionMatrix(arbolprediccion, prueba$tipo_auto)
```

RStudio: Notebook Output

```
Confusion Matrix and Statistics

          Reference
Prediction  Auto Camion Furgon   SUV
    Auto   54283    228   1885  5466
    Camion     0    210      0     0
    Furgon    22      0   1307   100
    SUV       81      0     97  3863

Overall Statistics

               Accuracy : 0.8833
                 95% CI : (0.8809, 0.8858)
    No Information Rate : 0.8052
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5391

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Auto Class: Camion Class: Furgon Class: SUV
Sensitivity               0.9981      0.479452       0.39739    0.40969
Specificity               0.4239      1.000000       0.99810    0.99694
Pos Pred Value            0.8775      1.000000       0.91463    0.95595
Neg Pred Value            0.9819      0.996614       0.97002    0.91235
Prevalence                0.8052      0.006485       0.04870    0.13960
Detection Rate            0.8037      0.003109       0.01935    0.05719
Detection Prevalence      0.9159      0.003109       0.02116    0.05983
Balanced Accuracy         0.7110      0.739726       0.69774    0.70332
```

Figura 14 Matriz de confusión Árbol de decisiones

En la figura 14 observamos el resultado de la matriz de confusión de tal forma podemos decir que:

- Si las ventas son de neumáticos de tipo auto pues el modelo predijo de manera correcta que 54283 casos serán del tipo auto y se ha equivocado en 228 casos los cuales predijo que serían del tipo Camión, 1885 casos los cuales predijo que serían del tipo Furgón y 5466 casos que serían del tipo SUV.
- Si las ventas son de neumáticos de tipo camión pues el modelo predijo de manera correcta que 210 casos que han sido del tipo camión y en este caso no ha predicho casos erróneos.
- Si las ventas son de neumáticos de tipo Furgón pues el modelo predijo de manera correcta que 1307 casos serán del tipo furgón y se ha equivocado en 22 casos los cuales predijo que serían del tipo auto, y 100 casos que ha dicho que serían del tipo SUV.
- Si las ventas son de neumáticos de tipo SUV pues el modelo predijo de manera correcta que 3863 casos serán del tipo SUV y se ha equivocado en 81 casos que predijo que serían de tipo auto y 97 casos que predijo que serían de tipo furgón.

Se obtuvo una exactitud (accuracy) de 88.33%, con un error del 11.67%. El modelo alcanzo la mayor sensibilidad en el tipo auto con un 0.9981, es decir, sugiere que la mayoría de ventas debe enfocarse en los neumáticos del tipo auto.

4.4.1.2 Naive Bayes

Creando el modelo

Para el entrenamiento del modelo utilizaremos la función del algoritmo "naiveBayes()" que pertenece al paquete e1071, que nos permitirá crear el modelo clasificatorio de probabilidades.

Al igual que en el modelo anterior también requiere de una formula en la cual indicaremos la variable objetivo "tipo_auto ~." y será aplicada sobre el dataset "Entrenamiento".

```
nB = naiveBayes(formula= tipo_auto~., data = entrenamiento)
```

Evaluando el modelo

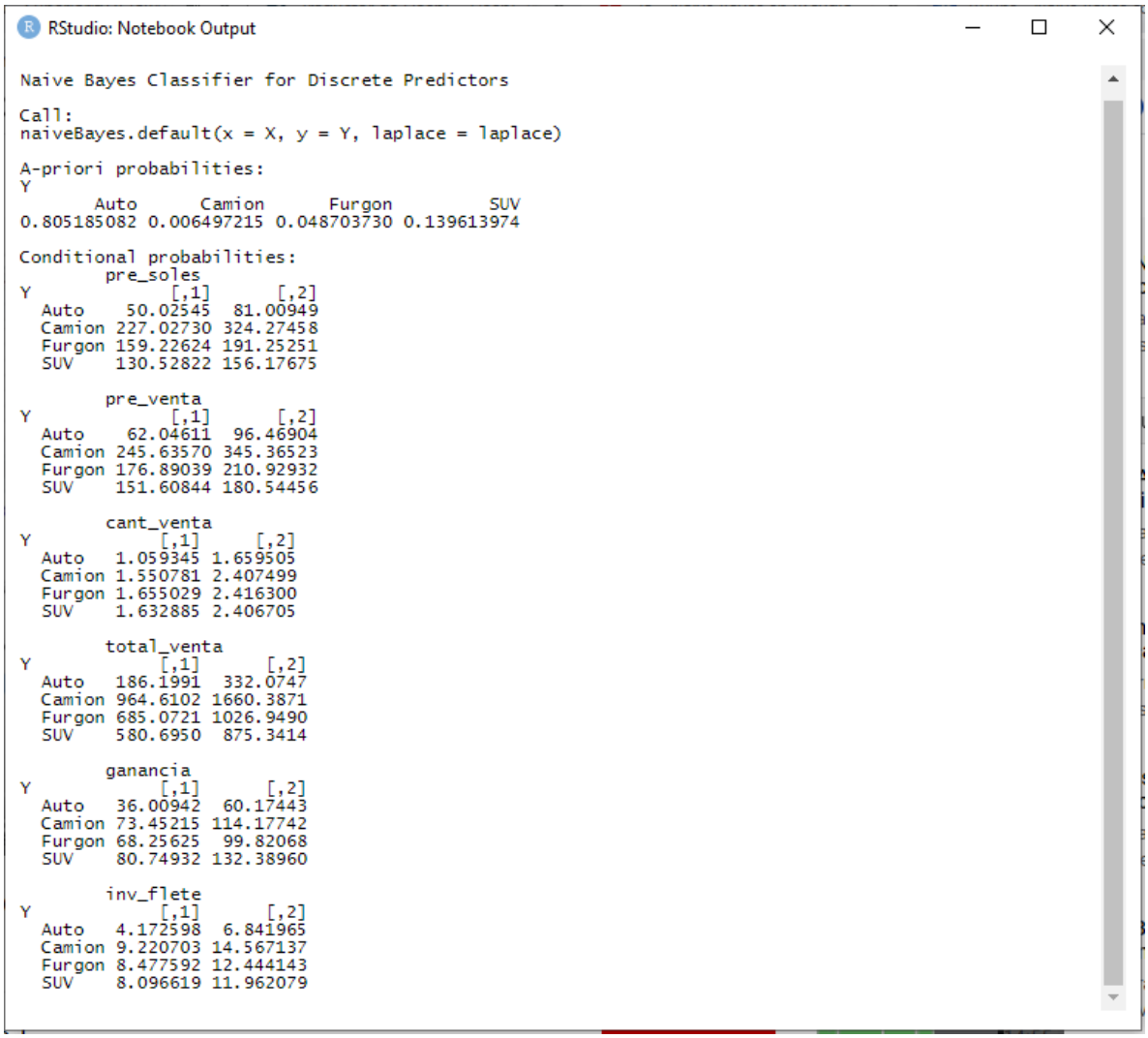


Figura 15 Resultados del modelo Naive Bayes

Analizando los resultados, observamos que al emplear el modelo de clasificación Naive Bayes ha realizado las *probabilidades a priori* de tal forma poder conocer el tipo de auto, de esa forma se obtuvo que el 80% pertenece a auto, el 65% a camión, para furgón el 4.87% y el 13.96% a SUV.

Comprobación del Modelo predictivo

Para poder determinar la eficacia de la predicción, utilizaremos la función "predict()" sobre la data prueba.

```
nBprediccion = predict(nB, newdata = prueba[-1])
```

Una vez realizada la predicción, procederemos a crear nuestra matriz de confusión utilizando la función "confusionMatrix()".

```
confusionMatrix(nBprediccion, prueba$tipo_auto)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  Auto Camion Furgon   SUV
    Auto   48428    228   1885  5666
    Camion     9    108    187   109
    Furgon   300      0    293   285
    SUV     5649    102    924  3369

Overall Statistics

               Accuracy : 0.7728
                 95% CI : (0.7696, 0.776)
    No Information Rate : 0.8052
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2636

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: Auto Class: Camion Class: Furgon Class: SUV
Sensitivity               0.8904      0.246575      0.089085    0.35730
Specificity               0.4087      0.995455      0.990895    0.88514
Pos Pred Value            0.8616      0.261501      0.333713    0.33542
Neg Pred Value            0.4744      0.995084      0.955058    0.89461
Prevalence                0.8052      0.006485      0.048696    0.13960
Detection Rate            0.7170      0.001599      0.004338    0.04988
Detection Prevalence      0.8322      0.006115      0.012999    0.14871
Balanced Accuracy         0.6496      0.621015      0.539990    0.62122
```

Figura 16 Matriz de confusión Naive Bayes

Interpretando la matriz de confusión, podemos decir que:

- Si las ventas realizadas son de neumáticos de tipo auto, el modelo predijo de manera correcta que 48428 casos serán del tipo auto y erró en 228 casos los cuales predijo que serían del tipo Camión, 1885 casos los cuales predijo que serían del tipo Furgón y 5666 casos que serían del tipo SUV.

- Si las ventas son de neumáticos de tipo camión pues el modelo predijo de manera correcta que 108 casos serán del tipo camión y se ha equivocado en 9 casos los cuales predijo que serían del tipo auto, 187 casos los cuales predijo que serían del tipo Furgón y 109 casos que serían del tipo SUV.
- Si las ventas son de neumáticos de tipo Furgón pues el modelo predijo de manera correcta que 293 casos serán del tipo furgón y se ha equivocado en 300 casos los cuales predijo que serían del tipo auto, y 285 casos que ha dicho que serían del tipo SUV.
- Si las ventas son de neumáticos de tipo SUV pues el modelo predijo de manera correcta que 3369 casos serán del tipo SUV y se ha equivocado en 5649 casos que predijo que serían de tipo auto, 102 casos que predijo que serían del tipo camión y 924 casos que predijo que serían de tipo furgón.

Se obtuvo una exactitud (accuracy) de 77.28%, con un error del 22.72%. Nos da un intervalo de confianza (95%) para la eficacia y nos dice que entre un 0.7696 y 0.776 nos va a acertar. También nos dice que el modelo alcanzo la mayor sensibilidad en el tipo auto con un 0.8904, es decir, sugiere que la mayoría de ventas debe enfocarse en los neumáticos del tipo auto.

4.4.1.3 Random Forest

Creando el modelo

Como en los modelos anteriores, empezaremos con el entrenamiento usando la función randomForest(). En esta función incluiremos una formula en la cual comenzaremos especificando a la variable objetivo (tipo_auto~.), también definiremos la data sobre la cual se hará el entrenamiento (entrenamiento) y para finalizar fijaremos que la cantidad de nodos utilizados sea 2 y de árboles 150.

```
RF = randomForest(tipo_auto~., data = entrenamiento, nodesize = 2,
ntree = 150)
```

Evaluando el Modelo

```
Call:
 randomForest(formula = tipo_auto ~ ., data = entrenamiento, nodesize = 2,      ntree = 150)
               Type of random forest: classification
                     Number of trees: 150
No. of variables tried at each split: 2

        OOB estimate of  error rate: 11.62%
Confusion matrix:
         Auto Camion Furgon  SUV class.error
Auto   126746      0      0  156 0.001229295
Camion    622    402      0    0 0.607421875
Furgon   4391      0   3051  234 0.602527358
SUV     12661      0    246 9097 0.586575168
```

Figura 17 Resultados del modelo Random Forest

Viendo el resumen del modelo nos dice que ha detectado que es de tipo clasificación, también vemos que el número de árboles que se utilizaron fueron 150 los cuales asignamos en la fórmula de entrenamiento, así también el número de variables que se empleó fueron 2 y nos damos cuenta que se la tasa de error fue del 11.62%.

La matriz de confusión nos dice que obtuvo una clasificación certera de 126,746 de 144,420 del tipo auto, los 402 del tipo camión fueron hallados correctamente, del tipo furgón 3,051 de 3,297 y de los 9,487 del tipo SUV clasifico 156.

Comprobación del modelo predictivo

Habiendo realizado el entrenamiento, con la función predict() evaluaremos eficiencia de la predicción. Esta será aplicada sobre la data “prueba”.

```
RFprediccion = predict(RF, prueba, type = "class")
```

Una vez aplicada la predicción se creará la matriz de confusión.

```
confusionMatrix(RFprediccion, prueba$tipo_auto)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  Auto Camion Furgon   SUV
    Auto   54318    228   1885  5479
    Camion     0    210      0     0
    Furgon     0      0   1341   123
    SUV       68      0     63  3827

Overall Statistics

               Accuracy : 0.8838
                 95% CI : (0.8814, 0.8862)
    No Information Rate : 0.8052
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5402

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Auto Class: Camion Class: Furgon Class: SUV
Sensitivity               0.9987      0.479452       0.40772    0.40588
Specificity               0.4229      1.000000       0.99809    0.99775
Pos Pred Value            0.8774      1.000000       0.91598    0.96690
Neg Pred Value            0.9879      0.996614       0.97052    0.91190
Prevalence                0.8052      0.006485       0.04870    0.13960
Detection Rate            0.8042      0.003109       0.01985    0.05666
Detection Prevalence      0.9166      0.003109       0.02168    0.05860
Balanced Accuracy         0.7108      0.739726       0.70290    0.70181
```

Figura 18 Matriz de confusión Random Forest

La matriz de confusión nos dice que:

- Del tipo ***auto*** se ha pronosticado de manera correcta 54,318 casos, así también erró prediciendo 228 clasificándolos como tipo camión, 1885 como tipo furgón y en SUV erro 5479.
- Del tipo ***camión*** ha pronosticado de manera correcta que 210 casos.
- Del tipo ***furgón*** se ha pronosticado de manera correcta 1,341 casos, así también erró prediciendo 123 casos clasificándolos como tipo SUV.
- Del tipo ***SUV*** ha pronosticado de manera correcta que 3,827 casos, errando con 68 casos clasificándolos como tipo auto y 63 como furgón.

Se obtuvo una exactitud (accuracy) de 88.38%, con un margen de error del 11.62%. El modelo alcanzo la mayor sensibilidad en el tipo auto con un 0.9987, es decir, sugiere que la mayoría de ventas debe enfocarse en los neumáticos del tipo auto.

4.4.1.4 SVM

Creando el modelo

Para la creación del modelo instalaremos el paquete e1071, del cual usaremos la función svm(). Al igual que los demás modelos también se requiere de una fórmula, que estará compuesta por la variable objetivo tipo_auto~. y la data a la cual será aplicada, que será entrenamiento.

```
svm_cla = svm(formula = tipo_auto~., data = entrenamiento)
```

Evaluando el modelo

```
Call:
svm(formula = tipo_auto ~ ., data = entrenamiento)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  46510
```

Figura 19 Resultados del modelo SVM

Vemos que se ha detectado que es un modelo de tipo clasificación, que el tipo de kernel que se utilizo es de tipo radial y la cantidad de números de vectores de soporte fueron 46,510.

Comprobación del modelo predictivo

```
Confusion Matrix and Statistics

          Reference
Prediction  Auto Camion Furgon   SUV
    Auto   53488    228   1899  5738
    Camion     0    108      0     0
    Furgon     0      0    276   148
    SUV      898    102   1114  3543

Overall Statistics

               Accuracy : 0.8501
                 95% CI : (0.8473, 0.8527)
    No Information Rate : 0.8052
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4156

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Auto Class: Camion Class: Furgon Class: SUV
Sensitivity               0.9835      0.246575      0.083916    0.37576
Specificity               0.4022      1.000000      0.997697    0.96362
Pos Pred Value            0.8718      1.000000      0.650943    0.62630
Neg Pred Value            0.8549      0.995106      0.955109    0.90489
Prevalence                0.8052      0.006485      0.048696    0.13960
Detection Rate            0.7919      0.001599      0.004086    0.05246
Detection Prevalence      0.9084      0.001599      0.006278    0.08376
Balanced Accuracy         0.6928      0.623288      0.540806    0.66969
```

Figura 20 Matriz de confusión de SVM

La matriz de confusión nos dice que:

- Del tipo ***auto*** se ha pronosticado de manera correcta 53,488 casos, así también erró prediciendo 228 clasificándolos como tipo camión, 1,899 como furgón y en SUV erro clasificando 5,738.
- Del tipo ***camión*** ha pronosticado de manera correcta que 108 casos.
- Del tipo ***furgón*** se ha pronosticado de manera correcta 276 casos, así también erró prediciendo 148 casos clasificándolos como tipo auto SUV.
- Del tipo ***SUV*** ha pronosticado de manera correcta que 3,543 casos, errando con 898 casos clasificándolos como tipo auto, 102 como tipo camión y 1114 como furgón.

Se obtuvo una exactitud (accuracy) de 85.01%, con un margen de error del 14.99%. El modelo alcanzo la mayor sensibilidad en el tipo auto con un 0.9835, es decir, sugiere que la mayoría de ventas debe enfocarse en los neumáticos del tipo auto.

4.4.2 Validación K-fold

En esta fase se realizará la comparación de los modelos a través de la técnica de validación cruzada (cross validation), debido a que es la más adecuada para evaluar modelos de Machine Learning en R-Studio.

Como primer paso comenzaremos definiendo los valores de la prueba, el primero lo nombraremos “control” en el cual utilizaremos la función traincontrol() que tendrá como atributos el tipo método y al número de k-fold. La segunda variable será “metric” que hará referencia al tipo de métrica la cual evaluaremos, en este caso será “Accuracy” para determinar la precisión de los modelos.

```
control = trainControl(method="cv", number=10)
metric = "Accuracy"
```

Una vez definido los valores de la prueba, comenzaremos con el entrenamiento de los modelos.

Para entrenar los modelos se utilizará la función “train()” del paquete Caret, se definirá la variable objetivo (tipo_auto~.) para luego nombrar la data que se utilizará. También tendremos que definir el atributo “method” para nombrar el método a utilizar en cada modelo, por último, llamaremos a las variables definidas anteriormente (“metric” y “control”).

SVM

```
set.seed(123456)
fit.svm = train(tipo_auto~., data=entrenamiento, method="svmRadial",
metric=metric, trControl=control)
```

Arboles de decisiones

```
set.seed(123456)
fit.cart = train(tipo_auto~., data=entrenamiento, method="rpart",
metric=metric, trControl=control)
```

Random Forest

En el caso de random forest, tendremos que definir los nodos y arboles a utilizar.

```
set.seed(123456)
fit.rf = train(tipo_auto~., data=entrenamiento, method="rf",
metric=metric, trControl=control,nodesize = 2, ntree = 150)
```

Naive Bayes

```
set.seed(123456)
fit.nb = train(tipo_auto~., data=entrenamiento, method="nb",
metric=metric, trControl=control)
```

Una vez realizado el entrenamiento de los modelos, procederemos a ejecutar la función resamples() para extraer las mejores métricas que se obtuvo utilizando la función train() y ser almacenadas en un dataframe. Se seleccionará la mejor métrica obtenida de cada modelo.

```
resultados = resamples(list(svm=fit.svm, ad=fit.cart, rf=fit.rf,
nb=fit.nb))
summary(resultados)

datafr <- data.frame(resultados)
Presicion < c(datafr[7,1], datafr[5,2], datafr[1,3], datafr[8,4])
Modelos <- c("SVM", "AD", "RF", "NB")
dataCV <- data.frame(modelos, Presicion)
```

A continuación, se mostrará una gráfica de los resultados del cross validation.

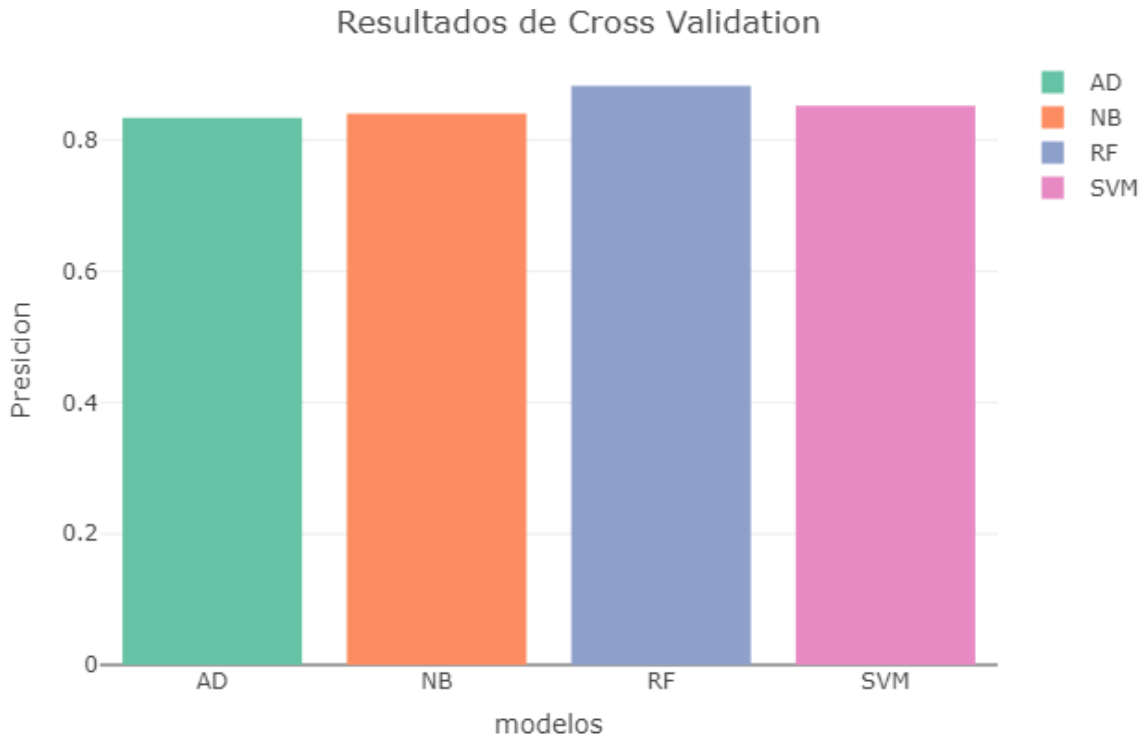


Figura 21 Resultados del Cross Validation

4.5. Interpretación y Evaluación

Los resultados obtenidos por el cross validation son los siguientes

Tabla 3 Resultados del Cross Validation

Medida	RF	SVM	NB	AD
Precisión	88.28	85.21	84.02	83.42
Error	11.72	14.79	15.98	16.58

Observamos que en primer lugar encontramos al modelo Random Forest con una precisión del 88.28%, seguido de SVM con un 85.21 %, a Naive Bayes con 84.02% y en el último lugar a Árbol de Decisiones con un 83.42%.

Los resultados que se obtuvieron en la primera parte son los siguientes:

Tabla 4 Resultados de los modelos aplicados de manera individual

MEDIDA	**RF**	**SVM**	**AD**	**NB**
Precisión	88.38	85	83.33	77.28
Error	11.62	15	11.67	22.72

Comparando las tablas, observamos que los valores de exactitud de la primera parte tienen el siguiente orden:

Random Forest, SVM, Arboles de decisiones y Naive Bayes.

A diferencia de la implementación del Cross Validation, se obtiene lo siguiente:

Random Forest, SVM, Naive Bayes y Arboles de decisiones.

De tal forma en ambas pruebas se obtiene como modelo con más precisión a Random Forest, el cual al aplicar el cross validation se obtuvo una precisión de 88.28% y aplicándolo independiente se obtuvo un 88.38% de precisión.

En el Cross Validation, se manifiesta una mejora del modelo Naive Bayes respecto a la primera parte, ganándole al modelo Árbol de Decisiones.

Habiendo mencionado los resultados obtenidos en las diferentes pruebas, podemos confirmar que el modelo Random Forest es el que tiene mejor precisión para predecir, por ende, aplicaremos ese modelo en la data "prueba" a través de la función predict().

```
predicciones = predict(fit.rf, prueba)
```

Para observar los resultados crearemos una matriz de confusión utilizando la función confusionMatrix().

```
matrizCV<-confusionMatrix(predicciones, prueba$tipo auto)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  Auto Camion Furgon   SUV
    Auto   54305    228   1885  5466
    Camion     0    210      0     0
    Furgon     0      0   1323   104
    SUV       81      0     81  3859

Overall Statistics

               Accuracy : 0.8839
                 95% CI : (0.8814, 0.8863)
    No Information Rate : 0.8052
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5407

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Auto Class: Camion Class: Furgon Class: SUV
Sensitivity               0.9985      0.479452       0.40225    0.40927
Specificity               0.4239      1.000000       0.99838    0.99721
Pos Pred Value            0.8775      1.000000       0.92712    0.95971
Neg Pred Value            0.9857      0.996614       0.97026    0.91231
Prevalence                0.8052      0.006485       0.04870    0.13960
Detection Rate            0.8040      0.003109       0.01959    0.05713
Detection Prevalence      0.9162      0.003109       0.02113    0.05953
Balanced Accuracy         0.7112      0.739726       0.70032    0.70324
```

Figura 22 Resultados de la predicción

5. Discusión de la Hipótesis

5.1. Formulación del Problema

¿En qué medida un modelo de análisis predictivo sobre la información del abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo influye en la toma de decisiones sobre la gestión de suministros?

5.2. Hipótesis

El desarrollo de un modelo de análisis predictivo sobre la información de abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo, permitirá tener una mejor toma de decisiones sobre la gestión de los suministros.

Variables

VI: Modelo de análisis predictivo en la empresa Top Llantas.

VD: Mejor toma de decisiones en la gestión de abastecimiento de llantas.

Tabla 5 Definición de variables

Variable	**Dimensión**	**Indicador**	**Unidad de medida**	**Instrumento de Investigación**
VI	Precisión	% de precisión de cada modelo	% precisión	Hoja de datos
VD	Satisfacción	Grado de satisfacción	Rango de satisfacción	Tabla de satisfacción

5.3. Población y muestra

5.3.1 Población

Empresas dedicadas a la venta de autopartes o accesorios automovilísticos en el departamento La Libertad.

5.3.2 Muestra

Registro de inventario y ventas del 2018 al 2019 de la empresa Top Llantas que se dedica a la compra venta de Llantas al público en general.

5.3.3 Unidad de Análisis

Abastecimiento de los productos.

5.4. Diseño pre experimental pre-test y post-test

- Diseño pre experimental pre-test (T_1): corresponde a la evaluación previa de la variable independiente al grupo experimental.
- Diseño pre experimental pre-test (T_2): corresponde a la evaluación posterior de la variable independiente al grupo experimental.

El diseño de la investigación presenta un solo grupo, al cual se le aplicó un cuestionario para la evaluación del antes y después de la variable independiente. Este se muestra de la siguiente forma:

$$G \quad T_1 \quad X \quad T_2$$

Donde:

X: Tratamiento de la variable independiente

T_1: Pre test

T_2: Post test

G: Grupo experimental

5.4.1 Cálculo de indicadores de la hipótesis

Tabla 6 Resumen del resultado de los modelos aplicados individualmente

MEDIDA	**RF**	**SVM**	**AD**	**NB**
Precisión	88.38	85	83.33	77.28
Error	11.62	15	11.67	22.72
Sensibilidad				
Automóvil	0.9987	0.9835	0.9981	0.8904
SUV	0.4059	0.3758	0.4165	0.3573
Furgón	0.4077	0.0839	0.3673	0.0890
Camión	0.4795	0.2466	0.4269	0.2466

Basándonos en los resultados de los análisis de los 4 modelos aplicados, podemos decir que el modelo con más precisión y que mejor se adecua a nuestra data es Random Forest.

5.4.2 Análisis Estadístico

- **Paso 1: Planteamiento de la hipótesis**

  $H_0 \rightarrow T_1 \geq T_2$

  $H_1 \rightarrow T_2 \geq T_1$

  Dónde:

  H_0 es la hipótesis Nula: "El desarrollo de un modelo de análisis predictivo sobre la información de abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo, no permitirá tener una mejor toma de decisiones sobre la gestión de los suministros"

  H_1 es la hipótesis Alternativa: "El desarrollo de un modelo de análisis predictivo sobre la información de abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo, permitirá tener una mejor toma de decisiones sobre la gestión de los suministros"

- **Paso 2: Nivel de significancia**

  Si p-valor es menor o igual al 5%, H_1 será aceptado y H_0 se rechazará. ($\alpha = 0,05$)

- **Paso 3: Prueba estadística**

  Por motivo a que la muestra es menor a 30, la técnica estadística a utilizar será t-student. En el cual encontramos dos momentos un antes y un después, donde el primer periodo sirve como un testigo para conocer los cambios que existen luego de aplicar la variable experimental.

- **Paso 4: Calculo de T**

$$\bar{D} = \frac{\Sigma D}{n}, \delta = \sqrt{\frac{\Sigma(Di - \bar{D})^2}{n-1}}, t_c = \frac{\bar{D}}{\frac{\delta}{\sqrt{n}}}$$

Donde:

T_c : T estadístico

δ : Desviación estándar

n : Tamaño de la muestra

$\bar{D}$: Promedio o media aritmética de las diferencias entre los momentos pre y post

Para hallar T se ha realizado un cuestionario de tal forma poder saber el grado de satisfacción de las personas involucradas con el modelo predictivo para la gestión de abastecimiento luego de su manipulación con este.

Tabla 7 Tabla de grado de satisfacción

Rango	Grado de Satisfacción
0 - 2	Muy insatisfecho
2.1 - 4	Insatisfecho
4.1 - 6	Neutral
6.1 - 8	Satisfecho
8.1 - 10	Muy Satisfecho

Tabla 8 Indicadores para el grado de satisfacción

Indicadores	Media Pre	Media Post	D
Se puede saber en qué neumáticos conviene hacer una mejor inversión	2.7	8.7	6
Se puede conocer los productos del tipo de auto con mayor demanda en mercado	3.3	8.7	5.3
Se puede conocer el tipo de auto que genere mayor ganancia	3.3	9.3	6
Se puede conocer las posibles pérdidas de una inversión según su el tipo de auto	3.3	9.3	6

Tabla 9 Resultados del T-student

$\overline{D}$ = 5.8	n = 4	t = 35	gl = 3	α = 0.05	j = 2.353	P-valor= 0.000026	δ = 0.33

Donde:

gl: grado de libertad

α: nivel de significancia

j: valor crítico

$$t_c = \frac{\overline{D}}{\frac{\delta}{\sqrt{n}}} = \frac{5.8}{\frac{0.33}{\sqrt{4}}} = \mathbf{35}$$

Interpretación:

De acuerdo a lo obtenido, P (0.00026) es menor al nivel de significancia (0.05), de tal forma que la hipótesis alternativa es aceptada y la hipótesis nula se rechaza, con esto se entiende que "El desarrollo de un modelo de análisis predictivo sobre la información de abastecimiento y ventas de neumáticos en la empresa Top Llantas de Trujillo, permite tener una mejor toma de decisiones sobre la gestión de los suministros"

6. CONCLUSIONES

- Se realizó la recolección de datos, los cuales eran almacenados en un Excel que contenía el inventario de los diferentes productos vendidos y las ventas realizadas diariamente, en dicho Excel se hizo una depuración, tratamiento y creación de variables; quedando un total de 14 columnas con 225148 registros de ventas.
- En la limpieza y procesamiento realizamos la depuración de valores innecesarios o perdidos (espacios en blancos, valores extraños, etc.), asignándoles también un tipo de dato (numeric, factor, date, etc.). Se decidió dividir la data en 2 grupos, uno con el 70% que se utilizará para el entrenamiento del modelo, y el segundo con el 30% para la prueba
- Se aplicó 4 modelos de aprendizaje supervisado como son: Árbol de decisiones, Naive Bayes, Random Forest y SVM empleando el lenguaje R. En la validación hold out se hizo una evaluación por cada uno de los modelos, obteniendo a Random Forest con una precisión del 88.38%, seguido de SVM con 85%, Árbol de decisiones con 83.33 % de precisión y por ultimo a Naive Bayes 77.28%. Para K-Fold, se realizó la validación cruzada (cross validation) donde encontramos a Random Forest con un 88.28% de precisión, a SVM con una precisión del 85.21%, con un 84.02 % tenemos a Naive Bayes y, por último, con un 83.42% de precisión a Arboles de decisiones. De esta manera, se concluyó que Random Forest es el modelo que mejor se acomoda para la predicción.
- Después de desarrollar y evaluar los cuatro modelos, logramos obtener que el modelo Random Forest nos da una precisión de 88.28% de datos clasificados correctamente, siendo el modelo que aplicaremos a nuestra data, para analizar la gestión de la venta de neumáticos.

7. RECOMENDACIONES

- ✓ Se recomienda definir la misma semilla con la misma longitud para todos los modelos, ya que de no hacerlo el sistema establecerá una de manera aleatoria para cada modelo generando variaciones entre los datos.
- ✓ Se recomiendo utilizar el lenguaje R ya que es idóneo para realizar estadística, permite una fácil maniobrabilidad de los datos y de manera precisa. También se destaca que puede manipular grandes de datos y ejecución de muchas plataformas, sobre todo que es gratuito. Además, que presenta una gran variedad de paquetes para realizar Machine Learning ya que tiene implementado una gran cantidad de algoritmos.
- ✓ Se recomienda utilizar el paquete dplyr ya que facilita la manipulación de datos de manera rápida; y en la presentación de gráficos ggplot y plotly que presentan una visualización con capacidades más avanzadas.

8. REFERENCIAS BIBLIOGRÁFICAS

Angeles Gonzales, E. I. (2017). *Analítica de negocios en la gestión de ventas de la empresa Inversiones Generales Fabrizio.* Tesis de titulo profesional, Universidad Nobert Wiener, Facultad de Ingeniería, Lima.

Apolaya Torres, C. H., & Espinosa Diaz, A. (2018). *Técnicas de inferencias, predicción y minería de datos.* Tesis para titulo profesional, Universidad Peruana de Ciencias Aplicadas (UPC), Facultad de Ingeniería, Lima. doi:10.19083/tesis/624497

Bahit, E. (2018). *Introducción al lenguaje Python.*

Berry, M., Yap, B. W., & Mohamed, A. (2020). *Supervised and Unsupervised Learning for Data Science.* Switzerland: Springer.

Bhatia, P. (2019). *Data Minig and Data Warehousing: Principles and Practical Techniques.* Cambridge: Cambridge University Press.

Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.* tesis para el grado de titulacion, Universitat Oberta de Catalunya, Catalunya. Obtenido de http://hdl.handle.net/10609/59565

Grández Márquez, M. A. (2017). *Aplicación de Minería de datos para determinar patrones de consumo Futuro en Clientes de una Distribuidora de Suplementos Nutricionales.* Tesis para título profesional, Universidad San Ignacio de Loyola (USIL), Facultad de Ingeniería, Lima. Obtenido de http://repositorio.usil.edu.pe/handle/USIL/2763

Grolemund, G., & Wickham, H. (2017). *R for Data Science.* O'Reilly.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning.*

Jiménez Chura, A. C. (2017). *Análisis Predictivo para los Procesos de Admisión de la Universidad Nacional del Altiplano - Puno.* Tesis de doctorado, Universidad Nacional

del Altiplano, Ciencias de la Computacion, Puno. Obtenido de http://repositorio.unap.edu.pe/handle/UNAP/6212

Jiménez, J. U. (2019). *Introducción a R y Rstudio.*

Morales Tabares, Z. E. (2016). *Modelo Multivariado de Predicción del Stock de Piezas de Repuesto para Equipos Médicos.* Tesis de doctorado, Universidad de las Ciencias Informatícas (UCI), Ingeniería y Gestión de Software, La Habana.

Sumiran, K. (2018). *An Overview of Data Mining Techniques and Their Application in Industrial Engineering.*

Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps.* Bangalore: apress.

Vaibhav, K., & M. L., G. (2018). *Predictive Analytics: A Review of Trends and Techniques.* Dehradun.

Valdez Alvarado, A. (2018). *INTRODUCCIÓN AL MACHINELEARNING.*

9. ANEXOS

Anexo 1 – Cuestionario

Preguntas	Muy Insatisfecho	Insatisfecho	Neutral	Satisfecho	Muy Satisfecho
Se puede saber en qué neumáticos conviene hacer una mejor inversión					
Se puede conocer los productos del tipo de auto con mayor demanda en mercado					
Se puede conocer el tipo de auto que genere mayor ganancia					
Se puede conocer las posibles pérdidas de una inversión según su el tipo de auto					