



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

Título:

Optimización de rutas en el transporte de contenedores dentro del centro de acopio a través de un modelo de clasificación basado en Machine Learning.

Autores:

Bohórquez Quisnia, Connie Scarlet
Chamaidan Morocho, Carlos David

**Trabajo de Integración curricular previo a la obtención de título de
LICENCIADO NEGOCIOS INTERNACIONALES**

Tutor:

Ing. Carrera Buri, Félix Miguel, Mgs.

Guayaquil, Ecuador
23 de agosto del 2024



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación fue realizado en su totalidad por **Bohórquez Quisnia, Connie Scarlet, y Chamaidan Morocho, Carlos David**, como requerimiento para la obtención del título de Licenciada en Negocios Internacionales.

TUTOR

f. _____
Ing. Carrera Buri, Félix Miguel, Mgs.

DIRECTORA DE LA CARRERA

f. Gabriela Hurtado Cevallos

Ing. Hurtado Cevallos, Gabriela Elizabeth Mgs.

Guayaquil, a los 23 del mes de agosto del año 2024



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Nosotros, **Bohórquez Quisnia, Connie Scarlet**
Chamaidan Morocho, Carlos David

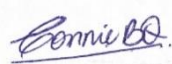
DECLARAMOS QUE:

El Trabajo de Titulación, **Optimización de rutas en el transporte de contenedores dentro del centro de acopio a través de un modelo de clasificación basado en Machine Learning** previo a la obtención del título de **Licenciados en Negocios Internacionales**, 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 23 del mes de agosto del año 2024

LOS AUTORES:

f. 

Bohórquez Quisnia, Connie Scarlet

f. 

Chamaidan Morocho, Carlos David



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES


AUTORIZACIÓN

Nosotros, **Bohórquez Quisnia, Connie Scarlet**
Chamaidan Morocho, Carlos David


Autorizamos a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Optimización de rutas en el transporte de contenedores dentro del centro de acopio a través de un modelo de clasificación basado en Machine Learning**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 23 del mes de agosto del año 2024

LOS AUTORES:

f. 

Bohórquez Quisnia, Connie Scarlet

f. 

Chamaidan Morocho, Carlos David



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

REPORTE COMPILATIO

CERTIFICADO DE ANÁLISIS
magister

Bohórquez Quisnia & Chamaidan Morochu

0% Textos sospechosos

16% Similitudes (ignorado)
< 1% similitudes entre comillas
8% entre las fuentes mencionadas
3% Idiomas no reconocidos (ignorado)

Nombre del documento: Bohórquez Quisnia & Chamaidan Morochu.docx
ID del documento: e2e0b6914e720f85bd189ac1bbcf73b49980c996
Tamaño del documento original: 13,95 MB
Autores: []

Depositante: Felix Miguel Carrera Buri
Fecha de depósito: 26/8/2024
Tipo de carga: interface
fecha de fin de análisis: 26/8/2024

Número de palabras: 30.928
Número de caracteres: 199.726

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

N°	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	aws.amazon.com ¿Qué es el machine learning? - Explicación sobre el machine le... 11 fuentes similares	3%		Palabras idénticas: 3% (882 palabras)
2	www.coursera.org 7 algoritmos de machine learning que hay que conocer: Guía ... 6 fuentes similares	2%		Palabras idénticas: 2% (720 palabras)
3	www.repsol.com ¿Qué es la inteligencia artificial y cómo nos ayuda? https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/inteligencia-artificial/index.cshml 8 fuentes similares	2%		Palabras idénticas: 2% (621 palabras)
4	www.descartes.com Optimización de rutas: conoce las ventajas y la importanci... https://www.descartes.com/es/resources/blog/optimizacion-de-rutas-conoce-la-estrategia-y-sus-ve... 2 fuentes similares	2%		Palabras idénticas: 2% (466 palabras)
5	www.redalyc.org Aplicación de algoritmos Random Forest y XGBoost en una bas... https://www.redalyc.org/journal/404/40471792003/html/ 5 fuentes similares	1%		Palabras idénticas: 1% (372 palabras)

Fuentes con similitudes fortuitas

N°	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	repositorio.ucsg.edu.ec http://repositorio.ucsg.edu.ec/bitstream/3317/17386/3/T-UCSG-PRE-ECO-ADM-606.pdf.txt	< 1%		Palabras idénticas: < 1% (33 palabras)
2	cybertesis.unmsm.edu.pe https://cybertesis.unmsm.edu.pe/bitstream/20.500.12672/2/1598/1/Chavez_ts.pdf	< 1%		Palabras idénticas: < 1% (30 palabras)
3	topics.libra.titech.ac.jp Transportation, storage, and disposal of radioactive mate... https://topics.libra.titech.ac.jp/en/recordID/catalog:libraA75458241	< 1%		Palabras idénticas: < 1% (31 palabras)
4	proceedings.mlr.press Proceedings of Machine Learning Research Proceedings... https://proceedings.mlr.press/v5/	< 1%		Palabras idénticas: < 1% (22 palabras)
5	Documento de otro usuario #651742 El documento proviene de otro grupo	< 1%		Palabras idénticas: < 1% (29 palabras)

Fuentes mencionadas (sin similitudes detectadas) Estas fuentes han sido citadas en el documento sin encontrar similitudes.

- <https://www.apd.es/algoritmos-del-machine-learning/>
- <https://doi.org/10.1016/j.trb.2011.02.004>
- <https://link.springer.com/book/9780387310732>
- https://doi.org/10.31570/Prosp_2020_01_3
- <https://doi.org/10.2991/mdsmes-19.2019.59>

Ing. Carrera Buri, Félix Miguel, Mgs.

AGRADECIMIENTO

No existen las palabras suficientes para agradecer a quienes hicieron posible este momento de mi vida, principal y fundamental Dios por darme la fuerza y fortaleza durante este inmenso recorrido. A mis padres, cuyo amor y apoyo incondicional que desde siempre se hizo presente, formó parte esencial para superar los miles de obstáculos que surgieron tanto en mi vida cotidiana como en mi vida académica. Atribuyo a sus constantes ánimos como la base de mi éxito, hoy y siempre.

A toda mi familia, mamita Amandina, mis tías, primos, quienes me apoyaron dentro de sus posibilidades, para que yo lograra llegar hasta este punto, y cumplir aquello por lo que tanto había luchado.

Agradezco a las personas que llegaron a mi vida para ayudarme a sobrellevarla y que la hicieron más divertida, aquellos amigos que estuvieron en mi infancia y aún siguen, amigos que se sumaron en esto cuatro años de carrera y que hicieron más ameno el recorrido.

Finalmente, quisiera darle demasiada gratitud a mi querida Sofia, quien con su amor ilumino de una forma tan bonita mi vida, quien llego en momentos difíciles y se convirtió en una fuente invaluable de motivación para mí, una compañía maravillosa.

- Chamaidan Morocho, Carlos David

AGRADECIMIENTO

La etapa universitaria ha sido cómo un tren y en cada parada encontrabas una sorpresa, una nueva experiencia, un sentimiento, una emoción y sin duda desbloqueando personajes nuevos. Por todo lo vivido y lo que falta vivir, quisiera agradecer la oportunidad otorgada por mis padres de poder estudiar, sonreír, equivocarme y aprender. Por todo el conocimiento adquirido y la guía ejemplar, quisiera agradecer a aquellos profesores que demostraron su amor hacia la docencia y su bondad en convertirnos profesionales del futuro. Por todas las fugas, copias y risas, quisiera agradecer a las personas encontradas en el camino, conocidos que se convirtieron en amigos, los que se fueron y se quedaron, siempre ocuparán un espacio dentro de mi corazón. Sin más ni menos, agradezco a todos los actores participes en esta etapa, espero nuestros caminos algún día vuelvan a cruzarse, y si no, espero el universo siempre esté cuidándolos.

- Quisnia Bohórquez, Connie Scarlet

DEDICATORIA

Este trabajo se lo dedico a mis padres, pilares de mi vida, quienes me brindaron un amor incondicional y que con su sacrificio día a día han hecho posible cada uno de mis logros. Desde siempre, me han acompañado, dándome su constante apoyo, llenándome de enseñanzas y demostrándome el valor del esfuerzo y perseverancia.

A mi madre Angela Morocho, quién ha sido un increíble modelo que seguir, una inspiración y un ejemplo de superación, una madre que me dio tanto amor desde mi nacimiento y sigue demostrándomelo hoy en día, siempre presente en mi vida, aquella que mostró lo maravilloso del mundo y la viva muestra del esfuerzo.

A mi padre, Carlos Chamaidan, mi héroe, un padre como ningún otro, quien a su manera me ha demostrado el mayor que se le puede tener un padre a su hijo. Recuerdo que en algún momento me dijo que no quería que fuera como él, sin embargo, a mi parecer él ha sido un ejemplo de fortaleza y de ingenio durante toda mi vida. Además, ya me parezco a él.

A los dos los amo y les doy las gracias por tanto en esta vida, es por eso por lo que este logro es tanto mío como suyo. Son los mejores padres que la vida me pudo haber dado.

Y una dedicatoria especial ese niño que fui, que tuvo grandes sueños y una imaginación inmensa, que visualizaba un futuro, rindo homenaje a tu esfuerzo y valentía, que, aunque estabas por ceder, nunca

lo hiciste, te abriste paso antes los desafíos. Mantén esto como un recordatorio de:

“Puedes y eres capaz de convertir tus sueños en metas, y de esas metas en una realidad”

- Chamaidan Morocho, Carlos David

DEDICATORIA

Con total alegría y confianza, quisiera dedicar la presente investigación a dos personas muy importantes. Querido Miguelito, espero el sol ilumine tu camino y las nubes te den paso para verme convertir en una profesional. Querida niña de ocho y dieciochos años que vivía aterrada y emocionada de cómo sería el futuro, y se cuestionaba si alguna vez fuese capaz de mucho, definitivamente, siempre has sido capaz, y siempre has tenido seguridad. Con certeza y el corazón en la mano pude, puedo y podré asegurarte de que la luz interna nunca te abandonará, eres tu propio eje y guía, creer en ti mismo es la prueba de amor propio más grande. Nunca dudaste de ti y no tenías por qué. Tranquila, todo salió increíble, tu mundo no se derrumbó cuándo no estudiaste lo que soñabas.

Cuando sientas que te estás ahogando recuerda:

“Dentro de 5 años, nada de esto importará, estaremos en mejores lugares, con mejores personas y en mejores vidas”

- Quisnia Bohórquez, Connie Scarlet



UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE ECONOMÍA Y EMPRESA
CARRERA DE NEGOCIOS INTERNACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. *Gabriela Hurtado*

Ing. Hurtado Cevallos, Gabriela Elizabeth, Mgs.

DIRECTORA DE CARRERA

f. *César Freire*

Lic. Freire Quintero César Enrique

COORDINADOR DEL ÁREA

f. _____

Mg. Franco Quiroga Santiago

OPONENTE

Índice de contenido

Capítulo 1: Generalidades de la Investigación	2
Introducción	2
Problemática.....	7
Justificación.....	10
Alcance.....	13
Objetivos	16
Objetivo General.....	16
Objetivos específicos	16
Capítulo 2: Fundamentos Teóricos	17
Marco Teórico	17
<i>Optimización de rutas</i>	17
<i>Transporte de Contenedores</i>	18
<i>Logística Terrestre en Centros de Acopio</i>	20
<i>Gestión de Rutas en Centros de Acopio</i>	21
<i>Machine Learning</i>	23
<i>Modelos de Clasificación en Machine Learning</i>	26
Marco Conceptual	32
<i>Centro de Acopio</i>	32
<i>Optimización de rutas</i>	32
<i>Machine Learning</i>	34
Algoritmos de Machine Learning.....	34
<i>Tipos de algoritmos de Machine Learning</i>	37
Inteligencia Artificial.....	39
<i>Tipos de inteligencia Artificial</i>	41
Marco Legal.....	43
Capítulo 3: Metodología.....	44
Metodología en el programa R studio.	45
<i>Instalación y carga de paquetes</i>	45
Metodología Árbol de Decisión.	47

Metodología Random Forest	49
Análisis de Resultados	51
Relación de Variables	51
Tablas de Frecuencia de variables objetivas.....	52
<i>Variable Ruta</i>	52
<i>Variable Caja</i>	54
<i>Variable Tarifa</i>	56
Gráficos de Frecuencia de variables objetivas	62
<i>Variable Viaje</i>	68
<i>Variable Tiempo</i>	72
Gráficos de relación entre las variables objetivas	74
<i>Variable Ruta x Variable Tiempo</i>	76
<i>Medidas de Tendencia Central</i>	80
<i>Medidas de Dispersión</i>	84
<i>Variable Viaje</i>	94
<i>Variable Distancia</i>	97
<i>Variable Tiempo</i>	99
Creación del Modelo de clasificación	101
<i>Árbol de Decisión</i>	101
<i>Bosque Aleatorio</i>	106
Discusión.....	114
Conclusiones	115
Referencias Bibliográficas	117
Anexos.....	123

Índice de Figuras

Figura 1	29
Figura 2	31
Figura 3	44
Figura 4	51
Figura 5	63
Figura 6	65
Figura 7	67
Figura 8	69
Figura 9	71
Figura 10	75
Figura 11	77
Figura 12	79
Figura 13	90
Figura 14	92
Figura 15	94
Figura 16	96
Figura 17	98
Figura 18	100
Figura 19	102
Figura 20	105
Figura 21	107
Figura 22	109
Figura 23	111
Figura 24	112

Índice de Tabla

Tabla 1	108
---------------	-----

Resumen

La optimización representa la parte más importante dentro de una empresa cuando se quiere reducir costos, incrementar el posicionamiento propio y mejorar la toma de decisiones dentro de una organización, para obtener resultados que brinden la mejor calidad y eficiencia al cliente final. Para ello, se debe reconocer que en la actualidad estos procesos pueden ser seguidos por herramientas que involucren la innovación y automatización junto con el análisis de datos de forma mucha más rápida y certera. Permitiendo así a las organizaciones, ser actores de primera fila en un mercado que es cada vez más rápido, más demandante, pero sobre todo más competitivo. A nivel empresarial es indispensable que las organizaciones opten por la aplicación de herramientas que involucren el crecimiento y aceleramiento de los procesos dentro de su cadena de suministro. La presente investigación se enfoca en el desarrollo y construcción de un modelo Bosque Aleatorio para la clasificación de rutas de transporte de forma más ágil y menos costosa, en la cuál se pretende obtener como resultado una ruta que sea favorable en cuestiones de tiempo y dinero para la entrega de cajas de banano hacia el Puerto de Machala. El objetivo es determinar cuáles son las variables a considerar y que influyen en los recorridos de las rutas al momento de elegir la más conveniente. La aplicación de Machine Learning se justifica porque permite clasificar de forma precisa y autónoma los factores que intervienen obteniendo una respuesta objetiva.

Palabras claves: Innovación, Análisis de datos, Inteligencia Artificial, Árbol de Clasificación, Random Forest para la clasificación, Rutas de Transporte

Abstract

Optimization represents the most important part within a company when you want to reduce costs, increase your own positioning and improve decision making within an organization, to obtain results that provide the best quality and efficiency to the end customer. To do this, it must be recognized that currently these processes can be followed by tools that involve innovation and automation along with data analysis in a much faster and more accurate way. Thus allowing organizations to be leading players in a market that is increasingly faster, more demanding, but above all more competitive. At the business level, it is essential that organizations opt for the application of tools that involve the growth and acceleration of processes within their supply chain. The present research focuses on the development and construction of a Random Forest model for the classification of transportation routes in a more agile and less costly way, in which the aim is to obtain as a result a route that is favorable in terms of time and money for the delivery of boxes of bananas to the Port of Machala. The objective is to determine which variables to consider and that influence the routes when choosing the most convenient one. The application of Machine Learning is justified because it allows the factors involved to be classified accurately and autonomously, obtaining an objective response.

Keywords : Innovation, Data Analysis, Artificial Intelligence, Classification Tree, Random Forest for classification, Transportation Routes

Résumé

L'optimisation représente l'élément le plus important au sein d'une entreprise lorsque l'on souhaite réduire les coûts, augmenter son propre positionnement et améliorer la prise de décision au sein d'une organisation, pour obtenir des résultats qui offrent la meilleure qualité et efficacité au client final. Pour ce faire, il faut reconnaître qu'actuellement ces processus peuvent être suivis par des outils qui impliquent l'innovation et l'automatisation ainsi que l'analyse des données de manière beaucoup plus rapide et précise. Permettant ainsi aux organisations d'être des acteurs de premier plan sur un marché de plus en plus rapide, plus exigeant, mais surtout plus compétitif. Au niveau commercial, il est essentiel que les organisations optent pour l'application d'outils qui impliquent la croissance et l'accélération des processus au sein de leur chaîne d'approvisionnement. La présente recherche se concentre sur le développement et la construction d'un modèle Random Forest pour la classification des itinéraires de transport de manière plus agile et moins coûteuse, dans lequel le but est d'obtenir comme résultat un itinéraire favorable en termes de temps et d'argent. pour la livraison de caisses de bananes au Port de Machala. L'objectif est de déterminer quelles variables prendre en compte et qui influencent les itinéraires lors du choix de celui qui convient le mieux. L'application du Machine Learning est justifiée car elle permet de classer les facteurs impliqués avec précision et de manière autonome, obtenant ainsi une réponse objective.

Mots-clés : Innovation, analyse de données, intelligence artificielle, arbre de classification, forêt aléatoire pour la classification, itinéraires de transport

Capítulo 1:

Generalidades de la Investigación

Introducción

Machala es una de las ciudades de la provincia de Guayas dónde la economía se mueve de forma rápida al igual que sus actividades dentro de su centro de acopio. Esto yace desde el siglo XX, cuando en la región costera del sur de Ecuador en la provincia del Oro, se comienza la actividad pionera de la cultivación y agricultura en una cartera amplia de productos (*La Economía de Machala Se Mueve al Ritmo de Sus Operaciones Portuarias, 2023*).

Debido a la producción masiva y la comercialización de una diversidad de materia prima en bienes de agricultura nacional, surge en la ciudad de Machala el resultado de un realce en la economía de las familias ecuatorianas parte de la clase obrera impulsando el PIB del país y a los empresarios, inversionistas y agricultores de la época. Concentrando así una demanda mayor de hasta el 65% de producción nacional (*La Economía de Machala Se Mueve al Ritmo de Sus Operaciones Portuarias, 2023*).

Dado que la producción de la cartera de productos varios ecuatorianos se incrementan, el gobierno decide comenzar procesos de exportación a nivel internacional de aquellos bienes más destacados para posicionar al país dentro de la lista de comerciantes en el mercado mundial y frente a sus competidores (*La Economía de Machala Se Mueve al Ritmo de Sus Operaciones Portuarias, 2023*).

El surgimiento de esta actividad pionera dentro de la industria es tan significativo debido a los visionarios que comenzaron a experimentar con el cultivo en tierras fértiles costeras de la región, además de ello, las condiciones climáticas y la ubicación estratégica cerca de uno de los puertos con mayor afluencia de comerciantes dentro de la zona, Puerto Bolívar. De esta manera, Machala comienza su postura cómo una ciudad perfecta con perfil exportador, comerciante y nicho económico.

Al pasar de los años, el Oro se hizo notorio en sus producciones y exportaciones, contando con seis empresarios que dejaron su legado en la provincia y país por contribuir en la actividad agrícola enmarcando su esfuerzo, dedicación y trabajo.

Por consiguiente, dicho primer éxito dentro de la región costera llamaría la atención de inversionistas extranjeros propuestos a expandir e incrementar la exportación y cultivación de productos peculiares para poder comercializarlo en mercados internacionales como los propios (*Consejo de la Judicatura, 2024*).

De esta forma, a la magnitud de producción se establece la necesidad de incentivar al gobierno en la construcción de un establecimiento que permita sobrellevar la logística de la cantidad de productos y empresas productoras dentro de la ciudad.

Por ello, se construyó el Centro de Acopio Portuario de Machala. Poco a poco este se convierte en la pieza fundamental dentro de la infraestructura de Machala, ya que no sólo le permitía la actividad comercial a nivel nacional, sino que promovía la actividad comercial internacional entre empresas, agricultores y países (*Pioneros de La Actividad Bananera En El Oro, 2020*).

A través de este Centro de Acopio, muchas de las transacciones comerciales se establecieron en una estimación de tiempo mínima y de una forma más cómoda entre comerciantes. Repartiéndose así la efectividad de intercambio de bienes e inaugurando sistemas logísticos novedosos en el comercio ecuatoriano (*Pioneros de La Actividad Bananera En El Oro, 2020*).

El primer Centro de Acopio Portuario de Machala se efectúa para la década de los años 70 en Ecuador, bajo la exigencia de comerciantes en búsqueda de satisfacer la necesidad de un puerto que establezca la eficiencia y rapidez en la exportación y comercialización de bienes agrícolas.

La ubicación geográfica de Machala fue elemental para el funcionamiento de un centro de acopio. Dado que el lugar escogido fue estratégico por su costa pacífica fijando así la entrada para el acceso de los diferentes transportes dentro del centro portuario accediendo a los mercados de forma más libre (*Consejo de la Judicatura, 2024*).

Durante 10 años el centro de acopio se mantuvo en sus primeras facetas debido a la falta de inversión, sin embargo, para los 80 y 90 se comenzaron a planear nuevas ideas de experimentación en relación con la expansión y modificación del acopio para mayor visibilidad

comercial.

Las modificaciones comenzaron con la modernización y ampliación del acopio, ya que, de esta forma, se permitiría un mayor manejo en el creciente volumen de exportaciones nacientes. Mejorando así las infraestructuras de la ciudad y su valor, como al mismo tiempo la globalización permitió el uso de tecnologías modernas que permitieron realizar operaciones mucho óptimas aprovechando el mayor tiempo posible.

Así mismo, esta etapa moderna ayudó con la expansión de bodegas y muelles cercanos a las áreas de embarque, además, el uso de sistemas de refrigeración en maquinarias que permitían el registro de los productos comercializados.

Como consecuencia de la modernización, la ciudad logra convertirse en el punto clave de actividades aduaneras, importación y exportación de bienes y de múltiples productos agrícolas y pesqueros nacionales. En esa misma línea, se potencializa la economía local en la clase obrera por el fortalecimiento de la balanza comercial del país.

Por otra parte, el centro de acopio fue la oportunidad esencial para el despegue de la empleabilidad. Disminuyendo la tasa de desempleo hasta un 20% durante la actividad comercial. Miles de puestos de trabajo directos e indirectos en el centro de acopio dentro de la región se crearon para cubrir las necesidades de personal por el flujo excesivo de bienes (Valverde, 2024).

Dichos puestos surgieron desde agentes portuarios a cargo de actividades de negociación y aduaneros hasta empleados a cargo de las industrias como la logística, servicios de warehouse, y procesamiento de los bienes con calificación alta, creándose una codependencia entre empleados y el acopio para un mejor funcionamiento de este.

En suma, este crecimiento aportó a la inversión de muchas otras áreas participes en el funcionamiento del centro de acopio para apoyar las necesidades requeridas durante el proceso de actividades. Tanto así que, el desarrollo de la infraestructura complementaria en la ciudad y región fue requerida para que las rutas se mantuvieran aptas y los sistemas de transporte y logísticos disminuyeran los posibles inconvenientes al momento de procesar sus actividades.

Adicionalmente, la inversión traspasó para programas de capacitación guiada múltiples en el desarrollo de habilidades, aptitudes y conocimientos de los trabajadores involucrados en las actividades portuarias. Como resultado, muchos empleados pudieron experimentar una eficiencia y eficacia en sus actividades de operatividad. (Valverde, 2024).

No obstante, el centro de acopio ha tenido que enfrentar desafíos debido a los diversos competidores dentro del estado nacional e internacional con puertos que ofrecen servicios mucho más modernos comparado con los propios. Por lo que, la etapa de mejoras se mantiene de forma continua en ámbitos de modernización y de regulaciones internacionales que permitan renovar los servicios actuales.

Inclusive, la demanda de grupos ambientalistas, influyó en la puesta en marcha de normativas y regulaciones que protejan al medio ambiente durante las actividades portuarias. Debiendo cumplir con normas cruciales que complementen los estándares internacionales de acción, y que permitan asegurar la relación de las actividades bajo el concepto de sostenibilidad.

A pesar de la modernización, el acopio ha tenido que sobrellevar las preocupaciones consecuentes de problemas ambientales. Estos guardan relación con las gestiones internas en la producción y eliminación de desechos, reducción de emisiones de carbón y salvaguardar la protección de los ecosistemas que rodean acopio costero.

Eventualmente, las prácticas de manejo ambiental responsable fueron las medidas que al acoger el centro de acopio potencializaron su aceptación del mercado y permitiéndole una mejor reputación dentro del círculo de competidores nacionales e internacionales. Esto a su vez, iniciando proyectos de conservación que busquen mitigar los impactos negativos que podrían causarse por medio de una actividad mal ejecutada.

Actualmente, el Centro de Acopio Portuario de Machala, en la peña de Oro, se mantiene con la exclusividad de considerarse así mismo como una pieza fundamental en el propulsor de la cultura, política, economía y sociedad ecuatoriana (*Consejo de la Judicatura*, 2024).

Con los avances de la industrialización y la globalización, la adquisición de tecnologías que elaboren avances instantáneos y autónomos es primordial. Por ello, el acopio se ha permitido

aumentar su capacidad redireccionando a procesos basados en la eficiencia.

La exploración de procesos abrió brecha a nuevos indicadores de inversión en futuros campos bodegueros que mantengan una diversificación de servicios acogedores a la necesidad de los actores. Se ha fortalecido la posición de superioridad en el mercado nacional frente a otros centros de acopio.

Por conclusión, el impacto ocasionado con la creación y la innovación del centro de acopio en todos sus ámbitos, externos e internos, de infraestructura, especialización y de personal asociado; han contribuido de forma significativa a un desarrollo nacional y no solo a Machala.

Este desarrollo social concedió un acceso a las mejoras en la calidad de vida de la población local que rodea la ciudad, pero también en el alza de PIB nacional frente a los múltiples y diversos países de América Latina.

Finalmente, es la visión y misión conjunta y estratégicamente orientada a un objetivo sostenible y creativo que da como efecto un centro de acopio apto para enfrentar futuros desafíos y competidores más veraces, y en el paso la continua contribución con el progreso de Ecuador.

Problemática

Una de las actividades logísticas que se realiza dentro y fuera de los centros de acopio es el transporte de contenedores, la cual es una parte fundamental para el funcionamiento de las cadenas de suministro y también para el desarrollo de la economía ecuatoriana, en especial para el comercio internacional.

Actualmente, en Ecuador este sector se encuentra atravesando constantemente diversos desafíos de suma importancia dentro de la optimización de rutas y su eficiencia

Algunos de los desafíos que este sector atraviesa vienen dados por la situación de inseguridad del país, mal estado en algunas de las vías de por las cuales circulan los transportes de cargas, ineficiencias en la planificación previa de rutas y falta en la integración de nuevas tecnologías (Calvopiña Llambo et al., 2023).

El aumento significativo de la inseguridad durante los últimos años en Ecuador se presenta como uno de los mayores desafíos por los que se enfrenta el sector logístico a nivel nacional, incluyendo el transporte de contenedores.

La alta posibilidad de que sucedan robos en el transcurso donde los contenedores se mueven de punto a punto, así también como el asalto a los choferes y las extorsiones o planeaciones para alterar las cargas, se ha convertido en una amenaza constante, no solo para las empresas, sino también para las personas que están detrás de todo este proceso (Calvopiña Llambo et al., 2023).

Además, esto trae como consecuencia un considerable aumento en costos de operación para las operaciones, como la implementación de medidas de seguridad, sistemas de rastreo satelital o en algunos casos contrataciones de personal de seguridad privada. Sin duda alguna, esto se ha convertido en una problemática de suma importancia para aumentar el nivel de eficiencia durante el transporte de contenedores en distintas áreas.

Otro de los temas que también tienen son las carreteras y el estado en los que estas se encuentran, en el país hoy en día aún existen carreteras en muy mal estado que afectan negativamente a las áreas de movimiento de mercancías, reduciendo su eficiencia. Las

condiciones y distintos problemas que estas presentan como baches, faltas de señalización, nulo mantenimiento y una pobre supervisión, son factores que no solo ralentizan el tráfico y circulación vehicular, sino que aumentan los tiempos de entregas de cargas, además, de generar un gran nivel de desgaste en los vehículos de transporte, lo que a su vez incrementa los costos de mantenimientos y reparaciones de estos.

Debemos de tener en cuenta que este tipo de problemáticas, no solo aumentan generan daños materiales, sino que pueden ser causas de accidentes y aumentar las posibilidades de estos, el mal estado de las vías no solo pone en riesgo las mercancías que se transportan, pone en riesgo la seguridad de los transportistas, así como las demás personas que se encuentren transitando las mismas carreteras y quienes usualmente son más propensos a recibir graves daños durante cualquier siniestro de tránsito que pueda ocurrir durante la circulación (Guaña Moya & Salgado Reyes, 2019).

En la logística de transporte en el país de Ecuador se ve un entorno que se enfrente a múltiples desafíos, y entre ellos, uno de los más importantes por el que tiene que pasar este sector es la falta de nuevas tecnologías, más avanzadas, para poder lidiar con el control y la optimización de las operaciones de transporte y la cadena de suministro de forma general.

Dentro del país la implementación de tecnología más avanzada en el sector de transporte de contenedores es muy escasa, la ausencia de sistemas informáticos y de comunicación con mayor eficiencia genera una limitación en la optimización de las rutas, además que hace mucho más difícil poder realizar una adecuada toma de decisiones dentro de los distintos tipos de operaciones se realicen en este sector.

La falta de datos precisos, un déficit en la gestión de rutas y una planificación de rutas poco eficiente, son consecuencia del impacto de un bajo nivel tecnológico presente dentro de este sector, esto limita de forma significativa las capacidades de predicciones y respuestas durante las apariciones de problemas operativos. Todo esto nos da a entender que la implementación de nuevas tecnologías y modernización de sistemas se vuelve un área crucial para un mayor desempeño dentro de la optimización de rutas para el transporte de los contenedores (Chicaiza & Sandaya, 2015).

Es necesario también mencionar que la capacidad de los centros de acopio se ven afectados, debido a la falta de una mejor infraestructura dentro de estos, teniendo varias veces un impacto negativo en la eficiencia de los procesos. Además de que, en distintas ocasiones, existe una falta de organización dentro de los centros, lo cual no permite sacar el máximo provecho o realizar un proceso más eficiente en los mismos (Riveros & Silva, 2007).

Llevar a cabo una optimización para el transporte de contenedores dentro y fuera de los centros de acopio es una compleja temática que enfrenta múltiples desafíos, pero que, a su vez, se convierte en un proceso fundamental para el desarrollo dentro del sector logístico y comercial del Ecuador.

Justificación

El transporte de contenedores es un componente vital en la cadena logística global. Con el constante crecimiento del comercio internacional, la eficiencia en el transporte de contenedores se ha vuelto crucial para las empresas que buscan reducir costos, mejorar la puntualidad y minimizar el impacto ambiental.

Este no solo facilita el movimiento de bienes a través de diferentes regiones y países, sino que también juega un papel esencial en la economía mundial, ya que aproximadamente el 90% de los bienes comercializados a nivel mundial se transportan por mar en contenedores (Rodríguez & Notteboom, 2015). Este método de transporte permite un mejor uso del espacio, una mayor seguridad de la carga y una mayor flexibilidad en la planificación de las rutas logísticas.

La constante complejidad durante la planificación de rutas y la gestión de contenedores presenta numerosos desafíos que requieren soluciones innovadoras. La evolución de la tecnología y la disponibilidad de grandes volúmenes de datos han abierto nuevas oportunidades para mejorar la eficiencia del transporte de contenedores mediante el uso de algoritmos de Machine Learning y otras técnicas avanzadas de análisis de datos (Casanova Lugo & Torres Anzola, 2020).

El sector del transporte de contenedores enfrenta múltiples desafíos que afectan su eficiencia y sostenibilidad. Entre los problemas más comunes se encuentran las rutas subóptimas, donde la planificación ineficiente de rutas puede llevar a recorridos más largos de lo necesario, aumentando los costos de combustible y el desgaste de los vehículos (Boldyrieva et al., 2019).

Además, la falta de coordinación puede provocar congestión en puertos y terminales, resultando en demoras y costos adicionales (Rodríguez & Notteboom, 2015). Las rutas ineficientes también contribuyen a mayores emisiones de CO₂, agravando el problema del cambio climático. (Bektaş & Laporte, 2011).

El uso de modelos basados en Machine Learning para la optimización de rutas ofrece una solución prometedora para abordar estos desafíos. Al aplicar algoritmos de clasificación, se pueden identificar patrones en los datos históricos de transporte, permitiendo la creación de rutas más eficientes y adaptativas (Bojic et al., 2020).

Un modelo optimizado puede reducir significativamente los tiempos de transporte y los costos operativos (Gendreau et al., 2018). Además, la optimización de rutas puede reducir las emisiones de gases de efecto invernadero, contribuyendo a la protección del medio ambiente (Bektaş & Laporte, 2011). Al mismo tiempo, ofrece una oportunidad para aumentar la precisión y eficacia en la planificación de rutas mediante la identificación de patrones complejos en los datos de transporte, algo que sería difícil de lograr con métodos tradicionales de optimización (Gendreau et al., 2018).

El presente proyecto busca generar un alto valor tanto académico como científico. Se amplía el cuerpo de conocimiento en la intersección entre logística y el transporte, como la aplicación del Machine Learning en dichas áreas. Además, de la intencionalidad de la creación y validación de modelos de clasificación que sean capaces de ofrecer nuevas metodologías aplicables y funcionales en el contexto de la optimización logística de transportes.

Implementar un modelo de clasificación basado en Machine Learning para la optimización de rutas en el transporte de contenedores puede generar varios beneficios prácticos. Al optimizar las rutas, se pueden reducir significativamente los costos asociados al combustible y al mantenimiento de los vehículos (Rodríguez & Notteboom, 2015). Rutas más eficientes resultan en una mayor puntualidad en las entregas, mejorando la satisfacción del cliente.

Sin mencionar que, la capacidad de anticipar y adaptarse a condiciones variables en tiempo real, como el tráfico o las condiciones climáticas, puede mejorar el nivel de resolución frente problemáticas operativas, así como la capacidad de respuesta del sistema logístico en su conjunto (Boldyrieva et al., 2019).

Para Eslava (2003) la disponibilidad de datos precisos y análisis avanzados facilita la toma de decisiones estratégicas y operativas, permitiendo a las empresas adaptarse rápidamente

a las demandas cambiantes del mercado.

Debido a la naturaleza del proyecto, es importante mencionar que es muy factible su elaboración gracias a disponibilidad de grandes volúmenes de datos históricos de transporte y rutas permite el entrenamiento efectivo de los modelos de Machine Learning. Existen numerosas herramientas y plataformas de Machine Learning que facilitan la implementación y ajuste de modelos de clasificación. (Schneider et al., 2016).

Las empresas de transporte y las instituciones académicas muestran un creciente interés en adoptar tecnologías avanzadas para mejorar la eficiencia y competitividad.

La optimización de rutas de transporte mediante el uso de Machine Learning es una temática de suma importancia la cual puede ponerse como una de las prioridades a resolver dentro del sector logístico y de transporte, ya sea por solución que ofrece a ciertas problemáticas existentes en el área o por los beneficios y avances de eficiencia, tanto de forma operativa como de sostenibilidad y competitividad. Este trabajo busca proporcionar soluciones prácticas y reales que generen un gran impacto, y que a su vez contribuyan con el desarrollo tecnológico y de innovación.

En conclusión, este proyecto busca generar una respuesta innovadora frente a los grandes desafíos a los cuales actualmente el sector logístico y de transporte se enfrentan de forma constante, además de generar una mejora significativa al momento de su implementación, siendo así un aporte muy valioso tanto para la industria como para la comunidad científica.

Alcance

La importancia de la optimización del tiempo en la entrega de un producto habla mucho del tipo de servicio que se maneja en una organización. Pero mucho más allá de ello, es el reflejo de la eficiencia integrada en la manipulación de los procesos. El buen funcionamiento de una entidad empresarial se encuentra en la administración del procesamiento de pedidos de sus rutas y almacenamiento. Por ello, son estos aspectos de carácter crítico que constan con mayor relevancia cuando se requiere adquirir una organización eficiente y eficaz (Manzanilla, 2022).

La inteligencia artificial emerge como diversos sistemas informáticos en el nuevo siglo totalmente revolucionaria, capaz de impulsar las mejoras de forma drástica con respecto a lo que denominamos eficiencia cuando de procesos rápidos nos referimos. Siendo capaz de desafiar constantemente a la industria en encontrar formas innovadoras de gestionar y optimizar las operaciones de entrega (Ramirez, 2023).

La práctica de la inteligencia artificial a través del Machine Learning durante los últimos años ha cambiado la forma en que las organizaciones gestionan el suministro y sus operaciones, ya que, esta se ha convertido en un instrumento productora de estrategias que ha permitido mantenerse y establecerse mucho más competitivo en un mercado que exige y demanda más (*Inteligencia Artificial en Logística*, 2024).

La implementación de estas metodologías es que puede ser capaces de analizar grandes cantidades de datos conocidos como Big Data, que proveen la identificación de patrones que predicen los diversos comportamientos en la toma de decisiones para mejorar la eficiencia y precisión en la etapa del proceso logístico como lo es la optimización de rutas y entregas en el centro de acopio.

Utilizando algoritmos avanzados más la recopilación de datos junto con métodos basados en Machine Learning que son esenciales para analizar las múltiples variables presentes, cómo el tráfico, el clima, los obstáculos y los patrones de demanda de productos que serán aquellos indicadores que determinarán y seleccionarán las rutas óptimas para el transporte de mercancías.

Además, debido a estos sistemas, los tiempos de entrega entre el acopio y el destino final

de los usuarios se acortaron hasta el punto de que podrían disminuir el costo del consumo de combustible excedente reduciendo costos operativos (*Inteligencia Artificial en Logística*, 2024).

Por ende, se iniciaría así una categorización que permita llevar un proceso más acertado en el menor tiempo posible, pero con la más grande efectividad en el bienestar de la entrega al cliente final. De tal manera, se ha mejorado el servicio al cliente y la percepción de la organización en el mercado, potencializando la preferencia del acopio de Machala en particular por los usuarios.

Por otro lado, aunque la aplicación de estas herramientas representa una inversión significativa dentro de las organizaciones, se dictamina como indispensable en el ámbito de modernización e innovación por su capacidad de minimizar errores comunes en la gestión capacitando al personal en el paso al mundo automatizado.

Por esta razón, la presente investigación surge para cubrir las incógnitas hacia estudiantes, profesionales, empresas y organizaciones que buscan despejar dudas sobre los beneficios del uso correcto del Machine Learning dentro de sus actividades comerciales y que deseen adquirir conocimiento de cómo aplicarlo dependiendo de la necesidad requerida a satisfacer.

Adicionalmente, se conoce que sólo el 20% de las empresas ecuatorianas en la industria han optado por la implementación de prototipos de IA dentro de sus operaciones. Apostado así, al desarrollo tecnológico de forma simultánea y continua en un intento de modernización y crecimiento (Ramirez, 2023).

Concretamente, para el 2025 se estima que el 40% de las empresas ecuatorianas adhieran herramientas tecnológicas de IA en la gestión de sus organizaciones, mientras que el 70% de la industria podría estar planeando y considerando en implementarlo de ser necesario y demostrar eficiencia (Ramirez, 2023).

Concluyendo, esta apreciación permite comprender el impacto que se le ha otorgado a la IA como una herramienta complementaria en actividades empresariales y en la vida cotidiana.

Múltiples líderes a nivel mundial reconocen el impulso de ventajas competitivas que estos sistemas proveen convirtiendo a sus colaboradores y empresa mucho más actualizados con el mundo que nunca.

Objetivos

Objetivo General

- ❖ Desarrollar un modelo de clasificación para la optimización de rutas del centro de acopio de Machala mediante el uso de Machine Learning

Objetivos específicos

- ❖ Indagar, analizar y comprender los conceptos teóricos y definiciones del Machine Learning aplicado en el área logística y de transporte.
- ❖ Identificar y explicar las aplicaciones de los métodos y algoritmos parte del Machine Learning aplicables al proyecto
- ❖ Analizar la base de datos del centro de acopio de Machala del año 2023 para el desarrollo de un modelo de clasificación

Capítulo 2:

Fundamentos Teóricos

Marco Teórico

Optimización de rutas

La optimización de rutas en el transporte juega un papel fundamental en la eficiencia y rentabilidad de las operaciones logísticas.

Gendreau et al (2018) destacan la relevancia de implementar un sistema de asignación de rutas, específicamente el Vehicle Routing Problem (VRP), con el objetivo principal de minimizar tiempos, costos y mejorar la calidad del servicio de entrega y recogida de mercancías.

En el contexto actual, las ciudades se caracterizan por ser entidades complejas con redes logísticas interconectadas que requieren una optimización constante para garantizar su sostenibilidad. La eficiente gestión de las rutas de transporte urbano de personas y mercancías se vuelve crucial para mantener el flujo de bienes y servicios de manera efectiva y competitiva en el mercado.

Valencia (2004) menciona que El transporte de carga es una actividad fundamental en la economía de un país, ya que permite la distribución de productos hasta el consumidor final, impulsando la economía y facilitando la circulación de bienes. No obstante, los costos relacionados con el transporte constituyen uno de los principales desafíos logísticos, afectando directamente las relaciones con proveedores, clientes y competidores.

La implementación de un modelo de optimización de rutas no solo busca reducir los costos operativos, sino también mejorar la satisfacción del cliente al garantizar entregas puntuales, seguras y en las cantidades requeridas. Además, al gestionar de manera eficiente el transporte, las organizaciones pueden aumentar su competitividad, lograr economías de escala y potencialmente reducir los precios de sus productos, lo que se conoce como una ventaja o beneficio de la misma forma para la organización y para sus clientes.

Transporte de Contenedores

El transporte de contenedores es una parte fundamental de la logística moderna, permitiendo el movimiento eficiente de mercancías a través de diferentes modos de transporte.

El contenedor es un equipo diseñado para contener mercancías durante su transporte, facilitando su manipulación y traslado entre diferentes modos de transporte. Una de las características clave de los contenedores es que permiten ser transportados fácilmente con medios mecánicos, lo que agiliza el proceso de carga y descarga. Además, los contenedores son apilables y resistentes, lo que les permite ser reutilizados múltiples veces sin comprometer la seguridad de las mercancías.

Otra característica importante es que los contenedores están provistos de elementos de anclaje interior, que facilitan la estiba de las mercancías y aseguran su estabilidad durante el transporte. Por último, los contenedores cumplen con normas internacionales de dimensiones y seguridad, lo que garantiza su uso global y facilita el comercio internacional al permitir una fácil interoperabilidad entre diferentes medios de transporte y países (Boldyrieva et al., 2019).

Los desafíos en el transporte de contenedores en tierra incluyen la congestión del tráfico, las ineficiencias en la planificación de rutas y los costos operativos elevados. Estos problemas pueden llevar a retrasos significativos y aumentos en los costos de transporte, afectando negativamente la competitividad de las empresas (Rodríguez et al., 2006).

Uno de los problemas más comunes en la logística terrestre es la congestión del tráfico. En muchas áreas urbanas y regiones industriales, la densidad del tráfico puede causar retrasos significativos, lo que afecta el tiempo de entrega y la fiabilidad del servicio de transporte (Rodríguez et al., 2006). Esta congestión no solo aumenta los tiempos de tránsito, sino que también incrementa los costos operativos debido al mayor consumo de combustible y al desgaste de los vehículos.

Otro desafío importante es la infraestructura insuficiente. En algunos países, la calidad de las carreteras y puentes no es adecuada para soportar el peso y el volumen del tráfico de contenedores, lo que puede causar daños a los vehículos y retrasos en el transporte (Smith et al., 2004). Además, la falta de áreas de estacionamiento adecuadas para camiones puede crear

cuellos de botella en las rutas de transporte, complicando aún más la planificación logística.

La variabilidad en los tiempos de tránsito también es un problema significativo. Factores como las condiciones meteorológicas, los accidentes de tráfico y las obras viales pueden introducir incertidumbre en los tiempos de entrega, lo que dificulta la planificación y la coordinación de la cadena de suministro (Boldyrieva et al., 2019). Esta variabilidad puede resultar en penalizaciones por retrasos y en una reducción de la satisfacción del cliente.

El costo del transporte de contenedores en tierra es uno de los componentes más importantes del costo total de la logística. Los costos de combustible, mantenimiento de vehículos y salarios de los conductores son factores que contribuyen significativamente al costo operativo total (Schneider et al., 2016). Con los precios del combustible fluctuando regularmente, las empresas enfrentan incertidumbre en sus costos de operación, lo que puede afectar sus márgenes de beneficio y su competitividad en el mercado.

La eficiencia operativa es otro aspecto crítico. La optimización de rutas es esencial para minimizar los costos y maximizar la utilización de los recursos. Sin embargo, lograr una eficiencia óptima puede ser complicado debido a la necesidad de equilibrar múltiples factores como la distancia, el tiempo, la capacidad de carga y las restricciones legales (Kang et al., 2019). Las ineficiencias en la planificación de rutas pueden resultar en viajes más largos de lo necesario, lo que aumenta el consumo de combustible y el desgaste de los vehículos, además de incrementar los tiempos de entrega.

Adicionalmente, la gestión de flotas es un desafío importante. Las empresas deben mantener un equilibrio entre tener suficientes vehículos disponibles para cumplir con la demanda y minimizar los costos asociados con el mantenimiento y la depreciación de la flota (Christopher, 1994). La tecnología de gestión de flotas, como los sistemas de seguimiento y telemetría, puede ayudar a mejorar la eficiencia, pero su implementación y mantenimiento también representan un costo significativo.

La relación con proveedores, clientes y competidores se ve afectada por la eficiencia y los costos del transporte de contenedores en tierra. Un desempeño logístico ineficiente puede llevar a rupturas en la cadena de suministro, afectando la disponibilidad de productos y la

satisfacción del cliente (Bojic et al., 2020). Además, los altos costos de transporte pueden reducir la competitividad de las empresas al aumentar el precio final de los productos, lo que puede resultar en una pérdida de cuota de mercado frente a competidores más eficientes.

Logística Terrestre en Centros de Acopio

Un centro de acopio se define como una instalación o infraestructura destinada a la recepción, almacenamiento, y posterior distribución de mercancías. Su principal función es consolidar productos de diversos orígenes para su distribución eficiente a múltiples destinos. Estos centros actúan como puntos intermedios en la cadena de suministro, facilitando la agrupación y redistribución de mercancías de manera que se optimicen los costos y se reduzcan los tiempos de entrega (Christopher, 1994).

El rol de los centros de acopio en la logística es multifacético. Sirven como puntos de consolidación donde las mercancías de distintos proveedores se agrupan antes de ser enviadas a los minoristas o directamente a los consumidores finales. Este proceso de consolidación permite una mayor eficiencia en el transporte, ya que las cargas se optimizan para maximizar la capacidad de los vehículos y reducir el número de viajes necesarios (Higginson & Bookbinder, 2005).

Los centros de acopio actúan como amortiguadores en la cadena de suministro, absorbiendo las variaciones en la oferta y la demanda. Esto es particularmente importante en industrias donde la demanda es altamente variable o donde los tiempos de entrega de los proveedores pueden ser impredecibles (Rodríguez et al., 2006).

Estos centros facilitan la gestión de inventarios y la distribución estratégica. Al centralizar el almacenamiento y la distribución, las empresas pueden tener un control más preciso sobre sus niveles de inventario, reducir los costos asociados con el almacenamiento descentralizado y mejorar la eficiencia de la distribución (Grigalunas et al., 2007). Además, la ubicación estratégica de los centros de acopio puede reducir significativamente los tiempos de transporte y los costos logísticos, proporcionando una ventaja competitiva en el mercado.

Los centros de acopio llevan a cabo una serie de procesos clave que son esenciales para su funcionamiento eficiente y efectivo. Entre estos procesos se incluyen la recepción de

mercancías, el almacenamiento, la gestión de inventarios, la preparación de pedidos y la distribución.

La recepción de mercancías es el primer proceso crítico en un centro de acopio. Este proceso implica la descarga de productos de los vehículos de transporte, la inspección de la mercancía para asegurar que cumpla con los estándares de calidad y cantidad, y la actualización de los registros de inventario (Smith et al., 2004). La eficiencia en este proceso es vital para evitar demoras y asegurar que las mercancías estén disponibles para el almacenamiento y la distribución lo antes posible.

El almacenamiento es otro proceso esencial que se realiza en los centros de acopio. Consiste en organizar y guardar las mercancías de manera que se optimice el espacio disponible y se facilite el acceso rápido a los productos cuando sea necesario (Košíček et al., 2012). Las estrategias de almacenamiento pueden variar desde el almacenamiento en estanterías hasta sistemas automatizados de almacenamiento y recuperación, dependiendo de las necesidades específicas del centro y de la naturaleza de los productos.

La distribución es el proceso de enviar los productos a sus destinos finales. Este proceso puede implicar la coordinación con transportistas externos, la planificación de rutas de entrega y la monitorización del progreso de los envíos para asegurar que se realicen de manera eficiente y puntual (Christopher, 1994). La optimización de la distribución es esencial para minimizar los costos de transporte y maximizar la satisfacción del cliente.

Gestión de Rutas en Centros de Acopio

Es indispensable poseer una gestión de rutas para la optimización logística del transporte y mejorar la eficiencia operativa. Sin embargo, existen aún muchas empresa y entidades que mantienen métodos tradicionales de planificación de rutas, como, por ejemplo:

- **Algoritmo de Clarke-Wright:** El algoritmo de Clarke-Wright es un método heurístico utilizado para resolver el problema del vendedor viajero (TSP, por sus siglas en inglés), donde el objetivo es encontrar la ruta más corta posible que pase por un conjunto de puntos y regrese al punto de origen. Este método busca combinar las entregas de múltiples destinos en una sola ruta para minimizar la

distancia total recorrida y optimizar el uso de los recursos disponibles (Mitchell, 1997).

- **Método del Vecino Más Cercano:** El método del vecino más cercano es otro enfoque heurístico utilizado para resolver problemas de optimización de rutas. Este método consiste en seleccionar el destino más cercano disponible en cada paso del recorrido, lo que simplifica la planificación de rutas pero puede no siempre resultar en la ruta más óptima en términos de distancia total recorrida (Hastie et al., 2009).
- **Planificación Manual:** Además de los algoritmos automatizados, muchas empresas aún dependen de la planificación manual de rutas, donde los gerentes logísticos utilizan su experiencia y conocimiento del terreno para diseñar rutas eficientes. Aunque este enfoque puede ser efectivo en ciertos casos, tiende a ser menos preciso y más propenso a errores humanos en comparación con los métodos computarizados (Maglogiannis, 2007).

La optimización de rutas tiene un impacto significativo en la eficiencia operativa de los centros de acopio, afectando directamente los costos y tiempos de entrega, así como los niveles de satisfacción de los consumidores y la competitividad que representa la empresa en el mercado nacional o extranjero.

Dentro de las ventajas o beneficios más significativos respecto a la optimización de rutas, se considera a la reducción de costos operativos como la más relevante. Al diseñar rutas más eficientes, las empresas pueden minimizar el consumo de combustible, reducir el desgaste de los vehículos y disminuir los costos asociados con el tiempo de conducción y mantenimiento (Chen et al., 2012). Esto se traduce en ahorros significativos para la empresa y una mejora en los márgenes de beneficio.

La optimización de rutas también mejora los tiempos de entrega de los productos. Al minimizar la distancia total recorrida y evitar rutas congestionadas o ineficientes, las empresas pueden cumplir con los plazos de entrega de manera más consistente y confiable (Jordan & Mitchell, 2015). Esto no solo aumenta la satisfacción del cliente al garantizar la puntualidad de

los pedidos, sino que también fortalece la reputación de la empresa en el mercado.

Otro beneficio clave es la capacidad de respuesta mejorada ante cambios inesperados en la demanda o en las condiciones del mercado. La optimización de rutas permite a las empresas adaptarse rápidamente a nuevas circunstancias, como cambios en la disponibilidad de recursos o variaciones en las condiciones del tráfico, sin comprometer la eficiencia o los costos operativos (Breiman, 2001).

Además de los beneficios económicos y operativos, la optimización de rutas también puede tener un impacto positivo en la sostenibilidad ambiental. Al reducir la distancia recorrida y minimizar las emisiones de carbono asociadas con el transporte, las empresas pueden contribuir a la reducción de su huella ambiental y cumplir con normativas ambientales más estrictas (Liaw & Wiener, 2001).

Machine Learning

Machine Learning, conocido también como aprendizaje automático, es un subcampo de las ciencias computacionales y una rama de la Inteligencia Artificial que permite a los sistemas aprender de los datos. En resumen, implica el procesamiento de grandes volúmenes de datos para desarrollar sistemas que aprenden de manera automática. Estos sistemas pueden clasificar, resolver y predecir situaciones o eventos futuros sin intervención humana directa. Esto se logra mediante la creación y el entrenamiento de algoritmos con extensas bases de datos, permitiéndoles identificar patrones y analizar los datos de manera efectiva (Alpaydin, 2020).

El pionero Arthur Samuel mencionaba que desde un punto de vista matemático que el Machine Learning es un conjunto de métodos dentro del campo de la inteligencia artificial que emplea técnicas estadísticas para descubrir patrones. Estos patrones son utilizados para desarrollar máquinas inteligentes que aprenden y toman decisiones basadas en datos empíricos recopilados de diversas fuentes (Wiederhold & McCarthy, 1992).

Para que una máquina pueda desarrollar un comportamiento de manera inteligente y establecerlo, esta debería de tener la capacidad de solución de problemas de la misma forma en que los humanos los resuelven, es decir, en base a la experiencia y el conocimiento. Por ende,

se implica que, dado a estos factores, podría ser capaz de auto modificar su comportamiento en base a la precisión exigida para obtener resultados precisos y que puedan ser comparados con los esperados.

En esta manera se puede encontrar a los tres más grandes grupos de algoritmos de Machine Learning:

- **Aprendizaje Supervisado:** En el aprendizaje supervisado, los algoritmos aprenden a partir de ejemplos etiquetados, es decir, conjuntos de datos donde cada ejemplo está asociado con una etiqueta o resultado deseado. El objetivo es aprender una función que mapee las entradas a las salidas correctas, permitiendo al modelo hacer predicciones precisas sobre datos nuevos. Ejemplos comunes incluyen la clasificación y la regresión (Bishop, 2006).
- **Aprendizaje No Supervisado:** En el aprendizaje no supervisado, los algoritmos intentan encontrar patrones o estructuras ocultas en datos no etiquetados. El objetivo principal es explorar la estructura intrínseca de los datos, como agrupar datos similares juntos (clustering) o reducir la dimensionalidad de los datos (PCA, por sus siglas en inglés). Este tipo de aprendizaje es útil cuando los datos no están etiquetados o cuando se busca descubrir nuevas relaciones entre variables (Hastie et al., 2009).
- **Aprendizaje Reforzado:** En el aprendizaje reforzado, los algoritmos aprenden a través de la interacción con un entorno dinámico, recibiendo retroalimentación en forma de recompensas o penalizaciones por las acciones realizadas. Su finalidad es aprender el comportamiento de una razón para que esta maximice la recompensa acumulada a lo largo del tiempo. Este tipo de aprendizaje es frecuentemente utilizado en problemas de control y toma de decisiones secuenciales (Fonseca-Reyna et al., 2018).

El Machine Learning ha transformado significativamente la industria logística, permitiendo mejoras sustanciales en la eficiencia operativa, la planificación de rutas y la gestión de inventarios.

El uso de Machine Learning en la optimización de rutas permite a las empresas mejorar la eficiencia del transporte al encontrar las rutas más rápidas y económicas. Los algoritmos

pueden analizar grandes volúmenes de datos históricos de tráfico, condiciones meteorológicas y horarios de entrega para predecir las mejores rutas en tiempo real. Esto no solo reduce los costos operativos y los tiempos de entrega, sino que también optimiza el uso de recursos como combustible y vehículos (Jordan & Mitchell, 2015).

Los modelos de Machine Learning pueden ayudar en la programación de transporte al prever la demanda futura y optimizar la asignación de recursos. Por ejemplo, pueden predecir la necesidad de capacidad adicional en determinadas rutas basándose en patrones históricos de demanda y eventos estacionales, lo que permite una planificación más precisa y eficiente (Chen et al., 2012).

Mediante el análisis de datos de sensores en tiempo real y registros de mantenimiento, los modelos de Machine Learning pueden predecir fallos en los vehículos y optimizar la programación de mantenimiento preventivo. Esto ayuda a reducir el tiempo de inactividad no planificado de la flota y garantiza que los vehículos estén en condiciones óptimas para la operación diaria (Liaw & Wiener, 2001).

El uso de Machine Learning en la logística proporciona una serie de beneficios significativos que mejoran tanto la eficiencia operativa como la competitividad de las empresas en el mercado global.

Al optimizar las rutas, la asignación de recursos y la gestión de inventarios, el Machine Learning ayuda a reducir los costos operativos totales. Esto incluye ahorros en combustible, mantenimiento de vehículos y mano de obra, así como una mejor utilización de los activos existentes (Breiman, 2001).

Los modelos de Machine Learning pueden hacer predicciones más precisas sobre la demanda, el comportamiento del consumidor y las condiciones del mercado. Esto ayuda a las empresas a tomar decisiones más informadas y a mejorar la precisión de sus operaciones logísticas, reduciendo errores y tiempos de respuesta (AMAZON, 2024).

La capacidad de adaptarse rápidamente a cambios en el entorno operativo es crucial en logística. Los algoritmos de Machine Learning permiten a las empresas ajustar dinámicamente sus operaciones y estrategias en respuesta a nuevas condiciones, minimizando los impactos

negativos y aprovechando oportunidades emergentes (Higginson & Bookbinder, 2005).

Mejorar la eficiencia logística a través del Machine Learning también tiene un impacto directo en la experiencia del cliente. Entregas más rápidas y precisas, así como una mayor visibilidad sobre el estado de los pedidos, contribuyen a aumentar la satisfacción del cliente y fortalecer las relaciones

comerciales (Rodriguez & Notteboom, 2015).

Modelos de Clasificación en Machine Learning

La tarea de clasificación supervisada es comúnmente realizada por sistemas inteligentes. Para ello, existen numerosos enfoques desarrollados tanto por la estadística, como la Regresión Logística y el Análisis Discriminante, así como por la inteligencia artificial, como las Redes Neuronales, la Inducción de Reglas, los Árboles de Decisión y las Redes Bayesianas. Estos paradigmas permiten llevar a cabo eficazmente las funciones necesarias para la clasificación (Sze et al., 2017).

En la práctica, los modelos de clasificación ayudan a resolver problemas como la detección de fraudes, la clasificación de imágenes, la predicción de riesgos crediticios y la segmentación de clientes. Al implementar estos modelos correctamente, las organizaciones pueden mejorar la precisión de sus decisiones y optimizar recursos significativamente. Existen varios modelos de clasificación que son aplicados a distintos fines, como lo son:

- **Regresión Logística:** La regresión logística es un modelo estadístico que se utiliza para la clasificación binaria, es decir, cuando la variable objetivo tiene dos posibles resultados. Este modelo calcula la probabilidad de que un evento pertenezca a una clase específica utilizando una función logística. Es ampliamente utilizado en problemas como la predicción de riesgos, el análisis de crédito y la detección de spam (Hastie et al., 2009).
- **Máquinas de Soporte Vectorial (SVM):** SVM es un algoritmo de aprendizaje supervisado que puede utilizarse tanto para problemas de clasificación como de regresión. El objetivo de SVM es encontrar un hiperplano en un espacio dimensional superior que mejor separe las clases de datos. Es eficaz en espacios de alta dimensión y

se utiliza en aplicaciones como el reconocimiento de escritura, la clasificación de imágenes y la detección de anomalías (Bishop, 2006).

- **Árboles de Decisión:** Los árboles de decisión son estructuras jerárquicas de decisiones que se representan en forma de árbol. Cada nodo interno representa una característica de los datos, cada rama representa una decisión basada en esa característica y cada nodo hoja representa la etiqueta de clasificación. Este algoritmo es útil para problemas donde se deben tomar múltiples decisiones secuenciales basadas en diferentes características, como la segmentación de mercado y el diagnóstico médico (Mitchell, 1997).
- **Random Forest (Bosques Aleatorios):** Random Forest es una técnica que combina múltiples árboles de decisión durante el entrenamiento y produce una predicción basada en la mayoría de los votos de los árboles individuales. Es útil para problemas de clasificación y regresión y proporciona una mayor precisión al reducir el sobreajuste inherente en los árboles de decisión individuales. Se aplica en áreas como la biometría, la detección de enfermedades y la evaluación de riesgos financieros (Breiman, 2001).
- **Redes Neuronales Artificiales:** Las redes neuronales artificiales son modelos computacionales inspirados en el funcionamiento del cerebro humano. Consisten en múltiples capas de neuronas interconectadas, cada una con pesos que se ajustan durante el proceso de entrenamiento para mejorar la precisión de las predicciones. Se utilizan en una amplia gama de aplicaciones, como reconocimiento de voz, reconocimiento de imágenes y procesamiento de lenguaje natural (Erhan et al., 2009).

Los árboles de clasificación incorporan un enfoque de clasificación supervisada, la idea surgió de la estructura de un árbol que se compone de una raíz, nodos (las posiciones donde las ramas se dividen), ramas y hojas; de manera similar, un árbol de clasificación se construye a partir de nodos que representan los círculos y las ramas son representadas por los segmentos que conectan los nodos. Un árbol de clasificación se inicia desde la raíz, se extiende hacia abajo y generalmente se dibuja de izquierda a derecha. El nodo inicial se llama nodo raíz, mientras los nodos en los extremos de la cadena se les conocen como nodos hoja. Dos o más ramas pueden extenderse desde cada nodo interno, es decir, desde un nodo que no es el nodo hoja (Breiman,

2001).

Los árboles de clasificación y regresión (CART) fueron desarrollados por Breiman, Freidman, Olshen y Stone en el libro *Classification and Regression Trees*, publicado en 1984. Para los árboles de clasificación, la bondad de una división se cuantifica por una medida de impureza; se dice que una división es pura si, después de la división, todas las instancias de la elección de una rama pertenecen a la misma clase. Para el nodo raíz, esto es N . N_m^i de N_m pertenecen a las clases C_i , donde $\sum_i N_m^i = N_m$. Dado que una instancia alcanza el nodo m , la estimación de la probabilidad de la clase C_i es:

$$I_m = \sum_{i=1}^K p_m^i \log p_m^i$$

El nodo m es puro si p_m^i para todo i son 0 o bien 1. Es 0 cuando ninguna de las instancias del nodo m son de Clase C_i , y es 1 si todos estos casos son de C_i . Si la división es pura, no es necesario dividir más y se puede añadir un nodo hoja etiquetado con la clase para la cual p_m^i es 1.

Random Forest es una combinación de árboles predictivos (clasificadores débiles); es decir, una modificación del Bagging, el cual trabaja con una colección de árboles incorrelacionados y los promedia (Hastie et al., 2009), en el cual se tiene que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque.

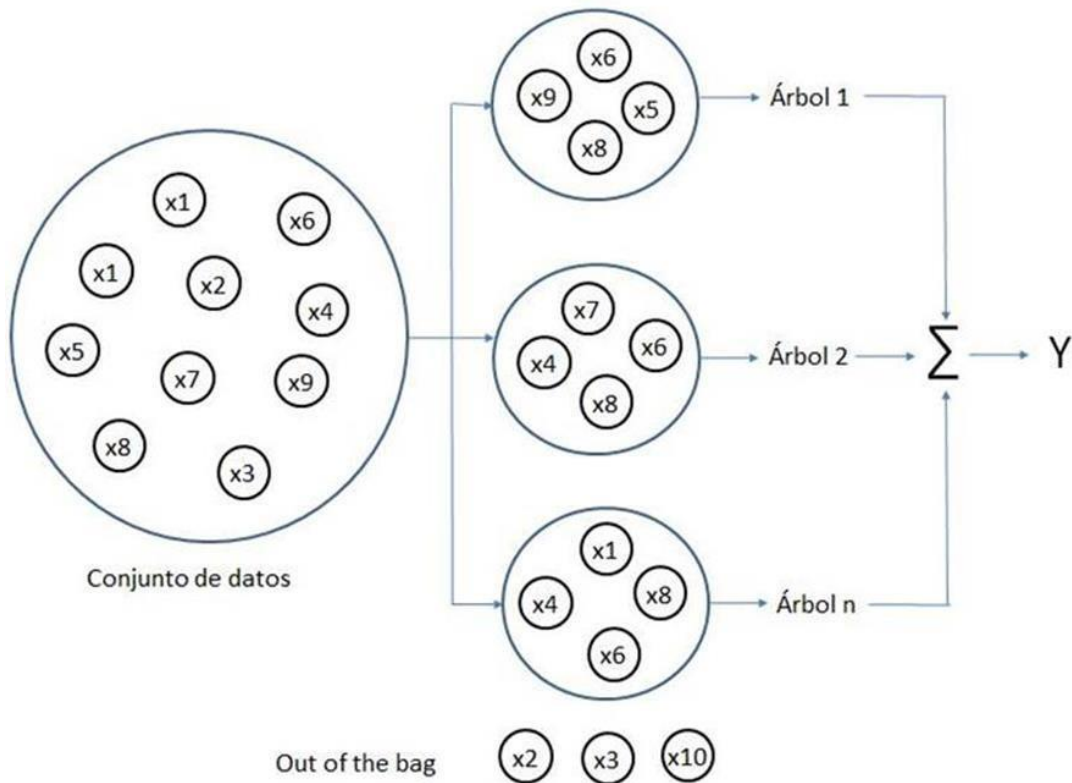
El algoritmo Random Forest es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado (Liaw & Wiener, 2001). Cada árbol se obtiene mediante un proceso de dos etapas:

- Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que $m < M$ (donde M = total de predictores).

- Cada árbol crece hasta su máxima extensión.

Figura 1

Algoritmo Random Forest



Nota: Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería, investigación y tecnología. (Espinosa-Zúñiga, 2020)).

Cada árbol generado por el algoritmo Random Forest contiene un grupo de observaciones aleatorias (elegidas mediante bootstrap, que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra). Las observaciones no estimadas en los árboles (también conocidas como “out of the bag”) se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final Y (conocida como ensamblado) que se obtiene mediante alguna regla (generalmente el promedio, cuando las salidas de los árboles del ensamblado son numéricas y, conteo de votos, cuando las salidas de los árboles del ensamblado son categóricas) (Espinosa-Zúñiga, 2020).

Las principales ventajas que presenta el algoritmo de Bosques Aleatorios son:

- Pueden usarse para clasificación o predicción: En el primer caso, cada árbol “vota” por una clase y el resultado del modelo es la clase con mayor número de “votos” en todos los árboles, de forma que cada nueva observación se presenta a cada uno de los árboles y se asigna a la clase más “votada”. En el segundo caso, el resultado del modelo es el promedio de las salidas de todos los árboles.
- El modelo es más simple de entrenar en comparación con técnicas más complejas, pero con un rendimiento similar.
- Tiene un desempeño muy eficiente y es una de las técnicas más certeras en bases de datos grandes.
- Puede manejar cientos de predictores sin excluir ninguno y logra estimar cuáles son los predictores más importantes, es por ello por lo que esta técnica también se utiliza para reducción de dimensionalidad.
- Mantiene su precisión con proporciones grandes de datos perdidos.

El método Random Forest se basa en un conjunto de árboles de decisión, es decir, una muestra entra al árbol y es sometida a una serie de prueba binarios en cada nodo, llamados split, hasta llegar a una hoja en la que se encuentra la respuesta. Esta técnica puede ser utilizada para dividir un problema complejo en un conjunto de problemas simples (Rigatti, 2017).

En la etapa de entrenamiento, el algoritmo intenta optimizar los parámetros de las funciones de split a partir de las muestras de entrenamiento.

$$\theta_k^* = \operatorname{argmax}_{\theta_j \in \tau_j} I_j$$

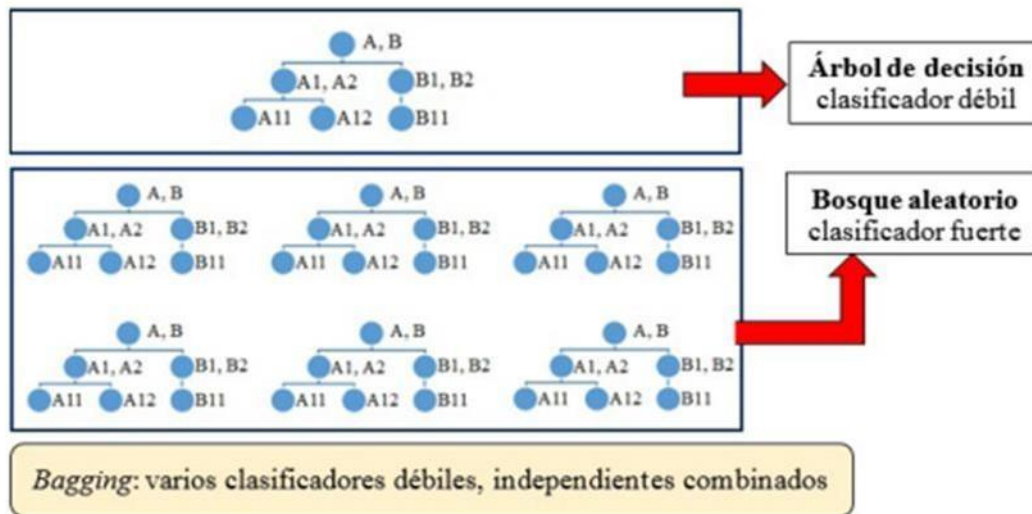
Para ello se utiliza la siguiente función de ganancia de información:

$$I_j = H(J) - \sum_{i \in 1,2} \frac{|S_j^i|}{|S_j|} H(S_j^i)$$

Donde S representa el conjunto de muestras que hay en el nodo por dividir, y Si son los dos conjuntos que se crean de la escisión. La función mide la entropía del conjunto, y depende del tipo de problema que abordamos(Breiman, 2001).

Figura 2

Árbol de decisión vs Bosque aleatorio



Fuente: Merino, R. F. M., & Chacón, C. I. Ñ. (2017). *Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python*

Marco Conceptual

Centro de Acopio

El Centro de acopio primario contiene espacios de acaparamiento, zonas de encuentro, zonas culturales y gastronómicas, que promueven emprendimientos y alternativas económicas en una ciudad. El diseño se basa en la sustentabilidad, utilizando y optimizando los materiales y recursos locales, vinculados al contexto natural en donde se encuentra ubicado, las áreas parten del análisis del volumen de la producción agrícola que tiene la parroquia, además se incorporarán áreas complementarias de las actividades y necesidades de las comunidades (Ayala Villareal & Nazate Salazar, 2020).

Entre los factores que han limitado el crecimiento de las exportaciones del sector uno de ellos es la falta de centros de acopio y verdaderas redes de valor que permitan acceder a la comercialización a los centros de alto consumo en el país, coadyuvando a mejorar los precios actuales en el mercado y al fortalecimiento e impulso de la actividad ganadera en la región, de ahí, yace su importancia (Fideicomiso de Riesgo, 2017).

Optimización de rutas

La optimización de rutas es un concepto relacionado al transporte de productos, se refiere a la elección del mejor camino reduciendo costos y es mucho más complejo que simplemente encontrar la distancia más corta entre dos puntos («¿Qué es la optimización de rutas y por qué es importante?», 2022).

Cuando se trata de exportaciones e importaciones, hay un factor crucial que impacta el valor de las transacciones: los costos logísticos. Por tanto, una de las principales responsabilidades de los gestores es optimizar las rutas. Sólo en la región latinoamericana, el 47% del PIB (producto interno bruto) se genera en las relaciones comerciales. Si se optimizan las rutas de entrega, el costo del flete disminuye y las ganancias aumentan tanto para quienes exportan como para quienes importan. De esta manera, las organizaciones que confían en el software de optimización de rutas suman una ventaja competitiva al reducir costos y aumentar la eficiencia (Grigalunas et al., 2007).

La optimización de rutas es una herramienta que utiliza algoritmos computacionales para encontrar la mejor ruta a un conjunto de destinos, teniendo en cuenta factores como la distancia, el tiempo de viaje, los costos y las restricciones específicas. Esta solución, utilizada frecuentemente por empresas de logística, transporte y servicios de entrega, maximiza la eficiencia de las operaciones y reduce los costos. Entre las principales características del software de optimización de rutas se encuentra la integración con sistemas de seguimiento de vehículos y comunicación en tiempo real (Grigalunas et al., 2007).

De esta forma, se garantiza la flexibilidad y agilidad en el proceso de entrega mediante el seguimiento y actualización de las rutas según sea necesario. En términos económicos, el software de optimización de rutas puede reducir los costos operativos al minimizar el tiempo de viaje, el consumo de combustible y los costos de mantenimiento de la flota. Además, al optimizar las rutas, las empresas aumentan la eficiencia de las entregas, mejoran la satisfacción del cliente e incluso amplían sus operaciones sin necesidad de inversiones adicionales en la flota (Grigalunas et al., 2007).

Del punto de vista ecológico, la optimización de las rutas contribuye a reducir las emisiones de gases de efecto invernadero -que alcanzarán un récord mundial en 2023, según un estudio noruego del Centro Internacional de Investigación del Clima- y la huella de carbono de las operaciones de transporte. Esto se debe a que, al reducir la distancia recorrida por los vehículos y evitar la congestión, el software de optimización de rutas ayuda a minimizar el impacto ambiental de las actividades de transporte, lo que promueve prácticas más sostenibles y alineadas con las demandas de la sociedad actual (Grigalunas et al., 2007).

Sin embargo, la importancia del software de optimización de rutas va más allá de los beneficios económicos y ambientales, ya que también influye directamente en la competitividad y adaptabilidad de las empresas en el mercado. Con la creciente complejidad de las cadenas de suministro y la demanda de entregas rápidas y eficientes, tener una solución que permita una planificación y ejecución de rutas optimizadas es esencial para seguir siendo competitivo y cumplir con las expectativas de los clientes (Grigalunas et al., 2007).

Machine Learning

El machine learning es la ciencia de desarrollo de algoritmos y modelos estadísticos que utilizan los sistemas de computación con el fin de llevar a cabo tareas sin instrucciones explícitas, en vez de basarse en patrones e inferencias. Los sistemas de computación utilizan algoritmos de machine learning para procesar grandes cantidades de datos históricos e identificar patrones de datos. Esto les permite generar resultados con mayor precisión a partir de un conjunto de datos de entrada (AMAZON, 2024).

Este permite que las empresas impulsen el crecimiento, generen nuevas fuentes de ingresos y resuelvan problemas complejos. Los datos son la fuerza que impulsa la toma de decisiones empresariales. Estos suelen tener diversos orígenes, como los comentarios de los clientes, los empleados y las finanzas. La investigación dedicada al machine learning automatiza y optimiza este proceso. Las empresas pueden obtener resultados más rápido con programas que analizan grandes volúmenes de datos a gran velocidad.

La idea central del machine learning es la existencia de una relación matemática entre cualquier combinación de datos de entrada y salida. El modelo de machine learning no conoce de antemano esta relación, pero puede adivinarla si se le dan suficientes conjuntos de datos. Esto significa que cada algoritmo de machine learning se crea en torno a una función matemática modificable (AMAZON, 2024).

Algoritmos de Machine Learning

Aprendizaje supervisado

Los científicos de datos suministran algoritmos con datos de entrenamiento etiquetados y definidos para evaluar las correlaciones. Los datos de muestra especifican tanto la entrada como la salida del algoritmo. Por ejemplo, las imágenes de cifras manuscritas están anotadas para indicar a qué número corresponden. Un sistema de aprendizaje supervisado puede reconocer los clústeres de píxeles y formas asociadas a cada número, si se dan suficientes ejemplos. Con el tiempo, reconocerá números escritos a mano y distinguirá de forma fiable entre los números 9 y 4 o 6 y 8 (AMAZON, 2024).

Los ideales del aprendizaje supervisado son la simplicidad y facilidad de diseño. Es útil para predecir un posible conjunto limitado de resultados, dividir los datos en categorías o combinar los resultados de otros dos algoritmos de machine learning.

Sin embargo, es un reto etiquetar millones de conjuntos de datos sin etiquetar. El etiquetado de datos es el proceso de categorizar los datos de entrada con sus correspondientes valores de salida definidos. Los datos de entrenamiento etiquetados son necesarios para el aprendizaje supervisado. Por ejemplo, habría que etiquetar millones de imágenes de manzanas y plátanos con las palabras “manzana” o “plátano”. Las aplicaciones de machine learning podrían utilizar estos datos de entrenamiento para adivinar el nombre de la fruta cuando se les dé una imagen de esta (AMAZON, 2024).

Aprendizaje No Supervisado

Los algoritmos de aprendizaje no supervisado se entrenan con datos no etiquetados. Analizan los nuevos datos con la intención de establecer conexiones significativas entre las entradas y salidas predeterminadas. Pueden detectar patrones y categorizar los datos. Los algoritmos no supervisados pueden agrupar artículos de noticias de diferentes sitios en categorías comunes como deportes, crimen, entre otros (AMAZON, 2024).

Pueden utilizar el procesamiento de lenguaje natural para comprender el significado y la emoción del artículo. En el sector minorista, el aprendizaje no supervisado puede encontrar patrones en las compras de los clientes y proporcionar resultados de análisis de datos como: es más probable que el cliente compre pan si también compra mantequilla.

El aprendizaje no supervisado es útil para el reconocimiento de patrones, la detección de anomalías y la agrupación automática de datos en categorías. Como los datos de entrenamiento no necesitan etiquetado, la configuración es fácil. Estos algoritmos también se pueden utilizar para automáticamente limpiar y procesar datos con vistas a su posterior modelado. La limitación de este método es que no puede ofrecer predicciones precisas. Además, no puede señalar de forma independiente tipos específicos de resultados de datos (AMAZON, 2024).

Aprendizaje semisupervisado

Este método combina el aprendizaje supervisado y el no supervisado. Para entrenar los sistemas, esta técnica se basa en el uso de una pequeña cantidad de datos etiquetados y de una gran cantidad de datos sin etiquetar. En primer lugar, los datos etiquetados se utilizan para entrenar parcialmente el algoritmo de machine learning (AMAZON, 2024).

Después, el propio algoritmo entrenado parcialmente etiqueta los datos no etiquetados. Este proceso se denomina pseudo etiquetado. A continuación, el modelo se vuelve a entrenar con la mezcla de datos resultante sin programarlo explícitamente.

El ideal de este método es que no necesita grandes cantidades de datos etiquetados. Resulta útil cuando se trabaja con datos como documentos largos que los humanos tardarían mucho en leer y etiquetar (AMAZON, 2024).

Aprendizaje por refuerzo

El aprendizaje por refuerzo es un método con valores de recompensa adjuntos a los diferentes pasos que debe dar el algoritmo. Así, el objetivo del modelo es acumular tantos puntos de recompensa como sea posible y alcanzar una meta final (AMAZON, 2024).

En la última década, la mayor parte de la aplicación práctica del aprendizaje por refuerzo se produjo en el ámbito de los videojuegos. Los algoritmos de aprendizaje por refuerzo más avanzados obtuvieron impresionantes resultados en videojuegos clásicos y modernos, a menudo superando de manera significativa a sus homólogos humanos (AMAZON, 2024).

Aunque este método funciona mejor en entornos de datos inciertos y complejos, rara vez se aplica en contextos empresariales. No es eficiente para tareas bien definidas y el sesgo del desarrollador puede afectar los resultados. El científico de datos puede influir en los resultados ya que diseña las recompensas (AMAZON, 2024).

Tipos de algoritmos de Machine Learning

Algoritmo Regresión Lineal

La regresión lineal es un algoritmo de aprendizaje supervisado que se utiliza para predecir y pronosticar valores dentro de un rango continuo, como cifras de ventas o precios. Procedente de la estadística, la regresión lineal asigna una pendiente constante utilizando un valor de entrada con una variable de salida para predecir un valor numérico o una cantidad. La regresión lineal usa datos etiquetados para hacer predicciones estableciendo una línea de mejor ajuste (Coursera, 2023).

Algoritmo Regresión Logística

La regresión logística es un algoritmo de aprendizaje supervisado utilizado para la clasificación binaria, como decidir si una imagen encaja en una clase u otra. Originaria de la estadística, la regresión logística predice técnicamente la probabilidad de que una entrada pueda clasificarse en una única clase primaria. En la práctica, sin embargo, puede emplearse para agrupar las salidas en una de dos categorías: “clase primaria” o “clase secundaria” Esto se consigue creando un rango para la clasificación binaria, de forma que cualquier salida entre 0 y 0,49 se incluya en un grupo y cualquier salida entre 0,50 y 1,00 se incluya en otro. Como resultado, la regresión logística en el aprendizaje automático se utiliza normalmente para la categorización binaria en lugar de para el modelado predictivo (Coursera, 2023).

Algoritmo Bayesiano

El clasificador bayesiano ingenuo es un conjunto de algoritmos de aprendizaje supervisado que se utilizan para crear modelos predictivos de categorización binaria o múltiple. Basado en el Teorema de Bayes, Naive Bayes opera con probabilidades condicionales, que son independientes entre sí, pero indican la probabilidad de una clasificación basada en sus factores combinados (Coursera, 2023).

Algoritmo de Árbol de Decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado utilizado para la clasificación y el modelado predictivo. Semejante a un diagrama de flujo gráfico, un árbol de

decisión comienza con un nodo raíz, que formula una pregunta concreta a los datos y luego los envía por una rama en función de la respuesta. Cada una de estas ramas conduce a un nodo interno, que a su vez formula otra pregunta a los datos antes de dirigirlos hacia otra rama en función de la respuesta. Esto continúa hasta que los datos llegan a un nodo final, también llamado nodo hoja, que no se ramifica más. Los árboles de decisión son habituales en el aprendizaje automático porque pueden manejar conjuntos de datos complejos con relativa sencillez (Coursera, 2023).

Algoritmo de Bosque Aleatorio

Un algoritmo de bosque aleatorio utiliza un conjunto de árboles de decisión para la clasificación y el modelado predictivo. En un bosque aleatorio, muchos árboles de decisión (a veces cientos o incluso miles) se entrenan utilizando una muestra aleatoria del conjunto de entrenamiento (un método conocido como bagging). Después, los investigadores introducen los mismos datos en cada árbol de decisión del bosque aleatorio y cuentan sus resultados finales. Luego se selecciona el resultado más común como el más probable para el conjunto de datos (Coursera, 2023).

Aunque pueden llegar a ser complejos y requerir mucho tiempo, los bosques aleatorios corrigen el problema común del “sobreajuste” que puede producirse con los árboles de decisión. Se habla de sobreajuste cuando un algoritmo se ajusta demasiado a su conjunto de datos de entrenamiento, lo que puede repercutir negativamente en su precisión cuando se introduce posteriormente en nuevos datos (Coursera, 2023).

Algoritmo K-Nearest Neighbor

Un algoritmo K-Nearest neighbor es un algoritmo de aprendizaje supervisado que se usa para la clasificación y el modelado predictivo. Fieles a su nombre, los algoritmos KNN clasifican una salida por su proximidad a otras salidas en un gráfico. Si una salida está más cerca de un grupo de puntos azules en un gráfico que de un grupo de puntos rojos, se clasificaría como miembro del grupo azul. Este enfoque significa que los algoritmos KNN pueden utilizarse tanto para clasificar resultados conocidos como para predecir el valor de resultados desconocidos (Coursera, 2023).

Algoritmo K-Means

K means es un algoritmo no supervisado que se emplea para la clasificación y el modelado predictivo. Al igual que KNN, K means utiliza la proximidad de un resultado a un conglomerado de puntos de datos para identificarlo. Cada uno de los conglomerados está definido por un centroide, un punto central real o imaginario del conglomerado. K means es útil en grandes conjuntos de datos, especialmente para la agrupación, aunque puede fallar cuando maneja valores atípicos (Coursera, 2023).

Algoritmo de Redes Neuronales

Una red neuronal artificial (RNA) comprende unidades dispuestas en una serie de capas, cada una de las cuales se conecta a las capas anexas. Las RNA se inspiran en los sistemas biológicos, como el cerebro, y en cómo procesan la información. Por lo tanto, son esencialmente un gran número de elementos de procesamiento interconectados, que trabajan al unísono para resolver problemas específicos. También aprenden con el ejemplo y la experiencia, y son extremadamente útiles para modelar relaciones no lineales en datos de alta dimensión, o donde la relación entre las variables de entrada es difícil de entender (APD, 2024).

Inteligencia Artificial

La inteligencia artificial, o IA, es tecnología que permite que las computadoras simulen la inteligencia y las capacidades humanas de resolución de problemas. "Es la ciencia e ingeniería de hacer máquinas inteligentes, especialmente programas informáticos inteligentes. Se relaciona con la tarea similar de usar equipos para comprender la inteligencia humana, pero la IA no tiene que ajustarse a los métodos biológicos observables" (IBM, 2023).

La inteligencia artificial (IA) se ha convertido en un término general para referirse a aplicaciones que realizan tareas complejas para las que antes era necesaria la intervención humana, como la comunicación en línea con los clientes o jugar al ajedrez. El término a menudo se usa indistintamente junto con los nombres de sus subcampos, el aprendizaje automático y el aprendizaje profundo. Es importante tener en cuenta que, aunque todo machine learning es IA, no toda la IA es machine learning (Oracle, 2024).

La inteligencia artificial hace referencia a sistemas informáticos que buscan imitar la función cognitiva humana a través de máquinas, procesadores y softwares con el objetivo de realizar tareas de procesamiento y análisis de datos. Se trata de máquinas diseñadas para razonar, aprender, realizar acciones y resolver problemas. La IA integra un diseño de programación que es capaz de almacenar información sobre determinada área para convertirla en conocimiento e implementarla en el día a día de la actividad humana (Ferrovia, 2024).

Algunas maneras de cómo se aplica la inteligencia artificial en diferentes sectores:

1. **Personal:** Asistencia a través de smartphones, tabletas y ordenadores.
2. **Informático:** Garantías de ciberseguridad.
3. **Productivo:** Ensamblaje y automatización en fábricas y laboratorios
4. **Financiero:** Permite detectar posibles fraudes (como el blanqueo de capitales), predecir el comportamiento de los mercados y aconsejar las operaciones y productos idóneos para cada cliente.
5. **Climático:** Reducción de la deforestación y el consumo energético.
6. **Sanitario:** Identificación de factores genéticos que anticipen la detección de enfermedades.
7. **De transporte y sector energético:** Fabricación de vehículos autónomos e inteligentes, ayuda no solo a optimizar las rutas tanto en tiempo como en consumo energético; también permite reducir los accidentes en carretera, anticiparse a posibles problemas al predecir la necesidad de mantenimiento del vehículo con antelación, y planificar las rutas de transporte según la demanda y la capacidad, entre otras ventajas. Además, ya es una parte fundamental de los vehículos eléctricos, de modo que permite gestionar y transmitir los datos entre distintos dispositivos conectados.
8. **Agrícola:** Anticipación de impacto ambiental y mejora del rendimiento agrícola.

9. **Comercial:** Pronóstico de ventas.
10. **Educativo:** Capaz de realizar propuestas personalizadas de cursos, mejorar las tutorías en línea y analizar las competencias de los estudiantes mediante el método learning analytics a fin de conocer cuáles son sus necesidades educativas.

Tipos de inteligencia Artificial

Machine Learning (aprendizaje automático)

Es la capacidad que tiene una inteligencia artificial para aprender por sí misma. Se basa en un ciclo de aprendizaje a partir de datos, entrenamiento y resultados. Existen varios subtipos en función de si su aprendizaje requiere la supervisión de un ser humano o se permite que la IA aprenda de forma autónoma, según unas reglas establecidas. Se suele utilizar en asistentes virtuales y chatbots, entre otros (Repsol, 2023).

Deep Learning (aprendizaje profundo)

Su objetivo es recrear la forma en la que aprenden los humanos a través de lo que se denominan redes neuronales, que consisten en nodos interconectados que emulan la red de neuronas de un cerebro humano. Se emplea, por ejemplo, en la búsqueda de productos basada en imágenes (Repsol, 2023).

Reinforcement Learning (aprendizaje por refuerzo)

Se inspira en la psicología conductista y su objetivo es permitir a la IA diseñar estrategias de manera automática. Es muy práctico para el mantenimiento predictivo o para personalizar las experiencias de los clientes (Repsol, 2023).

Generative Adversarial Networks (redes generativas antagónicas)

Son un tipo de algoritmos que se implementan por un sistema de dos redes neuronales. Estas dos redes compiten mutuamente. Sirve para generar objetos y experiencias a partir de muestras (por ejemplo, fotografías). (Repsol, 2023)

Natural Language Processing (procesamiento del lenguaje natural)

Investiga la manera en que las máquinas se comunican con las personas, con el objetivo de lograr que aquellas comprendan y extraigan la información relevante. Sus aplicaciones son múltiples, desde el análisis de sentimiento u opinión hasta la anonimización de documentos, pasando por el entrenamiento de chatbots (Repsol, 2023).

Computer Vision (visión artificial)

Enseña a los ordenadores a «ver» e interpretar el contenido de las imágenes digitales, a fin de que puedan producir información simbólica que se pueda interpretar. Se usa para el reconocimiento de objetos, la restauración de imágenes o la reconstrucción de escenas (Repsol, 2023).

Speech Recognition (reconocimiento de habla)

Su fin es hacer posible que los humanos puedan comunicarse con los ordenadores y viceversa, y es especialmente útil para los sistemas de navegación de vehículos controlados por voz, las aplicaciones de dictado o los sistemas para personas con discapacidad (Repsol, 2023).

Knowledge Graph (grafo de conocimiento)

El grafo es una manera de representar relaciones entre entidades y crear vínculos entre datos y metadatos. Cuando el contenido de los grafos se enriquece y se logra que realicen un procesamiento automático «inteligente» de los datos, se convierten en grafos de conocimiento. Son muy populares en sistemas de organización de la información (Repsol, 2023).

Augmented Reality (realidad aumentada)

Se trata de un conjunto de tecnologías que permiten que el usuario interactúe con el mundo real mediante dispositivos que añaden información gráfica virtual, de modo que el usuario ve al mismo tiempo el mundo que le rodea, pero con objetos virtuales superpuestos. Se utiliza en un amplísimo número de aplicaciones, desde operaciones hasta pruebas virtuales de colores de maquillaje o recreaciones de cómo quedará un mueble determinado en tu hogar (Repsol, 2023).

Marco Legal

Para hacer hincapié en el ámbito legal, podemos percibir que se han desarrollado diversos planes que se encuentran alineados a una normativa legal entre los que más resaltan cómo: El Plan Nacional de Telecomunicaciones y Tecnología de Información del Ecuador 2016-2021.

De esta misma manera, el Convenio Marco de Cooperación Interinstitucional se encuentra firmado, sellado y avalado con las diferentes instituciones como el Ministerio de Telecomunicaciones y de La Sociedad de la Información, La Secretaría de Educación Superior, Ciencia, Tecnología e Innovación.

Asimismo, se hace referencia a Los Planes Nacionales para el Desarrollo (2013-2017), junto con la Ley Orgánica de Telecomunicaciones. También, El Código Orgánico de La Economía Social de los Conocimientos, Creatividad e Innovación.

Adicionalmente, como los artículos que contemplan el desarrollo de la investigación científica junto con la innovación en la tecnología, de Las Tecnologías de la Información y las Comunicaciones (TIC) y distintas áreas del conocimiento que se encuentran en la Constitución de la República del Ecuador de los cuales se puede acentuar los (Art. 281, Art. 385, Art. 385, Art. 387, Art. 423, Art. 388, Art. 313) que dan la posibilidad de alinearse y ser parte de la adopción de estas tecnologías que están emergiendo para el cambio y la acogida de una cultura transformadora digital para un mejor desarrollo a nivel empresarial, en el ámbito nacional con proyección hacia lo internacional, lo que lo vuelve en un impacto positivo y cambiador.

Cabe resaltar que, la Constitución de la República del Ecuador (2008) en su Art. 313 menciona que “El Estado se reserva el derecho de administrar, regular, controlar y gestionar los sectores estratégicos, de conformidad con los principios de sostenibilidad ambiental, precaución, prevención y eficiencia” (pág. 133). Finalmente, la Constitución de la República del Ecuador (2008) adiciona que en el Art. 321 de la Carta Magna, “El Estado reconoce y garantiza el derecho a la propiedad en sus formas públicas, privadas, comunitarias, estatales, asociativas, cooperativas, mixta, y que deberá cumplir su función social y ambiental” (pág. 149).

Capítulo 3: Metodología

El actual trabajo es de corte transaccional correlacional – causal, debido a que se recolectarán los datos en un período específico de tiempo. La técnica de recolección a emplear para la investigación proviene de la data proporcionada por el Centro de Acopio de Machala en relación con el Puerto Aduanero de Machala, encargado de la recepción y distribución de las cajas de banano por diversos proveedores locales, empresas y fincas.

Los datos extraídos yacen bajo el período 2023, la cual cuenta con 5001 observaciones en total dentro de la base, con la finalidad de desarrollar un modelo de Bosques Aleatorios para la clasificación de rutas de forma óptima

Figura 3

Número de Observaciones

	RUTA	TARIFA	VI.TICOS	GASOLINA	VIAJE	TIEMPO..min.	DISTANCIA..km.	SALIDA	PRODUCTO
1	2	23.63	2.25	3.38	18	10	10	CENTRO DE ACOPIO MACHALA	FRESH BANAN
2	4	36.35	4.10	5.25	27	17	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
3	4	36.35	4.10	5.25	27	15	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
4	4	36.35	4.10	5.25	27	14	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
5	4	36.35	4.10	5.25	27	13	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
6	4	30.25	2.00	6.25	22	16	16	CENTRO DE ACOPIO MACHALA	FRESH BANAN
7	4	36.35	4.10	5.25	27	15	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
8	4	32.28	3.60	4.68	24	15	14	CENTRO DE ACOPIO MACHALA	FRESH BANAN
9	4	32.28	3.60	4.68	24	14	14	CENTRO DE ACOPIO MACHALA	FRESH BANAN
10	4	36.35	4.10	5.25	27	15	12	CENTRO DE ACOPIO MACHALA	FRESH BANAN
11	4	30.25	2.00	6.25	22	19	16	CENTRO DE ACOPIO MACHALA	FRESH BANAN
12	4	30.25	2.00	6.25	22	18	16	CENTRO DE ACOPIO MACHALA	FRESH BANAN
13	4	30.25	2.00	6.25	22	19	16	CENTRO DE ACOPIO MACHALA	FRESH BANAN
14	4	32.28	3.60	4.68	24	14	14	CENTRO DE ACOPIO MACHALA	FRESH BANAN
15	1	23.63	2.25	3.38	18	11	10	CENTRO DE ACOPIO MACHALA	FRESH BANAN
16	1	23.63	2.25	3.38	18	10	10	CENTRO DE ACOPIO MACHALA	FRESH BANAN
17	1	23.63	2.25	3.38	18	11	10	CENTRO DE ACOPIO MACHALA	FRESH BANAN

Showing 1 to 17 of 5,001 entries, 12 total columns

Nota: Elaboración propia de los autores en RStudio

La investigación es de corte transaccional correlacional – causal debido a que los datos han

sido recolectados durante un período de tiempo mensual en el cual surgieron análisis en relación con las variables interrelacionadas. De esta manera, el trabajo se categoriza como exploratorio por la búsqueda de la clasificación de rutas para la selección de la más conveniente entre estas.

La aplicación de bosques aleatorios para la clasificación es utilizada para segmentar las diferentes variables que se desean seleccionar y que serán de utilidad para conocer con mayor precisión los datos mediante la prueba y los predichos. Se determinará las variables que conllevan mayor relación con las rutas y permitirán optimización.

Metodología en el programa R studio.

En primera estancia se procede a convertir la base de datos de Excel a un formato csv para ser utilizado dentro del programa. Dicha base que ha sido construida mediante los datos proporcionados por el Centro de Acopio de Machala cuyo nombre determinado será *RutasT*.

La base de datos será cargada con la secuencia siguiente:

```
Rutas a1 <- read.csv2("../Data/rutasa1.csv")
```

Dado que el trabajo de investigación es de bosques aleatorios para la clasificación se requiere eliminar aquellas columnas que contienen datos categóricos o cualitativos con menos relevancia para la construcción del modelo.

Eliminamos con el siguiente comando:

```
RutasT <- Rutasa1[,c(-8, -9)]
```

Instalación y carga de paquetes

```
install.packages = ("name of package")
```

Los paquetes que abarcan colecciones de funciones para realizar tareas de ciencias de datos tales como: visualización, limpieza, programación y manipulación.

A continuación, los paquetes detallados por sus funciones dentro del programa:

- ❖ **Ggplot2:** Create Elegant Data Visualisations Using the Grammar of Graphics. Es un sistema que permite la visualización de datos.
- ❖ **Modeest:** Mode Estimation. Es un sistema que permite el uso de las medidas de tendencia central.
- ❖ **Moments:** Moments, Cumulants, Skewness, Kurtosis and related Tests. Es un sistema que permite calcular la distribución de datos, sus características y su dispersión.
- ❖ **Grid:** The Grid Graphics Package. Es un sistema que permite la reescritura de las capacidades de diseño de gráficos.
- ❖ **GridExtra:** Miscellaneous Functions for “Grid” Graphics. Es un sistema que permite la reescritura de diseños gráficos de forma múltiple.
- ❖ **Caret:** Classification and Regression Training. Es un sistema en donde existen funciones que facilitan la utilización de métodos complejos de clasificación, regresión y principalmente para la partición de datos.
- ❖ **Rpart:** Recursive Partitioning and Regression Trees. Es un sistema que permite la construcción de árboles de decisión y árboles de regresión por aprendizaje supervisado.
- ❖ **Rpart.plot:** Plot “rpart” Models: An Enhanced Version of “plot.rpart”. Es un Sistema que permite la mejora en la visualización de los árboles de decisión o regresión creados.
- ❖ **RandomForest:** Breiman and Cutler’s Random Forests for Classification and Regression. Es un sistema que permite la creación de bosques aleatorios a partir de un conjunto de árboles de decisión entrenados.

Por consiguiente, se procede a cargar las paqueterías ya instaladas.

A continuación, el comando que permite la actualización y llamado de asistencia de los

paquetes instalados para su uso:

```
library("ggplot2, modeest, moments, grid, gridextra, caret,  
rpart, rpart.plot, randomforest")
```

Metodología Árbol de Decisión.

El Árbol de Decisión permite conocer las diversas decisiones por la asignación de etiquetas a una observación dada basado en los valores de sus características especificados.

Para comenzar se debe plantar o sembrar el código denominado cómo semilla, existen diversos tipos de semillas, sin embargo, dentro del modelos usaremos el de nuestra preferencia. Esta será la encargada de la reproducibilidad de los resultados generando entrenamiento de modelos.

A continuación, el comando encargado de la réplica del modelo en forma aleatoria:

```
set.seed(123)
```

Se procede a realizar el entrenamiento de datos dónde se utiliza un conjunto de datos para construir un modelo predictivo de una categoría o valor de una variable específica a conocer.

A continuación, el código de entrenamiento:
Train = createDataPartition(DATA\$Variable Objetiva, p = 0.8, list = FALSE)

- **Train:** Es un vector que contendrá los datos clasificados del entrenamiento.
- **createDataPartition:** Es la función que permite la división de los datos de forma aleatoria en conjuntos u subconjuntos en relación con la variable objetivo.
- **Data:** Es el vector que contiene a la variable objetivo dentro de tu dataframe.
- **Variable Objetiva:** Es la variable que se predecirá en el modelo.
- **p=0.8:** Es el encargado de indicar la cantidad de datos que se desean utilizar para entrenar al modelo. En este caso es de 0.8 u 80%.

- **list = FALSE:** Esta función de lista en falso permite indicar que los resultados no deberán ser entregados en forma de lista sino como filas para una mejor manipulación y visualización de los datos.

Comienza la creación del árbol de clasificación dónde se registran las variables dependientes e independientes del dataframe.

A continuación, el código para la construcción del árbol:

```
Arbol = rpart(Variable Objetivo~., data = Clientes[Train,], method
              = class, control = rpart.control(minsplit = 300, cp = 0.01))
```

- **Arbol:** Es el variable que almacenará el modelo del árbol de decisión ya entrenado.
- **rpart:** Es el comando encargado de construir el modelo de árbol de decisión
- **Variable Objetivo:** Es la variable a la cuál se le va a predecir en el modelo y que dentro de este se presentará cómo una categoría.
- **~. :** Son los símbolos usados para indicar que la variable objetiva es a la cuál se le va a aplicar el modelo en función a las demás variables independientes predictoras dentro del dataframe.
- **data = DATA[Train,]:** Esta función específica de cuál conjunto de datos se obtendrá la información, en este caso se ha seleccionado el dataframe de DATA pero de los datos entrenados Train.
- **method = "class":** Esta función indica que el modelo deberá usar un método para clasificar una categoría o clase en el árbol.
- **control = rpart.control :** Esta función permite controlar los parámetros a utilizar por el modelo dentro de su construcción.
- **minsplit = 300:** Esta función permite la definición de la cantidad de observaciones que se utilizarán por categoría y para sus subdivisiones y el ajuste del modelo
- **cp=0.01:** Esta función se denomina como “complexity parameter” y controla las ramas del modelo de árbol para su mejor precisión; la cantidad de ajuste de precisión del modelo es de 0.01 o 1% hacia los datos que no tengan relación con las demás ramas.

A continuación, la variable Árbol que será llamada para la ejecución y visualización de las

coordenadas del modelo:

Arbol

Para graficar el árbol, se usarán los siguientes códigos:

```
rpart.plot(arbol, type = 1, digits = -1,  
extra = 0, cex = 0.7, nn = TRUE,  
fallen.leaves = TRUE)
```

- **rpart.plot:** Es la función encargada de mostrar una visualización de los árboles entrenados de forma detallada.
- **Arbol:** Es el modelo del árbol que fue creado y entrenado.
- **type = 1:** Esta función controla la cantidad de etiquetas y cómo se mostrarán dentro del árbol.
- **digits = -1:** Esta función permite determinar la cantidad de dígitos sin decimales que se mostraran en los nodos del modelo.
- **extra = 0:** Esta función permite indicar si se desea mostrar información, valores o categorías extras dentro del modelo.
- **cex = 0.7:** Esta función es la encargada de ajustar los textos de los gráficos por sus nodos.
- **nn= TRUE:** Esta función indica si se quiere mostrar o no lo números de los nodos establecidos.
- **fallen.leaves = TRUE:** Esta función permite indicar si se requiere o no el posicionamiento de las hojas de forma lineal dentro del modelo.

Metodología Random Forest.

El Random Forest es un algoritmo que permite clasificar y predecir datos de forma más específica y óptima. Se construyen múltiples árboles de decisión durante la fase de entrenamiento de datos para luego predecir el mejor de estos.

```
DATA$Variable Ojetiva = factor(DATA$Variable Objetiva)
```

- **DATA:** Es la dataframe a utilizar para extraer los datos
- **Variable Ojetiva :** Es la variable que usaremos para el modelo
- **factor:** Es una función que permite convertir una variable como factor dentro de

las demás variables que se trabajarán en el modelo.

- **DATA\$Variable Objetiva:** Mediante el factor se ha seleccionado la dataframe y la variable objetiva que será asignada como factor.

A continuación, se realiza el bosque aleatorio:

```
Bosque = randomForest(x = DATA[Train, 1: 8],  
                      y = DATA[Train, 9],  
                      ntree = 3000, keep.forest = TRUE)
```

- **Bosque:** Es la variable que almacenará el modelo de bosque aleatorio con los datos.
- **randomForest:** Es la función que se utilizará para construir el modelo específico de bosque aleatorio.
- **x:** Será la variable independiente de columnas que no se segmentarán
- **DATA[Train,1:8]:** El dataframe está otorgado por DATA, mientras que Train es el vector que contiene los datos a utilizar ya entrenados previamente y que indica usar en rango o columna de 1 a 8.
- **y:** es la variable dependiente objetiva que el modelo segmentará.
- **DATA[Train,9]:** El dataframe está otorgado por DATA, mientras que Train es el vector que contiene los datos a utilizar ya entrenados previamente y que indica usar en rango o columna 9
- **ntree = 3000:** Esta función permitirá indicar la cantidad de números de árboles que se construirán en el bosque, en este caso 3000 es el estimado.
- **keep.forest = TRUE:** Esta función indica si se requiere mantener los árboles creados dentro del modelo de bosque aleatorio.

Análisis de Resultados

Relación de Variables

Para poder empezar a construir el modelo, se debe tener en consideración las variables que cuentan con una relación entre las mismas para poder relacionarlas.

A continuación, código para demostrar la correlación entre las variables:

```
cor(RutasT)
```

cor: Es una función que permite visualizar la relación entre dos o más variables en un rango de +1 o -1.

RutasT: Es el nombre de la nueva dataframe con las variables objetivas a buscar relación.

Figura 4

Correlación en las variables objetivas

```
> cor(RutasT)
```

	RUTA	TARIFA	VI.TICOS	GASOLINA	VIAJE	TIEMPO..min.
RUTA	1.00000000	0.11971020	-0.31348064	0.57973283	0.06924667	0.411490928
TARIFA	0.11971020	1.00000000	0.78993418	0.60146438	0.99610137	0.578406743
VI.TICOS	-0.31348064	0.78993418	1.00000000	-0.01312791	0.83942473	0.057836069
GASOLINA	0.57973283	0.60146438	-0.01312791	1.00000000	0.52950160	0.883369233
VIAJE	0.06924667	0.99610137	0.83942473	0.52950160	1.00000000	0.514845099
TIEMPO..min.	0.41149093	0.57840674	0.05783607	0.88336923	0.51484510	1.000000000
DISTANCIA..km.	0.37434389	0.36539817	-0.18124060	0.86436807	0.28729323	0.850189206
PESO.BRUTO.KG.	0.59824460	-0.14194325	-0.40999596	0.24525678	-0.16512855	0.005546016
PESO.NETO..kg.	0.59479887	-0.08595645	-0.34986073	0.25688122	-0.10669208	0.015372697
CAJAS	0.72826281	0.14707500	-0.37265708	0.68866035	0.08917636	0.462410705

	DISTANCIA..km.	PESO.BRUTO.KG.	PESO.NETO..kg.	CAJAS
RUTA	0.3743439	0.598244598	0.59479887	0.72826281
TARIFA	0.3653982	-0.141943248	-0.08595645	0.14707500
VI.TICOS	-0.1812406	-0.409995965	-0.34986073	-0.37265708
GASOLINA	0.8643681	0.245256782	0.25688122	0.68866035
VIAJE	0.2872932	-0.165128555	-0.10669208	0.08917636
TIEMPO..min.	0.8501892	0.005546016	0.01537270	0.46241070
DISTANCIA..km.	1.0000000	-0.134913577	-0.14561635	0.38443670
PESO.BRUTO.KG.	-0.1349136	1.000000000	0.99785994	0.83206207
PESO.NETO..kg.	-0.1456164	0.997859938	1.00000000	0.83075132
CAJAS	0.3844367	0.832062074	0.83075132	1.00000000

Nota: Elaboración propia de los autores en el programa de RStudio.

Esta figura muestra la correlación que existe con tres variables objetivas dentro del dataframe, las cuáles serán consideradas para demostrar cómo la relación entre estas puede influir en la selección de una ruta óptima.

La variable “Ruta” tiene una correlación significativa con la variable “Cajas”. De esta forma se reconocerá cuál ruta será la encargada de entregar la mayor cantidad de cajas.

La variable “Tarifa” tienen una correlación significativa con la variable “Viaje”. De esta forma se reconocerá cuál tarifa influye más en el costo de viaje por cada ruta.

La variable “Tiempo” tiene una correlación significativa con la variable “Distancia”. De esta forma se reconocerá cuál es la distancia en kilómetros de las rutas que es más larga, pero podría ser la más rápida o lenta.

Tablas de Frecuencia de variables objetivas

Se realizarán tablas de frecuencia de las variables para conocer sus frecuencias relativas, acumuladas, y acumuladas relativas dentro de la categorización de cada uno de sus valores y reconocer cuál de estos tienen mayor impacto con las rutas. Las variables a analizar son las que tienen mayor correlación entre las mismas, es decir, “ruta”, “cajas”, “distancia”, “tiempo”, “tarifa” y “viaje”.

Variable Ruta

table (RutasT \$RUTA)

table: Es una función que permite la creación de una tabla de frecuencia sobre los datos de la variable objetiva.

RutaT: Es el nombre del dataframe.

\$RUTA: Es la selección de la variable dentro del dataframe.

Fre_ruta = as.data.frame(table(RutasT\$RUTA))

Fre_ruta: Es el nombre de la nueva variable creada

as.data.frame: Esta función permite convertir la tabla de frecuencias generadas por “table” en una dataframe

table(RutasT\$RUTA) : Es la función que contiene la tabla de frecuencias de la variable ruta.

Fre_ruta

Fre_ruta : Es el nombre de la nueva variable creada y la que se puede llamar para visualizar la tabla de frecuencia creada

```
Tab_Fre_ruta = transform(Fre_ruta,  
Frel = round(prop.table(Fre_ruta$Freq),3), FAcu  
= cumsum(Fre_ruta$Freq),  
FAcuR = cumsum(round(prop.table(Fre_ruta$Freq),3)))
```

Tab_Fre_ruta: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_ruta: la variable que contienen las frecuencias de ruta

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_ruta\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_ruta\$Freq: Es la variable que contiene las frecuencias

FAcuR: Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_ruta\$Freq: Es la variable que contiene las frecuencias

3: cantidad de decimales

Tab_Fre_ruta

Tab_Fre_ruta: Es el data frame resultante

Variable Caja

Para la variable caja se debe agrupar los datos por la regla de sturges, ya que permitirá crear intervalos de la cantidad de cajas existentes para una mejor visualización de la frecuencia de estas.

$$x1 = \text{RutasT\$CAJAS}$$

x1: Es la variable nueva creada

RutasT\$CAJAS: La variable “CAJAS” ha sido seleccionada para analizar dentro del dataframe de “RutasT”

$$k1 = \text{nclass.Sturges}(x1)$$

k1: Es la variable nueva creada .

nclass.Sturges: Es la función que calcula el numero óptimo de clases o intervalos para agrupar los datos en forma de histograma de la variable cajas.

x1: Variable que contiene la cantidad de cajas del intervalo

$$k1$$

k1: Es la variable que almacena la cantidad de clases de intervalos creados

$$\text{Int1} = \text{cut}(x1, \text{breaks} = k1)$$

Int1: Es un factor que indica a qué intervalo pertenece cada valor en x1

cut: Divide la variable x1, en este caso de variable cajas

x1: Variable que contiene la cantidad de cajas del intervalo

breaks: Especifica el corte de los números en que se basará el intervalo

k1: Variable que contienen los intervalos

$$\text{Int1}$$

Int1: Es la variable que almacena los cortes de intervalos creados por distribución

$$\text{Fre_caja} = \text{as.data.frame}(table(\text{Int1}))$$

Fre_caja: Es la nueva dataframe creada que contienen una columna de intervalos y frecuencia de cajas por intervalo

as.data.frame: Esta función permite convertir la table de frecuencias generadas en una

dataframe

table(Int1): Es la función que contiene la tabla de frecuencias de la variable intervalos de cajas

Fre_caja

Fre_caja: Es la variable que contiene la tabla de frecuencia de cajas por intervalos

$$\begin{aligned} Tab_caja &= transform(Fre_caja, \\ Frel &= round(prop.table(Fre_caja\$Freq),3), FAcu \\ &= cumsum(Fre_caja\$Freq), \\ FAcuR &= cumsum(round(prop.table(Fre_caja\$Freq),3))) \end{aligned}$$

Tab_Fre_caja: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_caja: la variable que contienen las frecuencias de cajas

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_caja\$Freq: Es la variable que contiene las frecuencias

3: cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_caja\$Freq: Es la variable que contiene las frecuencias

FAcuR: Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_caja\$Freq: Es la variable que contiene las frecuencias

3: cantidad de decimales

Tab_caja

Tab_caja: Es el data frame resultante

Variable Tarifa

table (RutasT \$TARIFA)

table: Es una función que permite la creación de una tabla de frecuencia sobre los datos de la variable objetiva

RutaT: Es el nombre del dataframe

\$TARIFA: es la selección de la variable dentro del dataframe

Fre_tarifa = as.data.frame(table(RutasT\$TARIFA))

Fre_tarifa: Es el nombre de la nueva variable creada

as.data.frame: Esta función permite convertir la table de frecuencias generadas por “table” en una dataframe

table(RutasT\$TARIFA): Es la función que contiene la tabla de frecuencias de la variable tarifa.

Fre_tarifa

Fre_tarifa: Es el nombre de la nueva variable creada y la que se puede llamar para visualizar la tabla de frecuencia creada

*Tab_Fre_tarifa = transform(Fre_tarifa,
Frel = round(prop.table(Fre_tarifa\$Freq),3), FAcu
= cumsum(Fre_tarifa\$Freq),
FAcuR = cumsum(round(prop.table(Fre_tarifa\$Freq),3)))*

Tab_Fre_tarifa: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_tarifa: La variable que contienen las frecuencias de tarifa

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_tarifa\$Freq: Es la variable que contiene las frecuencias

3: cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_tarifa\$Freq: Es la variable que contiene las frecuencias

FAcuR : Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_tarifa\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

Tab_Fre_tarifa

Tab_Fre_tarifa: Es el data frame resultante

Variable Viaje

table (RutasT \$VIAJE)

table: Es una función que permite la creación de una tabla de frecuencia sobre los datos de la variable objetiva

RutaT: Es el nombre del dataframe

\$VIAJE: Es la selección de la variable dentro del dataframe

Fre_viaje = as.data.frame(table(RutasT\$VIAJE))

Fre_viaje: Es el nombre de la nueva variable creada

as.data.frame: Esta función permite convertir la table de frecuencias generadas por “table” en una dataframe

table(RutasT\$VIAJE): Es la función que contiene la tabla de frecuencias de la variable viaje.

Fre_viaje

Fre_viaje: Es el nombre de la nueva variable creada y la que se puede llamar para visualizar la tabla de frecuencia creada

```
Tab_Fre_viaje = transform(Fre_viaje,  
  Frel = round(prop.table(Fre_viaje$Freq),3),FAcu  
  = cumsum(Fre_viaje$Freq),  
  FAcuR = cumsum(round(prop.table(Fre_viaje$Freq),3)))
```

Tab_Fre_viaje: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_viaje: La variable que contienen las frecuencias de viaje

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_viaje\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_viaje\$Freq: Es la variable que contiene las frecuencias

FAcuR: Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table : esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_viaje\$Freq : es la variable que contiene las frecuencias

3: cantidad de decimales

Tab_Fre_viaje

Tab_Fre_viaje: Es el data frame resultante

Variable Distancia

table (RutasT \$DISTANCIA)

table: Es una función que permite la creación de una tabla de frecuencia sobre los datos de la variable objetiva

RutaT: Es el nombre del dataframe

\$DISTANCIA: Es la selección de la variable dentro del dataframe

Fre_viaje = as.data.frame(table(RutasT\$DISTANCIA))

Fre_distancia: Es el nombre de la nueva variable creada

as.data.frame : Esta función permite convertir la table de frecuencias generadas por “table” en una dataframe

table(RutasT\$DISTANCIA) : es la función que contiene la tabla de frecuencias de la variable distancia.

Fre_distancia

Fre_distancia: Es el nombre de la nueva variable creada y la que se puede llamar para visualizar la tabla de frecuencia creada.

*Tab_Fre_distancia = transform(Fre_distancia,
Frel = round(prop.table(Fre_distancia\$Freq),3), FAcu
= cumsum(Fre_distancia\$Freq),
FAcuR = cumsum(round(prop.table(Fre_distancia\$Freq),3)))*

Tab_Fre_distancia: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_distancia: La variable que contienen las frecuencias de distancia

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en

relación con el total

Fre_distancia\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_distancia\$Freq: Es la variable que contiene las frecuencias

FAcuR: Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_distancia\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

Tab_Fre_distancia

Tab_Fre_distancia: Es el data frame resultante

Variable Tiempo

table (RutasT \$TIEMPO)

table: Es una función que permite la creación de una tabla de frecuencia sobre los datos de la variable objetiva

RutaT: Es el nombre del dataframe

\$TIEMPO: Es la selección de la variable dentro del dataframe

Fre_tiempo = as.data.frame(table(RutasT\$TIEMPO))

Fre_tiempo: Es el nombre de la nueva variable creada

as.data.frame : esta función permite convertir la table de frecuencias generadas por “table” en una dataframe

table(RutasT\$TIEMPO): Es la función que contiene la tabla de frecuencias de la variable

tiempo.

Fre_tiempo

Fre_tiempo: Es el nombre de la nueva variable creada y la que se puede llamar para visualizar la tabla de frecuencia creada

```
Tab_Fre_tiempo = transform(Fre_tiempo,  
Frel = round(prop.table(Fre_tiempo$Freq),3), FAcu  
= cumsum(Fre_tiempo$Freq),  
FAcuR = cumsum(round(prop.table(Fre_tiempo$Freq),3)))
```

Tab_Fre_tiempo: Es el nombre de la nueva variable creada

transform: Es una función que sirve para modificar o agregar columnas en un dataframe

Fre_tiempo: La variable que contienen las frecuencias de tiempo

Frel: Frecuencia relativa

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_tiempo\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

FAcu: Frecuencia acumulada

cumsum: Esta función calcula la suma acumulada de la frecuencia absoluta

Fre_tiempo\$Freq: Es la variable que contiene las frecuencias

FAcuR: Frecuencia acumulada relativa

cumsum: Esta función calcula la suma acumulada de las frecuencias relativas redondeadas

round: Redondear decimales

prop.table: Esta función calcula la proporción o frecuencia relativa de cada categoría en relación con el total

Fre_tiempo\$Freq: Es la variable que contiene las frecuencias

3: Cantidad de decimales

Tab_Fre_tiempo

Tab_Fre_tiempo: Es el data frame resultante

Gráficos de Frecuencia de variables objetivas

La construcción de gráficos de barras o histogramas se llevarán a cabo por las variables de frecuencias creadas cómo las frecuencias relativas, frecuencias acumuladas y frecuencias acumuladas relativas, además, el data frame de cada una de las variables objetivas segmentadas por la frecuencia de las observaciones

Variable Ruta

```
G1 <- ggplot(Tab_Fre_ruta, aes(x = Tab_Fre_ruta$Var1,
  y = Tab_Fre_ruta$Frel)) + geom_bar(stat = "identity", fill = "pink",
  colour = "black", size = 0.5) + geom_text(aes(label
  = paste0(Tab_Fre_ruta$Frel),
  position = position_stack(vjust = 0.8)) + labs(title
  = "Class por Ruta")
```

G1: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_Fre_ruta: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

Tab_Fre_ruta\$Var1: El dataframe presentado con la Var1 que contiene las categorías de rutas en el eje x

y: Eje dependiente

Tab_Fre_ruta\$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada categoría en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "pink": Establece el color de relleno de las barras, en este caso, rosa.

colour="black": Define el color del borde de las barras, en este caso, negro.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_Fre_ruta\$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra, pero cerca del borde superior.

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

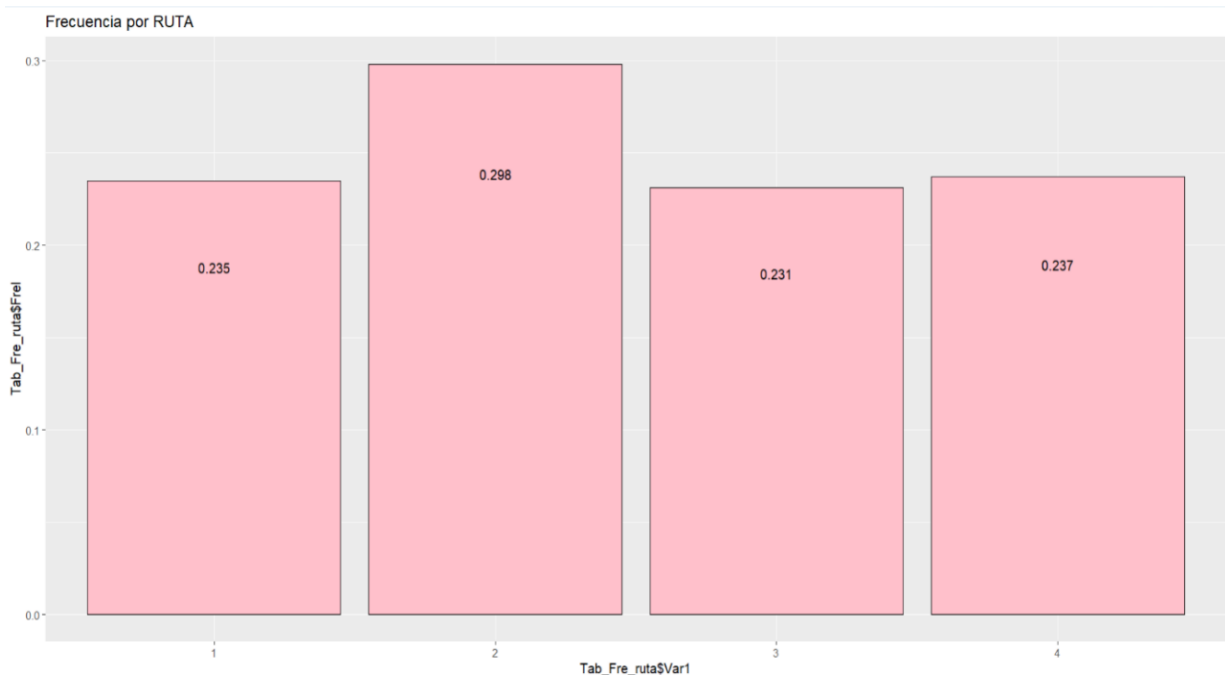
title = "Frecuencia por RUTA": Establece el título del gráfico.

G1

G1: Es la variable con la que se llamará al gráfico de frecuencias creado

Figura 5

Frecuencia de variable ruta



Nota: Elaboración propia de los autores en el programa de RStudio.

En esta figura se muestra la frecuencia del uso de cada una de las 4 rutas por los despachadores de cajas. Dando como resultado a la ruta 2, siendo la más frecuente en su uso mientras que la ruta 4 quedando en segundo lugar y la ruta 1 en tercero y finalmente la ruta 3 siendo la última por preferencia del conductor.

Variable Caja

```
G2 <- ggplot(Tab_caja, aes(x = Tab_caja$Int1, y = Tab_caja$Frel)) +  
  geom_bar(stat = "identity", fill = "skyblue", colour = "black", size = 0.5) +  
  geom_text(aes(label = paste0(Tab_caja$Frel)), position = position_stack(vjust  
    = 0.8)) +  
  coord_flip() + labs(title = "Frecuencia de CAJAS")
```

G2: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_caja: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

Tab_caja\$Int1: El dataframe presentado con la Int1 que contiene las categorías de cantidad de intervalos en el eje x

y: Eje dependiente

Tab_caja \$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada intervalo en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "skyblue": Establece el color de relleno de las barras, en este caso, celeste.

colour="black": Define el color del borde de las barras, en este caso, negro.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_caja \$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra, pero cerca del borde superior.

coord_flip: Ajuste de gráfica

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

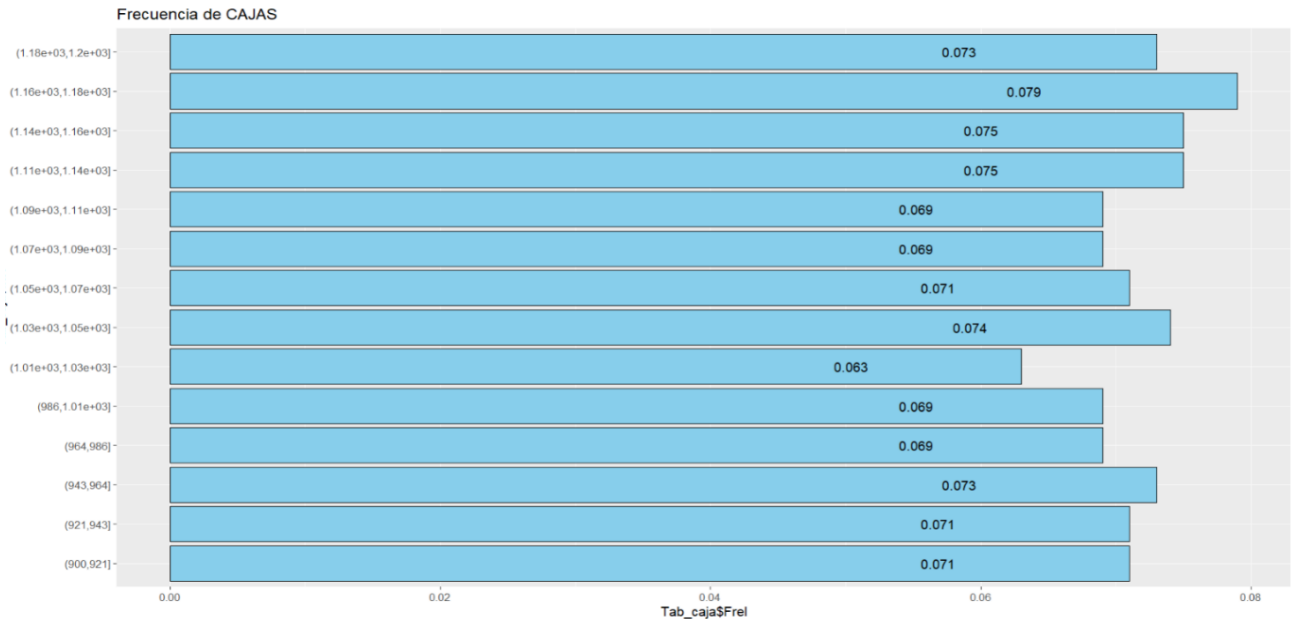
title = "Frecuencia por CAJAS": Establece el título del gráfico.

G2

G2: Es la variable con la que se llamará al gráfico de frecuencias creado

Figura 6

Frecuencia de variable caja



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se aprecian las frecuencias creadas por los intervalos obtenidos a través de strugers, cada intervalo clasificado con la cantidad de cajas y la frecuencia en cada uno de estos tiene para despachar la mayor o menor cantidad de cajas posibles. Cuya clasificación va desde 900 cajas hasta 1200 cajas.

Variable Tarifa

```
G3 <- ggplot(Tab_Fre_tarifa, aes(x = Tab_Fre_tarifa$Var1,
  y = Tab_Fre_tarifa$Frel)) + geom_bar(stat = "identity", fill
  = "skyblue",
  colour = "red", size = 0.5) + geom_text(aes(label
  = paste0(Tab_Fre_tarifa$Frel)),
  position = position_stack(vjust = 0.8)) + labs(title
  = "Frecuencia por TARIFA")
```

G3: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_Fre_tarifa: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

Tab_Fre_tarifa\$Var1: El dataframe presentado con la Var1 que contiene las categorías de tarifas en el eje x

y: Eje dependiente

Tab_Fre_tarifa\$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada categoría en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "skyblue": Establece el color de relleno de las barras, en este caso, celeste.

colour="red": Define el color del borde de las barras, en este caso, rojo.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_Fre_tarifa\$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra pero cerca del borde superior.

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

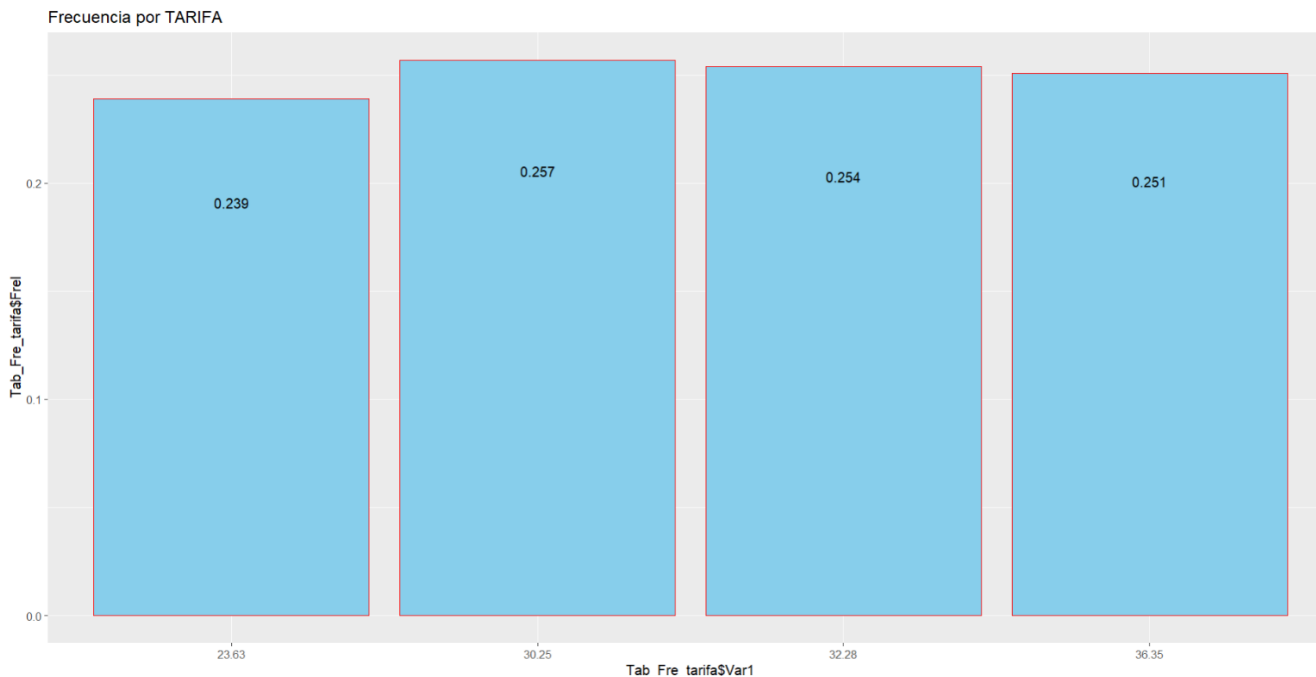
title = "Frecuencia por TARIFA": Establece el título del gráfico.

G3

G3: Es la variable con la que se llamará al gráfico de frecuencias creado

Figura 7

Frecuencia de variable tarifa



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se observa la frecuencia de las tarifas establecidas como frecuentes como gasto por la entrega de cajas y por las rutas. La tarifa de \$30,25 es la que tiene tendencia a repetirse

por los despachadores en el cobro de despacho de cajas. Siendo la tarifa menos costosa de \$23,63 la que menos es atractiva por los despachadores.

Variable Viaje

```
G4 <- ggplot(Tab_Fre_viaje, aes(x = Tab_Fre_viaje$Var1,
  y = Tab_Fre_viaje$Frel)) + geom_bar(stat = "identity", fill
  = "yellow",
  colour = "orange", size = 0.5) + geom_text(aes(label
  = paste0(Tab_Fre_viaje$Frel)),
  position = position_stack(vjust = 0.8)) + labs(title
  = "Frecuencia por VIAJE")
```

G4: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_Fre_viaje: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

Tab_Fre_viaje\$Var1: El dataframe presentado con la Var1 que contiene las categorías de tarifas en el eje x

y: Eje dependiente

Tab_Fre_viaje\$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada categoría en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "yellow": Establece el color de relleno de las barras, en este caso, amarillo.

colour="orange": Define el color del borde de las barras, en este caso, naranja.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_Fre_viaje\$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra pero cerca del borde superior.

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

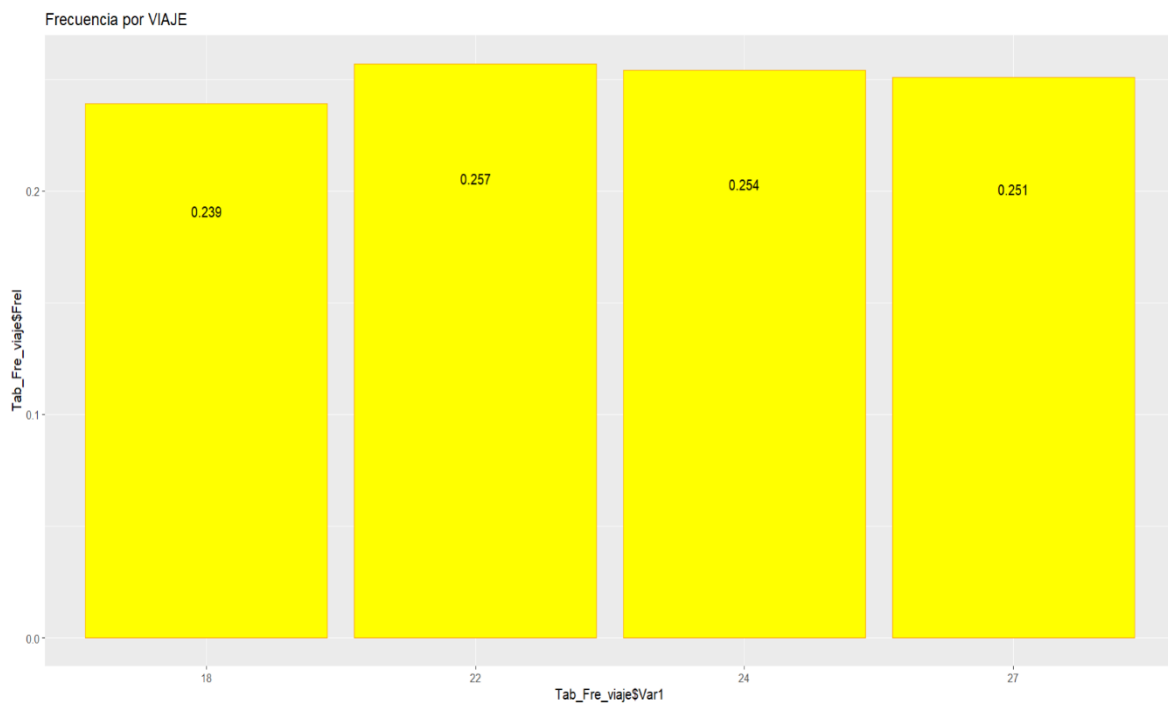
title = "Frecuencia por VIAJE": Establece el título del gráfico.

G4

G4: Es la variable con la que se llamará al gráfico de frecuencias creado

Figura 8

Frecuencia de variable viaje



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se presencia el valor del costo de viaje que tiende a ser más frecuente entre los conductores. En este caso, el viaje con costo de \$22 es el que usualmente tiene mayor afluencia por los conductores mientras que el viaje con costo de \$18 es el menos frecuente.

Variable Distancia

```
G5 <- ggplot(Tab_Fre_distancia, aes(x = Tab_Fre_distancia$Var1,
  y = Tab_Fre_distancia$Frel)) + geom_bar(stat = "identity", fill
  = "pink",
  colour = "purple", size = 0.5) + geom_text(aes(label
  = paste0(Tab_Fre_distancia$Frel)),
  position = position_stack(vjust = 0.8)) + labs(title
  = "Frecuencia por DISTANCIA")
```

G5: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_Fre_distancia: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

Tab_Fre_distancia\$Var1: El dataframe presentado con la Var1 que contiene las categorías de distancias en kilómetros en el eje x

y: Eje dependiente

Tab_Fre_distancia\$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada categoría en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "pink": Establece el color de relleno de las barras, en este caso, rosado.

colour="purple": Define el color del borde de las barras, en este caso, morado.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_Fre_distancia\$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra, pero cerca del borde superior.

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

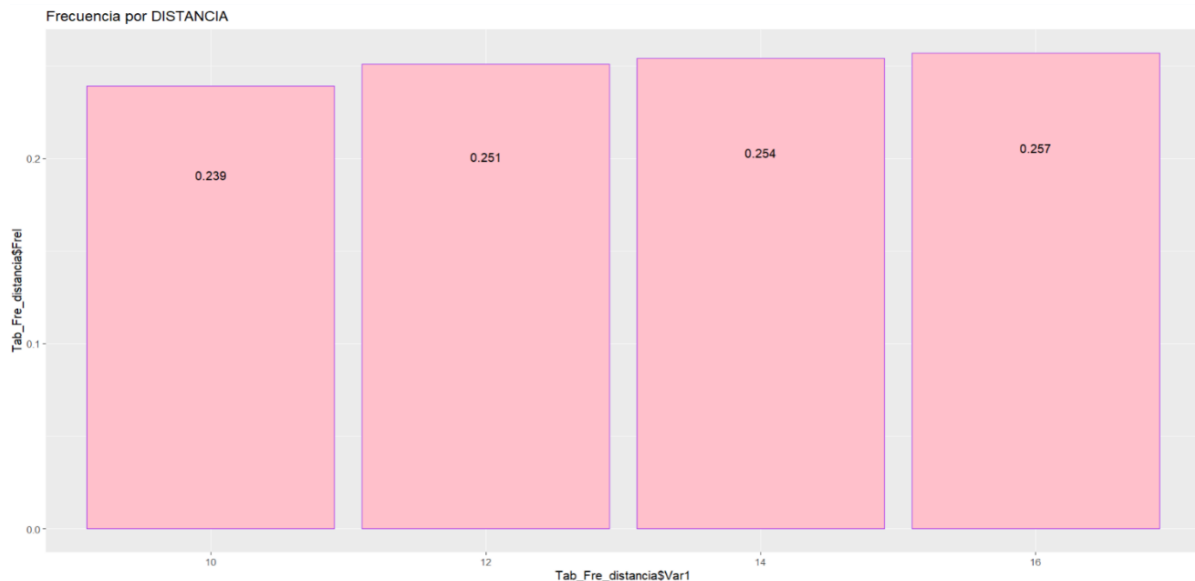
title = "Frecuencia por DISTANCIA": Establece el título del gráfico.

G5

G5: Es la variable con la que se llamará al gráfico de frecuencias creado

Figura 9

Frecuencia de variable distancia



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se demuestra la frecuencia de la distancia más corta o larga que ha sido la que mayor afluencia de rutas tiene. Dando como resultado a la distancia más larga siendo la que tiende a tener más recorridos, es decir, 16 kilómetros. Sin embargo, la distancia menor de 10 kilómetros es la menos frecuente a pesar de ser la más rápida.

Variable Tiempo

```
G6 <- ggplot(Tab_Fre_tiempo, aes(x = Tab_Fre_tiempo$Var1,  
  y = Tab_Fre_tiempo$Frel)) + geom_bar(stat = "identity", fill  
  = "green",  
  colour = "lightblue", size = 0.5) + geom_text(aes(label  
  = paste0(Tab_Fre_tiempo$Frel)),  
  position = position_stack(vjust = 0.8)) + labs(title  
  = "Frecuencia por TIEMPO")
```

G6: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

Tab_Fre_tiempo: Es el dataframe creado anterior para el uso de la construcción del gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: eje independiente

Tab_Fre_tiempo\$Var1: El dataframe presentado con la Var1 que contiene las categorías de tiempos en el eje x

y: Eje dependiente

Tab_Fre_tiempo\$Frel: El dataframe presentado con la Frel que contiene las frecuencias relativas de cada categoría en el eje y

geom_bar: Es la función que permite crear un gráfico de barras

stat = "identity": Es la función que permite indicar que las alturas de las barras deben corresponder directamente a los valores proporcionados por y o Frel

fill= "green": Establece el color de relleno de las barras, en este caso, verde.

colour="lightblue": Define el color del borde de las barras, en este caso, azul claro.

size=0.5: Especifica el grosor de las líneas que forman el borde de las barras.

geom_text: Agrega etiquetas de texto a las barras.

aes: Es la función que define las variables de los ejes x – y en el gráfico

label = paste0(Tab_Fre_tiempo\$Frel): Especifica que el texto de la etiqueta debe ser la frecuencia relativa (Frel) para cada barra.

position: Posición del texto

position_stack: Ajusta la posición vertical del texto en relación con la barra

vjust = 0.8: Mueve el texto ligeramente hacia abajo, posicionándolo dentro de la barra,

pero cerca del borde superior.

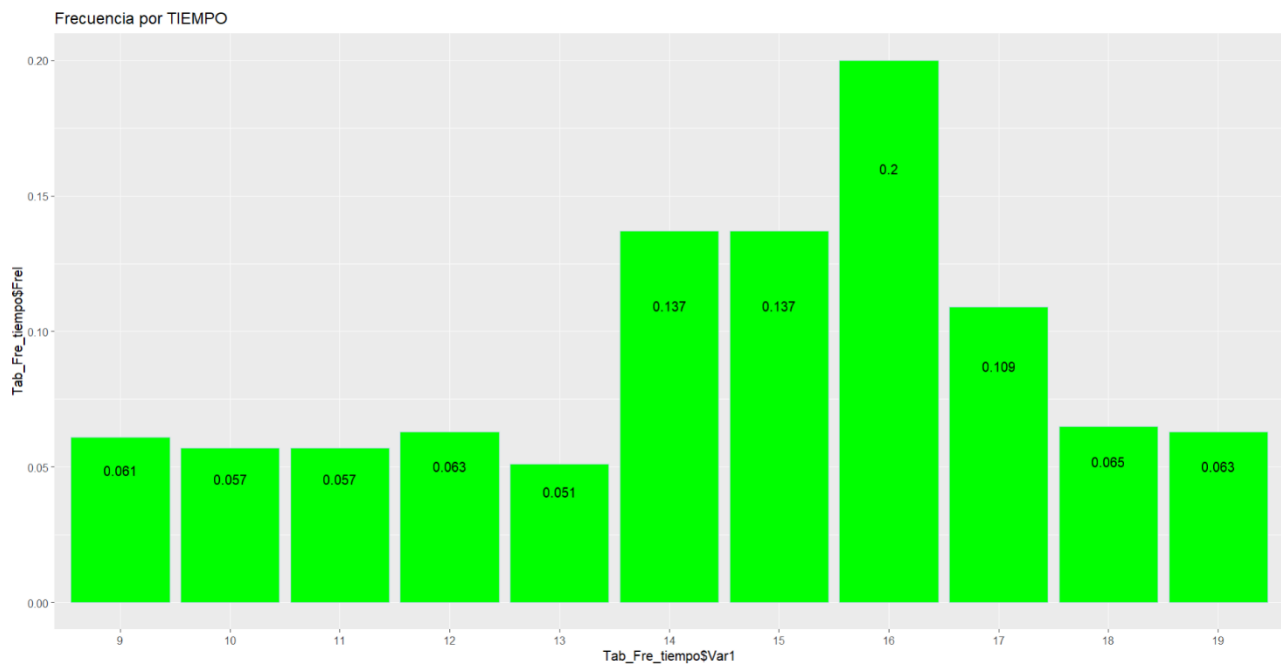
labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

title = "Frecuencia por TIEMPO": Establece el título del gráfico.

G6

G6: Es la variable con la que se llamará al gráfico de frecuencias creado

Frecuencia de variable tiempo



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se muestra la frecuencia del tiempo en que las cajas son entregadas por los repartidores durante las rutas. Un repartido tiene la tendencia de entregar cajas dentro de los 16 minutos, mientras que dentro de 14 a 15 minutos es la segunda tendencia en que demoran en realizar una entrega.

Gráficos de relación entre las variables objetivas

La construcción de gráficos de dispersión y boxplot para demostrar la relación que comparten las variables y cómo influyen entre ellas para la selección de una ruta específica que sea conveniente en el Centro de Acopio. Entre mayor sea la dispersión, mayor será la dependencia de las variables.

Variable Ruta x Variable Cajas

```
RutasT$RUTA = factor(RutasT$RUTA)
```

RutasT\$RUTA: El dataframe RutaT toma la variable Ruta para ser convertida en una variable categórica

factor: Es la función que permite convertir una variable numérica en categórica

RutasT\$RUTA: El dataframe RutaT toma la variable Ruta para ser convertida en una variable categórica

```
GB1 = ggplot(RutasT, aes(x = RutasT$RUTA, y = RutasT$CAJAS)) +  
geom_boxplot(fill = "pink", colours = "black") + labs(title = "RUTA X CAJA")
```

GB1: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$RUTA: Es la variable categórica que se está utilizando para agrupar los datos

y: Eje dependiente

RutasT\$CAJAS: Es la variable numérica que se está analizando, la distribución de cajas

geom_boxplot: Es la función que se utiliza para crear un gráfico de cajas

fill= "pink": Establece el color de relleno de las cajas, en este caso, rosa.

colours="black": Define el color del borde de las cajas, en este caso, negro.

labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

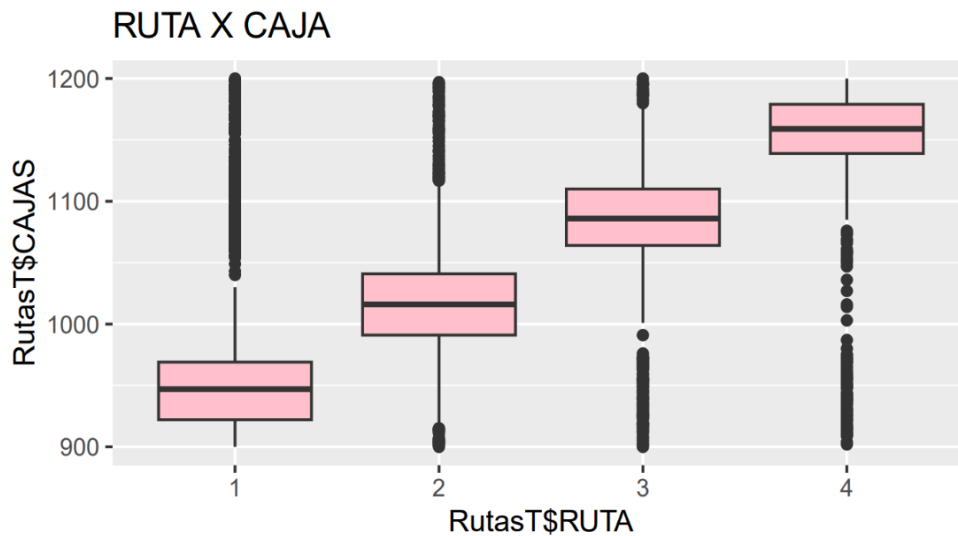
title = "RUTA X CAJA": Establece el título del gráfico.

GB1

GB1: Es la variable con la que se llamará al gráfico cajas creado

Figura 10

Boxplot de variables rutas x cajas



Nota: Elaboración propia de los autores en el programa de R studio.

En este gráfico de cajas o boxplot se muestra la distribución de la variable cajas para la categoría rutas en el dataset RutasT. En el eje de x están las 4 rutas mientras que en el eje y están los intervalos de cantidad de cajas de 900 a 1200. Se observa que cada ruta realiza una cantidad de entrega de cajas diferente causando una variabilidad de distribución. La ruta 1 es la que menor cantidad de cajas distribuye en un rango de 900 a 960 y tiene datos dispersos fuera de la caja por encima del tercer cuartil al 40% a partir del rango 1040 hasta 1200. La ruta 2 tiene un rango de entrega de cajas de 990 a 1045 y tienen datos dispersos fuera de la caja por debajo del primer cuartil en un 10% a partir del rango 900 hasta 910 y por encima del tercer cuartil a un 25% a partir del rango 1110 hasta 1200. La ruta 3 tiene un rango de entrega de cajas de 1060 a 1110 y tiene datos dispersos fuera de la caja por debajo del primer cuartil al 30% a partir del rango 900 hasta 990 y por encima del tercer cuartil en un 10% a partir del rango 1180 hasta 1200. La ruta 4 tiene

un rango de entrega de cajas de 1140 a 1170 y tiene datos dispersos fuera de la caja por debajo del primer cuartil al 40% en un rango de 900 hasta 1070.

Variable Ruta x Variable Tiempo

```
RutasT$RUTA = factor(RutasT$RUTA)
```

RutasT\$RUTA: El dataframe RutaT toma la variable Ruta para ser convertida en una variable categórica

factor: Es la función que permite convertir una variable numérica en categórica

RutasT\$RUTA: El dataframe RutaT toma la variable Ruta para ser convertida en una variable categórica

```
GB2 = ggplot(RutasT, aes(x = RutasT$RUTA, y = RutasT$TIEMPO..min.)) +  
geom_boxplot(fill = "blue", colours = "black") + labs(title = "RUTA X TIEMPO")
```

GB2: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$RUTA: Es la variable categórica que se está utilizando para agrupar los datos

y: Eje dependiente

RutasT\$TIEMPO..min.: Es la variable numérica que se está analizando, la distribución del tiempo

geom_boxplot: Es la función que se utiliza para crear un gráfico de cajas

fill= "blue": Establece el color de relleno de las cajas, en este caso, azul.

colours="black": Define el color del borde de las cajas, en este caso, negro.

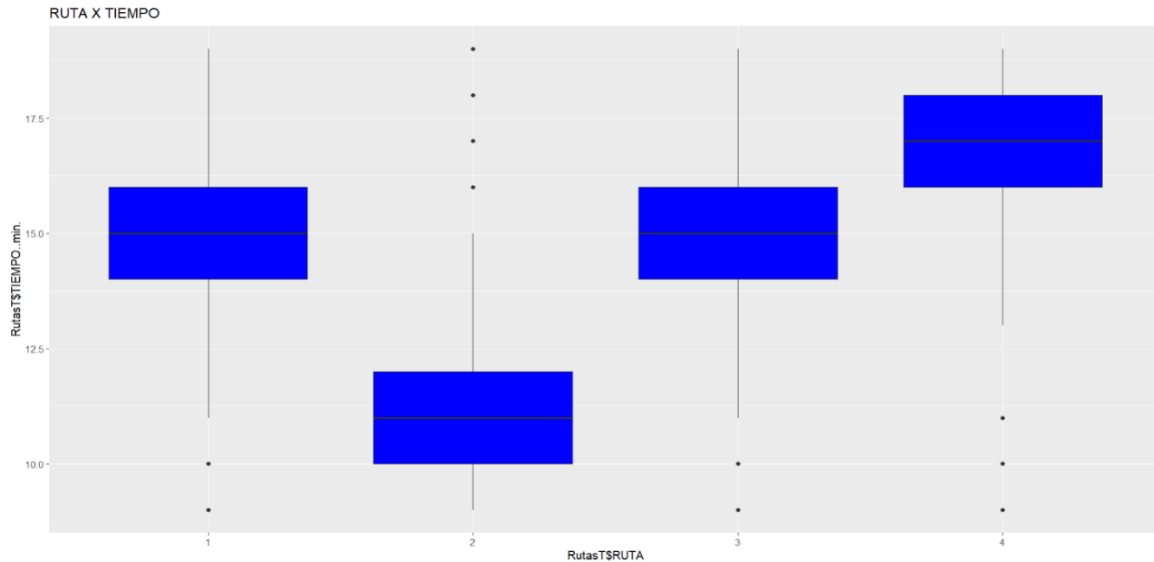
labs: Se utiliza para etiquetar diferentes partes del gráfico, como el título o los ejes,

title = "RUTA X TIEMPO": Establece el título del gráfico.

GB2: Es la variable con la que se llamará al gráfico de cajas creado

Figura 11

Boxplot de variables rutas x tiempo



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se observa en el eje x las categorías de 4 rutas mientras que en el eje y está el rango de tiempo de las rutas establecidas de 10 minutos a 17,5 minutos por entrega. La ruta 1 tiene un rango de entrega en un tiempo de 14 minutos a 16 minutos con una dispersión de datos fuera de su caja por debajo del primer cuartil en 9 minutos a 10 minutos. La ruta 2 tienen un rango de entrega en un tiempo de 10 minutos a 12 minutos con una dispersión de datos fuera de su caja por encima del tercer cuartil en 15,50 minutos hasta 19 minutos. La ruta 3 tiene un rango de entrega en un tiempo de 14 minutos a 16 minutos con una dispersión de datos fuera de su caja por debajo del primer cuartil en 9 minutos a 10 minutos, dando así una asimetría en la ruta 1 y 3. La ruta 4 tienen un tiempo de entrega de 16 minutos a 18,20 minutos con una dispersión de datos fuera de su caja por debajo del primer cuartil en 9 minutos a 11 minutos.

Variable Ruta x Variable Caja – Variable Distancia x Variable Tiempo

```
G7 = ggplot(RutasT, aes(x = RutasT$DISTANCIA..km.,  
                        y = RutasT$TIEMPO..min.,  
                        colour = RutasT$RUTA, size = RutasT$CAJAS)) + geom_point() +  
  labs(title = "Distancia x Tiempo")
```

G7: la nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$DISTANCIA: Es la variable categórica que se está utilizando para agrupar las distancias

y: Eje dependiente

RutasT\$TIEMPO..min.: Es la variable numérica que se está analizando, la distribución del tiempo

colours= RutasT\$RUTA: Define el color los que se clasificarán las rutas

size : RutasT\$CAJAS : Define el tamaño con los que se clasificarán las cantidades de rangos de cajas

geom_point: Permite crear un gráfico de puntos

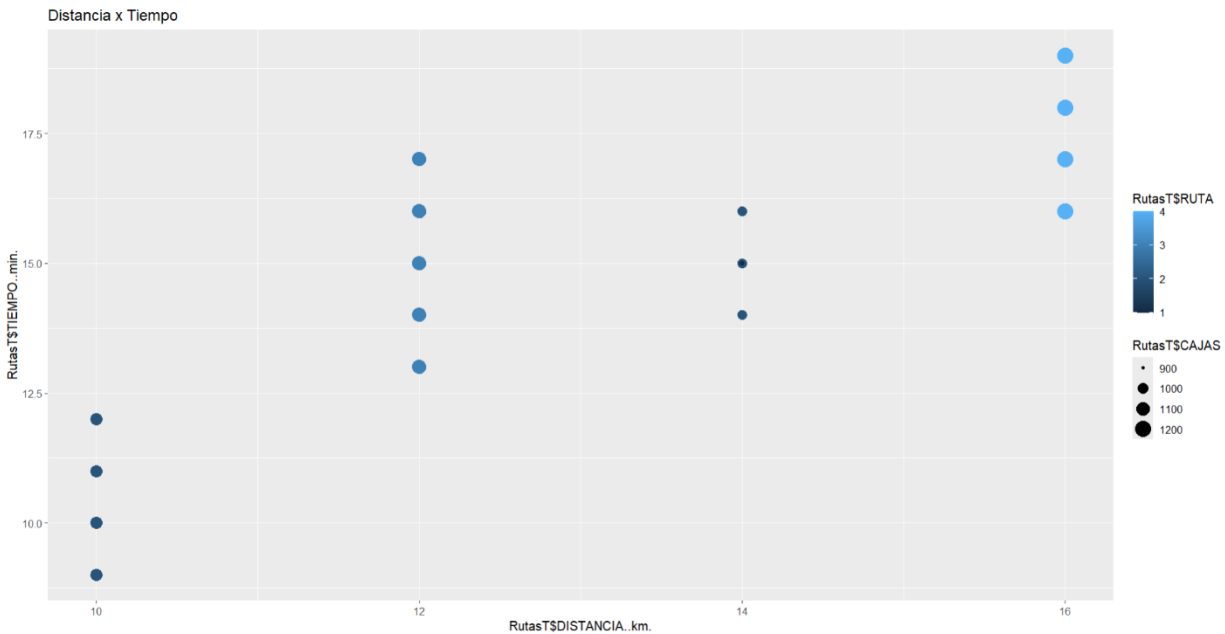
labs: Permite etiquetar los ejes del gráfico

title = "RUTA X TIEMPO": Establece el título del gráfico.

G7: Es la variable con la que se llamará al gráfico de puntos creado

Figura 12

Gráfico de variables rutas x caja – tiempo x distancia



Nota: Elaboración propia de los autores en el programa de RStudio.

Este gráfico permite visualizar la relación entre dos variables continuas, en este caso, DISTANCIA y TIEMPO, mientras se añaden capas adicionales de información mediante el color y el tamaño de los puntos que serán representados por la variable ruta y caja. En el eje x están los kilómetros de distancia de 10, 12 14 y 16 y en el eje de la y el rango de tiempo de 9 minutos a 19 minutos. La ruta 1 representada con color oscuro azul entrega entre 900 cajas a 970 cajas en un tiempo de 14 minutos a 16 minutos por 14 kilómetros de distancia. La ruta 2 representada con color medio oscuro azul entrega entre 1000 cajas a 1050 cajas en un tiempo de 9 minutos a 12 minutos por 10 kilómetros de distancia. La ruta 3 representada con color azul entrega entre 1060 cajas a 1100 cajas en un tiempo de 13 minutos a 17 minutos por 12 kilómetros de distancia. La ruta 4 representada con color celeste entrega entre 1110 cajas a 1200 cajas en un tiempo de 15,50 minutos a 19 minutos por 16 kilómetros de distancia.

Medidas de Tendencia Central

Las medidas de tendencia central conocidas como moda, media y mediana permiten conocer, la frecuencia, distancia y punto medio de los datos de cada variable. En este caso, se analizarán las medidas de las variables objetivas, ruta, caja, tarifa, viaje, distancia y tiempo.

Media

$$\text{mean}(RutasT\$RUTA)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$RUTA: Se obtiene el valor de distancia de la variable ruta

$$\text{mean}(RutasT\$CAJAS)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$CAJAS: Se obtiene el valor de distancia de la variable cajas

$$\text{mean}(RutasT\$TARIFA)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$TARIFA: Se obtiene el valor de distancia de la variable tarifa

$$\text{mean}(RutasT\$VIAJE)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$VIAJE: Se obtiene el valor de distancia de la variable viaje

$$\text{mean}(RutasT\$DISTANCIA..km.)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$DISTANCIA..km.: Se obtiene el valor de distancia de la variable distancia

$$\text{mean}(RutasT\$TIEMPO..min.)$$

mean: Función que permite obtener el valor entre la distancia de un dato al otro

RutasT\$TIEMPO..min.: Se obtiene el valor de distancia de la variable tiempo

Mediana

median(RutasT\$RUTA)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$RUTA: se obtiene el valor medio de la variable ruta

median(RutasT\$CAJAS)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$CAJAS.: Se obtiene el valor medio de la variable cajas

median(RutasT\$TARIFA)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$TARIFA.: Se obtiene el valor medio de la variable tarifa

median(RutasT\$VIAJE)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$VIAJE: Se obtiene el valor medio de la variable viaje

median(RutasT\$DISTANCIA..km.)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$DISTANCIA..km.: Se obtiene el valor medio de la variable distancia

median(RutasT\$TIEMPO..min.)

median: Función que permite obtener el valor en medio del conjunto de datos

RutasT\$TIEMPO..min.: Se obtiene el valor medio de la variable tiempo

Moda

mfv(RutasT\$RUTA)

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$RUTA: se obtiene el valor frecuente de la variable ruta

$$mfv(RutasT\$CAJAS)$$

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$CAJAS: Se obtiene el valor frecuente de la variable cajas

$$mfv(RutasT\$TARIFA)$$

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$TARIFA: Se obtiene el valor frecuente de la variable tarifa

$$mfv(RutasT\$VIAJE)$$

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$VIAJE: Se obtiene el valor frecuente de la variable viaje

$$mfv(RutasT\$DISTANCIA..min.)$$

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$DISTANCIA..km.: Se obtiene el valor frecuente de la variable distancia

$$mfv(RutasT\$TIEMPO..min.)$$

mfv: Función que permite obtener el valor más repetitivo de los datos
RutasT\$TIEMPO..min.: Se obtiene el valor frecuente de la variable tiempo

Coefficiente de Asimetría

$$skewness(RutasT\$RUTA)$$

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$RUTA: Se obtiene el valor de la variable ruta

skewness(RutasT\$CAJAS)

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$CAJAS: Se obtiene el valor de la variable caja

skewness(RutasT\$TARIFA)

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$TARIFA: Se obtiene el valor de la variable tarifa

skewness(RutasT\$VIAJE)

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$VIAJE: Se obtiene el valor de la variable viaje

skewness(RutasT\$DISTANCIA..km.)

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$DISTANCIA..km.: Se obtiene el valor de la variable distancia

skewness(RutasT\$TIEMPO..min.)

skewness: Función que permite conocer la distribución de los datos con respecto a su media

RutasT\$TIEMPO..min.: Se obtiene el valor de la variable tiempo

Medidas de Dispersión

Las medidas de dispersión permiten conocer la variabilidad que existen entre los datos y sus variables dentro del conjunto de datos y entre las mismas variables. Se tomarán en cuenta las variables objetivas para reconocer su dispersión e influencia entre ellas, es decir, variable ruta, cajas, tarifa, viaje, distancia y tiempo a través de la varianza, desviación estándar, coeficiente de variación y curtosis.

Varianza

$$\text{var}(\text{RutasT\$RUTA})$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$RUTA: Se obtiene el valor de la variable ruta

$$\text{var}(\text{RutasT\$CAJAS})$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$CAJAS: Se obtiene el valor de la variable cajas

$$\text{var}(\text{RutasT\$TARIFA})$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$TARIFA: Se obtiene el valor de la variable tarifa

$$\text{var}(\text{RutasT\$VIAJE})$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$VIAJE: Se obtiene el valor de la variable viaje

$$\text{var}(\text{RutasT\$DISTANCIA..km.})$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$DISTANCIA..km.: Se obtiene el valor de la variable distancia

$$var(RutasT$TIEMPO..min.)$$

var: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$TIEMPO..min.: Se obtiene el valor de la variable tiempo

Desviación estándar

$$sd(RutasT$RUTA)$$

Sd: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$RUTA: Se obtiene el valor de la variable ruta

$$sd(RutasT$CAJAS)$$

Sd: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$CAJAS: Se obtiene el valor de la variable caja

$$sd(RutasT$TARIFA)$$

Sd: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$TARIFA: Se obtiene el valor de la variable tarifa

$$sd(RutasT$VIAJE)$$

sd: Función que permite conocer la distancia que ocupan los datos entre ellos

RutasT\$VIAJE: Se obtiene el valor de la variable viaje

$$sd(RutasT$DISTANCIA..km.)$$

sd: Función que permite conocer la distancia que ocupan los datos entre ellos
RutasT\$DISTANCIA..km.: Se obtiene el valor de la variable distancia

$$sd(RutasT$TIEMPO..min.)$$

sd: Función que permite conocer la distancia que ocupan los datos entre ellos
RutasT\$TIEMPO..min. : Se obtiene el valor de la variable tiempo

Coefficiente de variación

$$sd(RutasT$RUTA)/mean(RutasT$RUTA)$$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media
RutasT\$RUTA: Se obtiene el valor de la variable ruta

$$sd(RutasT$CAJAS)/mean(RutasT$CAJAS)$$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media
RutasT\$CAJAS: Se obtiene el valor de la variable cajas

$$sd(RutasT$TARIFA)/mean(RutasT$TARIFA)$$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media
RutasT\$TARIFA: Se obtiene el valor de la variable tarifa

$$sd(RutasT$VIAJE)/mean(RutasT$VIAJE)$$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media
RutasT\$VIAJE: Se obtiene el valor de la variable viaje

$$sd(RutasT$DISTANCIA..km.)/mean(RutasT$DISTANCIA..km.)$$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media
RutasT\$DISTANCIA..km.: Se obtiene el valor de la variable distancia

$sd(RutasT\$TIEMPO..min.)/mean(RutasT\$TIEMPO..min.)$

sd / mean: Función que permite conocer la dispersión de los datos con respecto a la media

RutasT\$TIEMPO..min. : Se obtiene el valor de la variable tiempo

Curtosis

$kurtosis(RutasT\$RUTA)$

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$RUTA: Se obtiene el valor de distribución de la variable ruta

$kurtosis(RutasT\$CAJAS)$

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$CAJAS: Se obtiene el valor de distribución de la variable cajas

$kurtosis(RutasT\$TARIFA)$

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$TARIFA: Se obtiene el valor de distribución de la variable tarifa

$kurtosis(RutasT\$VIAJE)$

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$VIAJE: Se obtiene el valor de distribución de la variable viaje

$kurtosis(RutasT\$DISTANCIA..km.)$

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$DISTANCIA..km.: Se obtiene el valor de distribución de la variable viaje

kurtosis(RutasT\$TIEMPO,,min.)

kurtosis: Describe la forma de las colas de una distribución en relación con una distribución normal

RutasT\$TIEMPO..min. : Se obtiene el valor de distribución de la variable tiempo

Gráficas de normalidad

Los gráficos de normalidad empleados se usarán para analizar si la distribución de los datos por cada variable está de forma equitativa o dispersa a través de barras que representan la frecuencia de los valores dentro de intervalos específicos.

Variable Ruta

```
G9 = ggplot(RutasT, aes(x = RutasT$RUTA)) +  
  geom_histogram(aes(y = ..density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$RUTA)), color = "red", size = 1.5) +  
  geom_vline(aes(xintercept = median(RutasT$RUTA)), color = "blue", size  
    = 1.5, linetype = "dashed") +  
  geom_vline(aes(xintercept = mfv(RutasT$RUTA)),  
    color = "yellow", size = 1.5) +  
  labs(title = "HISTOGRAMA DE RUTAS")
```

G9: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$RUTA: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density: Muestra la densidad de las observaciones

fill= "green": Establece el color de relleno de las barras del histograma a verde.

colours="black": Define el color del borde del histograma, en este caso, negro.

geom_density: Añade una curva de densidad suavizada sobre el histograma

alpha = 2: Establece la transparencia de la curva de densidad

fill= "orange": Establece el color de relleno de la curva de densidad a naranja.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$RUTA): Establece la posición de la línea vertical en la media (promedio) de los valores de RUTA.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$RUTA): Establece la posición de la línea vertical en la mediana de los valores de RUTA.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$RUTA): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de RUTA.

color="yellow", size=1.5: Establece el color de la línea a amarillo y el grosor de la línea a 1.5.

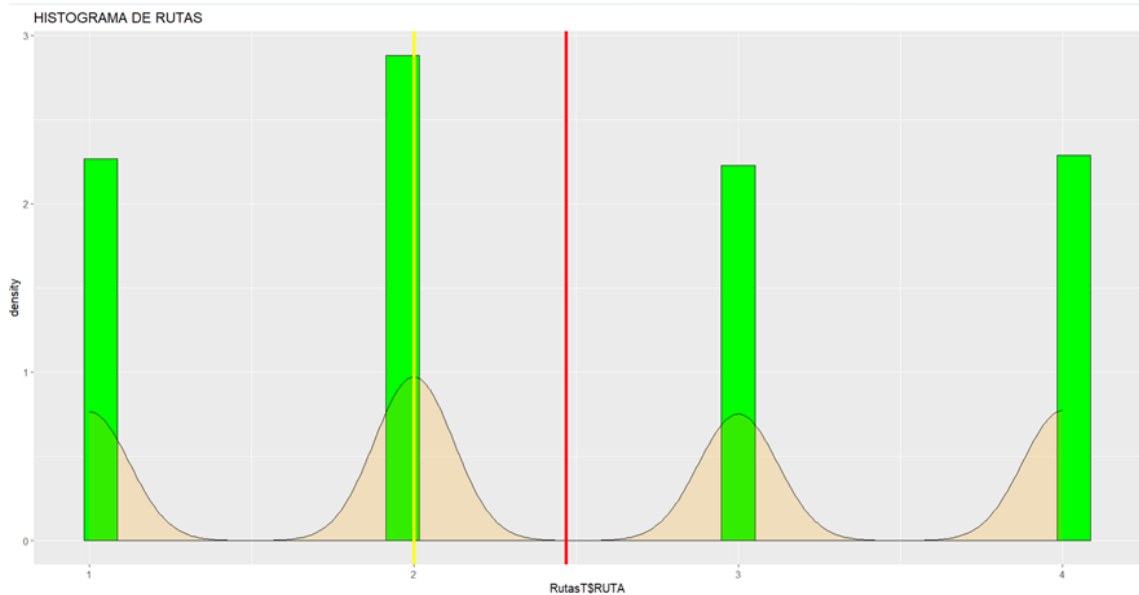
labs(): Añade etiquetas y títulos al gráfico

title= "HISTOGRAMA DE RUTAS": Establece el título del gráfico

G9: es la variable nueva creada que se llamará para visualizar

Figura 13

Gráfico de normalidad Variable Ruta - Histograma



Nota: Elaboración propia de los autores en el programa de R studio.

En este gráfico se observa las 4 rutas con sus frecuencias, la línea amarilla señala a la ruta 2 cómo la más frecuente o moda, la línea roja señala al punto medio de los datos entre la ruta 2 y 3 o media.

Variable Caja

```
G10 = ggplot(RutasT, aes(x = RutasT$CAJAS)) +  
  geom_histogram(aes(y = ..density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$CAJAS)), color = "red", size = 1.5) +  
  geom_vline(aes(xintercept = median(RutasT$CAJAS)), color = "blue", size  
    = 1.5, linetype = "dashed") +  
  geom_vline(aes(xintercept = mfv(RutasT$CAJAS)),  
    color = "yellow", size = 1.5) +  
  labs(title = "HISTOGRAMA DE CAJAS")
```

G10: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$CAJAS: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density: Muestra la densidad de las observaciones

fill= "green": Establece el color de relleno de las barras del histograma a verde.

colours="black": Define el color del borde del histograma, en este caso, negro.

geom_density: Añade una curva de densidad suavizada sobre el histograma

alpha = 2: Establece la transparencia de la curva de densidad

fill= "orange": Establece el color de relleno de la curva de densidad a naranja.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$CAJAS): Establece la posición de la línea vertical en la media (promedio) de los valores de CAJAS.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor de la línea a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$CAJAS): Establece la posición de la línea vertical en la mediana de los valores de CAJAS.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$CAJAS): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de CAJA.

color="yellow", size=1.5: Establece el color de la línea a amarillo y el grosor de la línea a 1.5.

labs(): Añade etiquetas y títulos al gráfico

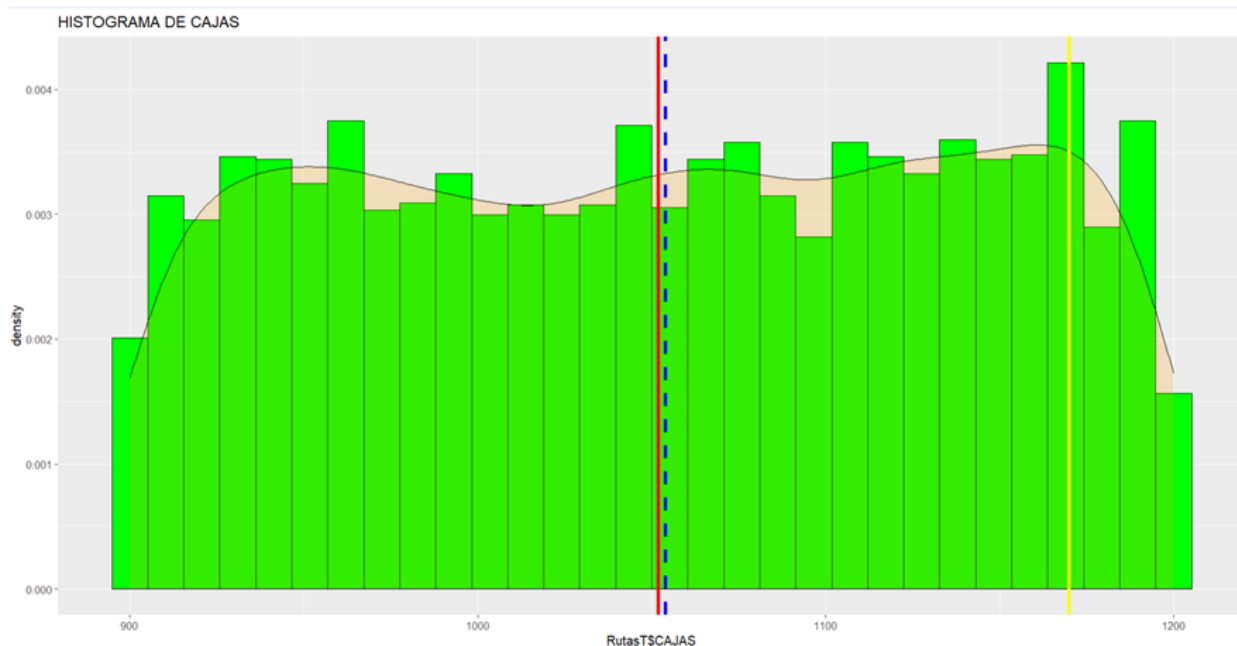
title= "HISTOGRAMA DE CAJAS": Establece el título del gráfico

G10

G10: Es la variable nueva creada que se llamará para visualizar

Figura 14

Gráfico de normalidad Variable Caja- Histograma



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se observa los 4 rangos de cantidad de cajas con sus frecuencias, la línea amarilla señala al rango de 1160 a 1170 de cajas cómo la frecuente o moda, la línea roja señala al punto medio de los datos entre el rango de 1050 cajas y las líneas azules señalan que la división equitativa de la muestra en entre las 1050 cajas.

Variable Tarifa

```
G11 = ggplot(RutasT, aes(x = RutasT$TARIFA)) +  
  geom_histogram(aes(y =..density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$TARIFA)), color = "red", size = 1.5) +
```

```
geom_vline(aes(xintercept = median(RutasT$TARIFA)), color = "blue", size
           = 1.5, linetype = "dashed") +
geom_vline(aes(xintercept = mfv(RutasT$TARIFA)),
           color = "yellow", size = 1.5) +
labs(title = "HISTOGRAMA DE TARIFA")
```

G11: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$TARIFA: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density: Muestra la densidad de las observaciones

fill= "green": Establece el color de relleno de las barras del histograma a verde.

colours="black": Define el color del borde del histograma, en este caso, negro.

geom_density: Añade una curva de densidad suavizada sobre el histograma

alpha = 2: Establece la transparencia de la curva de densidad

fill= "orange": Establece el color de relleno de la curva de densidad a naranja.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$TARIFA): Establece la posición de la línea vertical en la media (promedio) de los valores de TARIFA.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor de la línea a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$TARIFA): Establece la posición de la línea vertical en la mediana de los valores de TARIFA.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$TARIFA): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de TARIFA.

color="yellow", size=1.5: Establece el color de la línea a amarillo y el grosor de la línea a 1.5.

labs(): Añade etiquetas y títulos al gráfico

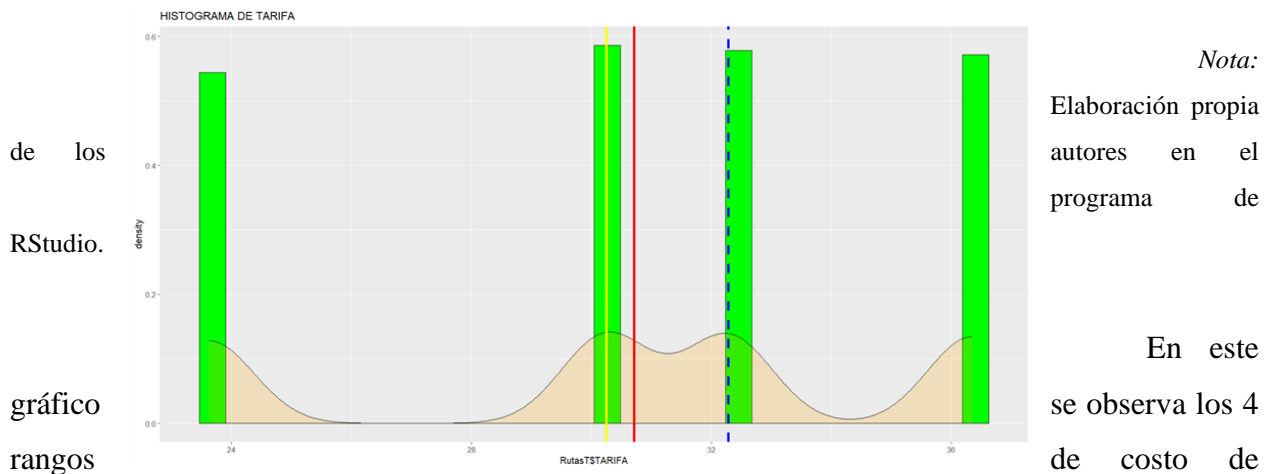
title= "HISTOGRAMA DE TARIFA": Establece el título del gráfico

G11

G11: Es la variable nueva creada que se llamará para visualizar

Figura 15

Gráfico de normalidad Variable Tarifa- Histograma



tarifa con sus frecuencias, la línea amarilla señala al rango de 29 a 30 dólares de costo de tarifa cómo la frecuente o moda, la línea roja señala al punto medio de los datos entre el rango de 31 dólares de costo de tarifa y las líneas azules señalan que la división equitativa de la muestra en 32 dólares de costo de tarifa.

Variable Viaje

```
G12 = ggplot(RutasT, aes(x = RutasT$VIAJE)) +  
  geom_histogram(aes(y =.. density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$VIAJE)), color = "red", size = 1.5) +
```

```
geom_vline(aes(xintercept = median(RutasT$VIAJE)), color = "blue", size
           = 1.5, linetype = "dashed") +
geom_vline(aes(xintercept = mfv(RutasT$VIAJE)),
           color = "yellow", size = 1.5) +
labs(title = "HISTOGRAMA DE VIAJE")
```

G12: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$VIAJE: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density : muestra la densidad de las observaciones

fill= "green" : establece el color de relleno de las barras del histograma a verde.

colours="black" : define el color del borde del histograma, en este caso, negro.

geom_density : añade una curva de densidad suavizada sobre el histograma

alpha = 2 : establece la transparencia de la curva de densidad

fill= "orange" : establece el color de relleno de la curva de densidad a naranja.

geom_vline: añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$VIAJE): Establece la posición de la línea vertical en la media (promedio) de los valores de VIAJE.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$VIAJE): establece la posición de la línea vertical en la mediana de los valores de VIAJE.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$VIAJE): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de VIAJE.

color="yellow", size=1.5: Establece el color de la línea a amarillo y el grosor de la línea a 1.5.

labs(): Añade etiquetas y títulos al gráfico

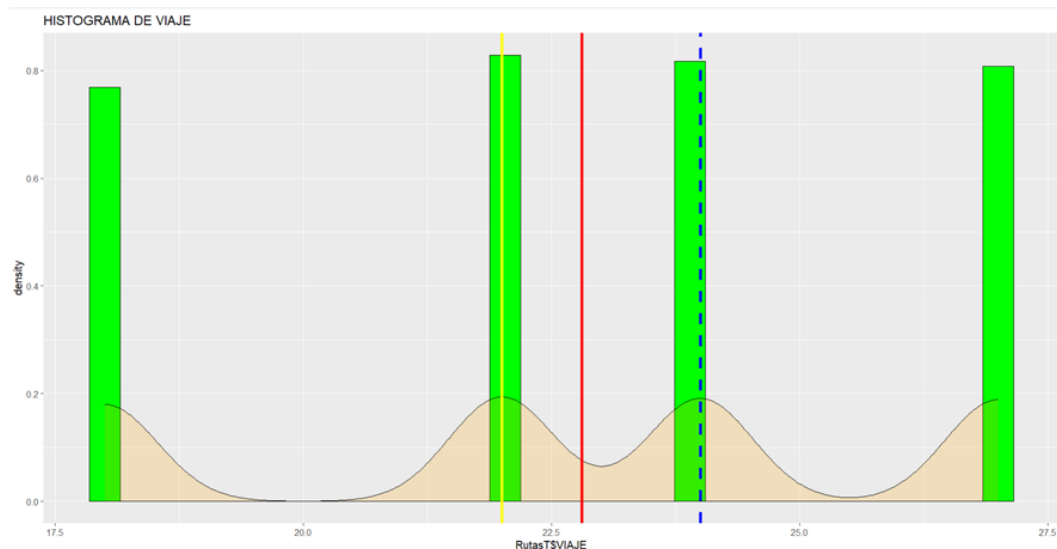
title= "HISTOGRAMA DE VIAJE": Establece el título del gráfico

G12

G12: es la variable nueva creada que se llamará para visualizar

Figura 16

Gráfico de normalidad Variable Viaje- Histograma



Nota: Elaboración propia de los autores en el programa de RSryudio.

En este gráfico se observa los 4 rangos de costo de viaje con sus frecuencias, la línea amarilla señala al rango de 21,50 dólares de costo de viaje cómo la frecuencia o moda, la línea roja señala al punto medio de los datos entre el rango de 23 dólares de costo de viaje y las líneas azules señalan que la división equitativa de la muestra en 24 dólares de costo de viaje.

Variable Distancia

```
G13 = ggplot(RutasT, aes(x = RutasT$DISTANCIA..km.)) +  
  geom_histogram(aes(y = ..density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$DISTANCIA..km.)), color = "red", size  
    = 1.5) +  
  geom_vline(aes(xintercept = median(RutasT$DISTANCIA..km.)), color = "blue", size  
    = 1.5, linetype = "dashed") +  
  geom_vline(aes(xintercept = mfv(RutasT$DISTANCIA..km.)),  
    color = "yellow", size = 1.5) +  
  labs(title = "HISTOGRAMA DE DISTANCIA")
```

G13: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$DISTANCIA: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density: Muestra la densidad de las observaciones

fill= "green": Establece el color de relleno de las barras del histograma a verde.

colours="black": Define el color del borde del histograma, en este caso, negro.

geom_density: Añade una curva de densidad suavizada sobre el histograma

alpha = 2: Establece la transparencia de la curva de densidad

fill= "orange": Establece el color de relleno de la curva de densidad a naranja.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$DISTANCIA): Establece la posición de la línea vertical en la media (promedio) de los valores de DISTANCIA.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$DISTANCIA): Establece la posición de la línea vertical en la mediana de los valores de DISTANCIA.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$DISTANCIA): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de DISTANCIA.

color="yellow", size=1.5: Establece color de la línea a amarillo y el grosor a 1.5.

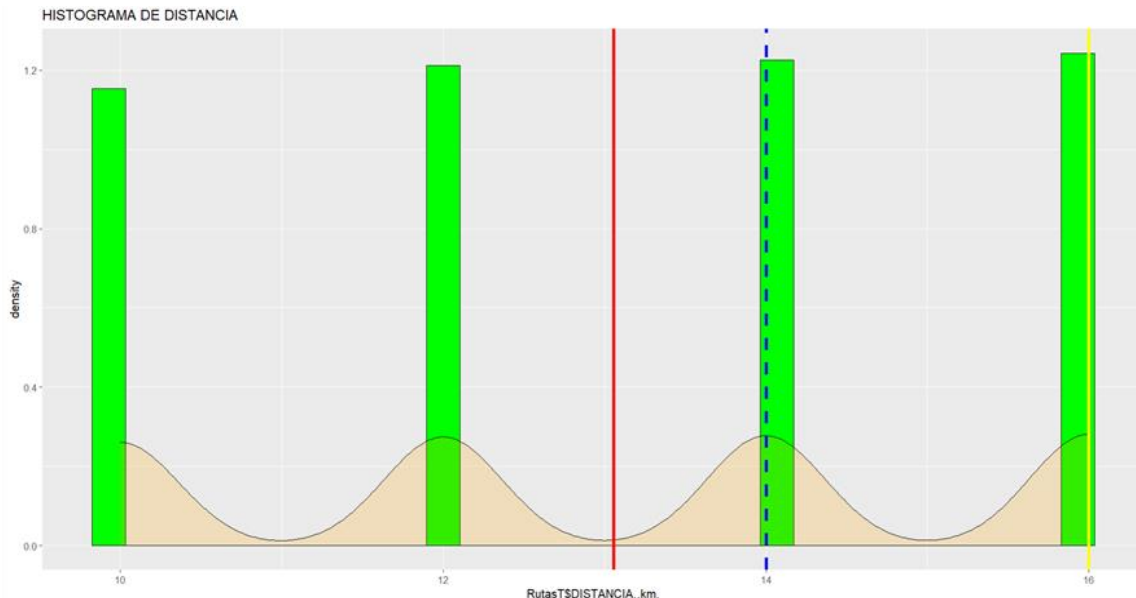
labs(): Añade etiquetas y títulos al gráfico

title= "HISTOGRAMA DE DISTANCIA": Establece el título del gráfico

G13

G13: Es la variable nueva creada que se llamará para visualizar **Figura 17**

Gráfico de normalidad Variable Distancia- Histograma



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se observa los 4 rangos de distancia por kilómetros con sus frecuencias, la

línea amarilla señala al rango de 16 kilómetros cómo la distancia frecuente o moda, la línea roja señala al punto medio de los datos entre el rango de 13 kilómetros de distancia y las líneas azules señalan que la división equitativa de la muestra es en 14 kilómetros de distancia.

Variable Tiempo

```
G14 = ggplot(RutasT, aes(x = RutasT$TIEMPO..min.)) +  
  geom_histogram(aes(y = ..density..), fill = "green", colour = "black") +  
  geom_density(alpha = .2, fill = "orange") +  
  geom_vline(aes(xintercept = mean(RutasT$TIEMPO..min.)), color = "red", size  
    = 1.5) +  
  geom_vline(aes(xintercept = median(RutasT$TIEMPO..min.)), color = "blue", size  
    = 1.5, linetype = "dashed") +  
  geom_vline(aes(xintercept = mfv(RutasT$TIEMPO..min.)),  
    color = "yellow", size = 1.5) +  
  labs(title = "HISTOGRAMA DE TIEMPO")
```

G14: La nueva variable creada de gráfica

ggplot: Es la función que permite crear un gráfico

RutasT: Es el data frame que contiene los datos que se van a graficar.

aes: Es la función que define las variables de los ejes x – y en el gráfico

x: Eje independiente

RutasT\$TIEMPO: Es la variable categórica que se está utilizando para agrupar los datos

geom_histogram: Realiza un histograma al gráfico

aes: Es la función que define las variables de los ejes x – y en el gráfico

y: Eje dependiente

density: Muestra la densidad de las observaciones

fill= "green": Establece el color de relleno de las barras del histograma a verde.

colours="black": Define el color del borde del histograma, en este caso, negro.

geom_density: Añade una curva de densidad suavizada sobre el histograma

alpha = 2: Establece la transparencia de la curva de densidad

fill= "orange": Establece el color de relleno de la curva de densidad a naranja.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mean(RutasT\$TIEMPO): Establece la posición de la línea vertical en la media (promedio) de los valores de DISTANCIA.

color="red", size=1.5: Establece el color de la línea a rojo y el grosor a 1.5.

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=median(RutasT\$TIEMPO): Establece la posición de la línea vertical en la mediana de los valores de DISTANCIA.

color="blue", size=1.5, linetype="dashed": Establece el color de la línea a azul, el grosor a 1.5 y el tipo de línea a discontinua (dashed).

geom_vline: Añade una línea vertical al gráfico.

aes: Es la función que define las variables de los ejes x – y en el gráfico

xintercept=mfv(RutasT\$TIEMPO): Establece la posición de la línea vertical en la moda (el valor más frecuente) de los valores de TIEMPO.

color="yellow", size=1.5: Establece el color de la línea a amarillo y el grosor de la línea a 1.5.

labs(): Añade etiquetas y títulos al gráfico

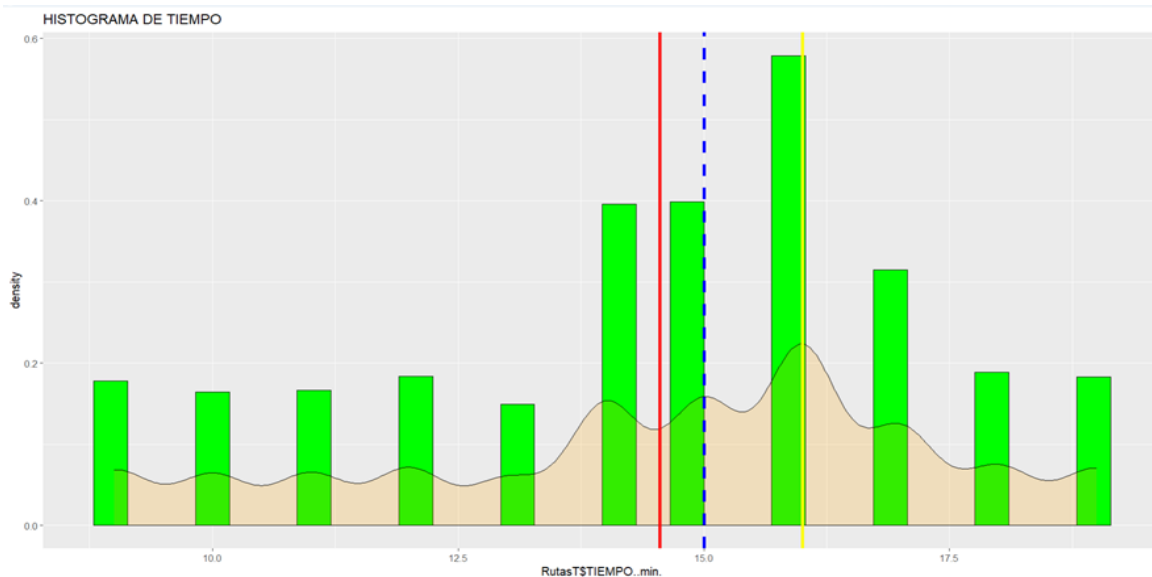
title= "HISTOGRAMA DE TIEMPO": Establece el título del gráfico

G14

G14: Es la variable nueva creada que se llamará para visualizar

Figura 18

Gráfico de normalidad Variable Tiempo- Histograma



Nota: Elaboración propia de los autores en el programa de RStudio.

En este gráfico se observa los 4 rangos de tiempo por minutos con sus frecuencias, la línea amarilla señala al rango de 16 minutos cómo el tiempo frecuente o moda, la línea roja señala al punto medio de los datos entre el rango de 14 minutos de tiempo y las líneas azules señalan que la división equitativa de la muestra es en 15 minutos de tiempo

Creación del Modelo de clasificación

```
set.seed(123)
```

set.seed(123): Es una función utilizada en R para establecer una semilla en la generación de números aleatorios. Esto asegura que los resultados de cualquier proceso aleatorio que se siga en dentro del script sean reproducibles. Este tipo de semilla es utilizada para la creación de modelos de clasificación.

```
Train = createDataPartition(RutasT$RUTA, p=.8, list = FALSE)
```

RutasT\$RUTA: Esto indica que estás accediendo a la columna RUTA de la data frame RutasT.

p = .8: Este argumento indica que el 80% de los datos se destinarán al conjunto de entrenamiento.

list = FALSE: Este argumento especifica que la salida será un vector en lugar de una lista. La función createDataPartition devuelve un vector de índices que corresponde a las filas seleccionadas para el conjunto de entrenamiento.

Árbol de Decisión

```
arbol = rpart(RUTA~., data = RutasT[Train, ],  
method = "class",  
control = rpart.control(minsplit = 300, cp = 0.01)
```

RUTA~.: Esto indica que RUTA es la variable dependiente que se está tratando de predecir, y el punto. significa que se utilizan todas las demás variables en RutasT como predictores.

data = RutasT[Train,]: Esta parte indica que los datos utilizados para entrenar el modelo

son aquellos en las filas especificadas por el índice `Train` del data frame `RutasT`. Esto es el subconjunto de entrenamiento que creaste anteriormente.

method = "class": Este argumento especifica que el modelo es un árbol de clasificación. Se utiliza cuando la variable dependiente (en este caso, `RUTA`) es categórica.

control = rpart.control(minsplit = 300, cp = 0.01): Este argumento define los parámetros de control para el ajuste del árbol:

minsplit = 300: Especifica el número mínimo de observaciones que deben estar presentes en un nodo para considerar la división del nodo.

cp = 0.01: Este es el parámetro de complejidad, que controla la poda del árbol. Un valor más alto de `cp` tiende a hacer que el árbol sea más simple (más podado), mientras que un valor más bajo permite que el árbol crezca más grande.

Este código está creando un árbol de decisión para clasificar la variable `RUTA` con base en las otras variables en el conjunto de datos `RutasT`, utilizando solo las filas seleccionadas en el subconjunto de entrenamiento.

Luego de este código escribimos `'arbol'` en la consola de R, de esta forma solicitamos la impresión del modelo que acabamos de crear. Esto mostrará un resumen del modelo de árbol de decisión, que incluye información como el número de divisiones (splits), los criterios utilizados para cada división, el número de observaciones en cada nodo, y más.

Lo cual nos otorga el siguiente resultado:

Figura 19

Resultado del árbol de decisión

```
n= 4003
node), split, n, loss, yval, (yprob)
* denotes terminal node
 1) root 4003 2810 2 (0.234574069 0.298026480 0.230577067 0.236822383)
 2) TARIFA< 26.94 946 28 2 (0.012684989 0.970401691 0.009513742 0.007399577) *
 3) TARIFA>=26.94 3057 2116 4 (0.303238469 0.089957475 0.298985934 0.307818122)
 6) TARIFA>=31.265 2038 1173 1 (0.424435721 0.105986261 0.422963690 0.046614328)
 12) PESO.NETO. .kg.< 20486.75 1013 253 1 (0.750246792 0.108588351 0.076011846 0.0651
53011) *
 13) PESO.NETO. .kg.>=20486.75 1025 240 3 (0.102439024 0.103414634 0.765853659 0.0282
92683) *
 7) TARIFA< 31.265 1019 173 4 (0.060843965 0.057899902 0.051030422 0.830225711) *
```

Nota: Elaboración propia de los autores en el programa de RStudio.

1. Nodo Raíz (Root):

n= 4003: El nodo raíz tiene 4003 observaciones.

2810: Es la cantidad de observaciones que se clasificarían incorrectamente si pararas en este nodo (sin hacer más divisiones).

yval=2: El valor predicho en este nodo es la clase 2.

(0.234574069 0.298026480 0.230577067 0.236822383): Estas son las probabilidades de pertenencia a cada clase en el nodo raíz, lo que significa que en este nodo, la clase 2 tiene la mayor probabilidad (alrededor de 29.8%).

2. Nodo 2 (TARIFA < 26.94):

Este nodo contiene 946 observaciones.

loss=28: Solo 28 observaciones se clasificarían incorrectamente.

yval=2: La clase 2 es la clase predicha.

(0.012684989 0.970401691 0.009513742 0.007399577): Alta probabilidad de que las observaciones pertenezcan a la clase 2 (97%).

*: El asterisco indica que es un nodo terminal, lo que significa que no hay más divisiones a partir de aquí.

3. Nodo 3 (TARIFA >= 26.94):

Contiene 3057 observaciones.

loss=2116: 2116 observaciones se clasificarían incorrectamente si se detuviera aquí.

yval=4: La clase 4 es la predicción en este nodo.

(0.303238469 0.089957475 0.298985934 0.307818122): Probabilidades de cada clase, siendo la clase 4 ligeramente la más probable.

Este nodo se divide en dos nodos hijos basados en otra condición (TARIFA >= 31.265).

4. Nodo 6 (TARIFA >= 31.265):

Contiene 2038 observaciones.

loss=1173: 1173 observaciones se clasificarían incorrectamente.

yval=1: La clase 1 es la predicción en este nodo.

(0.424435721 0.105986261 0.422963690 0.046614328): Probabilidad de pertenencia a cada clase, con la clase 1 siendo la más probable (42.4%).

Este nodo se divide en dos nodos hijos basados en PESO.NETO..kg.

5. Nodo 12 (PESO.NETO..kg. < 20486.75):

Contiene 1013 observaciones.

loss=253: 253 observaciones se clasificarían incorrectamente.

yval=1: La clase 1 es la predicción.

(0.750246792 0.108588351 0.076011846 0.065153011): Alta probabilidad (75%) de que las observaciones pertenezcan a la clase 1.

*: Este es un nodo terminal.

6. **Nodo 13 (PESO.NETO..kg. >= 20486.75):**

Contiene 1025 observaciones.

loss=240: 240 observaciones se clasificarían incorrectamente.

yval=3: La clase 3 es la predicción.

(0.102439024 0.103414634 0.765853659 0.028292683): Alta probabilidad (76.5%) de que las observaciones pertenezcan a la clase 3.

*: Este es un nodo terminal.

7. **Nodo 7 (TARIFA < 31.265):**

Contiene 1019 observaciones.

loss=173: 173 observaciones se clasificarían incorrectamente.

yval=4: La clase 4 es la predicción.

(0.060843965 0.057899902 0.051030422 0.830225711): Alta probabilidad (83%) de que las observaciones pertenezcan a la clase 4.

*: Este es un nodo terminal.

```
rpart.plot(arbol, type = 1, digits = -1,  
           extra = 0, cex = 0.7, nn = TRUE,  
           fallen.leaves = TRUE,)
```

arbol: Es el objeto de tipo rpart que contiene el árbol de decisión que se creó previamente.

type = 1: Este parámetro determina cómo se muestra la información en los nodos del árbol.

type = 1: muestra solo los nombres de las variables y los umbrales de decisión en los nodos internos, sin mostrar probabilidades o números de clase.

digits = -1: Este parámetro controla el número de decimales que se muestran en los números del gráfico.

digits = -1: indica que se muestran los números completos sin redondear.

extra = 0: Controla la información adicional que se muestra en los nodos.

extra = 0 significa que no se añade información adicional en los nodos, solo se muestran las divisiones.

cex = 0.7: Este parámetro ajusta el tamaño del texto en el gráfico.

cex = 0.7: reduce el tamaño del texto al 70% del tamaño predeterminado, lo que puede ser útil si el árbol es grande y necesitas que todo el texto quepa en la gráfica.

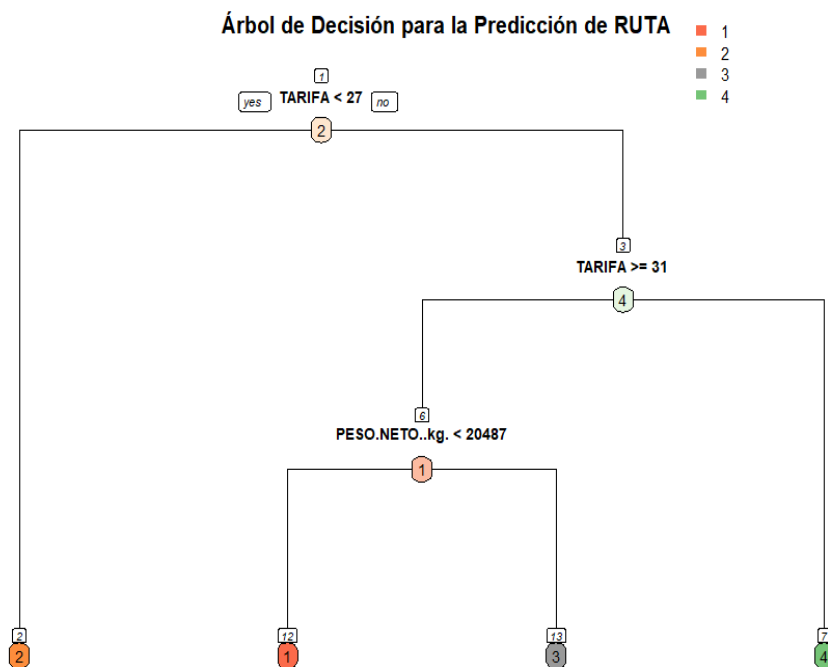
nn = TRUE: Esto agrega los números de nodo a los nodos del árbol, lo cual puede ser útil para referenciar nodos específicos.

fallen.leaves = TRUE: Hace que las hojas del árbol (los nodos terminales) estén alineadas en la parte inferior de la gráfica, lo que facilita su visualización.

Este comando generará un gráfico que visualiza el árbol de decisión con todas las divisiones y hojas alineadas en la parte inferior derecha de RStudio, con el tamaño del texto ajustado para facilitar la lectura.

Figura 20

Árbol de Decisión para predecir la ruta



Nota:
Elaboración propia de los autores en el programa de RStudio.

El título del árbol fue otorgado mediante la función:

title(main = "Árbol de Decisión para la Predicción de RUTA", cex.main = 1, col.main

= "black")

cex.main = 1: Usa el tamaño de texto predeterminado para el título, sin escalado adicional.

col.main = "black": El texto del título será de color negro, que es el color por defecto.

mean(RutasT\$TARIFA)

La función `mean(RutasT$TARIFA)` en R calcula el valor promedio de la columna TARIFA de la data frame RutasT

RutasT\$Prediccion: Aquí se crea una nueva columna llamada Prediccion en el data frame RutasT. Esta columna contendrá las predicciones generadas por el modelo.

predict(arbol, RutasT):

- `predict:` Esta función se utiliza para hacer predicciones basadas en un modelo que ya ha sido entrenado.
- `arbol:` Es el modelo de árbol de decisión que se creó anteriormente.
- `RutasT:` Es el nuevo conjunto de datos en el cual se está aplicando el modelo para hacer predicciones. Puede ser un conjunto de datos de prueba o un conjunto de datos completamente nuevo.

Bosque Aleatorio

Bosque = randomForest(x = RutasT[Train, 2: 10],

y = RutasT[Train, 1],

ntree = 10000, keep. forest = TRUE)

randomForest: Se utiliza para construir un modelo de bosque aleatorio.

x=RutasT[Train,2:10]: Especifica las variables predictoras para el modelo. Se están utilizando las columnas 2 a 10 del data frame RutasT para entrenar el modelo.

y=RutasT[Train,1]: Especifica la variable objetiva o de respuesta. En este caso, se utiliza la primera columna del data frame RutasT (es decir, RUTA) para predecir.

ntree = 10000: Establece el número de árboles en el bosque aleatorio. Es decir, se están construyendo 10,000 árboles.

keep.forest = TRUE: Mantiene el bosque aleatorio en el objeto resultante. Esto es útil para hacer análisis adicionales o visualizar el bosque después de la creación.

Bosque

Al momento de ejecutar el Bosque en R, se proporcionará un resumen del modelo, que incluye información sobre el rendimiento del modelo y los parámetros utilizados.

Figura 21

Resultado del Bosque

```
Call:
  randomForest(x = RutasT[Train, 2:10], y = RutasT[Train, 1], ntree = 10000,
    keep.forest = TRUE)
  Type of random forest: classification
    Number of trees: 10000
No. of variables tried at each split: 3

  OOB estimate of error rate: 17.34%
Confusion matrix:
  1  2  3  4 class.error
1 760 12 105 62 0.1906283
2 110 918 106 59 0.2305113
3 77 9 785 52 0.1495125
4 64 7 31 846 0.1075949
```

Nota: Elaboración propia de los autores en el programa de RStudio.

Type of random forest: classification: Indica que el modelo es utilizado para la clasificación.

Number of trees: 10000: El modelo utiliza 10,000 árboles en el bosque. Se utilizó un número alto de árboles para mejorar la estabilidad del modelo y la precisión.

No. of variables tried at each split: 3: En cada división de un árbol, se consideran 3 variables aleatorias para encontrar la mejor división.

OOB estimate of error rate: 17.34%: La tasa de error estimada utilizando la técnica Out-of-Bag. En este caso, el modelo tiene un error de clasificación del 17.34% en los datos no utilizados durante el entrenamiento de los árboles.

Confusion Matrix: Se refiere a la matriz de confusión, la cual muestra el rendimiento del modelo en términos de predicciones correctas e incorrectas para cada clase.

Tabla 1

Matriz de confusión

Predicciones	P1	P2	P 3	P4	Error de Clase
Clase 1	760	12	105	62	0.1906
Clase 2	110	918	106	59	0.2305
Clase 3	77	9	785	52	0.1495
Clase 4	64	7	31	846	0.1076

Nota: Elaboración propia de los autores en el programa de RStudio.

- **Valores Diagonales (760, 918, 785, 846):** Son las predicciones correctas para cada clase.
- **Valores Fuera de la Diagonal:** Representan los errores de clasificación, es decir, los casos donde una observación de una clase particular fue clasificada incorrectamente como otra clase.

Error de Clase:

- **Clase 1:** 19.06% de las observaciones de la clase 1 se clasificaron incorrectamente.
- **Clase 2:** 23.05% de las observaciones de la clase 2 se clasificaron incorrectamente.
- **Clase 3:** 14.95% de las observaciones de la clase 3 se clasificaron incorrectamente.
- **Clase 4:** 10.76% de las observaciones de la clase 4 se clasificaron incorrectamente.

ClassB= predict(Bosque, RutasT[-Train,])

predict(Bosque, ...): La función *predict* se utiliza para hacer predicciones utilizando el modelo Bosque que se ha entrenado.

RutasT[-Train,]: Este índice se refiere al subconjunto del data frame *RutasT* que no fue utilizado para el entrenamiento del modelo. *Train* es el índice de las filas utilizadas para entrenar el modelo, y *-Train* selecciona las filas restantes que no están en *Train*. Esto se utiliza para evaluar el rendimiento del modelo en datos que no se usaron durante el entrenamiento.

*MatrizTest = table(RutasT[-Train,"RUTA"],
ClassB,dnn = c("Actuales",*

"Predichos"))

table(...): La función table crea una tabla de contingencia, que en este caso se usa para crear una matriz de confusión.

RutasT[-Train, "RUTA"]: Selecciona la columna de clases verdaderas del conjunto de datos de prueba. Asume que "RUTA" es el nombre de la columna que contiene las clases verdaderas.

ClassB: Las clases predichas por el modelo para el conjunto de datos de prueba.

dnn = c("Actuales", "Predichos"): Etiqueta los ejes de la tabla. dnn especifica los nombres para las dimensiones de la tabla; en este caso, los ejes están etiquetados como "Actuales" (para las clases verdaderas) y "Predichos" (para las clases predichas).

MatrizTest

Esto mostrará la matriz de confusión que se ha creado:

Figura 22

Matriz de confusión del test

		Predichos			
Actuales		1	2	3	4
1	186	5	22	21	
2	33	235	19	11	
3	27	5	180	18	
4	11	2	6	217	

Nota: Elaboración propia de los autores en el programa de RStudio.

Valores Diagonales (186, 235, 180, 217): Son las predicciones correctas para cada clase.

Valores Fuera de la Diagonal: Representan los errores de clasificación, es decir, los casos donde una observación de una clase particular fue clasificada incorrectamente como otra clase.

Bosque\$confusion

Bosque\$confusion: Accede a la matriz de confusión asociada con el modelo Bosque, que es un objeto del tipo randomForest. La matriz ya se encuentra en la Figura 21.

```
par(mfrow = c(1,2))
```

par(mfrow=c(1,2)): Se utiliza para configurar el entorno gráfico para que muestre múltiples gráficos en una sola ventana. Específicamente, mfrow=c(1,2) configura la ventana gráfica para que muestre gráficos en una disposición de 1 fila y 2 columnas.

```
GME = mosaicplot(Bosque$confusion, type = n,  
  main = "Eficiencia del modelo – Entrenamiento",  
  color = c("yellow", "blue", "green", "red"))
```

mosaicplot: Se utiliza para crear un diagrama de mosaico, que es una visualización gráfica de una tabla de contingencia o matriz de confusión. Esto nos permite visualizar la relación entre las categorías de las filas y columnas en una matriz.

Bosque\$confusion: La matriz de confusión que se obtuvo del modelo Bosque.

type=n: Indica que se usarán los valores absolutos de la matriz en el diagrama de mosaico.

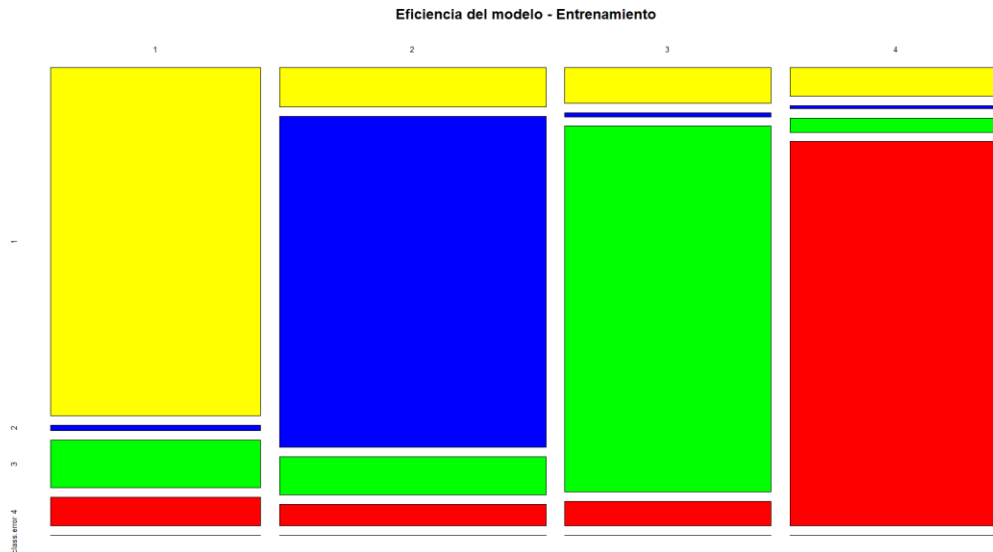
main = "Eficiencia del modelo - Entrenamiento": Le da el título del gráfico.

color = c("yellow", "blue", "green", "red"): Especifica los colores utilizados para las diferentes categorías en el diagrama de mosaico.

Al ejecutar el código nos muestra el siguiente gráfico:

Figura 23

Diagrama de mosaico del entrenamiento



Nota: Elaboración propia de los autores en el programa de RStudio.

```
GMT=mosaicplot(MatrizTest, type=n,  
main = "Eficiencia del modelo - Pruebas",  
color = c("yellow", "blue", "green", "red"))
```

MatrizTest: La matriz de confusión obtenida de tus predicciones en el conjunto de pruebas.

type=n: Indica que se deben usar los valores absolutos en el diagrama de mosaico. Esto muestra la frecuencia de las observaciones en cada celda de la matriz.

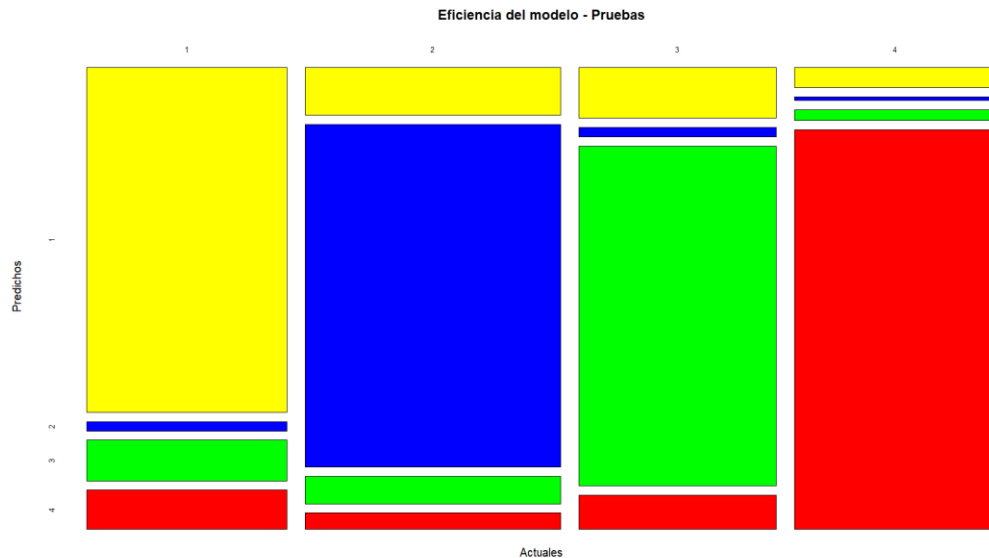
main = "Eficiencia del modelo - Pruebas": Título del gráfico que describe que el diagrama de mosaico representa la eficiencia del modelo en el conjunto de pruebas.

color = c("yellow", "blue", "green", "red"): Define los colores para las diferentes categorías en el diagrama de mosaico. Los colores ayudan a distinguir entre las diferentes clases.

Al ejecutar el código nos muestra el siguiente gráfico:

Figura 24

Diagrama de mosaico del test



Nota: Elaboración propia de los autores en el programa de RStudio.

`par(mfrow=c(1,1))`

par(mfrow=c(1,1)): Se utiliza para restablecer la configuración del entorno gráfico a una sola ventana, es decir, para mostrar un único gráfico en la ventana gráfica. Esto lo utilizamos, puesto que anteriormente usamos la función `par(mfrow=c(1,2))`, la cual que dividió la ventana en múltiples paneles.

`RutasT[, "prob"] = predict(Bosque, RutasT)`

RutasT[, "prob"]: Esto indica que se va a agregar una nueva columna llamada prob al dataframe RutasT.

predict(Bosque, RutasT): Este comando devuelve las predicciones del modelo para cada observación en RutasT. En el caso de clasificación, devolverá las clases predichas.

Bosque: Pertenece al modelo de bosque aleatorio creado con la función randomForest.

RutasT: Es el dataframe sobre el cual se desean hacer predicciones.

RutasT\$Probabilidad = predict(Bosque, RutasT,type = "prob")

RutasT\$Probabilidad: Esto agrega una nueva columna llamada Probabilidad al dataframe RutasT.

Bosque: Es el modelo de bosque aleatorio creado con la función randomForest.

RutasT: Es el dataframe sobre el cual estamos haciendo las predicciones.

type="prob": Indica que buscamos obtener las probabilidades de pertenecer a cada clase para cada observación, en lugar de solo las clases predichas.

Discusión

El script desarrollado abarca un análisis estadístico completo, desde el desarrollo del área descriptiva de las variables hasta la construcción del modelo de clasificación utilizando árboles de decisión y bosques aleatorios.

Iniciamos cargando nuestra base de datos denominada *RutasT*, luego realizamos una limpieza de datos retirando de esta forma variables que no nos servirían al momento de realizar nuestro modelo. Procedimos ejecutando el comando *cor(RutasT)* para poder realizar un análisis de correlación entre nuestras variables y comprobar la calidad de nuestros datos.

Se realizó un análisis de descriptivos con la finalidad de obtener una mejor visualización de nuestras observaciones y relaciones de variables, antes de poder realizar la construcción de nuestro modelo clasificador.

Tras definir nuestra variable de salida, plantamos la semilla que utilizaríamos para crear el modelo de clasificación, empezariamos con el entrenamiento de los datos antes de proceder a la creación de un árbol de decisión, visible en la Figura 19 el cual nos mostraría las variables utilizadas para al momento de clasificar. Tras de esto, generamos una nueva variable que nos mostraría la predicción realizada por el árbol para la asignación de la nueva clasificación de las rutas.

Creamos el bosque aleatorio, ya que la finalidad del trabajo es realizar la clasificación mediante el uso de Random Forest, para de esa forma obtener un modelo clasificador confiable y robusto. El modelo usó la variable RUTA como variable de salida, y tuvo una robustez de 10.000 árboles. El modelo tuvo un error de 17,34%, una tasa de error muy baja, y lo cual nos indicaba que este contaba con un margen de acierto bastante alto al momento de realizar su clasificación.

En pocas palabras, el modelo se desarrolló correctamente y posee la capacidad de clasificar de forma muy eficiente las rutas, con un margen de error bastante reducido, por lo cual presenta una alta confiabilidad en sus predicciones.

Conclusiones

Se puede dictaminar que efectivamente el uso de herramientas de Business Intelligence permiten la adquisición de modelos mucho más eficientes disminuyendo errores y mejorando la toma de decisiones a nivel empresarial en diversos ámbitos especialmente en económicos y logísticos, pero sobre todo óptimos.

En este caso de investigación, el enfoque principal fue optimizar rutas de transporte teniendo en consideración a sus indicadores cómo su tiempo, cantidad de entrega de cajas y costo por viaje, donde se usaron herramientas de Machine Learning por aprendizaje supervisado como Árbol de decisión y Bosques Aleatorios que permitirán la construcción de un modelo preciso y ajustado sobre la influencia de las múltiples variables objetivas para el resultado de la selección de solo una ruta favorable.

Además, el uso de herramientas de análisis descriptivo permitió reconocer las tendencias que surgen entre las variables objetivas con respecto a sus frecuencias de datos dentro de cada ruta y la dispersión presentada en razón de todas.

Como consecuente, el análisis descriptivo dio como resultado la existencia de 6 variables que crean una correlación de forma dependiente, es decir, variables objetivas de: rutas, cantidad de cajas, tiempo, distancia, costo de tarifa y viaje. Aquellas que fueron las analizadas para determinar cuál de las 4 rutas existentes representa una ruta eficiente y eficaz para el Centro de Acopio en cuestiones logísticas y económicas.

El análisis descriptivo presenta dos rutas como las posibles óptimas en función a todos los factores mencionados, la ruta 2 y ruta 4. La ruta 2 es considerada cómo la ruta más frecuente por los conductores debido a los siguientes factores: es la ruta que entrega cajas dentro del rango de 990 hasta 1050 por recorrido donde cada recorrido tiene una duración 12 minutos por una distancia de 10 kilómetros con un costo de viaje de 18 dólares por una tarifa de hasta 24 dólares. Es decir, la ruta presentaría una ventaja de preferencia frente a las demás rutas no solo por su tiempo corto, sino por el costo mínimo de recorrido de una tarifa total de 42 dólares

Mientras que, por otro lado, la ruta 4 es considerada la segunda mejor ruta por los siguientes factores: es la ruta que entrega cajas dentro del rango de 1150 hasta 1200 cajas por recorrido donde cada recorrido tiene una duración de hasta 18 minutos por una distancia de 16 kilómetros con un costo de viaje de hasta 22 dólares por una tarifa de hasta 30, 25 dólares. Es decir, la ruta presentaría únicamente una ventaja por cantidad de diferencia de cajas de 210 más que la ruta 2 durante el recorrido que realice. Dado que presenta un costo elevado de recorrido total de 52,25 dólares con una diferencia de 10 dólares hacia la ruta 2.

Por ello, este análisis descriptivo planteado en 1 hora y 12 minutos de recorrido de ambas rutas dictaminó que la ruta 2 es la más favorable, si se quiere disminuir costos de recorrido y obtener mayor cantidad de entrega de cajas. Durante 1 hora y 12 minutos de recorrido, la ruta 2 presenta el recorrido de 3 entregas de mercancía en la ida y vuelta desde el Centro de Acopio hasta el Puerto de Machala, en un tiempo de 24 minutos cada uno, y con un costo total por hora de 126 dólares y un total de cajas de 2970. Sin embargo, la ruta 4 permite realizar 2 recorridos de 36 minutos en ida y vuelta cada uno, con una entrega de mercancías de 2400 cajas, en diferencia a la ruta dos con 570 cajas, con un costo total de 104,50 dólares.

Por ello, se utilizó un modelo de clasificación más ajustado y óptimo que permita deducir que recorrido represente una mejor ventaja para el Centro de Acopio. El modelo de árbol de clasificación indica que, si queremos disminuir costos, la ruta 2 es la mejor, pero si queremos mayor cantidad de cajas en un solo recorrido, la ruta 4 sería la ideal. El bosque señala que el error de clasificación de la ruta 4 y 1 ha sido mayor, lo que podría crear un sesgo acerca la selección de ruta sin considerar otros factores a simple vista. Mientras que la ruta 2 cuenta con menos errores dando como resultado que sea la más eficaz y eficiente a nivel empresarial.

Se concluye, que el modelo no solamente permitió ajustar las rutas, sino que pudo derivar todos los errores de clasificación de rutas por cantidad de cajas en la capacidad que cada una tiene para la entrega. En cuestiones de logística y tiempo, la opción ideal será la ruta 2 incluso cuándo a primera impresión se crea que no entrega suficientes cajas. Ya que la optimización y estrategias en la cadena de valor se da por encontrar las ventajas que los demás factores externos pueden brindar hacia un solo factor deseado.

Referencias

- Alpaydin, E. (2020). *Introduction to Machine Learning, fourth edition*. MIT Press.
- AMAZON. (2024). *¿Qué es el machine learning? - Explicación sobre el machine learning empresarial - AWS*. Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/machine-learning/>
- APD. (2024, julio 18). *¿Cuáles son los tipos de algoritmos del machine learning? APD España*. <https://www.apd.es/algoritmos-del-machine-learning/>
- Ayala Villareal, J. G., & Nazate Salazar, S. A. (2020). *Centro de acopio rural sustentable de productos agrícolas para la parroquia de la Concepción del cantón Mira, provincia del Carchi* [Pontificia Universidad Católica del Ecuador Ibarra]. <https://repositorio.puce.edu.ec/handle/123456789/40279>
- Bektaş, T., & Laporte, G. (2011). The Pollution-Routing Problem. *Transportation Research Part B: Methodological*, 45(8), 1232-1250. <https://doi.org/10.1016/j.trb.2011.02.004>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1.^a ed.). Springer-Verlag. <https://link.springer.com/book/9780387310732>
- Bojic, S., Zrnica, N., Rajkovic, R., & Dragovic, B. (2020). Optimization of container transport routes. *Prosperitas*, 7, 31-42. https://doi.org/10.31570/Prosp_2020_01_3
- Boldyrieva, L., Zelinska, H., Krapkina, V., & Komelina, A. (2019). *Problems and Solutions of Transport Logistics*. 317-320. <https://doi.org/10.2991/mdsmes-19.2019.59>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Calvopiña Llambo, A., Cajilema Herrera, L., Ramírez Ramírez, B., & Guerrero Pinela, R. (2023). Crisis de seguridad en Ecuador y autorización de uso civil para tenencia y porte de armas. *Polo del Conocimiento: Revista científico - profesional*, 8(5), 373-384.

- Casanova Lugo, C., & Torres Anzola, M. Y. (2020). *Propuesta documental para un modelo de gestión del riesgo en los laboratorios de ambiental y suelos del CDTI (Centro de Desarrollo Tecnológico e Innovación) de la Facultad de Ingeniería, Universidad El Bosque Bogotá*. <https://alejandria.poligran.edu.co/handle/10823/2722>
- Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36, 1165-1188. <https://doi.org/10.2307/41703503>
- Chicaiza, J., & Sandaya, F. (2015). La Investigación en Logística y Transporte: Comparación entre los países de la región Andina; Retos y Oportunidades para su desarrollo en el Ecuador. *Congreso de Ciencia y Tecnología ESPE*, 10(1), Article 1. <https://doi.org/10.24133/cctespe.v10i1.60>
- Christopher, M. (1994). Logistics and Supply Chain Management: Strategies for Reducing Costs and Improving Services. *Journal of the Operational Research Society*, 46. <https://doi.org/10.1057/jors.1994.209>
- Consejo de la Judicatura. (2024). https://eloro.funcionjudicial.gob.ec/index.php?option=com_content&view=article&id=1150:judicatura-de-el-oro-cuenta-centro-de-acopio-para-damnificados-de-terremoto&catid=39:noticias-home
- Coursera. (2023, junio 15). *7 algoritmos de machine learning que hay que conocer: Guía para principiantes*. Coursera. <https://www.coursera.org/mx/articles/machine-learning-algorithms>
- Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P. (2009). The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 153-160.

- <https://proceedings.mlr.press/v5/erhan09a.html>
- Eslava, J. de J. (2003). *Análisis económico-financiero de las decisiones de gestión empresarial*. Esic Editorial.
- Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, 21(3). <https://doi.org/10.22201/fi.25940732e.2020.21.3.022>
- Ferrovial. (2024). *Qué es y para qué sirve la inteligencia artificial (IA)*. Ferrovial. <https://www.ferrovial.com/es/recursos/inteligencia-artificial/>
- Fideicomiso de Riesgo. (2017, junio 27). *¿Sabes qué es un centro de acopio de alimentos y mermas?* gob.mx. <http://www.gob.mx/firco/articulos/sabes-que-es-un-centro-de-acopio-de-alimentos-y-mermas?idiom=es>
- Fonseca-Reyna, Y. C., Martínez-Jiménez, Y., & Nowé, A. (2018). Aprendizaje reforzado aplicado a la programación de tareas bajo condiciones reales. *Ingeniería Industrial*, 39(1), 36-45.
- Gendreau, M., Laporte, G., & Potvin, J.-Y. (2018). 9. Vehicle routing: Modern heuristics. En E. Aarts & J. K. Lenstra (Eds.), *Local Search in Combinatorial Optimization* (pp. 311-336). Princeton University Press. <https://doi.org/10.1515/9780691187563-012>
- Grigalunas, T., Luo, M., Trandafir, S., Anderson, C., & Kwon, S. (2007). *Issues in Container Transportation in the Northeast: Background, Framework, Illustrative Results and Future Directions*.
- Guaña Moya, E. J., & Salgado Reyes, N. E. (2019). *Análisis de causas de accidentes de tránsito en el Ecuador utilizando Minería de Datos*. <http://pucedspace.puce.edu.ec:80/handle/23000/3538>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data*

Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics).

Higginson, J. K., & Bookbinder, J. H. (2005). Distribution Centres in Supply Chain Operations.

En A. Langevin & D. Riopel (Eds.), *Logistics Systems: Design and Optimization* (pp. 67-91). Springer US. https://doi.org/10.1007/0-387-24977-X_3

IBM. (2023, agosto 25). *What is Artificial Intelligence (AI)?* | IBM. IBM. <https://www.ibm.com/topics/artificial-intelligence>

Inteligencia Artificial en Logística: Qué Debes Saber. (2024, mayo 27). Simpliroute. <https://simpliroute.com/es/blog/inteligencia-artificial-en-logistica>

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260. <https://doi.org/10.1126/science.aaa8415>

Kang, Y., Lee, S., & Chung, B. D. (2019). Learning-based logistics planning and scheduling for crowdsourced parcel delivery. *Computers & Industrial Engineering*, 132, 271-279. <https://doi.org/10.1016/j.cie.2019.04.044>

Košiček, M., Tesař, R., Dařena, F., Malo, R., & Motycka, A. (2012). Route planning module as a part of Supply Chain Management system. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 60, 135-142. <https://doi.org/10.11118/actaun201260020135>

La economía de Machala se mueve al ritmo de sus operaciones portuarias. (2023, septiembre 24). Diario Correo | El Diario de Todos. <https://diariocorreo.com.ec/88799/ciudad/la-economia-de-machala-se-mueve-al-ritmo-de-sus-operaciones-portuarias>

Liaw, A., & Wiener, M. (2001). Classification and Regression by RandomForest. *Forest*, 23.

Maglogiannis, I. G. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies.* IOS Press.

- Manzanilla, V. H. (2022, octubre 3). Importancia de la Logística en una Empresa · [2024]. *Emprendedor Growth Model*. <https://metodoegm.com/emprendimiento/cual-es-la-importancia-de-la-logistica-en-una-empresa/>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Oracle. (2024). *What is Machine Learning?* Oracle. <https://www.oracle.com/artificial-intelligence/machine-learning/what-is-machine-learning/>
- Pioneros de la actividad bananera en El Oro*. (2020, junio 24). Pioneros de La Actividad Bananera En El Oro. <https://www.diariocorreo.com.ec/47296/ciudad/pioneros-de-la-actividad-bananera-en-el-oro>
- ¿Qué es la optimización de rutas y por qué es importante? (2022, mayo 20). *Yobel Ecuador somos líderes en Supply Chain Management*. <https://www.yobelscm.biz/ecuador/optimizacion-de-rutas-de-distribucion>
- Ramirez, G. (2023, noviembre 16). Las empresas ecuatorianas apuestan por la inteligencia artificial para impulsar su crecimiento. *IT ahora*. <https://itahora.com/2023/11/16/las-empresas-ecuatorianas-apuestan-por-la-inteligencia-artificial-para-impulsar-su-crecimiento/>
- Repsol. (2023). *¿Qué es la inteligencia artificial y cómo nos ayuda?* REPSOL. <https://www.repsol.com/es/energia-futuro/tecnologia-innovacion/inteligencia-artificial/index.cshtml>
- Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, 47(1), 31-39. <https://doi.org/10.17849/in-sm-47-01-31-39.1>
- Rodriguez, J.-P., Comtois, C., & Slack, B. (2006). *The geography of transport systems*. Routledge.
- Rodriguez, J.-P., & Notteboom, T. (2015). *Containerization, Box Logistics and Global Supply*

- Chains: The Integration of Ports and Liner Shipping Networks* (pp. 5-28).
https://doi.org/10.1057/9781137475770_2
- Schneider, W., Miller, T. M., Holik, W. A., & University of Akron. (2016). *Route optimization*. (FHWA/OH-2016-20). <https://rosap.ntl.bts.gov/view/dot/31681>
- Smith, A. C., Hafner, R. S., & Gupta, N. K. (2004). *Transportation, storage, and disposal of radioactive materials--2004: Presented at the 2004 ASME/JSME Pressure Vessels and Piping Conference : San Diego, California, USA, July 25-29, 2004 / sponsored by the Pressure Vessels and Piping Division, ASME ; principal editors, A.C. Smith, R.S. Hafner ; contributing editors, N.K. Gupta [and others]*.
- Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, 105(12), 2295-2329. *Proceedings of the IEEE*. <https://doi.org/10.1109/JPROC.2017.2761740>
- Valencia, Z. A. D. (2004). Regulación de los servicios de transporte en Colombia y Comercio Internacional. *Archivos de Economía*, Article 3443. <https://ideas.repec.org/p/col/000118/003443.html>
- Valverde, C. (2024). *UNA CIUDAD MEDIA DEL ECUADOR: MACHALA. PROPUESTA DIGNOSITIVA*.
<http://observatoriogeograficoamericalatina.org.mx/egal9/Geografiasocioeconomica/Geografiaurbana/02.pdf>
- Wiederhold, G., & McCarthy, J. (1992). Arthur Samuel: Pioneer in Machine Learning. *IBM Journal of Research and Development*, 36, 329-331. <https://doi.org/10.1147/rd.363.0329>

Anexos

```
1 Rutasa1=read.csv2("../TESIS RSTUDIO/DATA/rutasa1.csv")
2
3 RutasT <- Rutasa1[,c(-8,-9)]
4 cor(RutasT)
5
6 ##### TABLA DE FRECUENCIA DATOS NO AGRUPADOS #####
7
8 #VI RUTA
9 table (RutasT $RUTA)
10
11 Fre_ruta = as.data.frame(table(RutasT$RUTA))
12 Fre_ruta
13
14 ##### TABLA FRECUENCIA VARIABLE RUTA #####
15
16 Tab_Fre_ruta = transform(Fre_ruta,
17                           FreI =round(prop.table(Fre_ruta$Freq),3),
18                           FAcu = cumsum(Fre_ruta$Freq),
19                           FAcuR = cumsum(round(prop.table(Fre_ruta$Freq),3)))
20
21 Tab_Fre_ruta
22
23
24 ##### TABLA FRECUENCIA VARIABLE CAJA #####
25
26
27 x1 = RutasT$CAJAS
28 #aplico la regla de Sturges para D.A
29
30 k1 = nclass.Sturges(x1)
31 k1
32
33 #la regla de sturges es para hacer los intervalos que existen, en este caso 9
34 #calculo la amplitud luego de hacer los intervalos
35
36 Int1 = cut(x1, breaks = k1)
37 Int1
38
39 Fre_caja = as.data.frame(table(Int1))
40 Fre_caja
41
42 ##### TABLA FRECUENCIA CAJA #####
43
482:39 ## GRAFICAMOS EL ARBOL
```

Console

```
41
42 ##### TABLA FRECUENCIA CAJA #####
43
44 Tab_caja = transform(Fre_caja,
45                      Frel = round(prop.table(Fre_caja$Freq), 3),
46                      FAcu = cumsum(Fre_caja$Freq),
47                      FAcuR = cumsum(round(prop.table(Fre_caja$Freq), 3)))
48
49 Tab_caja
50
51
52
53 #coord_flip sirve para poner rangos de valores más específicos en el gráfico
54
55
56 #V3 TARIFA
57 table (RutasT $TARIFA)
58
59 Fre_tarifa = as.data.frame(table(RutasT$TARIFA))
60 Fre_tarifa
61
62 ##### TABLA FRECUENCIA VARIABLE TARIFA #####
63
64 Tab_Fre_tarifa = transform(Fre_tarifa,
65                            Frel = round(prop.table(Fre_tarifa$Freq), 3),
66                            FAcu = cumsum(Fre_tarifa$Freq),
67                            FAcuR = cumsum(round(prop.table(Fre_tarifa$Freq), 3)))
68
69 Tab_Fre_tarifa
70
71
72 #V4 VIAJE
73 table (RutasT $VIAJE)
74
75 Fre_viaje = as.data.frame(table(RutasT$VIAJE))
76 Fre_viaje
77
78 ##### TABLA FRECUENCIA VARIABLE VIAJE #####
79
80 Tab_Fre_viaje = transform(Fre_viaje,
81                           Frel = round(prop.table(Fre_viaje$Freq), 3),
82                           FAcu = cumsum(Fre_viaje$Freq),
83
482:39 # GRAFICAMOS EL ARBOL
Console
```

```
61 |
62 | ##### TABLA FRECUENCIA VARIABLE TARIFA #####
63 |
64 | Tab_Fre_tarifa = transform(Fre_tarifa,
65 |                           Frel =round(prop.table(Fre_tarifa$Freq),3),
66 |                           FAcu = cumsum(Fre_tarifa$Freq),
67 |                           FAcuR = cumsum(round(prop.table(Fre_tarifa$Freq),3)))
68 |
69 | Tab_Fre_tarifa
70 |
71 |
72 | #V4 VIAJE
73 | table (RutasT $VIAJE)
74 |
75 | Fre_viaje = as.data.frame(table(RutasT$VIAJE))
76 | Fre_viaje
77 |
78 | ##### TABLA FRECUENCIA VARIABLE VIAJE #####
79 |
80 | Tab_Fre_viaje = transform(Fre_viaje,
81 |                           Frel =round(prop.table(Fre_viaje$Freq),3),
82 |                           FAcu = cumsum(Fre_viaje$Freq),
83 |                           FAcuR = cumsum(round(prop.table(Fre_viaje$Freq),3)))
84 |
85 | Tab_Fre_viaje
86 |
87 |
88 | #V5 DISTANCIA
89 | table (RutasT $DISTANCIA..km.)
90 |
91 | Fre_distancia= as.data.frame(table(RutasT$DISTANCIA..km.))
92 |
93 | Fre_distancia
94 |
95 | ##### TABLA FRECUENCIA VARIABLE DISTANCIA #####
96 |
97 | Tab_Fre_distancia = transform(Fre_distancia,
98 |                               Frel =round(prop.table(Fre_distancia$Freq),3),
99 |                               FAcu = cumsum(Fre_distancia$Freq),
100 |                               FAcuR = cumsum(round(prop.table(Fre_distancia$Freq),3)))
101 |
102 | Tab Fre distancia
103 |
```

61:1 # TABLA FRECUENCIA CAJA : R Script

Console

```

101
102 Tab_Fre_distancia
103
104
105
106 #V6 TIEMPO
107 table (RutasT $TIEMPO..min.)
108
109 Fre_tiempo = as.data.frame(table(RutasT$TIEMPO..min.))
110 Fre_tiempo
111
112 ▾ ##### TABLA FRECUENCIA VARIABLE TIEMPO #####
113
114 Tab_Fre_tiempo = transform(Fre_tiempo,
115                             Fre1 =round(prop.table(Fre_tiempo$Freq),3),
116                             FAcu = cumsum(Fre_tiempo$Freq),
117                             FAcuR = cumsum(round(prop.table(Fre_tiempo$Freq),3)))
118
119 Tab_Fre_tiempo
120
121
122 ▾ ##### GRÁFICO DE BARRA POR NÚMEROS #####
123
124 library(ggplot2)
125 ▾ ##### GRÁFICA 1 DE FREQ DE V.RUTA #####
126
127
128 G1 <- ggplot(Tab_Fre_ruta, aes(x=Tab_Fre_ruta$Var1,
129                               y=Tab_Fre_ruta$Fre1))+geom_bar(stat = "identity",
130                               fill= "pink",
131                               colour="black", size=0.5)+geom_text(aes(label= paste0(
132                               position = position_stack(vjust = 0.8))+
133                               labs(title = "Frecuencia por RUTA")
134
135 G1
136
137 ▾ ##### GRÁFICA 2 DE FREQ DE V.CAJA #####
138
139 G2 <- ggplot(Tab_caja, aes(x=Tab_caja$Int1,
140                               y=Tab_caja$Fre1))+geom_bar(stat = "identity",
141                               fill= "skyblue", colour="black", size=0.5)+
142 geom_text(aes(label= paste0(Tab_caja$Int1,
143                               ":", Tab_caja$Fre1)))
144
145

```

61:1 # TABLA FRECUENCIA CAJA ▾ R Script ▾

Console

```
136
137 ◦ ##### GRÁFICA 2 DE FREQ DE V. CAJA #####
138
139 G2 <- ggplot(Tab_caja, aes(x=Tab_caja$Int1,
140   y=Tab_caja$Frel))+geom_bar(stat = "identity",
141   fill= "skyblue", colour="black", size=0.5)+
142   geom_text(aes(label= paste0(Tab_caja$Frel)),
143   position = position_stack(vjust = 0.8))+
144   coord_flip()+labs(title = "Frecuencia de CAJAS")
145 G2
146
147 ◦ ##### GRÁFICA 3 DE FREQ DE V. TARIFA #####
148
149
150 G3 <- ggplot(Tab_Fre_tarifa, aes(x=Tab_Fre_tarifa$Var1,
151   y=Tab_Fre_tarifa$Frel))+geom_bar(stat = "identity",
152   fill= "skyblue",
153   colour="red", size=0.5)+geom_text(aes(label= paste0(Tab_Fre_tarifa$Frel)),
154   labs(title = "Frecuencia por TARIFA")
155 G3
156
157
158 ◦ ##### GRÁFICA 4 DE FREQ DE V. VIAJE #####
159
160
161 G4 <- ggplot(Tab_Fre_viaje, aes(x=Tab_Fre_viaje$Var1,
162   y=Tab_Fre_viaje$Frel))+geom_bar(stat = "identity",
163   fill= "yellow",
164   colour="orange", size=0.5)+geom_text(aes(label= paste0(Tab_Fre_viaje$Frel)),
165   labs(title = "Frecuencia por VIAJE")
166 G4
167
168
169 ◦ ##### GRÁFICA 5 DE FREQ DE V. DISTANCIA #####
170
171 G5 <- ggplot(Tab_Fre_distancia, aes(x=Tab_Fre_distancia$Var1,
172   y=Tab_Fre_distancia$Frel))+geom_bar(stat = "identity",
173   fill= "pink",
174   colour="purple", size=0.5)+geom_text(aes(label= paste0(Tab_Fre_distancia$Frel)),
175   labs(title = "Frecuencia por DISTANCIA")
176 G5
177
178
```

61:1 # TABLA FRECUENCIA CAJA

R Script

Console

```
175     labs(title = "Frecuencia por DISTANCIA")
176 G5
177
178
179
180 ▾ ##### GRÁFICA 6 DE FREQ DE V. TIEMPO #####
181
182
183 G6 <- ggplot(Tab_Fre_tiempo, aes(x=Tab_Fre_tiempo$Var1,
184                               y=Tab_Fre_tiempo$Frel))+geom_bar(stat = "identity",
185                               fill= "green",
186                               colour="lightblue", size=0.5)+geom_text(aes(label= paste0(Tab_Fre_tiempo$Frel)),
187                               position = position_stack(vjust = 0.8))+
188     labs(title = "Frecuencia por TIEMPO")
189 G6
190
191
192 ▾ ##### G. NUMÉRICA CON NUMÉRICA RUTAS X CAJA #####
193
194 ✘ BOX PLOT 1
195 ✘ BOX PLOT 2
196 AGREGAR
197
198 ▾ ##### G. NUMÉRICA CON NUMÉRICA DISTANCIA X TIEMPO #####
199
200 G7 = ggplot(RutasT, aes(x= RutasT$DISTANCIA..km.,
201                       y= RutasT$TIEMPO..min.,
202                       colour= RutasT$RUTA,
203                       size= RutasT$CAJAS))+
204     geom_point()+
205     labs(title = "Distancia x Tiempo")
206
207 G7
208
209
210
211 ▾ ##### MEDIDAS DE TENDENCIA CENTRAL #####
212
213 #instalar paquete de moda
214
215 library(modeest)
216 library(moments)
217
```

61:1 # TABLA FRECUENCIA CAJA

Console


```
216 library(moments)
217
218 #la media con la dispersión permiten clasificar valores y hacer pronósticos
219 #no existe dispersión en una V. categorica
220
221 #MEDIA
222
223 #Media de los datos
224 #el valor que sale es la cantidad de distancia entre un dato al otro
225
226
227 mean(RutasT$RUTA) #2.469506
228 mean(RutasT$CAJAS) #1051.917
229 mean(RutasT$TARIFA) #30.71506
230 mean(RutasT$VIAJE) #22.80664
231 mean(RutasT$DISTANCIA..km.) #13.05859
232 mean(RutasT$TIEMPO..min.) #14.55389
233
234
235 #MEDIANA
236
237 #Mediana de los datos
238 #es el valor en medio del conjunto de datos
239
240 median(RutasT$RUTA) #2
241 median(RutasT$CAJAS) #1054
242 median(RutasT$TARIFA) #32,28
243 median(RutasT$VIAJE) #24
244 median(RutasT$DISTANCIA..km.) #14
245 median(RutasT$TIEMPO..min.) #15
246
247
248
249 #MODA
250
251 #moda de los datos
252
253 mfV(RutasT$RUTA) #2
254 mfV(RutasT$CAJAS) #1170
255 mfV(RutasT$TARIFA) #30,25
256 mfV(RutasT$VIAJE) #22
257 mfV(RutasT$DISTANCIA..km.) #13
258 mfV(RutasT$TIEMPO..min.) #14
```

61:1 # TABLA FRECUENCIA CAJA

Console

```
254 mfV(RutasT$TARIFA) #22
256 mfV(RutasT$VIAJE) #22
257 mfV(RutasT$DISTANCIA..km.) #16
258 mfV(RutasT$TIEMPO..min.) #16
259
260
261 #COEFICIENTE DE ASIMETRÍA
262
263 #coeficiente de asimetría
264 #este valor sale positivo 50 es asimetría positiva hacia la derecha por 1,02
265
266 skewness(RutasT$RUTA) #0.08303598
267 skewness(RutasT$CAJAS) #-0.03711119
268 skewness(RutasT$TARIFA) #-0.4170569
269 skewness(RutasT$VIAJE) #-0.2281459
270 skewness(RutasT$DISTANCIA..km.) # -0.03324476
271 skewness(RutasT$TIEMPO..min.) # -0.4547189
272
273
274 ##### MEDIDAS DE DISPERSIÓN #####
275
276 #VARIANZA
277
278 #varianza de los datos, #la distancia que hay entre estos
279
280 var(RutasT$RUTA) #1.19192
281 var(RutasT$CAJAS) #7658.765
282 var(RutasT$TARIFA) #20.6175
283 var(RutasT$VIAJE) #10.4512
284 var(RutasT$DISTANCIA..km.) #4.963167
285 var(RutasT$TIEMPO..min.) #7.482745
286
287
288 #DESVIACIÓN ESTÁNDAR
289
290 #desviación estándar de los datos
291
292 sd(RutasT$RUTA) #1.091751
293 sd(RutasT$CAJAS) #87.51437
294 sd(RutasT$TARIFA) #4.54065
295 sd(RutasT$VIAJE) #3.232832
296 sd(RutasT$DISTANCIA..km.) #2.227817
297
```

61:1 # TABLA FRECUENCIA CAJA

Console

```
296 sd(Rutas$DISTANCIA. .km.) #2.227617
297 sd(Rutas$TIEMPO. .min.) #2.735461
298
299 #COEFICIENTE DE VARIACIÓN
300
301 #coeficiente de variación
302 #se saca dividiendo coeficiente divido por la media
303 #los datos son confiables
304
305 sd(Rutas$RUTA)/mean(Rutas$RUTA) #0.4420928
306 sd(Rutas$CAJAS)/mean(Rutas$CAJAS) #0.08319515
307 sd(Rutas$TARIFA)/mean(Rutas$TARIFA) # 0.1478314
308 sd(Rutas$VIAJE)/mean(Rutas$VIAJE) #0.1417496
309 sd(Rutas$DISTANCIA. .km.)/mean(Rutas$DISTANCIA. .km.) #0.1706016
310 sd(Rutas$TIEMPO. .min.)/mean(Rutas$TIEMPO. .min.) #0.1879539
311
312
313 ###### CURTOSIS #####
314
315 #NO PUEDEN NUNCA APARECER DATOS ACHATADOS O PLATICURTICA CERCA DE LA BASE
316 #es un conjunto de datos confiable porque es positivo, la asimetría es buena,
317 #la curtosis es normal mesocurtica casi achatada
318
319 kurtosis(Rutas$RUTA) #1.708011
320 kurtosis(Rutas$CAJAS) #1.773133
321 kurtosis(Rutas$TARIFA) #2.013268
322 kurtosis(Rutas$VIAJE) #1.882072
323 kurtosis(Rutas$DISTANCIA. .km.) #1.650477
324 kurtosis(Rutas$TIEMPO. .min.) #2.38424
325
326
327 ###### GRÁFICA DE NORMALIDAD #####
328
329 #se da por histograma y por las medidas de tendencia central
330 #moda, media, mediana
331 #se ve la curtosis y asimetría
332
333 #se pone el gráfico para ver la media, mediana, moda juntas.
334 #rojo es la media, azul es la mediana, amarillo es la moda
335 #interpretación= la probabilidad está dividida de 0,00 a 0,10, es decir 3 cuadrantes
336 #la distribución tiende a la normalidad por la forma de la curva con asimetría
337 #si no había asimetría entonces se revisa por las MTC
338
61:1 # TABLA FRECUENCIA CAJA
R Script
Console
```

```
334 #rojo es la media, azul es la mediana, amarillo es la moda
335 #interpretación= la probabilidad está dividida de 0,00 a 0,10, es decir 3 cuadrantes
336 #la distribución tiende a la normalidad por la forma de la curva con asimetría
337 #si no había asimetría entonces se revisa por las MTC
338
339
340 ### RUTAS ###
341
342 G22 = ggplot(RutasT, aes(x=RutasT$RUTA))+
343   geom_histogram(aes(y=..density..), fill="green", colour="black")+
344   geom_density(alpha=.2, fill="orange")+
345   geom_vline(aes(xintercept=mean(RutasT$RUTA)),
346             color="red", size=1.5)+
347   geom_vline(aes(xintercept=median(RutasT$RUTA)),
348             color="blue", size=1.5, linetype="dashed")+
349   geom_vline(aes(xintercept=mvn(RutasT$RUTA)),
350             color="yellow", size=1.5)+
351   labs(title= "HISTOGRAMA DE RUTAS")
352
353
354 G22
355
356
357 ### CAJAS ###
358
359 G23 = ggplot(RutasT, aes(x=RutasT$CAJAS))+
360   geom_histogram(aes(y=..density..), fill="green", colour="black")+
361   geom_density(alpha=.2, fill="orange")+
362   geom_vline(aes(xintercept=mean(RutasT$CAJAS)),
363             color="red", size=1.5)+
364   geom_vline(aes(xintercept=median(RutasT$CAJAS)),
365             color="blue", size=1.5, linetype="dashed")+
366   geom_vline(aes(xintercept=mvn(RutasT$CAJAS)),
367             color="yellow", size=1.5)+
368   labs(title= "HISTOGRAMA DE CAJAS")
369
370
371 G23
372
373 ### TARIFA ###
374
375 G24 = ggplot(RutasT, aes(x=RutasT$TARIFA))+
376
```

61:1 # TABLA FRECUENCIA CAJA

R Script

Console

```
Tesis.R x  Untitled1* x  TESIS COCO.R* x  T.SCRIPT.R* x  T.SCRIPT (1).R* x
Source on Save  Run  Source
374
375 G24 = ggplot(RutasT, aes(x=RutasT$TARIFA))+
376   geom_histogram(aes(y=..density..), fill="green", colour="black")+
377   geom_density(alpha=.2, fill="orange")+
378   geom_vline(aes(xintercept=mean(RutasT$TARIFA)),
379             color="red", size=1.5)+
380   geom_vline(aes(xintercept=median(RutasT$TARIFA)),
381             color="blue", size=1.5, linetype="dashed")+
382   geom_vline(aes(xintercept=mfv(RutasT$TARIFA)),
383             color="yellow", size=1.5)+
384   labs(title= "HISTOGRAMA DE TARIFA")
385
386
387 G24
388
389 ### VIAJE ###
390
391 G25 = ggplot(RutasT, aes(x=RutasT$VIAJE))+
392   geom_histogram(aes(y=..density..), fill="green", colour="black")+
393   geom_density(alpha=.2, fill="orange")+
394   geom_vline(aes(xintercept=mean(RutasT$VIAJE)),
395             color="red", size=1.5)+
396   geom_vline(aes(xintercept=median(RutasT$VIAJE)),
397             color="blue", size=1.5, linetype="dashed")+
398   geom_vline(aes(xintercept=mfv(RutasT$VIAJE)),
399             color="yellow", size=1.5)+
400   labs(title= "HISTOGRAMA DE VIAJE")
401
402
403 G25
404
405
406 ### DISTANCIA ###
407
408
409 G26 = ggplot(RutasT, aes(x=RutasT$DISTANCIA..km.))+
410   geom_histogram(aes(y=..density..), fill="green", colour="black")+
411   geom_density(alpha=.2, fill="orange")+
412   geom_vline(aes(xintercept=mean(RutasT$DISTANCIA..km.)),
413             color="red", size=1.5)+
414   geom_vline(aes(xintercept=median(RutasT$DISTANCIA..km.)),
415             color="red", size=1.5)+
416
417
61:1 # TABLA FRECUENCIA CAJA
R Script
Console
```

```
413     color="red",size=1.5)+
414   geom_vline(aes(xintercept=median(RutasT$DISTANCIA..km.)),
415             color="blue",size=1.5,linetype="dashed")+
416   geom_vline(aes(xintercept=mfv(RutasT$DISTANCIA..km.)),
417             color="yellow",size=1.5)+
418   labs(title= "HISTOGRAMA DE DISTANCIA")
419
420
421 G26
422
423
424 ### TIEMPO ###
425
426 G27 = ggplot(RutasT, aes(x=RutasT$TIEMPO..min.))+
427   geom_histogram(aes(y=..density..),fill="green",colour="black")+
428   geom_density(alpha=.2, fill="orange")+
429   geom_vline(aes(xintercept=mean(RutasT$TIEMPO..min.)),
430             color="red",size=1.5)+
431   geom_vline(aes(xintercept=median(RutasT$TIEMPO..min.)),
432             color="blue",size=1.5,linetype="dashed")+
433   geom_vline(aes(xintercept=mfv(RutasT$TIEMPO..min.)),
434             color="yellow",size=1.5)+
435   labs(title= "HISTOGRAMA DE TIEMPO")
436
437
438
439 G27
440
441 ##### CONVERTIMOS A FACTOR LA VARIABLE RUTA #####
442
443 RutasT$RUTA=factor(RutasT$RUTA)
444
445
446 ##### REALIZAMOS UN BOXPLOT #####
447
448 GB1 = ggplot(RutasT, aes(x= RutasT$RUTA,
449                        y= RutasT$CAJAS))+
450   geom_boxplot(fill="pink", colours= "black")+
451   labs(title = "RUTA X CAJA")
452
453 GB1
454
455
```

61:1 # TABLA FRECUENCIA CAJA R Script

Console

```
453 GB1
454
455 GB2 = ggplot(RutasT, aes(x= RutasT$RUTA,
456                       y= RutasT$TIEMPO..min.))+
457   geom_boxplot(fill="blue", colours= "black")+
458   labs(title = "RUTA X TIEMPO")
459
460 GB2
461
462 ##### PLANTAMOS LA SEMILLA #####
463
464 set.seed(123)
465
466 ##### ENTRENAMIENTO DE DATOS #####
467
468 Train = createDataPartition(RutasT$RUTA,
469                             p=.8, list = FALSE)
470
471 ##### CREAMOS EL ÁRBOL DE DECISIÓN #####
472
473
474 arbol=rpart(RUTA~.,data = RutasT[Train,],
475            method = "class",
476            control = rpart.control(minsplit = 300,
477                                   cp=0.01))
478 arbol
479 ##### GRAFICAMOS EL ARBOL #####
480
481 rpart.plot(arbol,type = 1, digits = -1,
482           extra = 0, cex = 0.7, nn=TRUE,
483           fallen.leaves = TRUE,)
484
485 title(main = "Arbol de Decisión para la Predicción de RUTA", cex.main = 1, col.mair
486
487 mean(RutasT$STARIFA)
488
489 ##### PREDICCIÓN DE LOS VALORES DE LA REGRESIÓN #####
490
491 RutasT$Prediccion = predict(arbol, RutasT)
492
493 ##### BOSQUE ALEATORIO PARA LA REGRESIÓN #####
494
```

61:1 # TABLA FRECUENCIA CAJA R Script

Console

```
493 ##### BOSQUE ALEATORIO PARA LA REGRESIÓN #####
494
495 #CREAMOS EL BOSQUE ALEATORIO
496
497
498 Bosque=randomForest(x=RutasT[Train,2:10],
499                    y=RutasT[Train,1],
500                    ntree = 10000, keep.forest = TRUE)
501
502 Bosque
503
504 ClassB= predict(Bosque, RutasT[-Train,])
505
506 MatrizTest=table(RutasT[-Train,"RUTA",
507                ClassB, dnn=c("Actuales",
508                "Predichos"))
509 MatrizTest
510
511 Bosque$confusion
512
513
514 par(mfrow=c(1,2))
515
516 GME=mosaicplot(Bosque$confusion, type=n,
517                main = "Eficiencia del modelo - Entrenamiento",
518                color = c("yellow", "blue", "green", "red"))
519
520
521 GMT=mosaicplot(MatrizTest, type=n,
522                main = "Eficiencia del modelo - Pruebas",
523                color = c("yellow", "blue", "green", "red"))
524
525 par(mfrow=c(1,1))
526
527
528 RutasT[, "prob"]=predict(Bosque,RutasT)
529
530 RutasT$Probabilidad = predict(Bosque, RutasT,type ="prob")
531
532
533
534
```

61:1 TABLA FRECUENCIA CAJA R Script

Console

DECLARACIÓN Y AUTORIZACIÓN

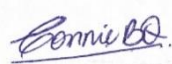
Nosotros, **Bohórquez Quisnia, Connie Scarlet**, con C.C: # **0926101601** y **Chamaidan Morocho, Carlos David**, con C.C: # **0706329463** autores del trabajo de titulación: **Optimización de rutas en el transporte de contenedores dentro del centro de acopio a través de un modelo de clasificación basado en Machine Learning** previo a la obtención del título de **Licenciado en Negocios Internacionales** 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, a los 23 del mes de agosto del año 2024

LOS AUTORES:

f. 

Bohórquez Quisnia, Connie Scarlet

f. 

Chamaidan Morocho, Carlos David

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TRABAJO DE INTEGRACIÓN CURRICULAR

TEMA Y SUBTEMA:	Optimización de rutas en el transporte de contenedores dentro del centro de acopio a través de un modelo de clasificación basado en Machine Learning	
AUTOR(ES)	Bohórquez Quisnia, Connie Scarlet Chamaidan Morocho, Carlos David	
REVISOR(ES)/TUTOR(ES)	Ing. Carrera Buri, Félix Miguel, Mgs.	
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil	
FACULTAD:	Economía y Empresas	
CARRERA:	Negocios Internacionales	
TÍTULO OBTENIDO:	Licenciado en Negocios Internacionales	
FECHA DE PUBLICACIÓN:	23 de agosto del 2024	No. DE PÁGINAS: 135
ÁREAS TEMÁTICAS:	Negocios, administración y contabilidad, Ciencias computacionales, Análisis de datos.	
PALABRAS CLAVES/ KEYWORDS:	Innovación, Análisis de datos, Inteligencia Artificial, Árbol de Clasificación, Random Forest para la clasificación, Rutas de Transporte	
RESUMEN/ABSTRACT:	<p>La optimización representa la parte más importante dentro de una empresa cuando se quiere reducir costos, incrementar el posicionamiento propio y mejorar la toma de decisiones dentro de una organización, para obtener resultados que brinden la mejor calidad y eficiencia al cliente final. Para ello, se debe reconocer que en la actualidad estos procesos pueden ser seguidos por herramientas que involucren la innovación y automatización junto con el análisis de datos de forma mucha más rápida y certera. Permitiendo así a las organizaciones, ser actores de primera fila en un mercado que es cada vez más rápido, más demandante, pero sobre todo más competitivo. A nivel empresarial es indispensable que las organizaciones opten por la aplicación de herramientas que involucren el crecimiento y aceleramiento de los procesos dentro de su cadena de suministro. La presente investigación se enfoca en el desarrollo y construcción de un modelo Bosque Aleatorio para la clasificación de rutas de transporte de forma más ágil y menos costosa, en la cual se pretende obtener como resultado una ruta que sea favorable en cuestiones de tiempo y dinero para la entrega de cajas de banano hacia el Puerto de Machala. El objetivo es determinar cuáles son las variables a considerar y que influyen en los recorridos de las rutas al momento de elegir la más conveniente. La aplicación de Machine Learning se justifica porque permite clasificar de forma precisa y autónoma los factores que intervienen obteniendo una respuesta objetiva.</p>	
ADJUNTO PDF:	<input checked="" type="checkbox"/> SI	<input type="checkbox"/> NO
CONTACTO CON AUTOR/ES:	Teléfono: +593992907821 +593987228461	E-mail: connie.bohorquez@cu.ucsg.edu.ec carlos.chamaidan@cu.ucsg.edu.ec
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UIC):	Nombre:	
	Teléfono:	
	E-mail:	
SECCIÓN PARA USO DE BIBLIOTECA		
N°. DE REGISTRO (en base a datos):		
N°. DE CLASIFICACIÓN:		
DIRECCIÓN URL (tesis en la web):		