UNIVERSIDAD DEL AZUAY

Departamento de Posgrados

Análisis de las características del rendimiento de combustible de vehículos de carga pesada de una empresa de transporte entre los destinos Cuenca - Guayaquil

Trabajo de titulación previo a la obtención del grado en Magíster en Estadística Aplicada

Autora:

María Belén Paredes Álvarez

Director:

Dr.Sc. Jonnatan Fernando Avilés González MSc.

Cuenca, Ecuador

2025

DEDICATORIA

A mis padres,
por enseñarme que los sueños se alcanzan con esfuerzo, paciencia y fe.

A mi hermano,
por su amor y compañía incondicional.

A familiares y amigos,
por estar en los momentos de duda y cansancio, gracias por recordarme siempre que sí puedo.

Este logro es también de ustedes,
porque cada palabra de aliento, cada gesto y cada muestra de afecto fueron el impulso que me permitió llegar hasta aquí.

AGRADECIMIENTO

A la empresa de transporte que me brindó los datos necesarios para la realización de este trabajo, por su constante apoyo y por la flexibilidad otorgada en los horarios laborales que me permitieron asistir a clases.

A mis profesores, por compartir su conocimiento, experiencia y compromiso con la excelencia académica. Cada enseñanza recibida aportó valor a este proyecto y enriqueció mi formación profesional, motivándome a seguir creciendo con responsabilidad y dedicación.

Finalmente, a mi familia y a todas las personas que me acompañaron en este proceso, gracias por su apoyo incondicional, por creer en mí y por ser el motor que impulsó cada paso hacia la culminación de esta meta.

RESUMEN

El presente estudio analiza los factores que influyen en el rendimiento de combustible en vehículos de carga pesada enfocándose en las rutas con más frecuencia, siendo estas Cuenca-Guayaquil y Guayaquil- Cuenca. Debido al incremento sostenido del precio del galón del diésel en Ecuador y su impacto en los costos logísticos, se aplicaron métodos estadísticos descriptivos, inferenciales y de aprendizaje automático para identificar factores críticos y predecir rendimientos bajos.

Para la investigación se utilizó variables obtenidas del dispositivo de rastreo satelital y complementando con datos sobre la información relacionada a la carga, tales como: tiempos de conducción, velocidad media, tiempo en ralentí y aceleraciones bruscas. Los resultados evidenciaron diferencias significativas entre marcas de los vehículos y las rutas analizadas, destacándose la de Cuenca – Guayaquil como prioritaria debido a su menor rendimiento de combustible y mayor variabilidad respecto a la ruta contraria. Además, se identificó que la velocidad promedio, el tiempo total de conducción y el peso de la carga son las variables más influyentes en el consumo de combustible.

Se recomienda implementar programas de capacitación en conducción eficiente y optimización de rutas, orientando la toma de decisiones hacia estrategias de conducción eficiente al planificar las cargas y mantenimientos preventivos, contribuyendo así a la sostenibilidad y mejora de la rentabilidad de la empresa.

Palabras clave:

Rendimiento de combustible, Transporte de carga pesado, Random Forest, XGBoost, Clustering, conducción eficiente.

ABSTRACT

This study analyzes the factors that influence fuel efficiency in heavy-duty vehicles, focusing on the most frequent routes, namely Cuenca–Guayaquil and Guayaquil–Cuenca. Due to the sustained increase in the price of diesel fuel in Ecuador and its impact on logistics costs, descriptive and inferential statistical methods, and machine learning were applied to identify critical factors and predict low fuel efficiency.

The research used variables obtained from satellite tracking devices, complemented with data related to cargo information, such as driving time, average speed, idling time, and sudden accelerations. The results showed significant differences between vehicle brands and the analyzed routes, with Cuenca–Guayaquil being prioritized due to its lower fuel efficiency and higher variability compared to the opposite route. In addition, it was identified that average speed, total driving time, and cargo weight are the most influential variables in fuel consumption.

It is recommended to implement training programs on efficient driving and route optimization, guiding decision-making toward Eco-Driving strategies in load planning and preventive maintenance, thus contributing to sustainability and improved company profitability.

Keywords:

Fuel efficiency, Heavy-Duty transport, Random Forest, Xgboost, Clustering, Efficient driving.

Índice de Contenido

Introducción **10**
Objetivo General **12**
1. Laboratorio: Enfoque Descriptivo e Inferencial **12**
1.1. Objetivos específicos: 12
1.2. Métodos: 12
1.3. Resultados y discusión: 13
1.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil: 13
1.3.1.1. Estadística descriptiva: 13
1.3.1.2. Pruebas de normalidad: 16
1.3.1.3. Correlación: 16
1.3.1.4. Prueba de hipótesis: 16
1.3.2. Resultados obtenidos de los viajes de Guayaquil a Cuenca: 17
1.3.2.1. Estadística descriptiva: 17
1.3.2.2. Pruebas de normalidad: 19
1.3.2.3. Correlación: 20
1.3.2.4. Pruebas de hipótesis: 20
1.4. Conclusiones parciales: 21
2. Laboratorio: Enfoque Machine Learning **21**
2.1. Objetivos específicos: 21
2.2. Métodos: 21
2.3. Resultados y discusión: 23
2.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil: 23
2.3.1.1. Selección de variables influyentes: 23
2.3.1.2. Métodos de clasificación supervisada: 24
2.3.1.3. Métodos de clasificación no supervisada: 27
2.3.1.4. Visualizaciones avanzadas: 29
2.3.2. Resultados obtenidos de los viajes de Cuenca a Guayaquil: 29
2.3.2.1. Selección de variables influyentes: 29
2.3.2.2. Métodos de clasificación supervisada: 30
2.3.2.3. Métodos de clasificación no supervisada: 33
2.3.2.4. Visualizaciones avanzadas: 35
2.4. Conclusiones parciales: 35
3. Laboratorio: Enfoque Toma de Decisiones **36**

3.1. Objetivos específicos: 36
3.2. Métodos: 36
3.3. Resultados y discusión: 36
3.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil: 36
3.3.2. Resultados obtenidos de los viajes de Guayaquil a Cuenca: 37
3.4. Conclusiones parciales: 38
4. Laboratorio: Enfoque Robustez y Validación 38
4.1. Objetivo específico: 38
4.2. Metodología: 38
4.3. Resultados y discusión: 39
4.3.1. Modelo de Random Forest para elección de variables: 39
4.3.2. Modelo de Random Forest para clasificación: 40
4.3.3. Modelo XGBoost para clasificación: 40
4.3.4. Análisis de Clustering: 41
4.4. Conclusiones parciales: 41
Conclusión general: 42
Bibliografía: 43

Índice de Figuras

Figura 1 Boxplot del rendimiento de combustible de las marcas de los vehículos 14
Figura 2 Rendimiento de combustible y distancia recorrida por mes 14
Figura 3 Rendimiento de combustible y peso cargado por mes.....15
Figura 4 Histograma de rendimiento de vehículos 15
Figura 5 Mapa de calor de correlación entre variables 16
Figura 6 Boxplot del rendimiento de combustible de las marcas de los vehículos 18
Figura 7 Rendimiento de combustible y distancia recorrida por mes 18
Figura 8 Rendimiento de combustible y peso cargado por mes.....19
Figura 9 Histograma de rendimiento de vehículos 19
Figura 10 Mapa de calor de correlación entre variables.....20
Figura 11 Distribución de variables continuas.....23
Figura 12 Importancia de variables según Random Forest (Regresión).....24
Figura 13 Árbol de decisión Cuenca - Guayaquil 25
Figura 14 Matriz de confusión – Árbol de decisión Cuenca - Guayaquil.....25
Figura 15 Matriz de Confusión – Random Forest Cuenca - Guayaquil 26
Figura 16 Matriz de confusión – XGBoost Cuenca - Guayaquil 26
Figura 17 Visualización de clusters mediante PCA – K-means (k = 3).....27
Figura 18 Visualización de clusters K-means (k = 2) 27
Figura 19 Visualización Clustering jerárquico 28
Figura 20 Visualización cluster PAM.....28
Figura 21 Mapa de calor entre velocidad media, peso de carga y rendimiento 29
Figura 22 Diagramas de caja de variables continuas 30
Figura 23 Importancia de variables – Random Forest (regresión).....30
Figura 24 Árbol de decisión Guayaquil - Cuenca 31
Figura 25 Matriz de confusión – Árbol de decisión Guayaquil - Cuenca.....31
Figura 26 Matriz de confusión – Random Forest Guayaquil - Cuenca.....32
Figura 27 Matriz de confusión – XGBoost Guayaquil - Cuenca 32
Figura 28 Visualización de clusters mediante PCA – K-means (k = 3).....33
Figura 29 Visualización de clusters K-means (k = 2) 33
Figura 30 Dendrograma – Clustering jerárquico.....34
Figura 31 Visualización MDS – PAM (k = 2).....34
Figura 32 Mapa de calor entre velocidad media, peso de carga y rendimiento 35

Índice de Tablas

Tabla 1 Descripción de las variables utilizadas en el estudio....12
Tabla 2 Estadística descriptiva por marca de vehículo 13
Tabla 3 Comparaciones post-hoc (Prueba de Dunn) 17
Tabla 4 Estadística descriptiva por marca de vehículo 17
Tabla 5 Comparaciones post-hoc (Prueba de Dunn) 21
Tabla 6 Comparación de métricas de evaluación para los tres modelos supervisados .26
Tabla 7 Comparación de métricas de evaluación para los tres modelos supervisados .33
Tabla 8 Promedio de variables operativas por cluster....37
Tabla 9 Promedio de variables operativas por cluster....37
Tabla 10 Resultados de los supuestos Random Forest....40
Tabla 11 Resultados de los supuestos Random Forest 40
Tabla 12 Resultados de los supuestos XGBoost 41
Tabla 13 Resultados supuestos clustering....41

Introducción

En los últimos años, el precio del galón de diésel en Ecuador ha experimentado un incremento sostenido que ha impactado profundamente al sector del transporte terrestre. En el año 2018, el galón de diésel costaba $1.037; a finales de 2020 subió a $1.30 y en octubre de 2025, debido a la eliminación del subsidio, alcanzó los $2.80 (EP Petroecuador, s. f.). Este aumento ha convertido al combustible en uno de los costos más altos de la operación logística, afectando directamente la rentabilidad de las empresas transportistas. La situación se agrava por la alta oferta de servicios de transporte, que limita el ajuste proporcional de las tarifas de flete.

Las condiciones sociales y económicas del país también influyen en este fenómeno. La eliminación progresiva del subsidio al diésel ha generado tensiones entre el gobierno y los gremios transportistas, quienes enfrentan dificultades para trasladar los costos a los usuarios finales. Además, el transporte de carga es esencial para la distribución de alimentos, medicinas y productos básicos, por lo que cualquier variación en el precio del combustible tiene efectos en cadena sobre la economía popular.

El problema central que se aborda es el elevado consumo de combustible en rutas específicas, agravado por el aumento sostenido del precio del diésel y la imposibilidad de ajustar proporcionalmente las tarifas de flete. Esta situación pone en riesgo la sostenibilidad financiera de las empresas de transporte, especialmente aquellas que operan con márgenes reducidos y alta competencia.

El fenómeno que se estudia se enmarca en el análisis del consumo de diésel en una empresa de transporte de carga que opera principalmente en las rutas Cuenca–Guayaquil y Guayaquil–Cuenca, utilizando las vías de Cañar, Zhud, Suscal, La Troncal y El Triunfo. Esta empresa forma parte de un sector industrial que moviliza carga pesada y cuya competitividad depende en gran medida de su capacidad para controlar costos logísticos. La flota está compuesta por 4 tipos de marcas de vehículos pesados que recorren distancias superiores a los 400 km por viaje.

El alcance del proyecto se limita al análisis de las dos rutas más frecuentes de la empresa objeto de estudio durante un periodo anual: Cuenca–Guayaquil y Guayaquil–Cuenca. Se estudiarán variables operativas y técnicas que inciden en el consumo de diésel, utilizando herramientas de análisis estadístico. Los resultados permitirán proponer recomendaciones concretas para mejorar la eficiencia energética y reducir costos operativos.

El fenómeno que se estudia se enmarca en el análisis del consumo de diésel en una empresa de transporte de carga que opera principalmente en las rutas Cuenca–Guayaquil y Guayaquil–Cuenca, utilizando las vías de Cañar, Zhud, Suscal, La Troncal y El Triunfo. Esta empresa forma parte de un sector industrial que moviliza carga pesada y cuya competitividad depende en gran

medida de su capacidad para controlar costos logísticos. La flota está compuesta por 4 tipos e marca de vehículos pesados que recorren distancias superiores a los 400 km por viaje.

Espinoza y Viteri (2025) analizaron la relación entre las importaciones de diésel y el crecimiento económico ecuatoriano, evidenciando la dependencia del país en este recurso para sectores estratégicos como el transporte. Este vínculo refuerza la necesidad de optimizar el uso del combustible en empresas logísticas para evitar impactos negativos en la economía. Por otro lado, Bermeo y Calderón (2025) demostraron que la optimización de rutas puede reducir significativamente el consumo de combustible y mejorar la eficiencia logística. Su estudio respalda la idea de que una planificación estratégica de itinerarios puede generar ahorros sustanciales en costos operativos.

El consumo de combustible está influenciado por factores relacionados con el vehículo, las condiciones del entorno y el estilo de conducción del chofer. Esta última ha cobrado relevancia por una técnica denominada “Eco-Driving”, la cual busca optimizar el uso de combustible mediante el manejo eficiente de variables que puede controlar el conductor, como la velocidad, la aceleración, la desaceleración y el cambio de marchas. Esta práctica permite reducir las pérdidas de energía y, en consecuencia, disminuir tanto el consumo de combustible como las emisiones contaminantes (Corcoba Magaña & Muñoz Organero, 2014).

Eco-Driving adquiere una relevancia particular en el ámbito del transporte, especialmente en empresas dedicadas a este sector. Una empresa de transporte en Bogotá implementó metodologías y prácticas orientadas a promover una conducción más eficiente y sostenible. Los resultados de la investigación evidenciaron, mediante análisis estadístico, el desempeño de los conductores en relación con variables asociadas a la conducción ecológica y al ahorro energético. Asimismo, esta técnica no solo contribuye a un ahorro de combustible, sino que también influye positivamente en el bienestar del conductor, mejorando aspectos como la seguridad y la salud laboral. Su principal ventaja radica en que su aplicación no depende de la tecnología del vehículo, aunque requiere que el conductor cuente con conocimientos y capacitación en técnicas de conducción eficiente (Espinoza Cuadrado et al., 2022).

Llanes-Cedeño et al. (2024) evidencian en su estudio que conducir durante los horarios pico y con el aire acondicionado encendido influye negativamente en el rendimiento de combustible en las rutas de la Amazonía ecuatoriana. Los autores destacan que estas condiciones incrementan el consumo energético del vehículo debido a mayores tiempos de ralentí, variaciones en la velocidad promedio y una mayor exigencia del motor para mantener la climatización.

Finalmente, Repsol Ecuador (2024) analizó estrategias de sostenibilidad empresarial aplicables al sector transporte, destacando que las prácticas responsables no solo mejoran la imagen corporativa, sino que también contribuyen a la eficiencia operativa y a la reducción del consumo energético.

Este proyecto es de vital importancia porque busca analizar variables que afectan en el consumo de combustible en rutas específicas, con el fin de optimizar recursos y mejorar la eficiencia operativa. En un contexto donde el costo del galón del diésel representa gran parte del presupuesto logístico de la empresa objeto de estudio, comprender los factores que afectan permite tomar decisiones basadas en datos y ofertar precios competitivos.

Objetivo General

Caracterizar el rendimiento de combustible de las rutas más comunes de los vehículos de una empresa de transporte de la ciudad de Cuenca, considerando factores como el peso de la carga, tiempo en ralentí y velocidades.

1. Laboratorio: Enfoque Descriptivo e Inferencial

1.1. Objetivos específicos:

2. Analizar el comportamiento de las variables importantes
3. Comprar los rendimientos entre marcas de vehículos

1.2. Métodos:

La base de datos utilizada en esta investigación incluyó múltiples variables, siendo el rendimiento de combustible la que se consideró como variable central para el análisis. En la Tabla 1 se presentó el listado completo de estas variables junto con su respectiva descripción.

Variables	Tipo	Descripción
Marca	Categórica	Marca del tráiler
Distancia	Numérica	Kilómetros recorridos
Viaje	Categórica	Ruta realizada
Peso carga	Numérica	Peso cargado en toneladas
Total de conducción	Numérica	Minutos conducidos
Velocidad media	Numérica	Velocidad media a la que recorrió
Velocidad máxima	Numérica	Velocidad máxima a la que llegó el vehículo
Tiempo en ralentí	Numérica	Minutos que el vehículo estaba encendido, pero no estaba circulando
Aceleraciones	Numérica	Cantidad de aceleraciones bruscas
Consumo	Numérica	Galones de diésel consumidos
Rendimiento	Numérica	Kilómetros recorridos con un galón de diésel

Tabla 1 Descripción de las variables utilizadas en el estudio

La base de datos se sustrajo del dispositivo del rastreo y adicional se colocó de manera manual el viaje asignado al vehículo ese día, como limpieza de datos no se encontraron datos duplicados y se eliminaron los valores vacíos puesto que, si hubo viaje y no hay distancia o rendimiento, se percibe como un daño en el sistema del rastreo.

Para garantizar la calidad de los datos, se aplicaron filtros de validación en la variable rendimiento, eliminando valores superiores a 15 km/gal e inferiores a 2 km/gal, considerados como anomalías del sistema de medición.

Es importante aclarar que dos vehículos no cuentan con rendimientos diarios puesto que al ser vehículos de marca Mercedes el dispositivo del rastreo no puede ingresar a la computadora para extraer estos datos, es por esto que se completó esta información de manera manual por viaje completo con los tickets de combustible entregados en la gasolinera. Sin embargo, estos vehículos frecuentan en su mayoría la ruta de Cuenca a Bajo Alto y tienen pocos viajes hacia las rutas escogidas para el análisis, es por esto que se eliminaron los viajes de estos vehículos.

Se analizaron las dos rutas escogidas de manera separada, primero se realizó el análisis descriptivo de las variables más importantes para una visión de la distribución y dispersión de los datos, adicional a través de diagramas como boxplot se revisaron los datos atípicos; de igual forma se representó con diagramas comparativos el rendimiento de combustible con respecto a las otras variables durante el año.

Para determinar la normalidad de los datos, se aplicaron las pruebas de Kolmogorov-Smirnov y Anderson-Darling, cuyos resultados orientaron la selección entre estadística paramétrica o no paramétrica según correspondiera.

Finalmente, se construyó un mapa de calor para evaluar las correlaciones entre la variable dependiente y las variables independientes con las que se cuentan en la base de datos, adicional se realizó una prueba de hipótesis para comparar si las medianas del rendimiento de combustible por marca son iguales con un análisis post hoc mediante la prueba de Dunn con el método de Bonferroni.

1.3. Resultados y discusión:

1.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil:

1.3.1.1. Estadística descriptiva:

Se evidencia en la Tabla 2 que las tres marcas de vehículos muestran una diferencia mínima entre la media y la mediana, lo que sugiere distribuciones relativamente simétricas. Sin embargo, los Kenworth presentan una mayor diferencia entre estos valores, lo que indica una distribución más sesgada. A pesar de esto, su variabilidad es menor en comparación con las otras marcas, lo que sugiere un rendimiento de combustible más consistente.

Marca	Media	Mediana	Desviación	Varianza	Mínimo	Máximo
Freightliner	4.87	4.87	1.21	1.46	2.20	8.03
International	5.07	5.00	1.29	1.67	2.30	11.40
Kenworth	4.68	4.52	0.86	0.75	3.08	9.39

Tabla 2 Estadística descriptiva por marca de vehículo

Los datos atípicos observados en los boxplot de la Figura 1 del rendimiento de combustible de cada marca de vehículo se conservarán, debido a que estos rendimientos pueden deberse a varios factores como pesos livianos con exportaciones, velocidades controladas, entre otros.

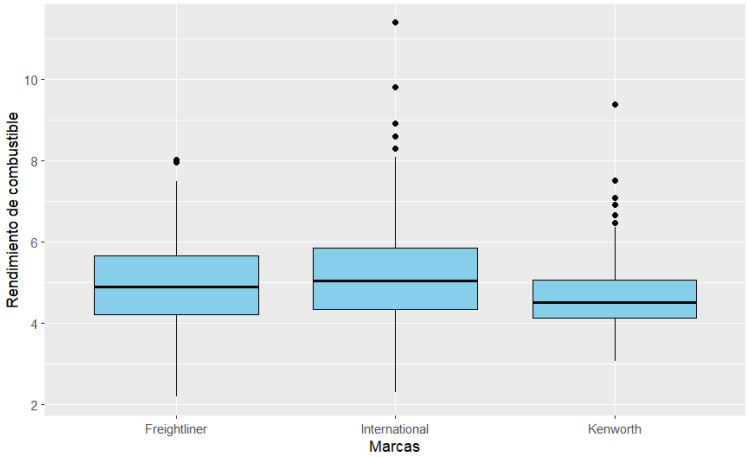

Figura 1 Boxplot del rendimiento de combustible de las marcas de los vehículos

Se observa en la Figura 2 que el rendimiento de combustible (barras) oscila entre los 5km/gal, mientras que la distancia recorrida (línea) tiene mayor variabilidad, alcanzando su punto máximo en marzo y en octubre presenta el total recorrido más bajo del año, lo que podría indicar el paro de vehículos por mantenimientos correctivos prolongados. Por otro lado, en los meses febrero y mayo la distancia recorrida aumenta y el rendimiento de combustible aumenta de igual forma.

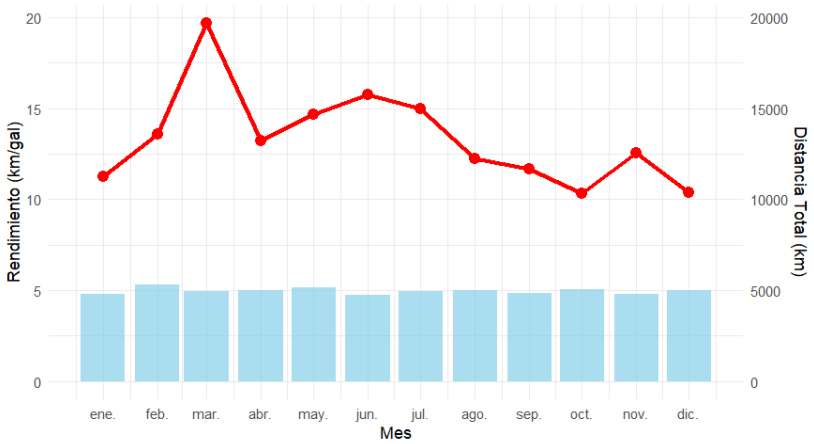

Figura 2 Rendimiento de combustible y distancia recorrida por mes

En la Figura 3 muestra el peso total cargado (línea) y el rendimiento de combustible (barras). En marzo, el peso transportado aumenta con respecto a febrero, lo que coincide con una disminución en el rendimiento de combustible, un patrón similar es observado en noviembre. En contraste, en mayo, tanto el peso transportado como el rendimiento de combustible aumentan. Por otro lado,

en junio y septiembre, la reducción en el peso transportado se acompaña de una disminución en el rendimiento de combustible.

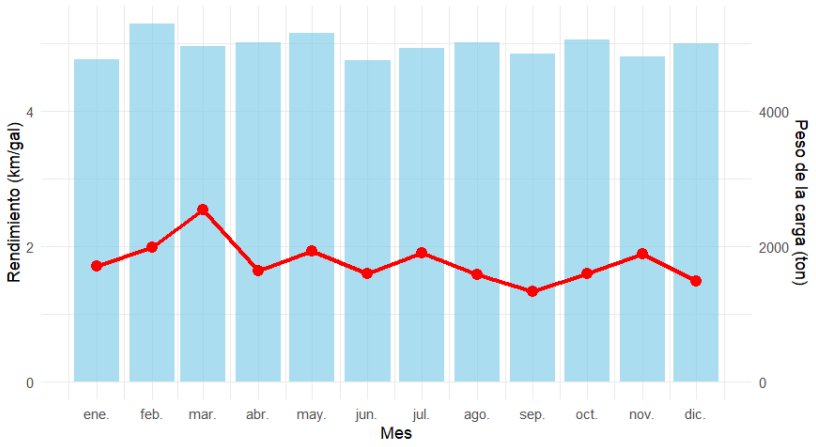

Figura 3 Rendimiento de combustible y peso cargado por mes

Se visualiza en la Figura 4 que la distribución de los datos que siguen el rendimiento de combustible en los viajes de Cuenca a Guayaquil es asimétrica hacia la derecha, con sesgo positivo, es unimodal y los datos están concentrados en un intervalo de 4.5 y 5.5.

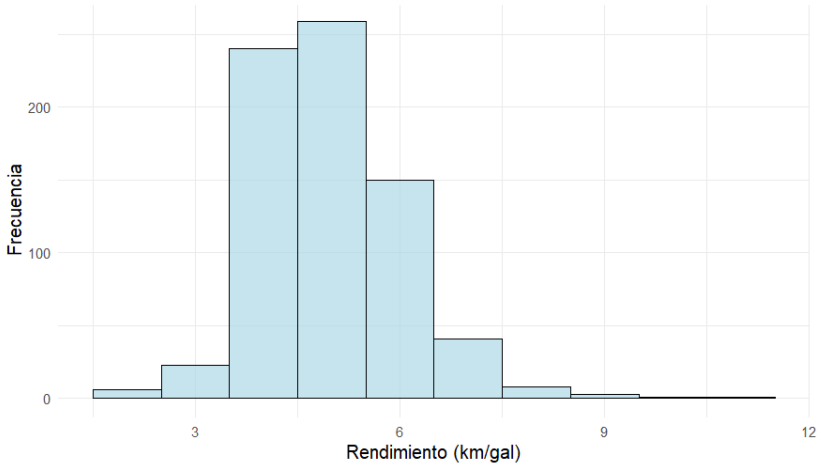

Figura 4 Histograma de rendimiento de vehículos

1.3.1.2. Pruebas de normalidad:

Las pruebas Kolmogorov-Smirnov y Anderson-Darling arrojaron un p-valor de 0.008 y 4.986e-15 correspondientemente, lo que es menor al nivel de significancia de 0.05, por lo que se concluye que los datos no siguen una distribución normal y se debe utilizar estadística no paramétrica.

1.3.1.3. Correlación:

Como se evidencia en la Figura 5, el análisis de correlación revela que únicamente la velocidad media presenta una correlación moderada con el rendimiento de combustible (0.33). Las demás variables independientes muestran coeficientes de correlación inferiores a ±0.3, indicando que existe correlación débil entre estas variables y el rendimiento.

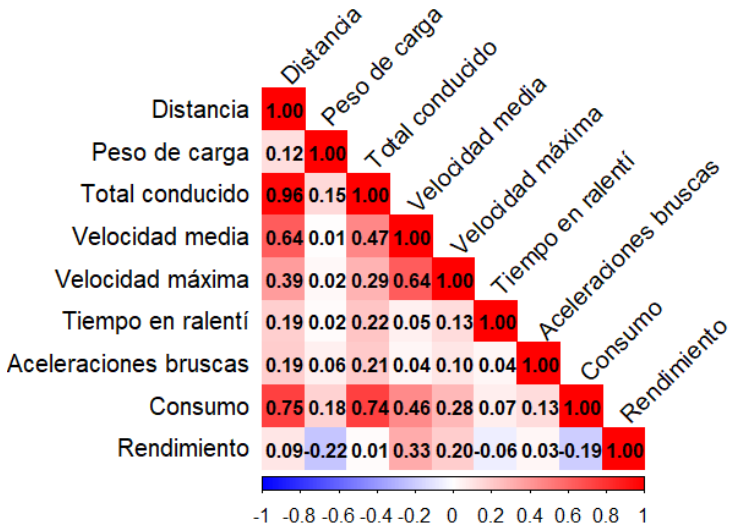

Figura 5 Mapa de calor de correlación entre variables

1.3.1.4. Prueba de hipótesis:

Para evaluar si existen diferencias significativas en el rendimiento entre las marcas de los vehículos se estableció la siguiente hipótesis estadística:

Hipótesis nula (H_0): Las medianas de la variable rendimiento de combustible son iguales entre las marcas Freightliner, International y Kenworth

Hipótesis alternativa (H_1): Al menos una marca tiene una mediana de rendimiento significativamente diferente.

Al realizar la prueba de Kruskal-Wallis para comparar los rendimientos entre las marcas, con un nivel de significancia de α = 0.05, se obtuvo un valor p<5.981e-06. Dado que el valor de p es menor a 0.05 se rechaza la hipótesis nula, concluyendo que existen diferencias significativas en los rendimientos entre las marcas.

Se realizó un análisis post hoc mediante la prueba de Dunn para identificar qué marcas presentan diferencias en sus rendimientos (Tabla 3). Los resultados muestran que no todas las comparaciones son estadísticamente significativas. La comparación entre Freightliner e International no presentan diferencias significativas (p=0.31), indicando que estas marcas tienen rendimientos similares. Por el contrario, se observan diferencias estadísticamente significativas entre Freightliner y Kenworth (p=0.0042), así como entre International y Kenworth (p=3.14e-06), lo que sugiere que Kenworth presenta un rendimiento significativamente diferente a las otras dos marcas.

Comparación	**Z**	**P.unadj**	**P.adj**
Freightliner – International	-1.63	0.1032	0.31
Freightliner – Kenworth	3.19	0.0014	0.0042
International – Kenworth	4.88	1.05e-06	3.14e-06

Tabla 3 Comparaciones post-hoc (Prueba de Dunn)

1.3.2. Resultados obtenidos de los viajes de Guayaquil a Cuenca:

1.3.2.1. Estadística descriptiva:

Conforme con los resultados expuestos en la Tabla 4, la marca Kenworth destaca por tener la media y media más alta con 6.81 y 7.05 respectivamente, además de una menor variabilidad, lo que indica un desempeño más estable. Los vehículos International presenta un rendimiento intermedio con una media de 6.54, con una mayor dispersión en los datos. Los Freightliner, por su parte, tiene la media más baja con 6.28 y una variabilidad similar a International, lo que indica que su desempeño es más variable.

Marca	**Media**	**Mediana**	**Desviación**	**Varianza**	**Mínimo**	**Máximo**
Freightliner	6.28	6.27	1.38	1.91	2.77	10.70
International	6.54	6.28	1.37	1.89	3.76	12.08
Kenworth	6.81	7.05	1.13	1.28	2.77	9.74

Tabla 4 Estadística descriptiva por marca de vehículo

A través de la Figura 6 se observa que los vehículos de marca International tienen más valores atípicos que los otros vehículos, esto puede darse por varios factores como el retorno liviano con contenedores vacíos y velocidades más estables, por lo que no se eliminaran los atípicos.

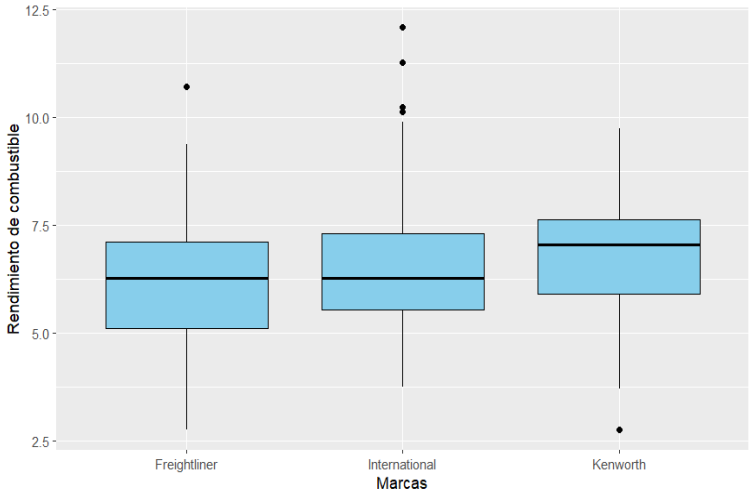

Figura 6 Boxplot del rendimiento de combustible de las marcas de los vehículos

En la Figura 7 se plasma la distancia recorrida en kilómetros (línea) y el rendimiento de combustible (barras), en donde se puede observar que la distancia recorrida tiene mayor variabilidad, alcanzando su punto máximo en abril y a partir del mes de junio empieza a disminuir con excepción del mes de noviembre, en cambio el rendimiento de combustible muestra poca variabilidad a lo largo de los meses con valores mayores a 5 Km/gal.

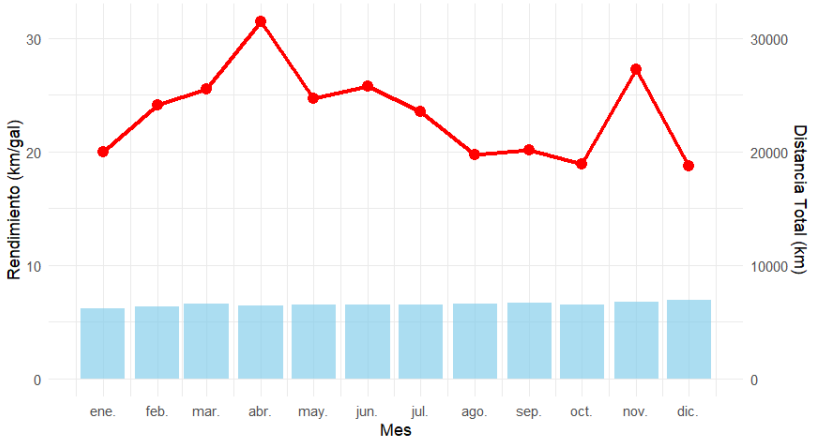

Figura 7 Rendimiento de combustible y distancia recorrida por mes

En el Figura 8 se analiza el peso de la carga (línea) y el rendimiento de combustible (barras). En marzo, tanto el peso como el rendimiento aumentan, mientras que en abril, a pesar del aumenta en el peso, el rendimiento disminuye. En diciembre, se observa el comportamiento contrario donde el peso transportado disminuye, pero el rendimiento aumenta. Por otro lado, en los meses de junio, julio y agosto, los rendimientos se mantienen similares; sin embargo, los pesos muestran una tendencia a la baja en cada mes.

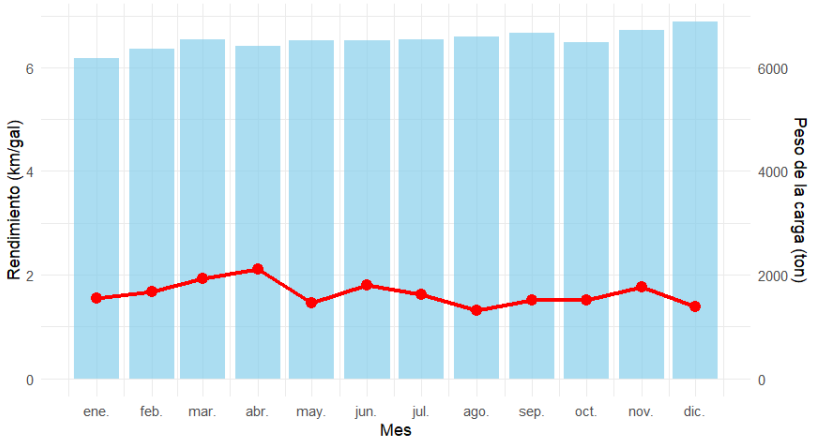

Figura 8 Rendimiento de combustible y peso cargado por mes

Se visualiza en la Figura 9 que la distribución de los datos que siguen el rendimiento de combustible en los viajes de Guayaquil a Cuenca es asimétrica hacia la derecha, con sesgo positivo, es unimodal y los datos están concentrados en un intervalo de 5.5 y 7.5.

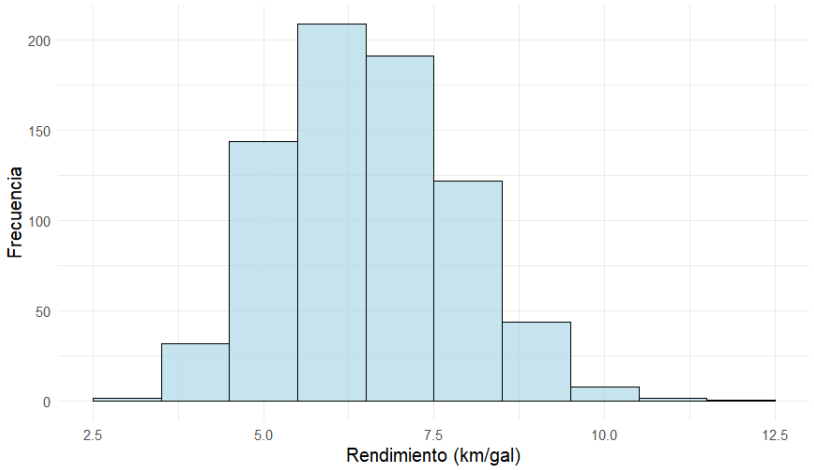

Figura 9 Histograma de rendimiento de vehículos

1.3.2.2. Pruebas de normalidad:

Las pruebas Kolmogorov-Smirnov y Anderson-Darling arrojaron un p-valor de 0.051 y 7.943e-05 correspondientemente, lo que da un empate entre las dos pruebas, para aceptar o rechazar normalidad se realizó una prueba adicional, Shapiro Wilk con un resultado 1.546e-13, rechazando la hipótesis nula dando como consecuencia que los datos no son normales y se debe utilizar estadística no paramétrica.

1.3.2.3. Correlación:

Como se evidencia en la Figura 10, el análisis de correlación para la ruta Guayaquil-Cuenca muestra que consumo tiene una correlación moderada negativa (-0.53), lo que es de esperar puesto que al aumentar el valor de esta variable el rendimiento disminuye. Las otras variables independientes presentan una correlación débil con el rendimiento de combustible, porque todos los coeficientes tienen valores inferiores a ±0.3. Los valores más altos corresponden a peso de carga, total conducido y velocidad media, pero aun así no alcanzan el umbral mínimo para considerar la existencia de correlación moderada.

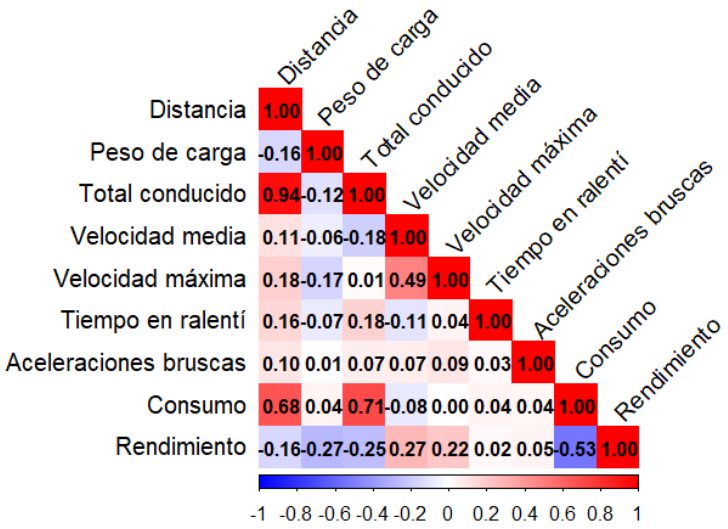

Figura 10 Mapa de calor de correlación entre variables

1.3.2.4. Pruebas de hipótesis:

Se realizó la misma prueba de hipótesis como en los viajes de Cuenca a Guayaquil, en donde se ejecutó la prueba de Kruskal-Wallis para comparar los rendimientos entre las marcas, con un nivel de significancia de $\alpha = 0.05$ y se obtuvo un valor $p < 5.017e\text{-}06$. Dado que el valor de p es menor a 0.05 se rechaza la hipótesis nula, concluyendo que existen diferencias en los rendimientos entre las marcas.

De igual forma, se realizó un análisis post hoc mediante la prueba de Dunn para identificar qué marcas presentan diferencias en sus rendimientos (Tabla 5). Los resultados muestran que no todas las comparaciones son estadísticamente significativas. La comparación entre Freightliner e International no presenta diferencias significativas (p=0.165), indicando que estas marcas tienen rendimientos similares. Por el contrario, se observan diferencias estadísticamente significativas entre Freightliner y Kenworth (p=2.78e-06), así como entre International y Kenworth

(p=2.85e-03), lo que sugiere que Kenworth presenta un rendimiento significativamente diferente a las otras dos marcas.

Comparación	Z	P.unadj	P.adj
Freightliner – International	-1.92	0.0551	0.165
Freightliner – Kenworth	-4.91	9.27e-07	2.78e-06
International – Kenworth	-3.31	9.49e-04	2.85e-03

Tabla 5 Comparaciones post-hoc (Prueba de Dunn)

1.4. Conclusiones parciales:

El rendimiento de combustible es mayor y presenta una mayor dispersión en la ruta de Guayaquil a Cuenca en comparación con la ruta de Cuenca a Guayaquil. Esto podría atribuirse a la diferencia en el peso de la carga transportada en cada viaje, ya que en Guayaquil las cargas suelen ser más livianas.

El análisis de correlación mostró que, en la mayoría de los casos, no existe una relación lineal significativa entre las variables, lo que sugiere la necesidad de utilizar otras herramientas de análisis.

Una posible limitación del estudio es que los datos corresponden únicamente a un año, lo que podría afectar la generalización de los resultados. Adicionalmente, no se cuenta con información sobre el estado de la vía ni sobre la pendiente de la carretera, factores que podrían influir en el rendimiento de combustible.

2. Laboratorio: Enfoque Machine Learning

2.1. Objetivos específicos:

1. Predecir el rendimiento de combustible de los vehículos utilizando técnicas de aprendizaje supervisado, considerando los factores más importantes del estudio.
2. Aplicar técnicas de clustering para clasificar los viajes en grupos con características similares.

2.2. Métodos:

Para revisar las relaciones entre las variables dependientes y la variable independiente (rendimiento de combustible), se generaron boxplot que mostraron la distribución de las variables: peso de la carga, tiempo total conducid, velocidad media, velocidad máxima, tiempo en ralentí y aceleraciones buscas frente al rendimiento de combustible. Estos gráficos se elaboraron por separado para cada ruta, este análisis visual ayudó a identificar cuáles variables podrían tener una mayor influencia en el rendimiento. Adicionalmente, se utilizó Random Forest para identificar

la importancia de cada variable con respecto al rendimiento de combustible. Se utilizó tanto variables categóricas como numéricas.

Para el análisis mediante técnicas de aprendizaje supervisado, se clasificó la variable de rendimiento de combustible en dos categorías, siguiendo las metas establecidas por la empresa objeto de estudio, siendo estas: rendimiento bajo (< 5.5 km/gal) y rendimiento adecuado (≥ 5.5 km/gal). Se aplicó el método ROSE (Random Over-Sampling Examples) para generar muestras sintéticas que ayuden a balancear el conjunto de datos. Posteriormente, se entrenaron tres modelos de clasificación: Árbol de Decisión, Random Forest y XGBoost, con el objetivo de predecir la categoría de rendimiento. Los datos fueron divididos en un 80 % para entrenamiento y un 20 % para prueba. La evaluación de los modelos se realizó utilizando las métricas: exactitud (accuracy), que indica el porcentaje de predicciones correctas, puntuación F1 (F1 score), que representa un balance entre la precisión y la sensibilidad del modelo, precisión (precision), que mide la proporción de predicciones positivas que fueron realmente correctas y sensibilidad (recall) que refleja la capacidad del modelo para identificar correctamente los casos positivos. Se seleccionó el modelo basándose en los resultados de estas métricas y en la matriz de confusión para reconocer cual estaba prediciendo de mejor manera los rendimientos bajos, puesto que se debe enfocar todos los esfuerzos en estos casos.

Finalmente, se aplicó técnicas de aprendizaje no supervisado para identificar patrones de comportamiento similares entre los vehículos sin necesidad de conocer previamente su categoría. Para esto, se realizó técnicas de clustering, utilizando los datos de forma separada para cada ruta. Previo al análisis, se seleccionaron variables numéricas y se normalizaron los datos puesto que las variables se encuentran en diferentes escalas y así se puede evitar posibles sesgos.

Se exploraron tres enfoques distintos de agrupamiento para comparar resultados y encontrar la técnica que mejor se adecue a los datos. En primer lugar, se aplicó el algoritmo k-means, seleccionando el número óptimo de clusters mediante la técnica de la curva del codo. Como segundo, se implementó un clustering jerárquico aglomerativo. Finalmente, se aplicó el método PAM (Partitioning Around Medoids), el cual emplea medoides en lugar de centroides y resulta menos sensible a valores atípicos. Se utilizó este enfoque debido a la naturaleza mixta de las variables. Para encontrar el método de clustering más adecuado, se empleó dos enfoques: uno visual, donde se revisó los gráficos de agrupamiento y otro cuantitativo con el índice de silueta, que evalúa la compactación y separación de los grupos.

2.3. Resultados y discusión:

2.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil:

2.3.1.1. Selección de variables influyentes:

El análisis comenzó evaluando cómo se distribuyen las variables continuas más relevantes a lo largo de la ruta. En la Figura 11, se pueden ver los diagramas de caja para las variables numéricas. Estas visualizaciones nos permiten observar la dispersión y la presencia de valores atípicos en cada variable, así como las diferencias en escala que justifican la normalización posterior en los análisis de clustering.

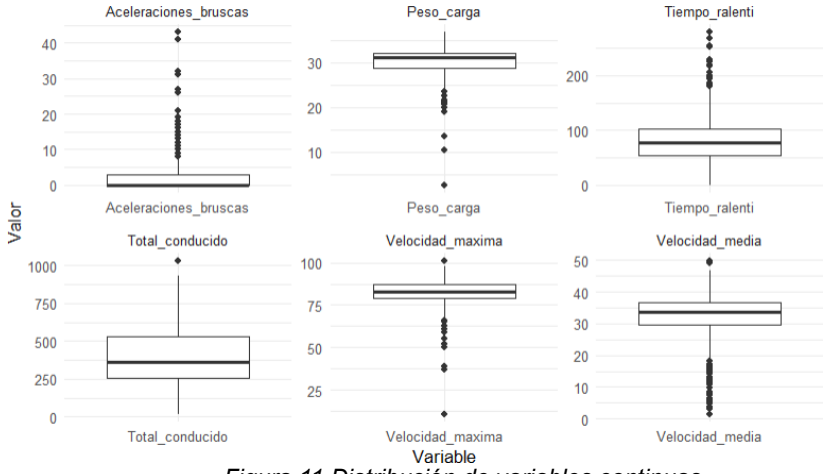

Figura 11 Distribución de variables continuas

Para poder identificar cuáles variables tienen un mayor impacto en el rendimiento del combustible, se utilizó un modelo de regresión basado en Random Forest, incorporando tanto variables continuas como categóricas. La Figura 12 representa los resultados del análisis de la importancia de las variables, presentados a través de dos métricas: el porcentaje de incremento del error cuadrático medio (%IncMSE), que indica que variables son críticas para la precisión general y el incremento de la pureza de los nodos (IncNodePurity), que muestra que tan eficiente es la variable para separar las clases en los diferentes árboles del modelo. En ambos casos, la velocidad media, total conducido y el peso de la carga se destacan como los predictores más significativos, lo que justifica su selección prioritaria para los modelos de clasificación y agrupamiento que se desarrollarán en etapas posteriores.

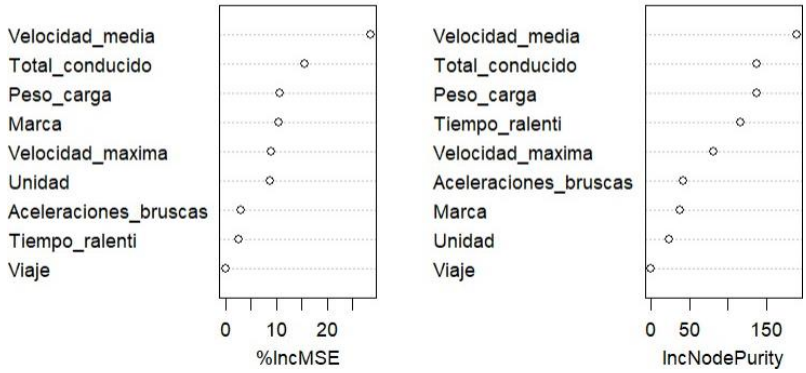

Figura 12 Importancia de variables según Random Forest (Regresión)

2.3.1.2. Métodos de clasificación supervisada:

- Árbol de decisión

En la Figura 13, se puede observar que variables como el peso de la carga y la velocidad media son las más influyentes. Por ejemplo, cuando el peso de la carga es inferior a 29 toneladas, el modelo predice un rendimiento bajo. En cambio, combinaciones de mayor peso con velocidades medias superiores a 28 km/h suelen asociarse con rendimientos adecuados.

A pesar de ser fácil de interpretar, el árbol de decisión mostró un rendimiento moderado: exactitud de 0.689, puntuación F1 de 0.477, precisión de 0.449 y sensibilidad de 0.509 (Tabla 6). La matriz de confusión (Figura 14) indica que el modelo tuvo problemas para identificar correctamente los casos de rendimiento adecuado, con un alto número de falsos negativos (120 casos que fueron predichos como “bajo” cuando en realidad eran “adecuados”).

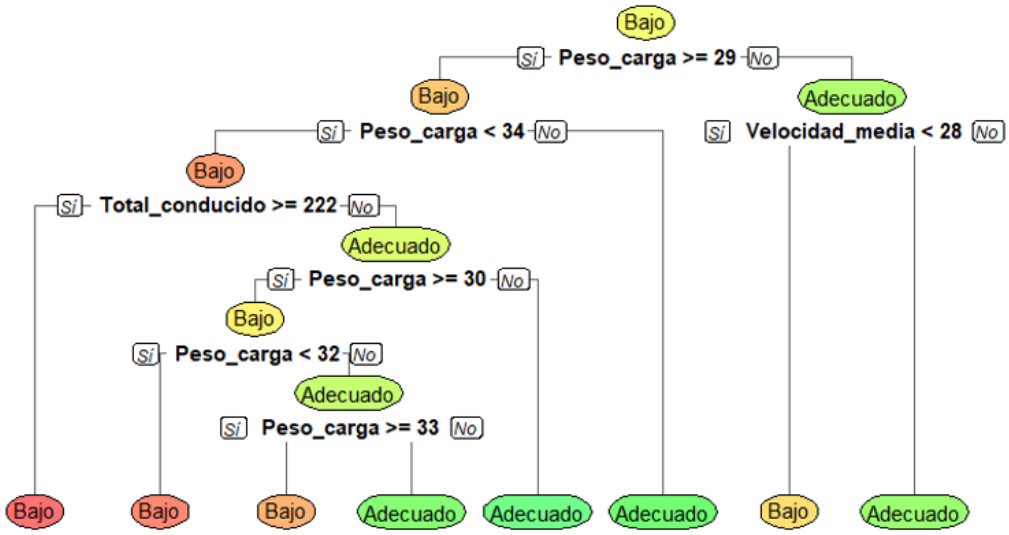

Figura 13 Árbol de decisión Cuenca - Guayaquil

Figura 14 Matriz de confusión – Árbol de decisión Cuenca - Guayaquil

- Random Forest

Al combinar múltiples árboles de decisión, mejoró ligeramente el rendimiento predictivo. Se alcanzó una exactitud de 0.712, una Puntuación F1 de 0.512, precisión de 0.485 y sensibilidad de 0.541, superando en todas las métricas al árbol individual (Tabla 6). La Figura 15 muestra la matriz de confusión del modelo, donde se observa una reducción leve en los errores de predicción para ambas clases, aunque aún persiste un sesgo hacia la clase “bajo”.

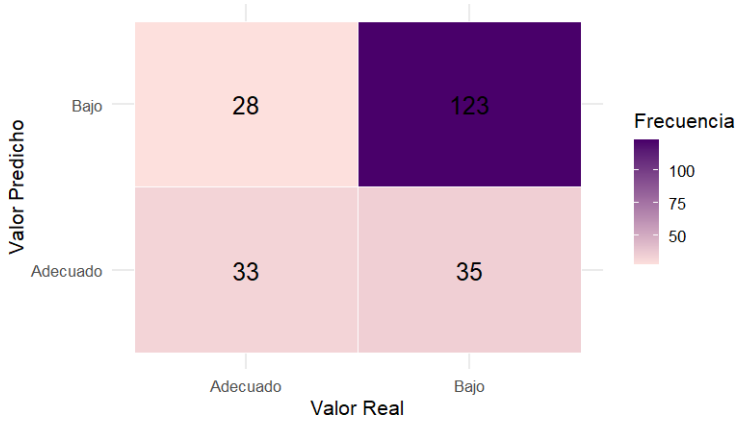

Figura 15 Matriz de Confusión – Random Forest Cuenca - Guayaquil

- XGBoost

Utiliza un enfoque de boosting donde empieza con un árbol de decisión simple y al generar uno nuevo corrige los errores del modelo anterior. En este caso su rendimiento fue un poco inferior al de Random Forest. Se logró una precisión de 0.699, un Puntuación F1 de 0.468, una precisión de 0.460 y sensibilidad de 0.475. La Figura 16 presenta su matriz de confusión, donde se pueden ver errores similares a los de los modelos anteriores, especialmente un alto porcentaje de falsos negativos.

Figura 16 Matriz de confusión – XGBoost Cuenca - Guayaquil

Modelo	Exactitud	Puntaje	Precisión	Sensibilidad
Árbol de decisión	0.6895	0.4769	0.4493	0.5082
Random Forest	0.7123	0.5116	0.4853	0.5410
XGBoost	0.6986	0.4677	0.4603	0.4754

Tabla 6 Comparación de métricas de evaluación para los tres modelos supervisados

Los tres modelos enfrentaron algunos retos al intentar clasificar correctamente el rendimiento, principalmente por la superposición de características entre las diferentes clases. A pesar de esto, Random Forest se destacó al obtener el mejor desempeño, gracias a su mayor precisión y Puntuación F1, puesto que es necesario que se identifique los casos con rendimiento bajo a los

que se desea estudiar.

2.3.1.3. Métodos de clasificación no supervisada:

- Clustering K-means (k = 3)

Aunque visualmente los grupos mostraron cierta separación (Figura 17), el índice de silueta fue 0.244, indicando clustering muy débil y con fuerte solapamiento entre clusters.

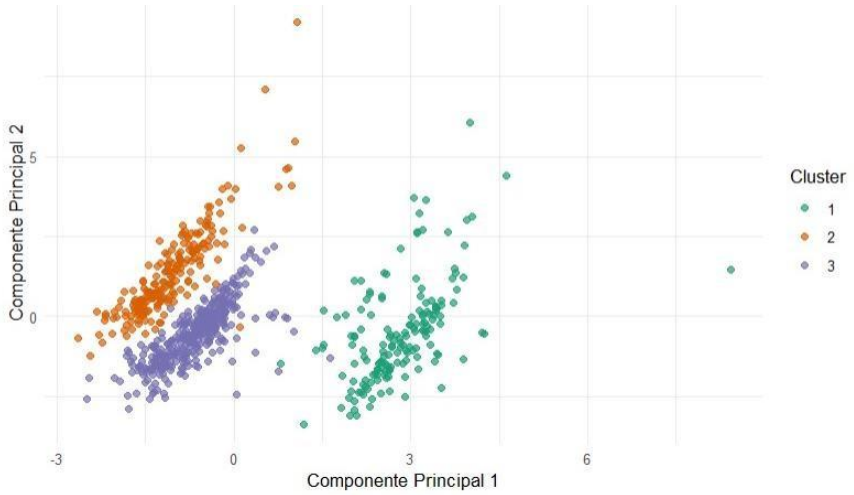

Figura 17 Visualización de clusters mediante PCA – K-means (k = 3)

- Clustering K-means (k = 2)

Reduciendo el número de grupos a dos (Figura 18), el índice de silueta aumentó a 0.33, indicando una mejora leve, pero aún con una estructura de cluster poco definida.

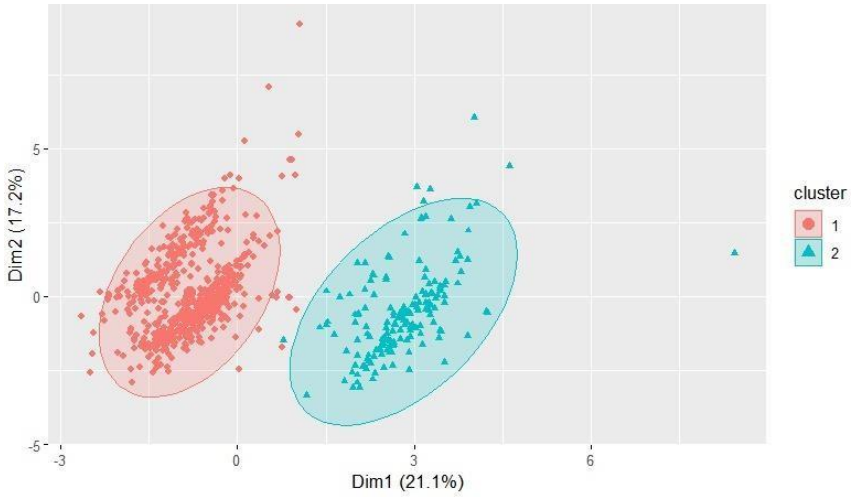

Figura 18 Visualización de clusters K-means (k = 2)

- Clustering Jerárquico

El dendrograma obtenido (Figura 19) permitió agrupar los datos en tres clusters con una separación más clara a nivel visual. El índice de silueta promedio fue 0.405, lo que representa un desempeño moderadamente mejor que K-means, aunque todavía dentro de la categoría de clustering débil con posible traslape.

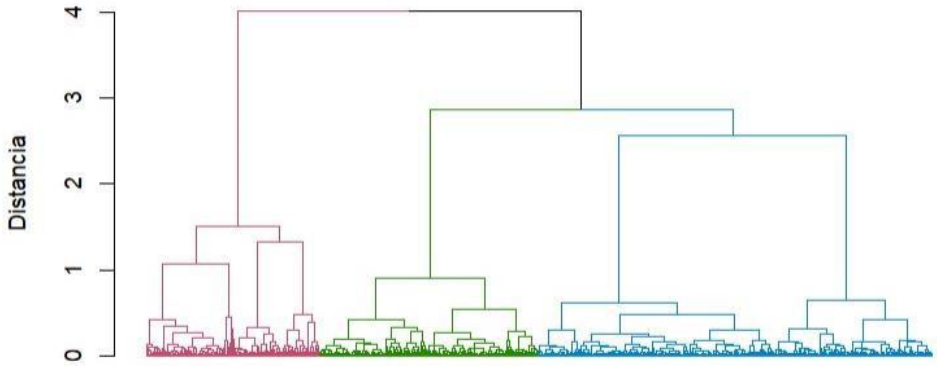

Figura 19 Visualización Clustering jerárquico

- Clustering con PAM

Finalmente, el método PAM logró el mejor resultado en el índice de silueta a comparación de los otros enfoques evaluados, alcanzando un valor de 0.494 (Figura 20). Aunque todavía es un resultado bajo, este valor indica que PAM ofrece la estructura de cluster más útil, permitiendo distinguir grupos con diferencias operativas relevantes, especialmente en variables como velocidad media y aceleraciones bruscas.

Los resultados indican que, si bien existe cierta estructura de segmentación en los datos, esta no es fuerte. El mejor desempeño lo mostró el método PAM con una métrica mayor a los otros modelos usados. Por lo tanto, se recomienda utilizar PAM. Sin embargo, se sugiere complementar con otros enfoques supervisados para una clasificación más robusta.

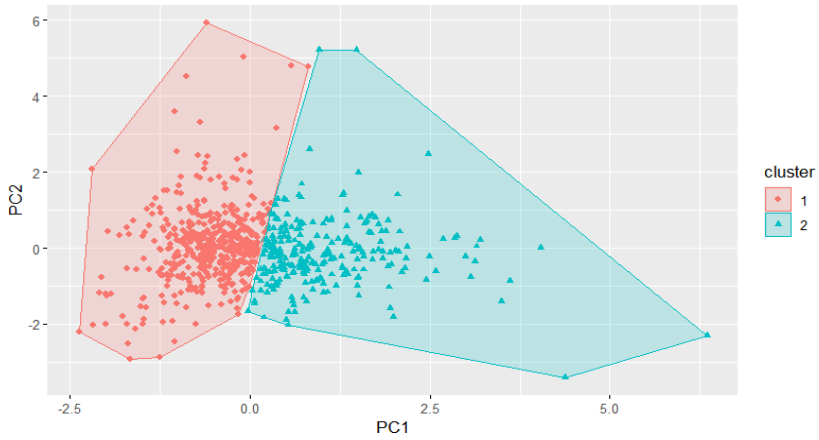

Figura 20 Visualización cluster PAM

2.3.1.4. Visualizaciones avanzadas:

Al analizar el rendimiento del combustible en la ruta Cuenca–Guayaquil, utilizando un mapa de calor (Figura 21), se puede observar que los niveles más bajos de eficiencia se concentran en las combinaciones de baja velocidad media (5–15 km/h) y cargas pesadas (26–30 toneladas), lo que muestra un impacto negativo de estas condiciones en el consumo de combustible. Por otro lado, los mejores rendimientos se logran en rangos de velocidad media entre 35 y 40 km/h, junto con cargas moderadas (10–20 toneladas). Esta tendencia sugiere que, dentro de ciertos límites operativos, aumentar la velocidad media puede mejorar la eficiencia del combustible, especialmente cuando no se transportan cargas excesivamente pesadas.

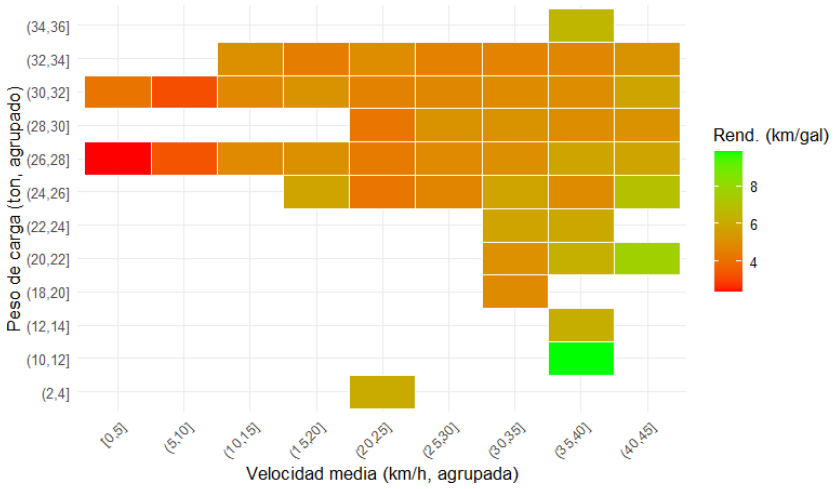

Figura 21 Mapa de calor entre velocidad media, peso de carga y rendimiento

2.3.2. Resultados obtenidos de los viajes de Cuenca a Guayaquil:

2.3.2.1. Selección de variables influyentes:

Los boxplot (Figura 22) muestran una mayor dispersión en variables como aceleraciones bruscas y tiempo en ralentí, y una velocidad media más alta en comparación con la ruta de ida. Esto sugiere condiciones operativas diferentes en el trayecto de retorno.

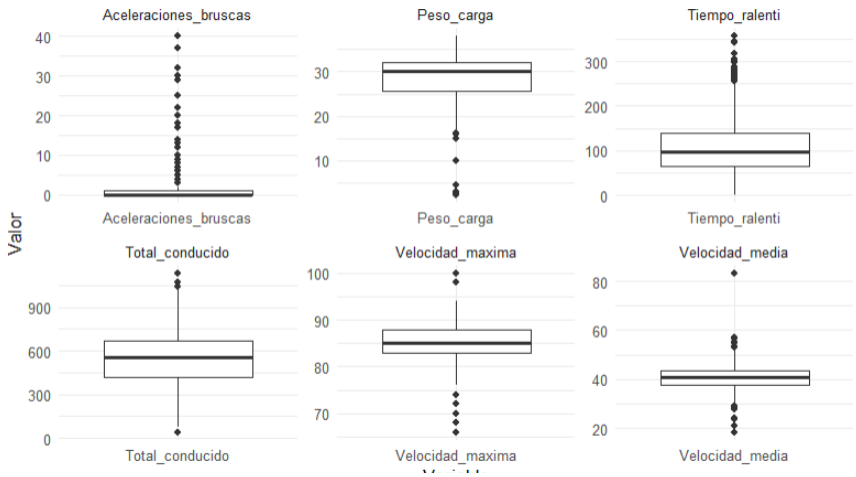

Figura 22 Diagramas de caja de variables continuas

La importancia de variables obtenida mediante Random Forest (Figura 23) indica que unidad, total conducido y velocidad media son las variables más relevantes para explicar el rendimiento de combustible. Variables como velocidad máxima y aceleraciones bruscas mostraron baja influencia.

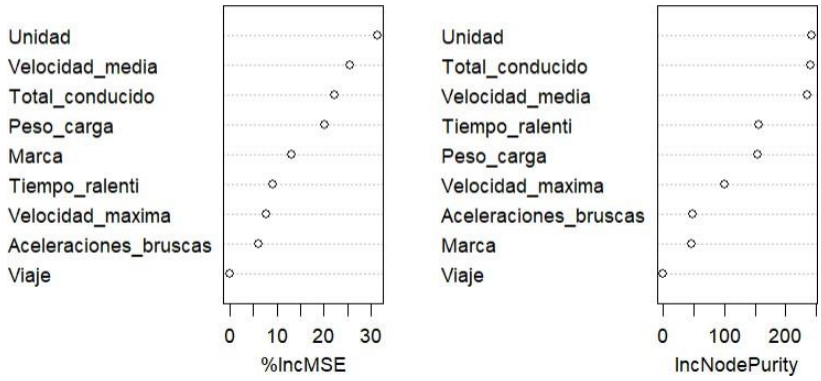

Figura 23 Importancia de variables – Random Forest (regresión)

En la ruta Guayaquil–Cuenca, el análisis de importancia de variables mediante Random Forest identificó a unidad, total conducido y velocidad media como las más influyentes en el rendimiento de combustible. No obstante, para mantener coherencia metodológica con la ruta Cuenca–Guayaquil, se decidió utilizar las mismas variables seleccionadas previamente.

2.3.2.2. Métodos de clasificación supervisada:

- Árbol de decisión

El este modelo (Figura 24) se generó una estructura de clasificación basada principalmente en las variables de peso de carga y velocidad media. Aunque es simple de interpretar, su matriz de

confusión (Figura 25) reveló limitaciones en la identificación correcta de rendimientos adecuados, alcanzando exactitud de 0.589 y sensibilidad de 0.549 (Tabla 7).

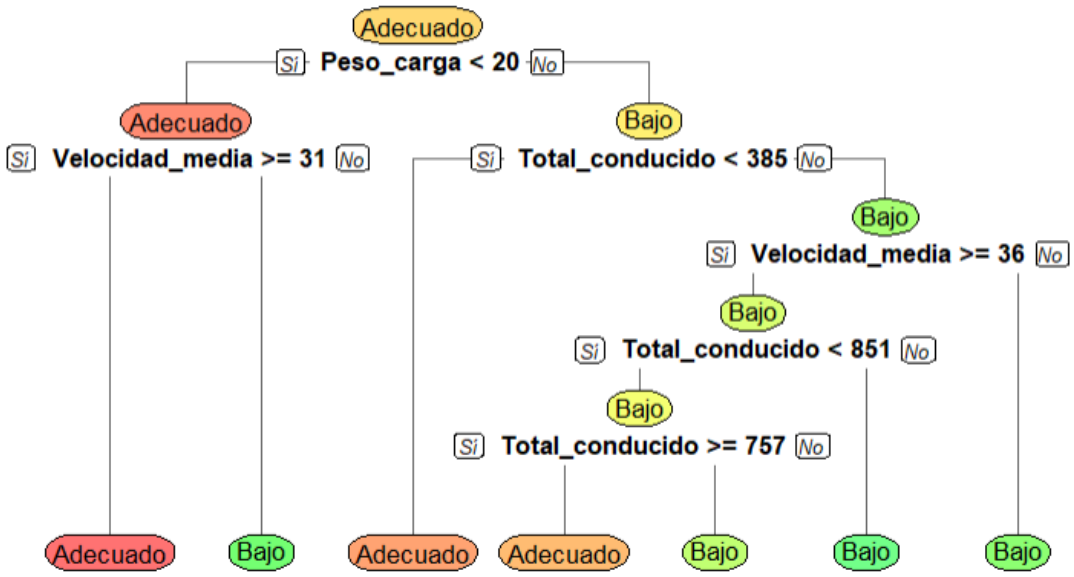

Figura 24 Árbol de decisión Guayaquil - Cuenca

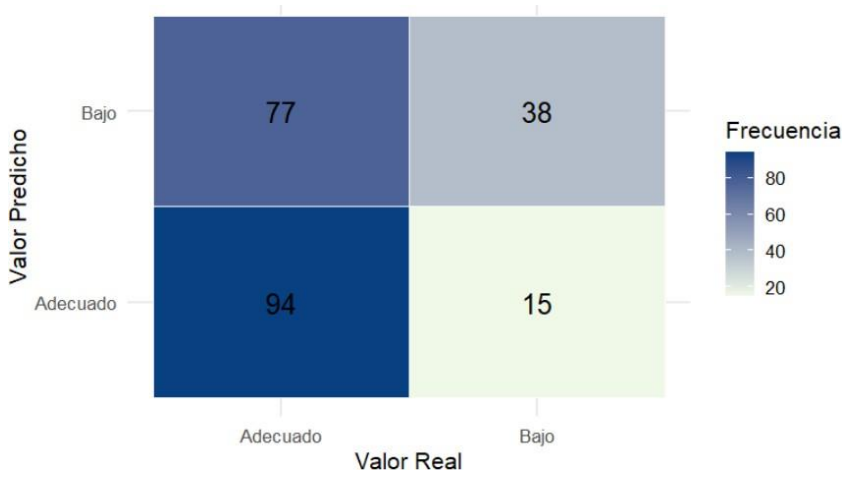

Figura 25 Matriz de confusión – Árbol de decisión Guayaquil - Cuenca

- Random Forest

En este modelo mejoró considerablemente el desempeño (Figura 26), alcanzando una exactitud de 0.669 y un F1- score de 0.76 (Tabla 7), lo que indica un mejor balance entre precisión y sensibilidad, en comparación con el árbol.

Figura 26 Matriz de confusión – Random Forest Guayaquil - Cuenca

- XGBoost

Fue el modelo con mejor rendimiento (Figura 27), logrando un valor de exactitud de 0.716, precisión de 0.797 y sensibilidad de 0.843 (Tabla 7), siendo el más robusto en la clasificación de ambas categorías de rendimiento.

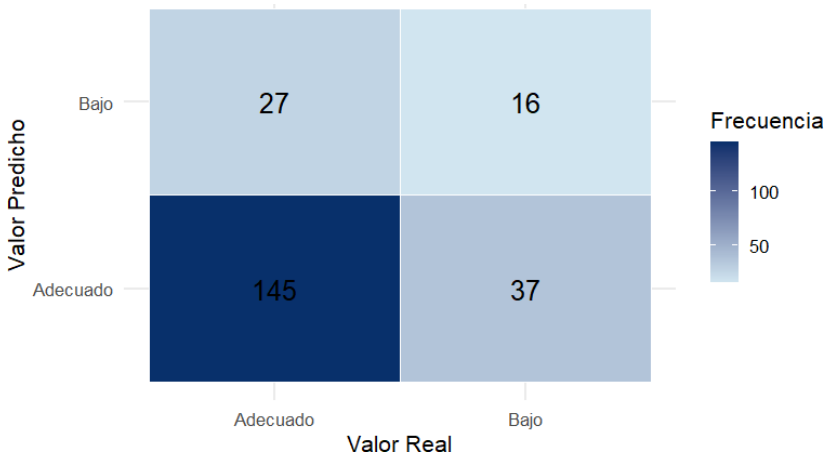

Figura 27 Matriz de confusión – XGBoost Guayaquil - Cuenca

En contraste con la ruta Cuenca - Guayaquil, el modelo XGBoost presentó el mejor desempeño general para la ruta estudiada, evidenciando valores superiores en exactitud y puntaje F1 (Tabla 7) comparado con los demás modelos evaluados, adicional tiene un adecuado equilibrio entre precisión y sensibilidad, lo que indica su capacidad para predecir adecuadamente la mayoría de los casos de consumo de combustible bajo.

Modelo	Exactitud	Puntaje F1	Precisión	Sensibilidad
Árbol de decisión	0.6239	0.7176	0.8438	0.6243
Random Forest	0.5752	0.6643	0.8407	0.5491
XGBoost	0.7156	0.8192	0.7967	0.8430

Tabla 7 Comparación de métricas de evaluación para los tres modelos supervisados

2.3.2.3. Métodos de clasificación no supervisada:

- Clustering K-means (k = 3)

Aunque visualmente los grupos mostraron cierta separación (Figura 28), el índice de silueta fue 0.168 lo que indica una agrupación muy débil y poco estructurada.

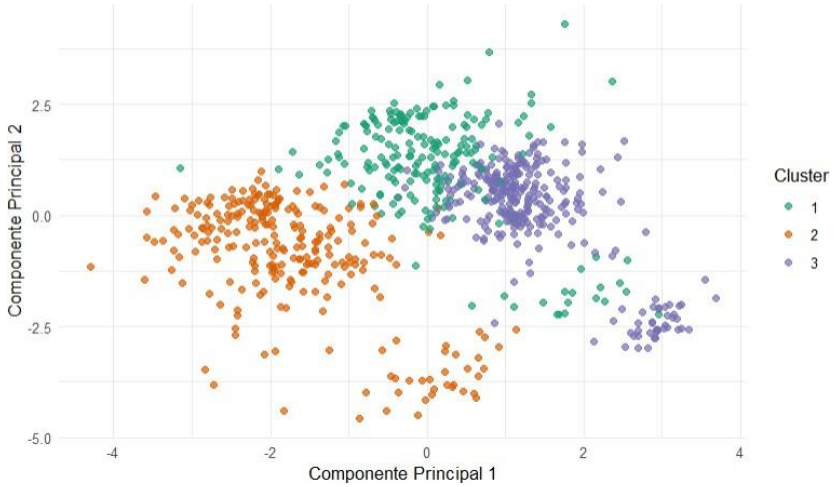

Figura 28 Visualización de clusters mediante PCA – K-means (k = 3)

- Clustering K-means (k = 2)

Reduciendo el número de grupos a dos (Figura 29), el índice de silueta descendió a 0.149, confirmando una segmentación aún más débil.

Figura 29 Visualización de clusters K-means (k = 2)

- Clustering Jerárquico

El dendrograma obtenido (Figura 30) permitió agrupar los datos en tres clusters con una separación más clara a nivel visual. El índice de silueta fue de 0.313, lo que representa un desempeño moderadamente mejor que K-means, aunque todavía dentro de la categoría de clustering débil con posible traslape.

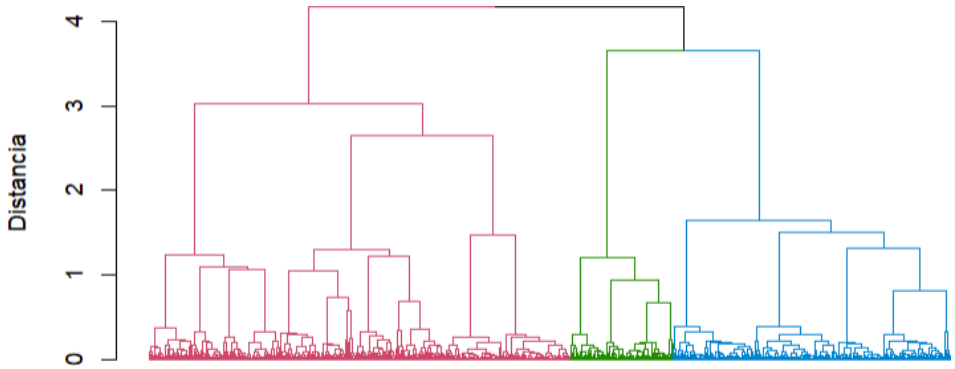

Figura 30 Dendrograma – Clustering jerárquico

- Clustering con PAM

El método PAM presentó el mejor desempeño con un índice de 0.337 (Figura 31). Aunque sigue siendo bajo, pero es mejor en comparación de los otros modelos utilizados.

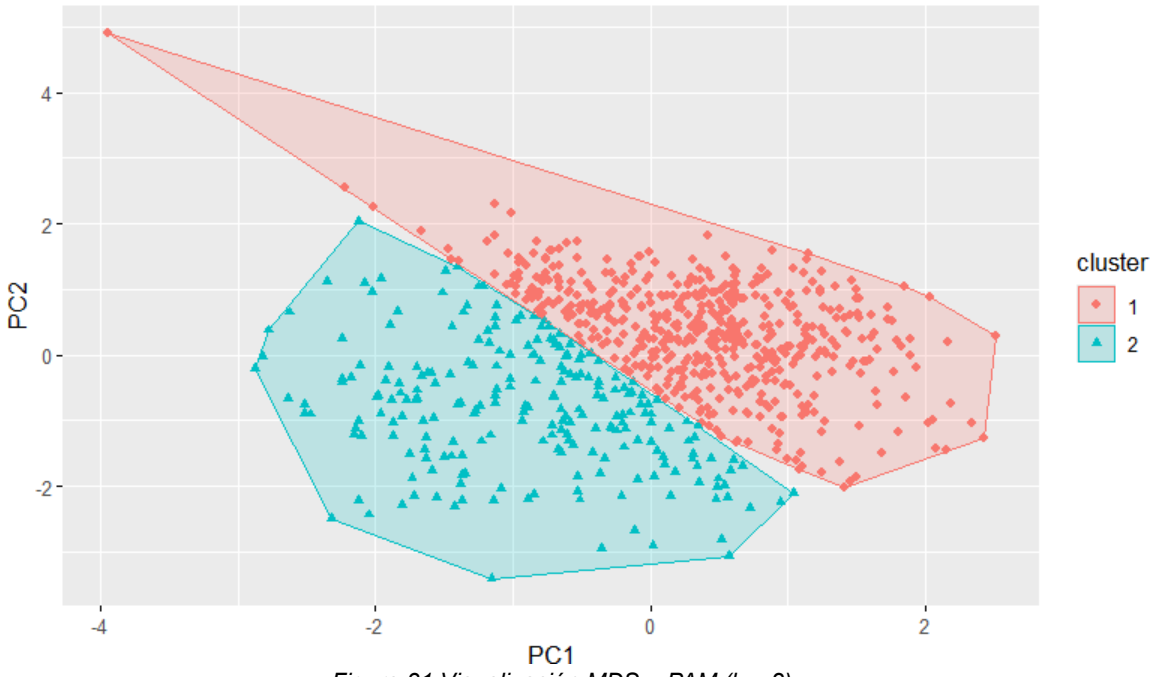

Figura 31 Visualización MDS – PAM (k = 2)

En general, los modelos utilizados no muestran una modera agrupación de los datos. Se recomienda probar otros métodos o mejorar el preprocesamiento de variables para una mejor diferenciación.

2.3.2.4. Visualizaciones avanzadas:

Al analizar el rendimiento del combustible en la ruta Guayaquil- Cuenca, utilizando un mapa de calor (Figura 32). Se puede observar que los niveles más bajos de eficiencia se concentran en las combinaciones de baja velocidad media y cargas pesadas, lo que muestra un impacto negativo de estas condiciones en el consumo de combustible. Por otro lado, los mejores rendimientos se logran en rangos de velocidad media desde 45 y 50 km/h, junto con cargas livianas. Esta tendencia sugiere que, dentro de ciertos límites operativos, aumentar la velocidad media puede mejorar la eficiencia del combustible, especialmente cuando no se transportan cargas excesivamente pesadas.

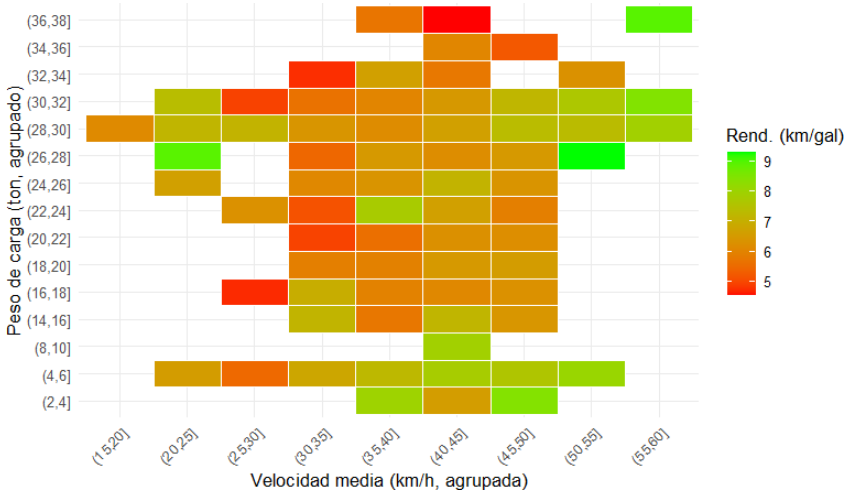

Figura 32 Mapa de calor entre velocidad media, peso de carga y rendimiento

2.4. Conclusiones parciales:

El análisis mostró que las variables que más impactan en el rendimiento del combustible cambian según la ruta. En el trayecto de Cuenca a Guayaquil, los factores más destacados son la velocidad media, el tiempo de conducción y el peso de la carga. Por otro lado, en el recorrido de Guayaquil a Cuenca, lo que más influye son la unidad, el tiempo de conducción y la velocidad media. También se notaron diferencias operativas entre ambos trayectos, como un mayor tiempo en ralentí y aceleraciones bruscas en el regreso. Estos hallazgos respaldan la efectividad de modelos como Random Forest para identificar los factores clave que afectan la eficiencia del transporte.

Los modelos de clasificación han sido clave para identificar los factores que influyen en el

rendimiento de combustible en ambas rutas. En el trayecto de Cuenca a Guayaquil, Random Forest se destacó al ofrecer un mejor equilibrio entre precisión y sensibilidad. En la ruta de Guayaquil a Cuenca, el modelo XGBoost se destacó como el mejor, logrando altos niveles de precisión y sensibilidad. Estos resultados indican que la elección del modelo ideal depende del contexto de cada ruta, y es fundamental considerar tanto la capacidad predictiva como la interpretabilidad para su aplicación práctica en la gestión del rendimiento de la flota.

El análisis de clustering reveló que la segmentación en ambas rutas no fue muy sólida. El método PAM proporcionó la mejor separación relativa para ambas rutas. No obstante, en la ruta Guayaquil – Cuenca su resultado fue menor. Por lo tanto, se sugiere mejorar el preprocesamiento o considerar otros métodos para conseguir agrupaciones más claras y útiles

3. Laboratorio: Enfoque Toma de Decisiones

3.1. Objetivos específicos:

1. Identificar los tipos de modelos de conducción a través de los modelos de clasificación en las rutas Cuenca – Guayaquil y Guayaquil – Cuenca
2. Generar propuestas para mejorar los modelos de conducción bajos a modelos eficientes

3.2. Métodos:

Se utilizó el clustering con PAM realizado en el laboratorio anterior para las dos rutas por sus valores mas altos en el índice de silueta, donde el número de grupos es K=2. En la ruta Cuenca – Guayaquil, se identificó a los viajes como eficiente o ineficiente y se entrenó un modelo de clasificación con el 80% de los datos y el 20% para la validación utilizando el modelo KNN (K-Nearest Neighbors). En cambio, para la ruta Guayaquil – Cuenca se analizaron los clusters sin aplicar una clasificación adicional, dado que los resultados indicaron un buen rendimiento de combustible.

3.3. Resultados y discusión:

3.3.1. Resultados obtenidos de los viajes de Cuenca a Guayaquil:

En la Figura 20 que se mostró en el laboratorio anterior se muestra el gráfico de dispersión de los clusters obtenidos, donde se observan dos grupos, el cluster 1, representado por puntos rojos y el cluster 2 representado por triángulos azules. Esta visualización demuestra la segmentación realizada.

A su vez, al analizar los valores medios de cada grupo (Tabla 8), se observa que el Cluster 1 presenta un menor rendimiento de combustible con un 4.52 Km/gal, mayor peso de carga con 31.22 toneladas, menor velocidad media 31.11 Km/h, mayor tiempo en ralentí con 90.85 minutos y más aceleraciones bruscas con 2.69. A diferencia del Cluster 2, que tiene un rendimiento de

combustible significativamente mayor con 5.89 km/gal, menos peso de carga con 28.35 toneladas velocidad media superior en 35.61 km/h, menor tiempo en ralentí 70.59 minutos y menos aceleraciones bruscas con 1.55.

Cluster	Rendimiento	Peso de carga	Velocidad media	Tiempo en ralentí	Aceleraciones bruscas	n
1	4.52	31.22	31.11	90.85	2.69	459
2	5.89	28.35	35.61	70.59	1.55	258

Tabla 8 Promedio de variables operativas por cluster

Una vez obtenido los clusters, se agruparon los viajes como eficientes o ineficientes. Posteriormente, se entrenó un modelo de clasificación utilizando el algoritmo KNN. Se obtuvo como resultado un 0.9407 de precisión, lo que indica una óptima concordancia entre la predicción y los datos reales. Adicional, se obtuvo un resultado de 0.9891 de sensibilidad y una especificidad del 0.8544, lo que demuestra su capacidad para reconocer viajes ineficientes.

3.3.2. Resultados obtenidos de los viajes de Guayaquil a Cuenca:

En la Figura 31 se presenta el gráfico de dispersión de los clusters obtenidos, donde se identificaron dos grupos, el Cluster 1 representados por puntos rojos y el Cluster 2 por los triángulos azules. A pesar de que el valor del índice de silueta fue moderado la visualización muestra una separación aceptable entre los grupos.

Al analizar los valores medios de cada Cluster (Tabla 9), se evidencia que ambos grupos presentan características relacionadas a un buen rendimiento de combustible. En el Cluster 1, presenta un rendimiento medio de 6.11 km/gal, un peso de carga de 30.03 toneladas, velocidad media de 40.38 km/h, tiempo en ralentí de 92.27 minutos y 1.23 aceleraciones bruscas por viaje. En cambio, el Cluster 2 tiene un rendimiento superior con 7.51 km/gal, un peso de carga significativamente menor de 17.42 toneladas, velocidad media de 41.49 km/h, tiempo en ralentí considerablemente mayor con 146.16 minutos y más aceleraciones bruscas con 1.93 por viaje. Se puede inferir que debido al cargar menos peso el rendimiento de combustible aumenta sin importar que el tiempo en ralentí sea mayor.

Cluster	Rendimiento	Peso de carga	Velocidad media	Tiempo en ralentí	Aceleraciones bruscas	n
1	6.11	30.03	40.38	92.27	1.23	526
2	7.51	17.42	41.49	146.16	1.93	226

Tabla 9 Promedio de variables operativas por cluster

Sin embargo, al analizar las medias de los clusters, se evidenció que ambos grupos presentaban un buen rendimiento de combustible en esta ruta. Por esta razón, no se aplicó un modelo de clasificación, ya que no se identificaron patrones de conducción ineficientes que justificaran una intervención.

3.4. Conclusiones parciales:

El análisis realizado permitió identificar diferencias relevantes entre las rutas Cuenca – Guayaquil y Guayaquil – Cuenca. Se concluye que la primera ruta mencionada representa un área crítica para el rendimiento de combustible ya que se identificaron viajes ineficientes relacionados a mayores tiempos en ralentí, mayor peso de carga, menor velocidad media y mayor número de aceleraciones bruscas por viaje. Por lo contrario, en la ruta Guayaquil – Cuenca mantiene un comportamiento más estable y eficiente con un rendimiento de combustible más elevado en los dos clusters.

Al utilizar el método de clasificación KNN se clasifico los viajes ineficientes en la ruta Cuenca – Guayaquil. Sin embargo, es importante aclarar algunas limitaciones del estudio, puesto que el índice de silueta obtenido para ambas rutas fue moderado, lo que señala que la separación entre los grupos no es la óptima. Además, no se consideran variables externas que puede influir en el estudio como el tráfico, condiciones climáticas o comportamientos de conducción, las cuales podrían afectar los resultados.

Con base a los resultados obtenidos, se recomienda principalmente enfocar todos los esfuerzos en mejorar el rendimiento de la ruta Cuenca – Guayaquil, ya que en esta se detectaron viajes ineficientes con características operativas que afectan al consumo de combustible. Para mejorar el rendimiento se sugiere identificar los viajes que sean ineficientes para implementar un plan de acción que incluya la capacitación a los conductores en técnicas de conducción eficiente, horarios para reducir los tiempos en ralentí, revisión de la planificación de la carga, con el fin de determinar causas específicas y tomar medidas correctivas.

4. Laboratorio: Enfoque Robustez y Validación

4.1. Objetivo específico:

Evaluar la validez estadística de los modelos aplicados al rendimiento de combustible, para determinar la confiabilidad en los análisis realizados.

4.2. Metodología:

El estudio se centró en la validación estadística de los métodos predictivos seleccionados en los laboratorios anteriores. En cada caso se evaluaron los aspectos posibles, incluyendo análisis de residuos y verificación de supuestos estadísticos, siendo estos los siguientes:

1. Modelo de Random Forest para elección de variables:

    a. **Validación de Residuos:**

        - Pruebas de normalidad: Shapiro-Wilk y Lilliefors
        - Prueba de ortogonalidad: Durbin-Watson

- Prueba de homocedasticidad: test de Levene

b. **Métricas de Error:**

- Error Cuadrático Medio (RMSE)
- Error Absoluto Medio (MAE)
- Error Porcentual Absoluto Medio (MAPE)
- Coeficiente de Determinación (R^2)

2. Modelos de Random Forest y XGBoost para clasificación:

a. **Validación de Residuos:**

- Pruebas de normalidad: Shapiro-Wilk y Lilliefors
- Prueba de ortogonalidad: Durbin-Watson
- Prueba de homocedasticidad: test de Levene

b. **Métricas de Clasificación:**

- Matriz de confusión
- Exactitud
- Puntuación F1
- Precisión
- Sensibilidad
- Prueba de Chi-cuadrado

3. Análisis de Clustering:

- Prueba de Chi-cuadrado
- Índice de silueta

Para garantizar la reproducibilidad y validez de los resultados, los conjuntos de datos fueron divididos aleatoriamente en subconjuntos del 80% para el entrenamiento y el 20% para validación mediante el establecimiento de una semilla aleatoria fija. Esta partición aleatoria elimina la necesidad de realizar pruebas adicionales de aleatoriedad. Además, todas las pruebas estadísticas se realizaron considerando un nivel de significancia de $\alpha = 0.05$.

4.3. Resultados y discusión:

4.3.1. Modelo de Random Forest para elección de variables:

Según los resultados presentados en la Tabla 10, se evidencia que los datos no siguen una distribución normal, son ortogonales (no presentan autocorrelación) y muestran heterocedasticidad. En cuanto a los errores, estos resultan bajos, lo que refleja una alta precisión predictiva del modelo. Asimismo, el coeficiente de determinación indica que se explica aproximadamente el 90% de la variabilidad de los datos; sin embargo, dado que su valor es cercano a 1, se puede decir que el modelo podría estar presentando sobreajuste.

Supuesto	Prueba	Resultado	Conclusión
Normalidad	Shapiro-Wilk	Pvalor = 9.043e-15	No son normales
Normalidad	Lilliefors	Pvalor = 4.17e-07	No son normales
Ortogonalidad	Durbin-Watson	D = 1.98	Son ortogonales
		Pvalor = 0.792	
Homocedasticidad	Levene	Pvalor = 0.01304	No hay homocedasticidad
Errores	MAE	0.3718	Bajo nivel de error
	RMSE	0.5051	Bajo nivel de error
	MAPE	6.6763 %	Bajo nivel de error, buen ajuste
	R^2	0.9097	Alta capacidad explicativa
	R^2 ajustado	0.9075	Alta capacidad explicativa

Tabla 10 Resultados de los supuestos Random Forest

4.3.2. Modelo de Random Forest para clasificación:

Para los viajes en la ruta de Cuenca – Guayaquil se escogió el modelo de Random Forest y según lo expuesto en la Tabla 11, se evidencia que los datos no siguen una distribución normal, son ortogonales (no presentan autocorrelación), muestran heterocedasticidad y las variables predictoras no son estadísticamente independientes entre sí. El modelo tiende a ser conservador, clasificando más casos como "Bajo rendimiento" de lo que realmente corresponde, lo cual es positivo para la empresa puesto que se presente analizar a este grupo.

Supuesto	Prueba	Resultado	Conclusión
Normalidad	Shapiro-Wilk	Pvalor = 7.755e-11	No son normales
Normalidad	Lilliefors	Pvalor = 2.2e-16	No son normales
Ortogonalidad	Durbin-Watson	D = 2.1	Son ortogonales
		Pvalor = 0.502	
Homocedasticidad	Levene	Pvalor = 0.00015	No hay homocedasticidad
Independencia	Chi-cuadrado	Pvalor = 1.64e-05	Las variables no son independientes
Clasificación	Exactitud	0.712	Buen desempeño general
	Puntuación F1	0.512	Equilibrio moderado entre precisión y sensibilidad
	Precisión	0.485	Casi la mitad de las predicciones positivas fueron correctas
	Sensibilidad	0.541	Detecta correctamente más de la mitad de los casos positivos

Tabla 11 Resultados de los supuestos Random Forest

4.3.3. Modelo XGBoost para clasificación:

Para los viajes en la ruta de Guayaquil – Cuenca se escogió el modelo de XGBoost y según lo expuesto en la Tabla 12, se evidencia que los datos no siguen una distribución normal, son ortogonales, muestran heterocedasticidad y las variables predictoras no son estadísticamente independientes. Los resultados de clasificación indica que presenta un desempeño satisfactorio para la clasificación de los rendimientos de combustible.

Supuesto	Prueba	Resultado	Conclusión
Normalidad	Shapiro-Wilk	Pvalor = 0.0013	No son normales
Normalidad	Lilliefors	Pvalor = 1.013e-05	No son normales
Ortogonalidad	Durbin-Watson	D = 1.94 Pvalor = 0.544	Son ortogonales
Homocedasticidad	Levene	Pvalor = 0.0252	No hay homocedasticidad
Independencia	Chi-cuadrado	Pvalor = 0.00592	Las variables no son independientes
Clasificación	Exactitud	0.715	Buen desempeño general
	Puntuación F1	0.819	Buen equilibrio entre precisión y sensibilidad
	Precisión	0.796	Alta proporción de predicciones positivas correctas
	Sensibilidad	0.843	El modelo detecta la mayoría de los casos positivos

Tabla 12 Resultados de los supuestos XGBoost

4.3.4. Análisis de Clustering:

Según los resultados presentados en la Tabla 13, los valores observados en ambas rutas son similares, lo que indica que las variables no son independientes y presentan cierta dependencia entre sí. Además, se evidencia una cohesión moderada en los agrupamientos, aunque la separación entre los grupos no es completamente nítida.

Ruta	Supuesto	Prueba	Resultado	Conclusión
Cuenca – Guayaquil	Independencia	Chi-cuadrado	Pvalor = 2.2e-16	Las variables no son independientes
	Cohesión de grupos	Indice de silueta	0.3619	Agrupamiento moderado
Guayaquil – Cuenca	Independencia	Chi-cuadrado	Pvalor = 2.22e-16	Las variables no son independientes
	Cohesión de grupos	Indice de silueta	0.35	Agrupamiento moderado

Tabla 13 Resultados supuestos clustering

4.4. Conclusiones parciales:

En todos los modelos analizados, los datos presentan características comunes: no son normales, son ortogonales, muestran heterocedasticidad y las variables predictoras no son completamente independientes.

Los modelos Random Forest para clasificación y XGBoost son altamente replicables debido a la consistencia de características estadísticas y sus métricas de desempeño. Ambos muestran comportamiento conservador operativamente favorable. No se recomienda el Random Forest inicial utilizado para selección de variables por evidencias de sobreajuste, lo que compromete su confiabilidad para predicción directa. Los modelos finales de clasificación proporcionan herramientas confiables para optimización de combustible.

Se podría incorporar nuevas variables como el estado de la vía, el clima, revoluciones por minuto, entre otras para mejorar la separación entre grupos y la interpretabilidad de los resultados.

Conclusión general:

El presente estudio caracterizó el rendimiento de combustible de una empresa de transporte de carga pesada con datos tomados de un año en las rutas Cuenca–Guayaquil y Guayaquil–Cuenca, donde se identificó las variables más influyentes y se evaluó distintos modelos estadísticos y de aprendizaje automático para su análisis y predicción.

Los resultados demostraron diferencias significativas entre ambas rutas, puesto que la ruta Cuenca-Guayaquil presentó un rendimiento menor con mayor variabilidad, mientras que la ruta Guayaquil-Cuenca alcanzó un desempeño superior con comportamientos más estables. Adicional, se evidencio que las variables: velocidad media, tiempo total de conducción y peso de la carga determinan en mayor proporción el rendimiento de combustible.

En modelado supervisado, Random Forest demostró ser el mejor modelo para determinar los viajes con bajo rendimiento en la ruta Cuenca – Guayaquil y para la segunda ruta fue XGBoost que presentó mejor desempeño predictivo. Para el análisis no supervisado en las dos rutas el modelo que permitió una clasificación óptima fue clustering con PAM donde se segmentó en dos grupos.

A partir de estos grupos en la ruta Cuenca–Guayaquil se diferenciaron claramente viajes eficientes e ineficientes, siendo estos últimos caracterizados por mayor peso de carga, menor velocidad media, mayor tiempo en ralentí y más aceleraciones bruscas. Sin embargo, en Guayaquil–Cuenca ambos clusters mostraron buen rendimiento debido principalmente a cargas más livianas.

Se concluye que la optimización del rendimiento de combustible requiere intervenciones específicas en la ruta Cuenca–Guayaquil, priorizando la capacitación en conducción eficiente, la gestión de tiempos en ralentí y la planificación de cargas. No obstante, las principales limitaciones fueron la falta de variables que podrían atribuirse a un alto consumo de combustible como: estado de la vía, clima, tránsito, condiciones topográficas, entre otras.

Bibliografía:

EP Petroecuador. (s. f.). *Histórico de precios a nivel de terminal*. EP Petroecuador. Recuperado el 05 de octubre de 2025, de https://www.eppetroecuador.ec/?p=20421

Espinoza, J., & Viteri, M. (2025). *Las importaciones de combustible diésel en el crecimiento económico del Ecuador*. *Revista de Economía Aplicada*, *22*(2), 232–245. https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0798-10152025000200232

Bermeo, A., & Calderón, R. (2025). *Optimización de medios de transporte para el abastecimiento eficiente*. *Revista de Ingeniería Logística*, *14*(3), 175–190. https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S2739-00632025000300175

Corcoba Magaña, V., & Muñoz Organero, M. (2014). *Eco-driving: ahorro de energía basado en el comportamiento del conductor* [Tesis doctoral, Universidad Carlos III de Madrid]. CORE. https://core.ac.uk/download/pdf/29406639.pdf

Espinoza Cuadrado, J., Pantoja Villacís, D., Castro Herrera, C., Sangovalin Chuilisa, J., & Villamarín Molina, J. (2022). *Consumo de combustible frente a la eco conducción y tráfico en una ruta mixta en la ciudad de Quito*. *Revista Científica y Tecnológica UPSE (RCTU)*, *9*(2), 85–96. https://doi.org/10.26423/rctu.v9i2.708

Llanes-Cedeño, E. A., Grefa Shiguango, S. F., Molina-Osejos, J. V., & Rocha-Hoyos, J. C. (2024). *Incidencia del aire acondicionado automotriz en el índice de consumo de combustible en vehículo de encendido provocado en una ruta de la Amazonía ecuatoriana*. *Ingenius. Revista de Ciencia y Tecnología*, *(31)*, 115–126. https://doi.org/10.17163/ings.n31.2024.10

Repsol Ecuador. (2024). *Estrategias de responsabilidad social y sostenibilidad: un estudio comparativo*. *Revista de Sostenibilidad Empresarial*, *10*(4), 8–22. https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S2665-01692024000400008