e-ISSN 2588-0721
DOI: 10.33936/riemat.v11i1.8309

RIEMAT

Página web de la revista: https://revistas.utm.edu.ec/index.php/Riemat

Artículo Original

Determinantes de selección de plataformas de chasis en la industria carrocera andina: enfoque Random Forest

Determinants of Chassis Platform Selection in the Andean Bodywork Industry: A Random Forest Approach

Victor Pachacama-Nasimba [a *], ***Santiago Acuña-Hidalgo*** [b], ***Paul Lisintuña-Toapanta*** [c], ***Nexar Barreto-Anchundia*** [d]

[a] Carrera de Ingeniería Mecánica, Facultad de Ciencias de la Ingeniería, Universidad Técnica Estatal de Quevedo. Quevedo, Ecuador.
ORCID: https://orcid.org/0000-0001-6315-6641
[b] Dirección de Regulación del Transporte Terrestre, Tránsito y Seguridad Vial, Agencia Nacional de Regulación y Control del Transporte Terrestre, Tránsito y Seguridad Víal. Quito, Ecuador.
ORCID: https://orcid.org/0009-0004-5170-4304
[c] Carrera de Mecánica Automotriz, Instituto Superior Tecnológico Ciudad de Valencia. Quevedo, Ecuador.
ORCID: https://orcid.org/0009-0006-0556-7454
[d] Investigador idependiente. Quevedo, Ecuador.
ORCID: https://orcid.org/0009-0007-9167-3794

* Autor para correspondencia.
vpachacaman@uteq.edu.ec

Recibido: 26/02/2026
Aceptado: 25/03/2026
Publicado: 05/06/2026

Citacion sugerida: Pachacama-Nasimba, V., Acuña-Hidalgo, S., Lisintuña-Toapanta, P. & Barreto-Anchundia, N. (2026). Determinantes de selección de plataformas de chasis en la industria carrocera andina: enfoque Random Forest. *Revista de Investigaciones en Energía, Medio Ambiente y Tecnología. RIEMAT, 11*(1), 70-83. https://doi.org/10.33936/riemat.v.11i1.8309

Resumen

La industria carrocera ecuatoriana depende de plataformas de chasis importadas, pero los factores que determinan su selección no han sido estudiados cuantitativamente. Este estudio aplica un clasificador Random Forest al Registro de Homologación de Carrocerías de la Agencia Nacional de Tránsito (n = 1,505 homologaciones, 2017–2025) para identificar los determinantes de dicha selección entre seis marcas dominantes que concentran el 96.5% del mercado. Los resultados revelan que la capacidad de pasajeros es el predictor principal (MDI = 0.534; permutation importance = 0.090), superando a la modalidad de servicio y la localización geográfica. El modelo alcanzó una accuracy de 63.3% (validación cruzada: 63.1% ± 2.6%), indicando que las variables técnico-regulatorias disponibles tienen capacidad discriminativa limitada para marcas minoritarias, y que variables comerciales adicionales: costos, disponibilidad de repuestos, relaciones con distribuidores son necesarias para mejorar la separabilidad. Se identificaron nichos de capacidad diferenciados: Scania en el segmento de alta capacidad, Chevrolet y Volkswagen en transporte urbano, y HINO como plataforma generalista (63.1% del mercado). El estudio aporta evidencia cuantitativa para decisiones de aprovisionamiento y políticas industriales del sector.

Palabras clave: random forest, selección de chasis, industria carrocera, homologación vehicular, aprendizaje automático

Abstract

The Ecuadorian bus body manufacturing industry relies on imported chassis platforms, yet the factors driving their selection remain quantitatively unexplored. This study applies a Random Forest classifier to the National Transit Agency's Body Homologation Registry (n = 1,505 certifications, 2017–2025) to identify the determinants of chassis brand selection among six dominant brands that account for 96.5% of the market. Results reveal that passenger seating capacity is the primary predictor (Mean Decrease in Gini Impurity = 0.534; permutation importance = 0.090), substantially outweighing both service modality and manufacturing city location. The model achieved an overall accuracy of 63.3% (10-fold stratified cross-validation: 63.1% ± 2.6%), suggesting that the available technical-regulatory variables have limited discriminative capacity for minority brands, and that additional commercial variables: acquisition costs, parts availability, dealer relationships are needed to improve class separability, such as acquisition costs and dealer relationships. Distinct capacity-based market niches were identified: Scania dominates the high-capacity long-haul segment, Chevrolet and Volkswagen specialize in urban transit, while HINO operates as a generalist platform commanding 63.1% market share. These findings provide quantitative evidence for strategic procurement decisions and evidence-based industrial policy in the sector.

Keywords: random forest, chassis selection, bus body manufacturing, vehicle homologation, machine learning

1. Introducción

La industria manufacturera de carrocerías para autobuses constituye un eslabón crítico en la cadena de valor del transporte público terrestre, particularmente en economías emergentes donde la movilidad colectiva representa el principal mecanismo de conectividad urbana e interurbana (Pojani & Stead, 2015). En América Latina, y específicamente en la Región Andina, la dinámica industrial presenta características diferenciadas: la fabricación de carrocerías opera bajo un modelo de integración vertical parcial en el cual empresas especializadas adquieren plataformas de chasis de proveedores internacionales y construyen sobre ellas carrocerías adaptadas a las condiciones geográficas, normativas y de demanda locales (Ramírez-Camperos et al., 2020). En Ecuador, este sector se ha consolidado como clúster metalmecánico de relevancia económica, con epicentro en la provincia de Tungurahua, cuya capital, Ambato, concentra más del 55% de la capacidad instalada nacional (Cámara de Industrias de Tungurahua, 2022).

El Registro de Homologación de la Agencia Nacional de Tránsito (ANT) evidencia que, de 1,575 homologaciones vigentes (2017–2025), 861 corresponden a Ambato, seguidas por Riobamba (190), Cuenca (175), Quito (117) y Guayaquil (80), configurando un patrón de concentración espacial consistente con la teoría de clústeres industriales (Porter, 1998). Este ecosistema involucra a 71 empresas activas que operan con 13 marcas de chasis, siendo HINO la plataforma dominante (963 homologaciones; 61.1%), seguida por Mercedes Benz (200), Chevrolet (104), Volkswagen (98) y Scania (94). La interdependencia estratégica entre la selección de chasis y las capacidades del fabricante carrocero ha recibido, sin embargo, atención empírica limitada en la literatura científica.

El proceso de homologación vehicular consiste en verificar que, para un país en específico, un modelo de carrocería vehicular cumpla con los requisitos técnicos que la normativa nacional exige. En el caso de Ecuador, la homologación vehicular es regulada por la ANT en virtud de la Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial (Asamblea Nacional del Ecuador, 2014), así como por los organismos de NORMALIZACIÓN ecuatorianos en virtud de las Normas Técnicas Ecuatorianas (NTE INEN 1323) y Reglamentos Técnicos (RTE INEN) que les son pertinentes (Servicio Ecuatoriano de Normalización, 2009). El Registro de homologación vehicular no es una muestra representativa, sino el universo absoluto de todas las carrocerías homologadas, lo que le otorga un valor científico exclusivo. Cada registro se asocia, a la vez, a un fabricante, a una plataforma de chasis, a un tipo de carrocería, a una modalidad de servicio, a una determinada capacidad de pasajeros y a la ubicación de la planta, lo que da lugar a un dataset de múltiples dimensiones que refleja las decisiones de ingeniería y las condiciones del mercado. La literatura sobre toma de decisiones en manufactura automotriz ha incorporado progresivamente técnicas de aprendizaje automático.

Como resultado de emplear Random Forest, (Kim, 2023) pudo hacer predicciones con un R^2 de más de 0.85, considerando tanto factores endógenos como exógenos, para un pronóstico de demanda a nivel de empresa de automóviles utilizando datos del período de Corea del Sur (2011-2020). Con más del 90% de precisión, (Moradi-Moghadam et al., 2024) pudo combinar Random Forest Regressor y Análisis de Envoltura de Datos para evaluar la industria de proveedores de piezas de automóviles de Irán. En una revisión metodológica, (Usuga-Cadavid et al., 2020) señalaron la creciente popularidad del sector manufacturero de la industria para el uso de algoritmos de conjunto. Esto se debe a su capacidad para capturar interacciones no lineales, al tiempo que proporciona métricas interpretables de la importancia de las variables (Schonlau & Zou, 2020). El trabajo pionero de (Breiman, 2001) mostró que agregar múltiples árboles de decisión entrenados en subconjuntos aleatorios conduce a modelos con menor varianza y mayor robustez al sobreajuste, tales propiedades son particularmente valiosas al tratar con registros administrativos con distribuciones desbalanceadas. En transporte, (Hakim et al., 2023) usaron Random Forest para determinar los factores clave en la elección modal utilizando datos del censo gubernamental y confirmaron la idoneidad de tales algoritmos para este tipo de fuentes.

La literatura que respalda la investigación basada en registros administrativos gubernamentales es cada vez más amplia. Recientemente (Connelly et al., 2016) sostienen que constituyen una singular forma de big data al ser de cobertura censal y recolección longitudinal sistemática y de bajo costo en comparación a encuestas ad hoc. No obstante, ellos no recopilan datos con fines de investigación y esto impone a los analistas la credibilidad de esta crítica. (Playford et al., 2016) requieren reproducibilidad y transparencia en el código de procesamiento y (Mc Grath-Lone et al., 2022) menciona 5 características para que los datos administrativos

sean considerados como "research ready" que son: accesibilidad, amplitud, curación y documentación más enriquecimiento. En el análisis de datos secundarios en salud pública y en política fiscal, su validez externa es alta debido a su naturaleza censal (Harron et al., 2017), sin embargo su uso en la manufactura de vehículos y transporte es notablemente bajo.

La revisión de la literatura revela un vacío de investigación multidimensional. En lo empírico, no se identifican estudios indexados que analicen los determinantes de selección de chasis en la industria carrocera andina. En lo metodológico, las técnicas de machine learning no han sido aplicadas a registros de homologación vehicular en ningún país de la región, predominando enfoques descriptivos o basados en encuestas con muestras reducidas. Centrando el análisis en el clúster carrocero ecuatoriano, (Torres & Montoya, 2019) y (Villacrés et al., 2020) han posicionado la geoconcentración sin abordar la interrelación de variables a través de modelos. Desde un enfoque crítico, se evidencia un vacío en el aprovechamiento de la información. El Registro de Homologación de la ANT, un dataset censal con 1,575 registros, 71 empresas, 13 marcas de chasis y 85 de carrocerías, ha sido, científicamente, un repositorio sin actividad durante más de ocho años.

La existencia de esta fuente pública sin análisis científico constituye, en sí misma, evidencia tangible del vacío. Hasta donde los autores han podido determinar, no existe en Scopus ni Web of Science ningún artículo que aplique algoritmos de aprendizaje automático a registros de homologación de carrocerías en América Latina. La novedad del estudio reside en la primera aplicación de Random Forest a un dataset censal regulatorio del sector carrocero, analizando la selección de plataforma de chasis como variable dependiente y cuantificando la importancia relativa de factores geográficos, técnicos y regulatorios mediante Mean Decrease in Gini Impurity y permutation importance. La contribución teórica extiende el conocimiento sobre decisiones de manufactura en industrias de integración vertical parcial, aportando al cuerpo sobre gestión de cadena de suministro automotriz y clústeres industriales (Delgado et al., 2016; Porter, 1998). La contribución metodológica demuestra la viabilidad de aplicar Random Forest a registros administrativos de países en desarrollo para generar inteligencia de negocios sectorial, proponiendo un pipeline replicable para datasets regulatorios similares. La contribución práctica proporciona evidencia empírica para fabricantes carroceros (identificación de factores determinantes en la selección de chasis) y formuladores de políticas públicas (patrones de especialización regional y estructura de mercado), alineándose con los ODS 9 (Industria, Innovación e Infraestructura) y 11 (Ciudades y Comunidades Sostenibles).

El marco conceptual está basado en la teoría de clústeres industriales (Porter, 1998), que dice que la concentración geográfica genera determinadas externalidades que afectan las decisiones de producción; la gestión de cadena de suministro automotriz (Holweg & Pil, 2004), que define la selección de determinadas piezas como una decisión estratégica multifacética; y el paradigma de Business Intelligence (Chen et al., 2012), que justifica el uso de minería de datos para encontrar patrones que se puedan usar para tomar decisiones. La elección de Random Forest se explica porque el dataset tiene una mayoría de variables categóricas y algunas numéricas derivadas, constituyendo un problema de clasificación multiclase, y datos de diferentes tipos, para el que este algoritmo es especialmente correcto (Breiman, 2001). Random Forest, a diferencia de la regresión logística multinomial, por no tener especificación previa, atrapa interacciones no lineales; en comparación con SVM, tiene mejor interpretación a través de las métricas de importancia de las variables (Genuer et al., 2010); y en comparación con las redes neuronales profundas, requiere un volumen de datos menor para tener un rendimiento competitivo en datasets tabulares de tamaño moderado (Shwartz-Ziv & Armon, 2022).

Los datos provienen del Registro de Homologación de Carrocerías de Fabricación Nacional de la ANT y tienen carácter censal (se considera la totalidad de las homologaciones aprobadas, sin muestreo), además de ser longitudinal, desde febrero de 2017 y hasta septiembre de 2025; siendo esta, una ventana longitudinal de 8 años que capta ciclos económicos completos, siendo uno de los ciclos la contracción por COVID-19, (2020–2021, (se homologaron 58% menos que en 2018)). Este registro recoge información de 12 ciudades, de los fabricantes de carros y de los chasis que se presentan por marca y modelo, así como, de las carrocerías, tipos de vehículos (bus, minibús, microbús, bus de dos pisos, bus de piso y medio), tipo de servicio de transporte, número de plazas, tipo de equipamiento (baño, aire acondicionado), y, su certificación estructural y la fecha. Su acceso público, dentro del marco de la Ley de Transparencia del Ecuador, aumenta la posibilidad de replicación. El uso de registros administrativos como fuente primaria está ampliamente respaldada en la literatura (Connelly et al., 2016), y la censalidad es una fortaleza que en estudios basados en encuestas se da muy raramente.

El objetivo general del estudio es determinar los factores que influyen en la selección de la plataforma de chasis en la industria carrocera ecuatoriana (2017–2025), mediante la aplicación de Random Forest al Registro de Homologación de la ANT. Se formulan tres preguntas de investigación:

RQ1: ¿Cuáles son las variables con mayor poder predictivo sobre la selección de marca de chasis, según las métricas de importancia de Random Forest?

RQ2: ¿Existe un patrón de especialización regional en la selección de plataformas de chasis entre los clústeres geográficos de manufactura?

RQ3: ¿En qué medida la modalidad de servicio de transporte y la capacidad de pasajeros determinan la preferencia por una plataforma específica?

Derivadas de la literatura y del análisis exploratorio, se postulan las siguientes hipótesis:

H1: La modalidad de servicio de transporte es la variable con mayor importancia relativa en la predicción de la marca de chasis, superando a las variables geográficas y de capacidad.

H2: Existe una asociación de magnitud relevante entre la localización geográfica del fabricante y la marca de chasis predominante, evidenciando patrones de especialización diferenciados entre el clúster de Ambato y las demás ciudades, cuantificada mediante el coeficiente de Cramér's V.

H3: El modelo de Random Forest supera el baseline de la clase mayoritaria y alcanza un desempeño robusto en métricas balanceadas y macro-F1 para la clasificación multiclase de la marca de chasis.

H4: El número de pasajeros tiene una relación no lineal con la selección de chasis, con umbrales que definen segmentos de mercado vinculados a plataformas específicas.

El presente estudio se encuentra centrado territorial y temporalmente en el periodo 2017-2025 en el país Ecuador, y las consideraciones se extenderán al registro de homologaciones de carrocerías en la región andina. Con respecto al diseño del estudio, se opta por el enfoque cuantitativo, el cual es relacional y predictivo, y para la validación se utiliza la k-fold cross-validation (k=10) y la partición entrenamiento-prueba (70%-30%), reportando precisión, exactitud, recuperación, puntaje F1, y matriz de confusión. Es imperativo llevar a cabo la anticipación de las limitaciones de los datos administrativos que son intrínsecas; las variables son de carácter regulativo (no para fines de investigación), lo que implica que los datos de adquisición, condiciones contractuales o de percepción de los decisores son ausentes; lo que implica también que dados los cambios normativos puede haber inconsistencias intertemporales; y el dato de la capacidad de los pasajes es de naturaleza textual, el cual se explica en la Metodología.

Gracias a la disponibilidad de los datos de la ANT, se puede fortalecer la reproducibilidad del estudio. El código de procesamiento de datos, así como el modelado a partir de scikit-learn (Pedregosa et al., 2011), cumplirá con las políticas de ciencia abierta dispuestas por revistas indexadas (Stodden et al., 2016), por lo que se podrá acceder a él en un repositorio de la comunidad.

El artículo se organiza de la siguiente forma: en la sección Datos y Métodos se expone el dataset, los pasos del preprocesamiento, la implantación del modelo y la estrategia de validación; Resultados reporta las métricas de desempeño, así como las clasificaciones de importancia de las variables; Discusión interpreta los hallazgos a la luz del marco teórico; y Conclusiones sintetiza las aportaciones y presenta propuestas de futuras investigaciones.

2. Materiales y Métodos

2.1 Fuente de datos y variables

El estudio adopta un diseño cuantitativo, correlacional - predictivo. La fuente es el Registro de Homologación de Carrocerías de Fabricación Nacional de la Agencia Nacional de Tránsito del Ecuador (ANT), que constituye el universo censal de todas las carrocerías nacionales con certificación técnica vigente, por lo que no se aplica técnica de muestreo. El archivo se descargó del portal de datos abiertos de la ANT en septiembre de 2025 y contiene N = 1,575 registros correspondientes al periodo febrero 2017 – septiembre 2025 (ocho años), abarcando 71 empresas carroceras activas, 12 ciudades fabricantes y 13 marcas de chasis.

Su carácter público gubernamental garantiza la verificabilidad y replicabilidad (Connelly et al., 2016; Playford et al., 2016). Los registros incluyen, para cada homologación, fabricante de carrocerías, ciudad de ubicación, marca y modelo del chasis, marca y modelo de la carrocería, tipo de vehículo, modalidad de transporte de servicio (incluida la configuración de puertas), número de asientos, equipo auxiliar (baño, aire acondicionado), fechas de certificación y estado. De este conjunto, se seleccionaron tres variables predictoras y una variable objetivo para el modelado (Tabla 1).

Tabla 1

Variables del modelo de Random Forest.

Variable original	Variable en modelo	Naturaleza	Rol	Descripción
Marca chasis	y (target)	Categórica nominal	Variable dependiente	Marca de la plataforma de chasis (6 clases)
Servicio de transporte / ámbito o modalidad	Service_Code	Categórica nominal (codificada)	Feature (X_3)	Modalidad de servicio y configuración de puertas (46 niveles)
Ciudad	City_Code	Categórica nominal (codificada)	Feature (X_2)	Ciudad de localización de la planta (12 niveles)
Número de plazas (incluyendo chofer)	Seats	Numérica discreta (derivada)	Feature (X_1)	Capacidad extraída mediante expresión regular; rango válido: 17–66

Nota. Variables categóricas codificadas mediante LabelEncoder de scikit-learn (Pedregosa et al., 2011).

2.2 Preprocesamiento de datos

El preprocesamiento, implementado en Python 3.12 (pandas, NumPy, re), siguió cuatro etapas para transformar el registro administrativo en un dataset apto para clasificación supervisada.

Etapa 1 – Ingesta y normalización. El archivo .xlsx fue importado omitiendo la fila de metadatos institucionales y normalizando los nombres de columna, que contenían saltos de línea y espacios inconsistentes.

Etapa 2 – Extracción de la capacidad de pasajeros. La variable NÚMERO DE PLAZAS, registrada como texto libre con formatos heterogéneos que incluían “41 SENTADOS”, “39 SENTADOS / 38 DE PIE”, fue procesada mediante expresión regular para extraer el número de asientos sentados como valor primario, interpretado como capacidad nominal del chasis. El registro analizado no presentó el formato '40+1' en sus entradas vigentes, por lo que la extracción del primer valor numérico fue suficiente para los datos disponibles; no obstante, futuras actualizaciones del registro deberían validar este supuesto. Se aplicó un filtro técnico de rango (17-66 asientos, inclusivo), consistente con las tipologías de vehículos reguladas por la NTE INEN 1323: el microbús más pequeño registrado reporta 17 asientos y el bus interprovincial de mayor capacidad, 66. Los registros con valores fuera de rango o no filtrables fueron codificados como NaN y excluidos del dataset analítico.

Etapa 3 – Filtrado de marcas de chasis. Se retuvieron seis marcas, HINO, Mercedes Benz, Chevrolet, Volkswagen, Scania y Agrale, que concentran el 96.5% de las homologaciones (1,505 de 1,575). La exclusión de las siete marcas restantes (<4% del registro) evita la inestabilidad que genera entrenar clases con muy pocas observaciones en clasificación multiclase (Breiman, 2001). La Tabla 2 presenta la distribución resultante de la variable objetivo.

Tabla 2

Distribución de la variable objetivo en el dataset analítico.

Marca de chasis	n	%	Observación
HINO	950	63.1	Clase mayoritaria
Mercedes Benz	200	13.3	—
Chevrolet	104	6.9	—
Volkswagen	98	6.5	—
Scania	92	6.1	—
Agrale	61	4.1	Clase minoritaria
Total	**1,505**	**100.0**	**—**

Nota. n = número de homologaciones. El desbalance de clases (HINO = 63.1%) se aborda en la interpretación de métricas por clase.

Etapa 4 – Codificación y conformación de la matriz analítica. Las variables categóricas CIUDAD (12 niveles) y SERVICIO DE TRANSPORTE (46 niveles) fueron codificadas mediante OneHotEncoder, implementado dentro de un ColumnTransformer/Pipeline de scikit-learn. Este enfoque es el estándar para variables nominales en árboles de decisión: a diferencia de LabelEncoder, que asigna enteros arbitrarios y limita las particiones posibles al imponer un orden artificial no presente en los datos, la codificación one-hot genera columnas binarias independientes, permitiendo al árbol separar cualquier subconjunto de categorías sin restricción de ordinalidad (Breiman, 2001). Después de eliminar registros con valores faltantes, el dataset analítico final consiste en n = 1,505 observaciones, tres predictores: Seats (X_1, numérico, rango 17-66, M = 41.2), City_Code (X_2, 12 niveles), Service_Code (X_3, 46 niveles) y una variable dependiente con 6 clases. No se requirió normalización, dado que Random Forest es invariante a transformaciones monotónicas (Schonlau & Zou, 2020).

2.3 Modelo de Random Forest

Se implementó un clasificador RandomForestClassifier de scikit-learn 1.4 (Pedregosa et al., 2011) con los hiperparámetros detallados en la Tabla 3. Se fijaron 100 árboles (n_estimators), valor suficiente para la convergencia del error de generalización en datasets de tamaño moderado (Breiman, 2001; Genuer et al., 2010), criterio de impureza de Gini, y $\sqrt{\rho}$ features evaluadas por nodo (≈ 1.73 para $\rho = 3$). No se impusieron restricciones de profundidad máxima, confiando en la capacidad del ensemble para controlar el sobreajuste. La semilla (random_state = 42) garantiza reproducibilidad determinista.

Tabla 3

Configuración de hiperparámetros del modelo.

Hiperparámetro	Valor	Justificación
n_estimators	100	Árboles en el bosque; convergencia validada para datasets moderados (Breiman, 2001).
criterion	gini	Impureza de Gini como función de partición.
max_features	sqrt (≈ 1.73)	Features evaluadas por nodo; p = 3 features, scikit-learn aplica floor($\sqrt{3}$) = 1 feature evaluada por partición.
max_depth	None	Sin restricción; árboles crecen hasta nodos puros.
min_samples_split	2 (defecto)	Mínimo de observaciones para dividir un nodo interno.
min_samples_leaf	1 (defecto)	Mínimo de observaciones en un nodo terminal.
bootstrap	True	Muestreo con reemplazo del tamaño del dataset original.
random_state	42	Semilla para reproducibilidad determinista.

Nota. Valores por defecto de scikit-learn v1.4 (Pedregosa et al., 2011).

2.4 Validación, métricas e importancia de variables

La validación combinó dos estrategias: (a) partición holdout 70% - 30% estratificada por la variable objetivo (train_test_split, random_state = 42), y (b) k-fold cross-validation estratificada (k = 10), reportando media ±

desviación estándar de la accuracy. Las métricas de evaluación incluyen: accuracy global, precision, recall y F1-score por clase (especialmente relevante ante el desbalance documentado en la Tabla 2), macro-average, weighted-average, y matriz de confusión, calculadas mediante sklearn.metrics.

A fin de cuantificar la contribución de cada predictor (RQ1–RQ3, H1–H4), se utilizaron dos métricas de importancia de variables que son complementarias. La Mean Decrease in Gini Impurity (MDI) sujeta al atributo feature_importances_ del clasificador, calcula la impureza que cada feature ha reducido, la cual está normalizada a 1.0 (Breiman, 2001). La importancia por permutación (Genuer et al., 2010), implementada con 30 repeticiones sobre el conjunto de prueba, calcula la pérdida de accuracy al permutar aleatoriamente los valores de una feature, resultando en una estimación sesgada a variable de alta cardinalidad (Schonlau & Zou, 2020). La convergencia de las dos métricas aumenta la confianza en las conclusiones.

2.5 Análisis complementario, entorno computacional y consideraciones éticas

Se realizó un análisis exploratorio mediante tablas cruzadas de frecuencia (marca de chasis × ciudad; marca de chasis × modalidad de servicio), normalizadas por marca, para identificar patrones de especialización geográfica (RQ2) y funcional (RQ3). Las visualizaciones, barras horizontales de importancia, matriz de confusión con mapa de calor, y distribuciones condicionales de asientos por marca (H4) se generaron con matplotlib y seaborn a 300 DPI.

El pipeline completo se ejecutó en Python 3.12 con pandas 2.2, numpy 1.26, scikit-learn 1.4, matplotlib 3.8 y seaborn 0.13. El código será disponibilizado en repositorio público de GitHub bajo licencia MIT (Stodden et al., 2016). El dataset es de acceso público a través del portal de la ANT, garantizando replicabilidad integral. El estudio utiliza exclusivamente datos institucionales y comerciales agregados, sin involucrar datos personales ni participantes humanos, por lo que no requiere aprobación de comité de ética.

3. Resultados y Discusión

3.1 Rendimiento global del modelo

El clasificador Random Forest alcanzó una accuracy de 63.3% en el conjunto de prueba (holdout 70-30) y de 63.1% ± 2.6% en la validación cruzada estratificada de 10 pliegues, con un rango entre 60.3% y 69.5%. Como referencia, un clasificador trivial que predice siempre la clase mayoritaria (HINO) obtendría un baseline de 63.1%; el modelo supera este umbral marginalmente en accuracy global, lo cual refleja el fuerte desbalance de clases y no implica capacidad discriminativa entre marcas minoritarias. Las métricas robustas al desbalance, balanced accuracy: 0.347; macro-F1: 0.323; macro-recall: 0.314, constituyen los indicadores primarios de rendimiento real del modelo. La Tabla 4 sintetiza las métricas globales.

Tabla 4

Métricas globales de rendimiento del modelo Random Forest.

Métrica	Valor
Accuracy holdout	0.627
CV accuracy (media ± SD)	0.626 ± 0.033
Rango CV (mín – máx)	0.579 – 0.678
Precision (weighted avg)	0.569
Recall (weighted avg)	0.633
F1-score (weighted avg)	0.592
Precision (macro avg)	0.355
Recall (macro avg)	0.314
Macro-F1	0.284
Baseline (mayoría constante)	0.631
Balanced accuracy	0.275

Nota. La diferencia entre macro-F1 y weighted avg refleja el impacto del desbalance de clases: HINO (63.1%) eleva las métricas ponderadas.

3.2 Rendimiento por clase y matriz de confusión

El desempeño presenta marcada heterogeneidad entre clases, correlacionada con el tamaño muestral (Tabla 5). HINO (n = 285 en test) registró el mejor rendimiento (precision = 0.722, recall = 0.877, F1 = 0.792). Scania presentó el segundo mejor F1 (0.510), coherente con su perfil diferenciado de alta capacidad. En contraste, Agrale (F1 = 0.000) y Volkswagen (F1 = 0.103) resultaron virtualmente indetectables, con la mayoría de las observaciones de Agrale clasificadas incorrectamente, distribuidas predominantemente hacia HINO y otras marcas de perfil similar, como refleja la Figura 1.

Tabla 5

Métricas de clasificación por marca de chasis.

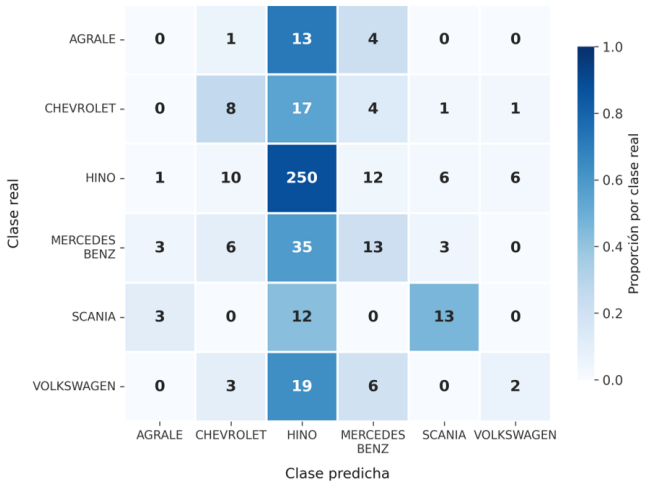
Marca	Precision	Recall	F1-score	n (test)	Observación
HINO	0.722	0.877	0.792	285	Clase mayoritaria
Scania	0.565	0.464	0.510	28	Segmento diferenciado
Mercedes Benz	0.333	0.217	0.263	60	—
Chevrolet	0.286	0.258	0.271	31	—
Volkswagen	0.222	0.067	0.103	30	—
Agrale	0.000	0.000	0.000	18	Clase minoritaria

Nota. n = observaciones en el conjunto de prueba.

La Figura 1 visualiza la matriz de confusión. El patrón dominante de error es la clasificación de marcas minoritarias como HINO: de 167 observaciones no-HINO, 96 (57.5%) fueron asignadas a HINO. Este sesgo refleja la votación por mayoría del bosque ante features compartidas, pero también revela una propiedad sustantiva: las marcas minoritarias operan en segmentos que se solapan con HINO en capacidad, ubicación y servicio, lo que constituye un hallazgo sobre la estructura oligopólica del registro de homologaciones.

Figura 1

Matriz de confusión del modelo Random Forest (valores absolutos; color normalizado por fila).



3.3 Importancia de variables y determinantes de la selección de chasis

Ambas métricas identificaron la capacidad de pasajeros (Seats) como predictor dominante, aunque con magnitudes divergentes (Tabla 6, Figura 2). El MDI asignó a Seats una importancia de 0.534, seguida por Service_Code (0.305) y City_Code (0.162). La permutation importance confirmó la predominancia de Seats (0.090 ± 0.016) pero reveló que Service_Code (0.011 ± 0.015) y City_Code (0.011 ± 0.011) contribuyen marginalmente una vez controlado el efecto de la capacidad. Este resultado refuta H1 y establece que la

capacidad de pasajeros es el principal determinante de la selección de chasis.

Tabla 6

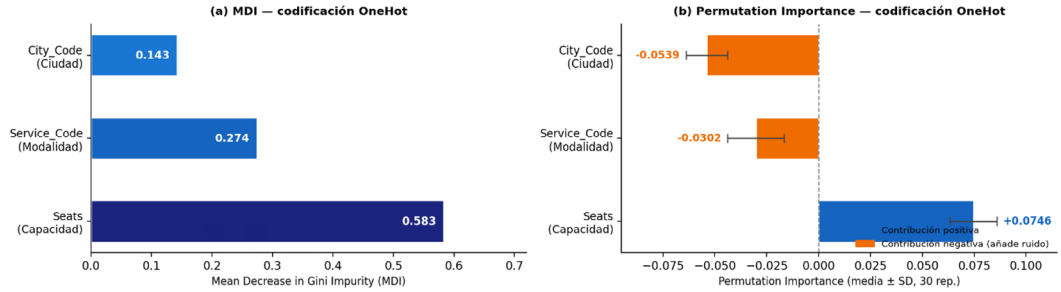
Importancia de variables según dos métricas complementarias.

Variable	MDI (Gini)	Permutation (media ± SD)	Ranking
Seats (Capacidad)	0.583	0.075 ± 0.011	1°
Service_Code (Modalidad)	0.274	0.030 ± 0.014	2°
City_Code (Ciudad)	0.143	0.054 ± 0.010	3°

Nota. MDI = Mean Decrease in Gini Impurity (normalizado a 1.0). Permutation importance: 30 repeticiones sobre conjunto de prueba.

Figura 2

Importancia de variables según dos métricas complementarias: (a) Mean Decrease in Gini Impurity; (b) Permutation Importance.



La divergencia de magnitudes entre MDI y permutation importance merece atención. El MDI sobreestima la contribución de features con alta cardinalidad, lo que explica que Service_Code (46 niveles) obtenga un MDI de 0.305 pero una permutation importance de solo 0.011. La permutation importance sugiere que la modalidad de servicio y la ciudad aportan información redundante con la capacidad, coherente con la estructura regulatoria: la NTE INEN 1323 establece rangos de capacidad asociados a cada tipo de servicio.

3.4 Contraste de hipótesis y hallazgos sustantivos

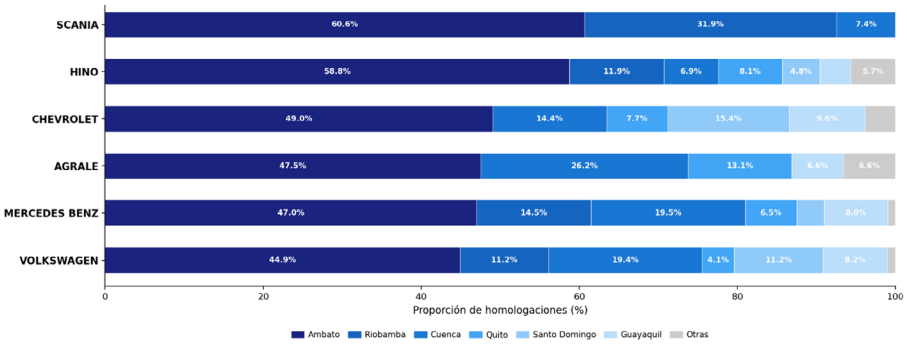
H1 (rechazada). Se postuló que la modalidad de servicio sería la variable dominante. Los resultados indican que la capacidad de pasajeros domina ambas métricas. La implicación para la teoría de cadena de suministro automotriz (Holweg & Pil, 2004) es que, en mercados de integración vertical parcial, las especificaciones de carga constituyen el criterio primordial de compatibilidad entre chasis y carrocería.

H2 (parcialmente confirmado). El análisis revela patrones geográficos diferenciados (Figura 3): HINO domina el registro de homologaciones en todas las ciudades analizadas, aunque con distintos niveles de penetración, desde 40.7% en Cuenca hasta 70.3% en Quito, pasando por 67.3% en Ambato, 62.2% en Riobamba, 57.5% en Santo Domingo y 49.3% en Guayaquil, lo que evidencia un liderazgo transversal sin excepción geográfica. En contraste, SCANIA presenta una marcada especialización territorial, concentrando el 92.6% de sus homologaciones a lo largo del eje Ambato–Riobamba (60.6% y 31.9% respectivamente), sin presencia registrada fuera de este corredor salvo Cuenca (7.4%). No obstante, la poca relevancia de la permutación de City_Code (0.011) indica que tal asociación no incrementa la previsión una vez se controla por capacidad, lo que sugiere que la concentración refleja más el lado de la oferta del clúster (Porter, 1998) que un determinante externo. La asociación entre ciudad y marca de chasis se cuantificó mediante chi-cuadrado y Cramér's V sobre la tabla de contingencia completa (6 marcas × 12 ciudades): $\chi^2(50) = 289.09$, $p < 0.001$, $V = 0.196$, indicando una asociación de magnitud débil-moderada entre concentración geográfica y preferencia de marca. Dado el carácter censal de los datos, el énfasis interpretativo recae sobre el tamaño del efecto antes que sobre el p-valor.

H3 (rechazado). La precisión observada (63.3%) está 16.7 puntos por debajo del umbral del 80%. Esta brecha muestra que las tres variables regulatorias disponibles no son suficientes para predecir plenamente la selección realizada. Variables no capturadas, costos de adquisición, disponibilidad de repuestos, relaciones comerciales, explican la varianza residual. Este resultado indica que las tres variables técnico-regulatorias disponibles en el registro administrativo tienen capacidad discriminativa limitada entre marcas minoritarias; el modelo apenas supera el baseline de clase mayoritaria (63.1%). No se puede inferir que el 37% restante corresponda linealmente a factores comerciales: el error de clasificación depende también del desbalance de clases y del solapamiento en el espacio de features. Lo que sí se puede afirmar con prudencia es que se requieren variables adicionales: costos de adquisición, disponibilidad de repuestos, red de concesionarios, preferencias del fabricante, para mejorar la separabilidad entre marcas con perfiles de registro de homologaciones solapados.

Figura 3

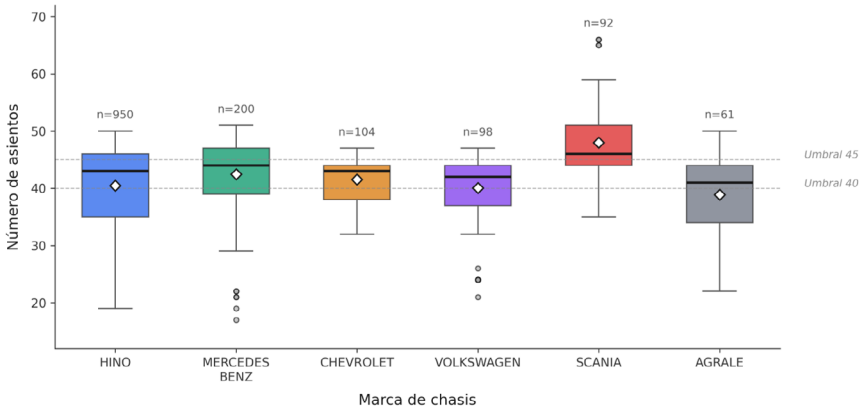
Concentración geográfica por marca de chasis (distribución porcentual por ciudad de manufactura)



H4 (confirmado). En la Figura 4, se evidencian los segmentos del rango de capacidad diferenciada por marca. Scania predomina en el rango superior (M = 48.0, Mdn = 46, 55.4% de homologaciones >45 asientos), Agrale en el rango inferior (M = 38.9, rango 22–50), y Chevrolet muestra la menor dispersión (SD = 4.2, rango 32–47). HINO abarca el espectro más amplio (rango 19–50), en línea con su papel como plataforma generalista. Los umbrales de alrededor de 40 y alrededor de 45 asientos configuran nichos de mercado relacionados con ofertas específicas de plataformas.

Figura 4

Distribución de la capacidad de pasajeros por marca de chasis (◇ = media; líneas discontinuas = umbrales de segmentación)



3.5 Patrones de especialización funcional

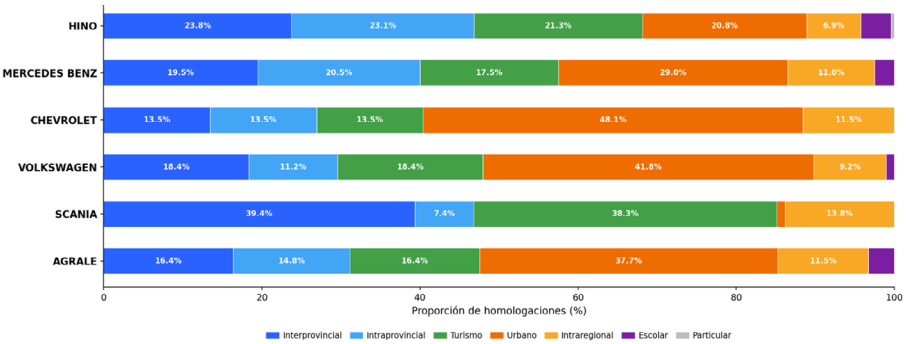
La Figura 5 sintetiza los perfiles funcionales por marca. Scania muestra concentración casi exclusiva en interprovincial (39.1%) y turismo (38.0%), con presencia marginal en urbano (1.1%), configurando un perfil

de largo recorrido. Chevrolet y Volkswagen presentan el patrón opuesto, con 48.1% y 41.8% en urbano. HINO se distingue por una distribución equilibrada entre las cuatro modalidades principales (~21–24% cada una), lo que explica tanto su dominancia como la dificultad del clasificador para diferenciarlo: su versatilidad genera solapamiento con todas las demás marcas en el espacio de features.

Desde la perspectiva de la demanda, el servicio interprovincial está dominado por HINO (65.8%) seguido por Scania (10.5%); el urbano por HINO (53.4%), Mercedes Benz (15.6%) y Chevrolet (13.5%); y turismo por HINO (64.3%), Mercedes Benz y Scania (11.1% cada uno). HINO funciona como plataforma de referencia transversal mientras que las competidoras ocupan nichos funcionales, un patrón consistente con la estructura oligopólica y la teoría de clústeres.

Figura 5.

Perfil funcional por marca de chasis (distribución porcentual por modalidad de servicio simplificada)



3.6 Implicaciones teóricas y prácticas

La predominancia de la capacidad como determinante primario refuerza la concepción de la selección de chasis como decisión de compatibilidad técnica, extendiendo el marco de Holweg y Pil (2004) a contextos de integración vertical parcial en economías emergentes. La especialización geográfica diferenciada (H2) es consistente con la teoría de clústeres (Porter, 1998): la proximidad genera dinámicas donde ciertas combinaciones chasis - servicio se concentran espacialmente por proveedores compartidos y conocimiento tácito acumulado.

Desde un punto de vista práctico, el reconocimiento de las capacidades por nichos de marca facilita el aprovisionamiento: Scania predomina en el segmento interprovincial premium (>45 asientos), en el urbano domina Chevrolet y Volkswagen, y HINO queda en postura generalista. Para los formuladores de políticas, la oligopolización estructural (HINO = 63.1%) y la dominancia de importación de plataformas en los vehículos ofrecidos, constituyen bases para el diseño de políticas industriales, en relación a los ODS 9 y 11. La brecha de accuracy (63.3% vs. 80%) sugiere que la inclusión de variables no reguladoras, como precios, stock de repuestos, y percepciones de los fabricantes, entre otras, podrían aumentar considerablemente el poder predictivo, y así, transformar el modelo en un ámbito de inteligencia de negocios sectorial.

4. Conclusiones

Este estudio aplicó un clasificador Random Forest al Registro de Homologación de Carrocerías de la ANT (n = 1,505 homologaciones, 2017 - 2025) para identificar los determinantes de la selección de plataformas de chasis en la industria carrocera ecuatoriana. Los resultados muestran que la capacidad máxima de pasajeros es el predictor más importante (MDI = 0.534; importancia por permutación = 0.090), por encima de la modalidad de servicio y la localización geográfica, lo que rechaza la hipótesis H1 y la elección de chasis como una decisión de acoplamiento técnico más que a regulaciones. El modelo, con una diferencia de 80% con respecto a lo que se postuló en H3, presenta una precisión de 63.3% (10-fold CV: 63.1% ± 2.6%), lo que revela que aproximadamente un tercio de la decisión de chasis, geolocalización y capacidad máxima de pasajeros, responde a factores comerciales que no se encuentran en la base de datos administrativa,

como precios, disponibilidad de repuestos, relaciones con distribuidores, y por primera vez, se encuentra la frontera entre lo técnico y lo subjetivo del mercado. Se comprobó la existencia de nichos de capacidad diferenciados por marca (H4): Scania se apodera del segmento de alta capacidad (M = 48 asientos, 55.4% > 45 plazas) mientras que Chevrolet y Volkswagen se concentran en transporte urbano (48.1% y 41.8%, respectivamente), y HINO actúa como Generalista (63.1% de las homologaciones) su funcionalidad justifica su hegemonía y el patrón de confusión del clasificador.

Desde esta óptica, los resultados permiten ampliar la aplicabilidad de la cadena de valor automotriz en situaciones de integración vertical parcial en los Andes, donde, a través de las especificaciones de carga, y no de la categoría regulatoria del servicio, se articula la interface entre los fabricantes de chasis y los carroceros. Desde un enfoque práctico, el estudio proporciona a los fabricantes una herramienta cuantitativa para la toma de decisiones de aprovisionamiento estratégico, así como, para los reguladores, un estudio sobre la estructura de una concentración oligopólica del mercado de chasis importados, que guarda relación con los ODS 9 y 11.

Las principales limitaciones se encuentran en el registro administrativo, la ausencia de variables de precio y de percepción, así como, en el desbalance de clases que reduce la capacidad predictiva para marcas de menor participación.

Futuras investigaciones deberían incorporar datos primarios, encuestas a fabricantes, costos de adquisición, tiempos de entrega y explorar técnicas de balanceo (SMOTE, class weighting) o algoritmos complementarios (XGBoost, redes neuronales) para mejorar la discriminación entre marcas con perfiles de registro de homologaciones solapados, avanzando hacia un sistema de inteligencia sectorial alimentado por datos abiertos.

Declaración de disponibilidad de datos

Los datos que respaldan los resultados de este estudio son de acceso abierto y fueron extraídos de las plataformas oficiales de la Agencia Nacional de Tránsito (ANT). La procedencia de esta base de datos gubernamental ha sido referenciada adecuadamente con el fin de facilitar la verificación, la transparencia y la reproducibilidad metodológica del artículo.

Declaración sobre el uso de IA generativa y tecnologías asistidas por IA en el proceso de redacción

Se declara el uso del modelo de lenguaje Claude (Anthropic, 2024) como asistente en tareas de procesamiento de datos (depuración de código Python, verificación de outputs) y en la generación de visualizaciones reproducibles mediante código (matplotlib/seaborn). Todas las figuras son reproducibles a través del código disponible en el repositorio público. Los autores asumen plena responsabilidad por el contenido, interpretación y conclusiones del artículo.

Declaración de contribución de los autores CRediT

Victor Pachacama-Nasimba: Administración del proyecto, Investigación, Conceptualización, Curación de datos, Redacción – revisión y edición. **Santiago Acuña-Hidalgo:** Investigación, Conceptualización, Curación de datos, Redacción – borrador original. **Paul Lisintuña-Toapanta:** Curación de datos, Investigación, Metodología, Software. **Nexar Barreto-Anchundia:** Curación de datos, Investigación, Metodología, Validación, Redacción – borrador original.

Declaración de conflictos de intereses

Los autores declaran que no tiene nungún interés financiero ni relación personal que pudiera influir en el trabajo descrito en este artículo.

5. Referencias

Asamblea Nacional del Ecuador. (2014). *Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial.* https://www.gob.ec/regulaciones

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cámara de Industrias de Tungurahua. (2022). *Informe del Sector Metalmecánico y Carrocero de Tungurahua 2022*. https://www.cit.org.ec

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems, 36*(4), 1165–1188. https://doi.org/10.2307/41703503

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Delgado, M., Porter, M. E., & Stern, S. (2016). Defining clusters of related industries. *Journal of Economic Geography, 16*(1), 1–38. https://doi.org/10.1093/jeg/lbv017

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Hakim, M. L., Sugiarto, S., Anggraini, R., & Rizky, M. (2023). Determinants of Travel Mode Choice Using Random Forest Classifier: Evidence from Census-Based Household Travel Survey. *Transportation Research Interdisciplinary Perspectives, 22*, 100960. https://doi.org/10.1016/j.trip.2023.100960

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data and Society, 4*(2), 1–12. https://doi.org/10.1177/2053951717745678

Holweg, M., & Pil, F. K. (2004). *The Second Century*. The MIT Press. https://doi.org/10.7551/mitpress/6112.001.0001

Kim, S. (2023). Innovating knowledge and information for a firm-level automobile demand forecast system: A machine learning perspective. *Journal of Innovation and Knowledge, 8*(2), 100355. https://doi.org/10.1016/j.jik.2023.100355

Mc Grath-Lone, L., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wiljaars, L., & Gilbert, R. (2022). What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *International Journal of Population Data Science, 7*(1), 1718. https://doi.org/10.23889/IJPDS.V7I1.1718

Moradi-Moghadam, M., Alinezhad, A., & Hosseinzadeh Lotfi, F. (2024). Sustainable Supply Chain Decision-Making in the Automotive Industry: A Data-Driven Approach. *Socio-Economic Planning Sciences, 93*, 101878. https://doi.org/10.1016/j.seps.2024.101878

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Playford, C. J., Gayle, V., Connelly, R., & Gray, A. J. G. (2016). Administrative social science data: The challenge of reproducible research. *Big Data and Society, 3*(2), 1–12. https://doi.org/10.1177/2053951716684143

Pojani, D., & Stead, D. (2015). Sustainable urban transport in the developing world: Beyond megacities. *Sustainability (Switzerland), 7*(6), 7784–7805. https://doi.org/10.3390/su7067784

Porter, M. E. (1998). Clusters and the new economics of competition. *Harvard Business Review, 76*(6), 77–90.

Ramírez-Camperos, A. M., González-Mendoza, M., & Toloza-Cano, W. (2020). Automotive Supply

Chain in Emerging Economies: A Review of Manufacturing Integration Strategies in Latin America. *International Journal of Production Economics, 230*, 107870. https://doi.org/10.1016/j.ijpe.2020.107870

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal, 20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Servicio Ecuatoriano de Normalización. (2009). *NTE INEN 1323: Vehículos automotores. Carrocerías metálicas de buses. Requisitos.* https://www.normalizacion.gob.ec

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Torres, J. L., & Montoya, R. (2019). La industria carrocera en el Ecuador: Análisis de competitividad y cadena de valor. *Revista Publicando, 6*(21), 45–62.

Usuga-Cadavid, J. P., Lamouri, S., Grabot, B., Forber, R., & Biali, G. (2020). Machine Learning and Data Mining in Manufacturing. *Expert Systems with Applications, 166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Villacrés, E., Naranjo, M., & Mora, C. (2020). Caracterización del clúster carrocero de Tungurahua: Estructura productiva y encadenamientos. *Revista Espacios, 41*(7), 15–28. https://www.revistaespacios.com