Análisis de tendencias y predicción de la emisión de facturas para los años 2026-2028 basado en el crecimiento del parque automotor y los cambios en las características de los vehículos

Deyanira Valenzuela Ramos

Asesor

Rafael Gaitán Ospina

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

El desarrollo del proyecto se llevó a cabo mediante un proceso sistemático para estimar la tendencia en la emisión de facturas del impuesto vehicular en Bogotá para el período 2026–2028. Avanzando progresivamente en la consecución de los objetivos específicos, se obtuvo los siguientes resultados.

Consolidar un dataframe totalmente gestionado y depurado, conformado por 8.150.296 registros y ocho variables. Este dataframe corresponde al conjunto de datos de las emisiones de facturas durante los períodos 2022–2025. Con base en estos datos, se realizó el análisis estadístico descriptivo, utilizando tablas cruzadas para explorar las variables relevantes, lo que permitió identificar patrones y tendencias significativas, que son detallados en las conclusiones y recomendaciones del presente estudio.

Finalmente, se construyó un modelo de predicción basado en redes neuronales LSTM (Long Short-Term Memory), como técnica especializada en el análisis de series temporales. Este modelo se desarrolla capa por capa para predecir la emisión anual de facturas para los años 2026, 2027 y 2028. Para fortalecer el modelo, se agruparon algunas variables conforme a sus características claves identificadas en el análisis descriptivo, como la clase de combustible y el tipo de vehículo, de esta manera se logra las predicciones cuantitativas para la emisión de las facturas del impuesto vehicular en Bogotá para el período 2026-2028.

***Palabras claves*:** Análisis, predicción, redes neuronales, series de tiempo, impuesto sobre vehículos.

Abstract

The project was developed through a systematic process to estimate trends in vehicle tax invoice issuance in Bogotá for the period 2026–2028. Progressively advancing toward achieving the specific objectives, the following results were obtained:

A fully managed and refined dataframe was consolidated, consisting of 8,150,296 records and eight variables. This dataframe corresponds to the invoice issuance dataset for the periods 2022–2025. Based on these data, a descriptive statistical analysis was performed, using cross-tabs to explore the relevant variables. This allowed for the identification of significant patterns and trends, which are detailed in the conclusions and recommendations of this study.

Finally, a prediction model was built based on LSTM (Long Short-Term Memory) neural networks, a specialized technique for time series analysis. This model is developed layer by layer to predict annual invoice issuance for the years 2026, 2027, and 2028. To strengthen the model, some variables were grouped according to their key characteristics identified in the descriptive analysis, such as fuel type and vehicle type. This provides quantitative predictions for vehicle tax invoice issuance in Bogotá for the period 2026-2028.

Keywords: Analysis, prediction, neural networks, time series, vehicle tax.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La Secretaría Distrital de Hacienda de Bogotá cumple un rol fundamental como entidad encargada de la política fiscal y la gestión eficiente de los recursos públicos de la ciudad. En este contexto, la gestión tributaria representa una fuente importante de ingresos para Bogotá, por tanto, un objetivo prioritario es la optimización de los procesos de recaudo mediante la gestión y análisis eficiente de los datos. Esto permite anticipar el comportamiento tributario y facilitar la toma de decisiones informadas a nivel directivo.

Bajo estas circunstancias el presente proyecto se enfoca en el diseño e implementación de un modelo predictivo para estimar la emisión de facturas del impuesto vehicular para las vigencias 2026 a 2028, utilizando los datos históricos disponibles entre los años 2022 y 2025. El desarrollo del trabajo se estructuró de la siguiente manera: se consolidó un conjunto de datos completamente depurado, que permitió realizar un análisis estadístico descriptivo orientado a identificar patrones y tendencias a partir de información histórica. Posteriormente, se desarrolló y entrenó un modelo predictivo basado en técnicas de ciencia de datos y aprendizaje automático, utilizando redes neuronales recurrentes del tipo LSTM.

Este documento presenta un informe detallado de los resultados obtenidos en las etapas de análisis y modelado, así como las conclusiones y recomendaciones derivadas del proceso. Se espera que este trabajo sirva como línea base para futuros desarrollos analíticos, y que contribuya a la mejora continua en la gestión del impuesto vehicular en la ciudad.

Planteamiento del Problema

La Secretaría Distrital de Hacienda de Bogotá, responsable del recaudo del impuesto sobre vehículos, basa su proceso de liquidación en datos consolidados de los automotores, como marca, placa, línea, carrocería, cilindraje, modelo y el avalúo comercial establecido por el Ministerio de Transporte. Con esta información, desde la Dirección Distrital de Impuestos se emiten anualmente las facturas correspondientes a los vehículos matriculados en la ciudad. Sin embargo, la imprecisión o ausencia de estos datos impide la emisión de facturas, conllevando a que los propietarios realicen una declaración manual en la Oficina Virtual de la Secretaría para registrar los datos faltantes y generar el Recibo Oficial de Pago (ROP), un documento tributario equivalente a la factura.

El impuesto vehicular es una de las fuentes claves de ingresos para la ciudad, pero su proceso de emisión se ve afectado por factores como cambios en el estado de las matrículas (nuevas, cancelaciones o traslados) y la evolución del parque automotor, influenciada por el tipo de vehículo y la clase de combustible. De acuerdo con información emitida por la Secretaría de Hacienda en 2023, de 2.305.955 vehículos obligados a pagar el impuesto, solo 2.023.442 recibieron factura, dejando 282.513 casos sin facturación debido a datos incompletos, no homologados con las tablas del Ministerio de Transporte o porque el contribuyente ya había declarado manualmente. (Bogotá, 2023). Y para 2025, la entidad reportó la entrega de 2 millones de facturas físicas, destacando la importancia de garantizar el acceso oportuno a la información tributaria para que los contribuyentes aprovechen descuentos por pronto pago. (Bogotá, 2025).

Estos datos reflejan la necesidad de optimizar la gestión tributaria mediante la anticipación de tendencias en el parque automotor, en este sentido, un modelo predictivo permitiría a la secretaría tomar decisiones estratégicas relacionadas con la emisión de facturas, la

logística de entrega y/o el diseño de incentivos en el proceso de matrículas, en beneficio de los contribuyentes.

En este contexto, se formula la siguiente pregunta de investigación: ¿Qué modelos predictivos y de datos se pueden utilizar para estimar con mayor precisión la emisión de facturas del impuesto vehicular en Bogotá para las vigencias 2026-2028?

Justificación

El impuesto sobre vehículos representa una de las fuentes esenciales del recaudo fiscal en Bogotá, por lo que resulta prioritario innovar en los procesos de planificación y gestión tributaria, con el fin de optimizar la emisión de facturas durante el período 2026–2028. Este recaudo depende directamente del número de facturas generadas, un indicador que se ve afectado por variables como la cancelación o traslado de matrículas y la evolución del parque automotor, relacionada con el tipo de vehículo y la clase de combustible.

Dado lo anterior, se identifica una oportunidad para analizar de forma sistemática el comportamiento del parque automotor y el estado de las matrículas a lo largo del tiempo. Ya que la emisión de facturas registra datos históricos organizados por vigencia y fecha de matrícula, resulta pertinente aplicar el modelo de red neuronal recurrente (RNN), diseñado para trabajar con datos temporales para realizar predicciones basadas en patrones aprendidos, para este caso, relacionado con las emisiones de facturas históricas. Esta elección busca anticipar el número de facturas que podrán ser emitidas en años futuros, permitiendo una mejor gestión tributaria.

Desde el ámbito académico, este proyecto aporta al desarrollo de la analítica predictiva aplicada a la gestión pública, al integrar ciencia de datos con políticas fiscales. Su implementación favorece a la entidad al permitir una planificación más eficiente y centrada en el ciudadano, facilitando la toma de decisiones estratégicas, la diversificación de incentivos, el fortalecimiento de la educación tributaria y la promoción del uso de herramientas tecnológicas para lograr una gestión tributaria más efectiva, moderna y transparente.

Objetivos

Objetivo General

Construir un modelo predictivo que permita estimar la emisión de facturas del impuesto vehicular en Bogotá para el período 2026-2028, mediante el uso de técnicas de análisis de datos y modelado predictivo.

Objetivos Específicos

Consolidar los datos históricos sobre la emisión de facturas del impuesto vehicular en Bogotá, correspondientes al período 2022-2025.

Analizar los datos históricos según variables relevantes de las características de los vehículos, con el fin de identificar patrones, tendencias y factores que influyen en la emisión de facturas.

Evaluar un modelo predictivo óptimo para proyectar la emisión de facturas del impuesto vehicular en Bogotá durante el período 2026-2028, empleando técnicas de machine learning.

Metodología

Este proyecto se orienta en la aplicación de la analítica predictiva, una rama fundamental de la ciencia de datos, ya que tiene por objetivo “analizar tendencias y predecir la emisión de facturas para los años 2026-2028, basándose en datos históricos”.

Tras la revisión bibliográfica, se identificó que la metodología más recomendada para el desarrollo de proyectos de análisis de datos es CRISP-DM (Cross Industry Standard Process for Data Mining). Este modelo proporciona un enfoque estructurado y flexible para el análisis de datos y el modelado predictivo, lo que lo hace especialmente adecuado para este estudio gracias al detalle y claridad de sus fases. A continuación, se presenta una descripción de este modelo:

CRISP-DM (Cross Industry Standard Process for Data Mining)

En este modelo se identifica seis fases así:

Fase de comprensión del negocio o problema: Se inicia la comprensión de los objetivos y requisitos del proyecto desde una perspectiva institucional, resaltando la importancia de problema para realizar la recolección de los datos y la interpretación de resultados.

Fase de comprensión de los datos: Se recopilan los datos de la emisión de facturas generado desde el sistema de información de la Secretaría Distrital de Hacienda.

Fase de preparación de los datos: se realiza la preparación de los datos como validación de valores nulos, imputación de datos, selección de variables y limpieza de datos.

Fase de modelado: En esta fase se selecciona las variables predictoras mediante el análisis de correlación de Pearson, para ello se obtiene una matriz de correlación entre las variables transformadas y el número de facturas emitidas por vigencia, permitiendo identificar qué características temporales y categóricas presentaban mayor asociación con el comportamiento histórico de la emisión de facturas.

Fase de Evaluación del modelo: Para evaluar las predicciones del modelo se aplica la métrica MSE (Error Cuadrático Medio) para medir el promedio de los cuadrados de los errores entre las predicciones y los valores reales y MAE (Error Absoluto Medio) para medir el promedio de los valores absolutos entre las predicciones y los valores reales

Fase de implementación: Es la transformación del conocimiento se presenta un basado en el modelo y los resultados.

Consolidación de los Datos Históricos de la Emisión Facturas del Período 2022-2025

Teniendo como base la metodología de CRISP-DM (Cross Industry Standard Process for Data Mining), se inicia con el desarrollo de las fases.

Información del Negocio y Problema

Para iniciar la fase de comprensión del negocio, se identifica que la construcción del modelo predictivo se desarrolla en la Secretaría Distrital de Hacienda, que de acuerdo con el Decreto 601 de 2014 y sus modificaciones, es la entidad del distro de Bogotá que tiene en las funciones, la siguiente:

Formular, orientar, coordinar y ejecutar las políticas tributarias, presupuestales, contables y de tesorería

Esta función que es una de las muchas que tiene definida la entidad, la responsabiliza de llevar a cabo:

La gestión eficaz y eficiente de los recursos públicos, mediante el recaudo de los impuestos, entre los que se encuentra el impuesto sobre vehículos automotores.

Innovar y optimizar sus procesos para garantizar el cumplimiento de la planificación y gestión del recaudo.

En esta misma línea, la administración ha enmarcado en los objetivos estratégicos 2024 – 2027, el siguiente:

Procesos y tecnología: Optimizar el modelo de operación institucional soportado en procesos eficientes y tecnologías de vanguardia, que faciliten la apropiación de la cultura de innovación y gestión de la información.

Figura 1

Objetivos Plan Estratégico Institucional 2024-207

Nota. Tomado de (Secretaria Distrital de Hacienda, 2024)

Adicionalmente la entidad se proyectó con la siguiente visión:

"Visión: En 2030, la Secretaría Distrital de Hacienda será reconocida por su gestión innovadora y eficiente, que afianza la cultura tributaria, impulsa la calidad del gasto público y fortalece la relación de confianza con el ciudadano." Tomado de (Secretaria Distrital de Hacienda, 2024)

De este modo se evidenció que la secretaria distrital de hacienda está soportando e innovando la operación en la entidad, entre otras, con tecnología de vanguardia para contar con una cultura tributaria eficiente, para ello, la aplicación de un modelo predictivo que anticipe

tendencias significativas en el comportamiento del parque automotor contribuye en la planificación y optimización de la emisión de facturas en el período 2026 – 2028. Es así como las tecnologías y modelos de machine learning permiten optimizar la gestión tributaria.

En congruencia, el planteamiento para la construcción del modelo predictivo para estimar la emisión de facturas del impuesto vehicular en Bogotá para el período 2026-2028, se determinó así:

Recopilar y consolidar los datos históricos de la emisión de factura del impuesto vehicular para la vigencia 2022 a 2025.

Realizar análisis descriptivo los datos para identificar patrones, tendencias y factores que influyen en la emisión de facturas.

Aplicar el modelo de redes neuronales recurrentes (RNN) para proyectar la emisión de facturas del impuesto vehicular en Bogotá durante el período 2026-2028

Recopilación de los Datos

Continuando con la metodología, se recopila el conjunto de datos correspondiente a la emisión de facturas de las vigencias 2022 a 2025, en que se obtiene la siguiente información: Clase de combustible por vigencia, tipo de vehículo por vigencia, estado matrícula por vigencia, fecha de matrícula en Bogotá.

La información está contenida en cuatro (4) archivos, todos con extensión txt, llamados: Vigencia2022.txt, Vigencia2023.txt, Vigencia2024.txt y Vigencia2025.txt

Identificación de los Campos

Los cuatro archivos comprendían datos estructurados con los siguientes campos:

VIGEN: Año de emisión factura, para esta variable solo deben estar en los registros los datos: 2022, 2023, 204 y 2025 y su formato debe ser entero, dado que se componen de valores numéricos.

FOING: Forma ingreso de los datos, la variable debe ser texto, por tanto, puede tener el formato string.

FE_ING_TRA: Fecha Ingreso de Tránsito, esta variable debe tener el formato de fecha o su homologo según la herramienta utilizada, y no tiene restricción frente a que fechas debe aparecer.

FE_REMATRI: Fecha matrícula en Bogotá, igualmente esta variable debe tener el formato de fecha o su homologo según la herramienta utilizada, y no tiene restricción frente a que fechas debe aparecer.

ID_COM: Identificador combustible, su formato debe ser entero, dado que se componen de valores numéricos y los únicos valores validos son: 1,2,3,4,5,6,7,8,9,10,11 y 12

DE_COM: Descripción clase combustible, la variable debe ser texto, por tanto, puede tener el formato string, los únicos valores validos son: gasolina, diesel- acpm, gas natural vehicular, gas-gasollina, etanol, eléctrico, diesel-electrico, biodisel, glp, gaso elec., hidrogeno y tracción animal.

ID_TVH: Identificador Tipo de vehículo, su formato debe ser entero, dado que se componen de valores numéricos y los únicos valores validos son: 1,2,3,4,5,6,7,8,9 y10.

DE_TVH: Descripción tipo de vehículo, la variable debe ser texto, por tanto, puede tener el formato string, los únicos valores validos son: automóviles, camperos y camionetas, camionetas doble cabina, eléctricos, motos y motocarros, pasajeros, carga, ambulancias, híbridos y gas.

ID_EST: Identificador estado, su formato debe ser entero, dado que se componen de valores numéricos y los únicos valores validos son: 1,2,7,8 y 9

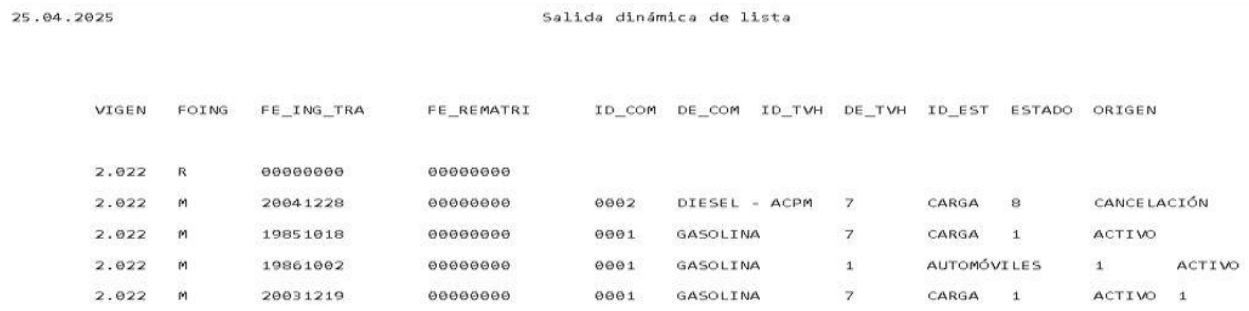
ESTADO: Descripción Estado, la variable debe ser texto, por tanto, puede tener el formato string, los únicos valores validos son: activo, inhabilitado, cancelación y trasladado

ORIGEN: Origen de emisión (factura, declaración), su formato debe ser entero, dado que se componen de valores numéricos y los únicos valores validos son: 1,2 y 3

A continuación, la visualización inicial de uno de los archivos, denominado vigencia 2022.txt.

Figura 2

Imagen de Datos en el Archivo Vigencia2022.txt



25.04.2025 Salida dinámica de lista

VIGEN	FOING	FE_ING_TRA	FE_REMATRI	ID_COM	DE_COM	ID_TVH	DE_TVH	ID_EST	ESTADO	ORIGEN
2.022	R	00000000	00000000							
2.022	M	20041228	00000000	0002	DIESEL - ACPM	7	CARGA	8	CANCELACIÓN	
2.022	M	19851018	00000000	0001	GASOLINA	7	CARGA	1	ACTIVO	
2.022	M	19861002	00000000	0001	GASOLINA	1	AUTOMÓVILES	1	ACTIVO	
2.022	M	20031219	00000000	0001	GASOLINA	7	CARGA	1	ACTIVO	1

Nota. Tomado de Jupyter Notebook.

Luego se sube cada uno de los cuatro (4) archivos a la herramienta jupyter notebook, cada archivo tiene la misma presentación de datos, tal cual muestra la **Figura *2***.

Identificación de los Registros

Finalizado el cargue de los cuatro (4) archivos se logra tener un único dataframe, la cantidad de registros por cada uno de los archivos y el total de registros en el dataframe final quedó así:

Tabla 1

Consolidado de Registros por Archivo

Nombre Archivo por vigencia	Total de registros
Vigencia 2022.txt	2.611.284
Vigencia 2023.txt	2.667.751
Vigencia 2024.txt	2.675.002
Vigencia 2025.txt	2.641.025
Dataframe final	10.595.062

Nota. La tabla muestra por cada archivo el total de registros.

Reconocimiento de Datos

Al consolidar el dataframe en uno solo, se procede a reconocer la cantidad de datos que lo conforman, validando la completitud de estos, que se espera tenga el dataframe en caso de que se presente un dataframe completos, esto sería así:

Tabla 2

Total de Datos Esperados por Vigencia

VIG	FE_REMAT	FE_ING_TR	FOING	ID_COM	DE_COM	ID_TVH	DE_TVH	ID_EST	ESTADO	ORIGEN	TOTALES
2022	2544860	2582564	2544773	2581339	2581338	2579428	2579428	2582445	2578534	2297149	25451858
2023	2640961	2665204	2640860	2664957	2664956	2653643	2653643	2665183	2660568	2502582	26412557
2024	2643257	2667371	2643154	2667122	2667120	2658712	2658712	2667351	2662696	2462877	26398372
2025	2619026	2630917	2618980	2628751	2628856	2627466	2627466	2630872	2630582	2545229	26188145

Nota. Esta tabla muestra el total datos para un dataframe completo.

Inicia la verificación del estado de los datos, previo a determinar el trabajo de preparación, para ello se hizo las siguientes actividades:

Inspección errores tipográficos, cómo símbolos o palabras erróneas, en esta actividad se realizó lo siguiente:

Reemplazar los símbolos '*****', que se encontraban en la variable DE_COM por 'TRACCIÓN ANIMAL'

Revisión de valores únicos en las columnas tipo texto: FOING, DE_COM, DE_TVH y ESTADO, para determinar si los valores estaban dentro de los rangos establecidos para cada variable categórica.

Revisión de valores únicos en las columnas tipo entero: VIGEN, ID_COM, ID_TVH, ID_EST y ORIGEN, para determinar si los valores estaban dentro de los rangos establecidos para cada variable numérica.

Revisión de valores únicos en las columnas: FE_REMATRI, FE_ING_TRA, para determinar si los valores estaban dentro de los rangos establecidos para cada variable de fechas.

Validación de fecha mínima y máxima para FE_REMATRI - FE_ING_TRA para identificar el rango lógico de las fechas.

Inspección de registros con fechas anómalas en FE_ING_TRA, FE_REMATRI para identificar fechas que no son lógicas.

Verificación del formato de los datos, a continuación, imagen de cómo se encontraban.

Figura 3

Validación Formato de Datos del Dataframe Final

Variable	Dtype
VIGEN	Float 64
FOING	Object
FE_ING_TRA	Float 64
FE_REMATRI	Float 64
ID_COM	Float 64
DE_COM	Object
ID_TVH	Object
DE_TVH	Object
ID_EST	Object
ESTADO	Object
ORIGEN	Float64

Nota. Tomado de Jupyter Notebook.

En la verificación de la cantidad de valores nulos o faltantes, se encontró lo siguiente.

Tabla 3

Reconocimiento Valores Nulos o Faltantes

VIGE	FOIN	FE_ING_T	FE_REMA	ID_COM	DE_CO	ID_TVH	DE_TVH	ID_EST	ESTADO	ORIGEN	TOTAL VIGE
2022	66511	1562933	2609411	29945	29946	32849	31856	28840	32750	314135	4739176
2023	26891	439767	2665664	2794	2795	14779	14108	2568	7183	165169	3341718
2024	31848	119384	2672906	7880	7882	16958	16290	7651	12306	212125	3105230

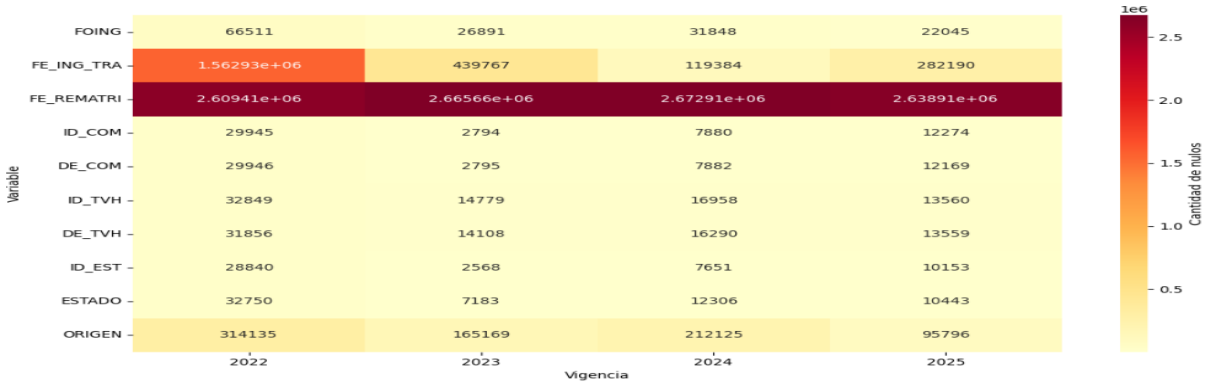
VIGE	FOIN	FE_ING_T	FE_REMA	ID_COM	DE_CO	ID_TVH	DE_TVH	ID_EST	ESTADO	ORIGEN	TOTAL VIGE
2025	22045	282190	2638909	12166	12169	13560	13559	10153	10443	95796	3111098
TOTAL	147295	2404274	10586.890	52785	52792	78146	75813	49212	62682	787.225	14297222

Nota. Tomado de Jupyter Notebook

El dataframe final en el que se consolidó todos los registros recopilados de la vigencia 2022 al 2025, presenta un total de 10.595.062 de registros y dado que se agrupa por vigencia, de este modo se determina que existen 10 variables, excluyendo la variable VIGEN que se utiliza para la agrupación, es decir, en el supuesto que todos los datos del dataframe estuvieran completos, debería existir 105.950.620 datos, sin embargo, se evidencia que hay 14.297.222 datos nulos o inválidos en todo el dataframe, esta cantidad de registros nulos representa un 13.49% aproximadamente del conjunto de datos. Para tener un detalle sobre la variable que más afectación presenta por falta de dato o datos inválidos se hizo el mapa de calor, en el que se presenta las variables vs la vigencia.

Figura 4

Variable vs Vigencia con Datos Nulos



Nota. Tomado de Jupyter Notebook

El gráfico de calor presenta la cantidad de valores nulos por variable y vigencia (2022 a 2025) en un dataset relacionado con la emisión de facturas de los impuestos de vehículos. A continuación, se identifica las oportunidades de mejora conforme a la información consolidada.

Variable FE_REMATRI: Es la variable con más datos nulos o inválidos, con un total de diez millones quinientos ochenta y seis mil ochocientos noventa (10.586.890), es decir, que del total de nulos, esta variable representa el 74%, sin embargo, no hay inconveniente latente con esta situación, dado que los datos que ahí reposan o deben reposar corresponde a la fecha de registro del vehículo en Bogotá, aquí se identifica la fecha de rematriculado o matriculado de un vehículo en la ciudad, es una variable importante pero la ausencia de los datos es imputable con la variable FE_ING_TR, dado que cuando un vehículo ha sido matriculado en Bogotá desde el inicio de su adquisición (nuevo) y no presenta cambio matrícula, la relación de la obligación tributaria se mantiene en Bogotá, para ello, con este campo se hará un proceso de imputación de datos a partir de los datos que se encuentra en el campo fecha de ingreso a tránsito.

Variable "ORIGEN": Es la segunda variable con mayor ausencia en la completitud de los datos, esta situación afecta innumerables análisis y razonamientos respeto al origen de la emisión para el pago del impuesto de vehículo, situación que permanece en el tiempo para las vigencias objeto de estudio, la importancia de esta información radica en que la variable identifica sí el contribuyente se le emitió factura o se le generó un recibo oficial de pago (ROP) que pudo darse a través de punto de atención o el contribuyente desde su oficina virtual (WEB) presentó una declaración. Indudablemente es una variable vital para analizar, dado que posibles resultados impactarían el rastreo sobre la emisión de la obligación tributaria para el pago del impuesto de vehículo. Esta variable no será tenida en cuenta en este trabajo y es eliminado del dataframe.

Variable “FOING”: Representa la forma de ingreso de los datos al sistema, esta variable tiene menos impacto que la variable anterior, ya que los datos son suministrados por la secretaria distrital de movilidad mediante comunicación entre los dos sistemas, en este sentido, los datos muestran la forma cómo llegan los datos a la entidad, por tanto, el análisis que se puede hacer estaría dirigido a evaluar la comunicación entre estos dos sistemas.

En conclusión, la calidad en cuanto a completitud en una variable es relevante para realizar análisis o predicciones con mayor precisión, la falta de datos y/o datos nulos afecta el estudio qué sobre la variable se pueda realizar, a causa de posibles sesgos en los resultados, de ahí la importancia de actualizar los datos para futuros análisis. Esta variable no será tenida en cuenta en este trabajo y es objeto de eliminación del dataframe.

Variables ID_COM, DE_COM, ID_TVH, DE_TVH, ID_EST, ESTADO: Son variables importantes para el presente trabajo, por tanto, se realizó todo el proceso de preparación de datos con estas variables para las imputaciones y limpieza de datos.

Preparación de los Datos

A partir de la identificación de las variables, registros y datos, se encontró las siguientes situaciones: datos con formatos erróneos, valores nulos o faltantes, errores tipográficos, variables con bajo impacto para el estudio, por tanto, se inició el proceso de preparación de los datos, en la que incluyó:

Ajuste de Formato de Datos

El dataframe final comprendía dentro de un mismo campo, diferentes tipos de datos, esto sucedió con las siguientes variables numéricas: FE_REMATRI, FE_ING_TRA, ID_COM, ID_TVH, ID_EST y ORIGEN.

Igualmente se realizó el ajuste de formato para las siguientes variables categóricas: FOING, DE_COM, DE_TVH y ESTADO.

A continuación, la imagen con la información de formato corregido.

Figura 5

Corrección Formato de Datos del Dataframe Final

Variable	Dtype
VIGEN	Int 64
FOING	String
FE_ING_TRA	Datetime 64[ns]
FE_REMATRI	Datetime 64[ns]
ID_COM	Int 64
DE_COM	String
ID_TVH	Int 64
DE_TVH	String
ID_EST	Int 64
ESTADO	String
ORIGEN	Int 64

Nota. Tomado de Jupyter Notebook

Imputación de Datos

Las imputaciones se realizaron con el fin de completar los datos vacíos con las variables de este dataframe, esto se hizo con las siguientes variables: de DE_COM desde ID_COM, de ID_COM desde DE_COM, de ID_COM y DE_COM desde ID_TVH, de DE_TVH desde

ID_TVH, de ID_TVH desde DE_TVH, de ESTADO desde ID_EST, de ID_EST desde ESTADO, de ID_EST Y ESTADO desde ORIGEN, de FE_REMATRI desde FE_ING_TRA

Figura 6

Resultado Imputaciones en DE_COM y ID_COM

Variable	Valore nulos
DE_COM	36.633
ID_COM	36.630

Nota. Tomado de Jupyter Notebook

En la tabla de Reconocimiento Valores Nulos o faltantes se presentó para DE_COM un total de 52.792 datos nulos y luego de las imputaciones se redujo en 36.633, en términos numéricos la cantidad bajo en 16.159 nulos para esta variable, resultado de las imputaciones utilizando las variables ID_COM que es el identificador de combustible y ID_TVH que es el identificador de tipo de vehículo, igual proceso se hizo con ID_COM que tenía 52.785 nulos y redujo a 36.630 nulos que corresponde a 16.155, ahora la diferencia entre los datos ID_COM y DE_COM se debe a que en ID_COM, se presentó datos que no pertenecían al rango de la variable, por tanto quedaron identificados para la eliminación de estos datos.

Figura 7

Resultado Imputaciones de Datos en DE_TVH y ID_TVH

Variable	Valore nulos
DE_TVH	75.813
ID_TVH	75.813

Nota. Tomado de Jupyter Notebook

En la *Tabla 3* de reconocimiento valores nulos o faltantes se identificó para ID_TVH que la cantidad de nulos ascendía a 78.146 y se redujo a 75.813 después de la imputación hecha desde DE_TVH, sin embargo, esta última variable no tuvo reducción dado que la única variable para imputar era ID_TVH, y variable contenía una mayor cantidad de nulos, por tal razón, la única variable en la que se redujo los nulo fue ID_TVH.

Figura 8

Resultado Imputaciones de Datos en ESTADO y ID_EST

Variable	Valore nulos
ESTADO	6.140
ID_EST	101

Nota. Tomado de Jupyter Notebook

La imputación se hizo entre las dos variables, la variable ID_EST tenía 49.212 datos nulos y después de las imputaciones desde ESTADO se redujo a 101 datos nulos, mientras que la variable ESTADO tenía 62.682 y solo se logró reducir a 6.140 datos nulos.

Figura 9

Resultado Imputaciones de Datos en FE_REMATRI y FE_ING_TR

Variable	Valore nulos
FE_REMATRI	2.402.275
FE_ING_TRA	2.404.274

Nota. Tomado de Jupyter Notebook

Para estas dos variables, se imputo los datos de FE_ING_TR hacia FE_REMATRI y viceversa, aunque la variable FE_REMATRI presentaba el mayor número de nulos con 10.586.890, mientras que FE_ING_TR tenía 2.404.274, al ejecutar la imputación, las dos variables, quedaron con cantidad de nulos cercanos, esto es 2.402.275 para FE_REMATRI que es la variable relevante para este proyecto.

Finalizado el proceso de imputaciones con el fin de consolidar el mayor número de datos en el dataframe, se continúa con el consolidado de datos, para determinar el estado del dataframe frente a los datos nulos y seguir con las depuraciones de datos y variables, para consolidar el dataframe final.

Tabla 4

Resultado de Valores Nulos Después de las Imputaciones

VIGE	FOIN	FE_ING_T	FE_REMAT	ID_CO	DE_CO	ID_TV	DE_TV	ID_EST	ESTAD	ORIGE	TOTA
2022	66511	1562933	1561290	24080	24080	31856	31856	54	1392	314135	3618187
2023	26891	439767	439563	1901	1901	14108	14108	13	734	165169	1104155
2024	31848	119384	119338	4969	4970	16290	16290	14	3955	212125	529183
2025	22045	282190	282084	5680	5682	13559	13559	20	59	95796	720674
TOTAL	147295	2404274	2402275	36630	36633	75813	75813	101	6140	787225	5972199

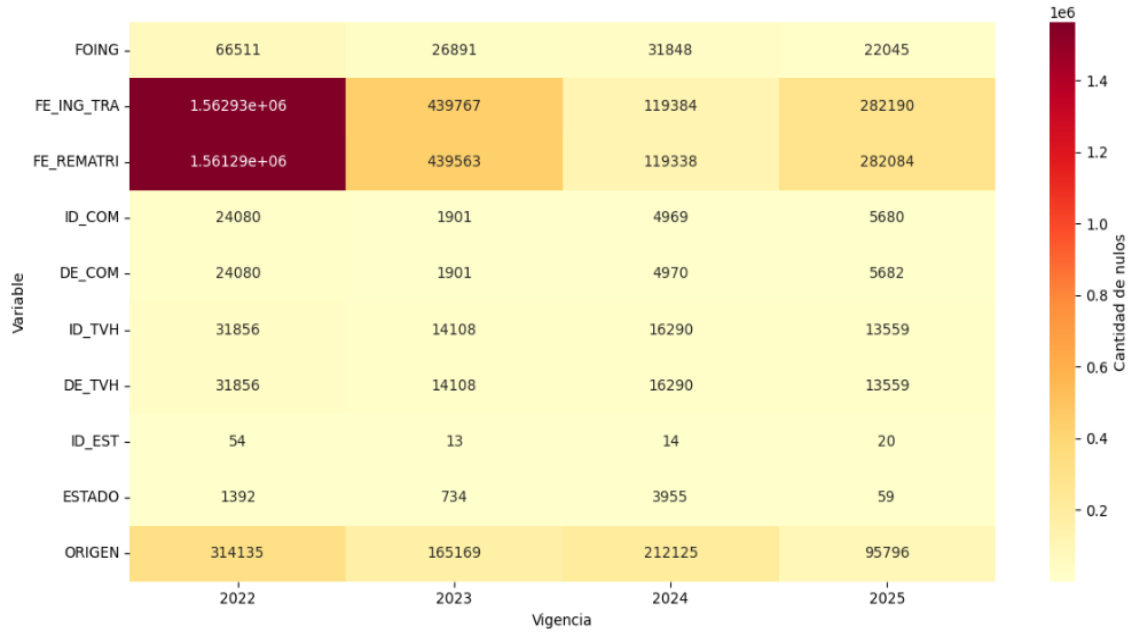
Nota. Tomado de Jupyter Notebook.

Teniendo en cuenta el tratamiento de datos a partir de las imputaciones con las variables del dataframe final, se logra avanzar hacia la completitud de los datos, especialmente para las variables DE_COM (descripción de la clase de combustible), la variable ESTADO y finalmente la variable FE_REMATRI , requerida para el modelo conforme a la secuencia de tiempo. Para tener un detalle sobre las variables que siguen afectadas por falta de dato o datos inválidos, se

hizo el mapa de calor, en el que se presenta las diferentes variables vs la vigencia después de las imputaciones realizadas.

Figura 10

Variables vs Vigencia con Datos Nulos Después de Imputaciones



Nota. Tomado de Jupyter Notebook.

Ya teniendo toda la información, se procede a realizar el consolidado comparativo conforme a los resultados de valores nulos antes y después de las imputaciones, así:

Tabla 5

Comparativo Registros Nulos Antes y Después de las Imputaciones

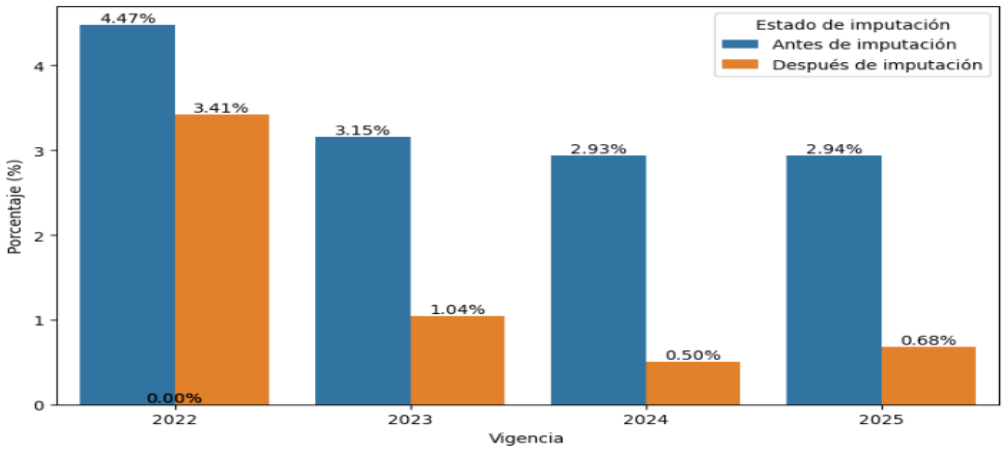
VIGEN	Datos nulos antes de imputaciones	Datos nulos después de imputaciones
2022	4739176	3618187
2023	3341718	1104155
2024	3105230	529183

VIGEN	Datos nulos antes de imputaciones	Datos nulos después de imputaciones
2025	3111098	720674
TOTAL	14297222	5972199

Nota. Para una mejor comprensión de los datos se realiza la siguiente gráfica.

Figura 11

Consolidado de Nulos por Vigencia (Antes y Después de las Imputaciones)



Nota. Tomado de Jupyter Notebook.

El gráfico de la *Figura* **11**, muestra que se logró reducir la cantidad de datos nulos o inválidos mediante imputaciones. Inicialmente, se tenían 14.297.222 datos nulos, y tras las imputaciones, esta cifra se redujo a 5.972.199, pasando de un 13.49% a un 5.63% del total de datos del dataframe final, lo que representa una mejora del 7.86%. Además, se identificó que la vigencia 2022 fue la más incompleta en datos, posiblemente debido a que en 2021 se estaba implementando el sistema BOGDATA, el ERP de SAP, y se presume que los datos provenientes del sistema anterior, SITII, presentaron dificultades técnicas. Aunque las vigencias posteriores

muestran una ligera mejora en la completitud de los datos, la persistencia de datos incompletos o de baja calidad en la vigencia 2025 revela una oportunidad de mejora.

Eliminación de Variables

Teniendo en cuenta que el objetivo de este proyecto es la emisión de facturas, se procedió a la eliminación de variables del dataframe irrelevantes para el análisis estadístico y modelo predictivo. Por tanto, se eliminaron las siguientes:

FE_ING_TRA: Esta variable no hace parte del objeto de estudio del proyecto para el análisis y más aún cuando se realizó imputación con la variable FE_REMATRI para determinar el comportamiento de vehículos matriculados o rematriculados en Bogotá, de tal modo que esta variable es eliminada del dataframe final.

FOING: Tal como se indicó, la variable representa la forma de ingreso de los datos al sistema, y dado que se realiza análisis del comportamiento del parque automotor y sus características para predecir la tendencia en la emisión de facturas con base en estos datos, esta variable tiene una significancia nula.

ORIGEN: Es una variable muy importante y pudo contribuir en el análisis de los datos históricos, sin embargo, el alto porcentaje en la completitud sugiere evitar su aplicación, pero en el informe de sugerencias se tendrá en cuenta para que la entidad actualice este dato que es importante para análisis y predicciones frente a esta variable.

Limpieza de Datos

A partir del análisis de los datos, se identificó que existían valores fuera del rango establecido para las variables. La eliminación de estos valores se realizó con los siguientes resultados:

ID_EST: Se eliminaron 6.140 datos fuera del rango de ID_EST, estos valores estaban como: 0, 2 y los vacíos.

ID_COM: Se eliminaron 35.675 datos que contenían valores que no pertenecían al rango de ID_EST, estos valores estaban como: 11, 155 y los vacíos

FE_REMATRI: Se eliminaron 2.366.359 datos nulos más 8 datos con fecha anómala, para un total de 2.366.367.

ID_TVH: Se eliminaron 36.585registros con datos vacíos

El total de datos eliminados fueron: **2.444.767,** este valor frente al total de registros iniciales que correspondió inicialmente a **105.950.620**, representa tan solo un: **2.30%,** de ese modo, el total de datos para realizar el proceso de análisis de datos históricos y posteriormente el modelo redes neuronales recurrentes es: **8.150.295**

Figura 12

Total Nulos en el Dataframe Final

Variable	Valores nulos
ID_EST	0
ESTADO	0
ID_COM	0
DE_COM	0
ID_TVH	0
DE_TVH	0
FE_REMATRI	0

Nota. Tomado de Jupyter Notebook

Consolidación del Dataframe Final

El dataframe inicial estaba conformado por 11variables y 10.595.062 registros, con estos datos se hizo un proceso de comprensión y preparación de datos, y dada su finalización, el dataframe quedó así:

Cantidad de Registros por Vigencia

El consolidado del dataframe final queda en total con **8.150.295** registros y por vigencia queda así:

Tabla 6

Consolidado de Registros por Vigencia

VIGENCIA	REGISTROS
2022	1.043.379
2023	2.216.278
2024	2.538.785
2025	2.351.853
Total	8.150.295

Nota. Tomado de Jupyter Notebook.

Total de Variables

De las variables que componían el conjunto de datos, tras finalizar el proceso de preparación, el dataframe quedó con las siguientes variables relevantes para el análisis estadístico y el modelo predictivo:

Variables categóricas: DE_COM, DE_TVH y ESTADO

Variables numéricas: VIGEN, ID_COM, ID_TVH y ID_EST

Variables fecha: FE_REMATRI

Cantidad de Datos Finales

El dataframe final queda sin registros nulos.

Tabla 7

Consolidado de Datos Nulos

VIGENCIA	FE_REMATRI	ID_COM	DE_COM	ID_TVH	DE_TVH	ID_EST	ESTADO
2022	0	0	0	0	0	0	0
2023	0	0	0	0	0	0	0
2024	0	0	0	0	0	0	0
2025	0	0	0	0	0	0	0

Nota. Tomado de Jupyter Notebook.

Análisis Estadístico Datos Históricos Períodos 2022-2025

En cumplimiento del segundo objetivo de este proyecto, se procede al análisis de los datos históricos relacionados con la emisión de facturas en el período 2022–2025. Una vez finalizada la preparación de los datos y con el consolidado correspondiente, se aplican técnicas de análisis estadístico descriptivo mediante tablas cruzadas. Para ello, se consideran las variables relevantes asociadas a las características de los vehículos que conforman el parque automotor matriculado en Bogotá, con el propósito de identificar posibles patrones y/o tendencias.

Frecuencias Relativas Cruzadas con la Variable Combustible

Clase de Combustible y Estado

Con el dataset procesado, se analizaron las variables:

Estado: Indica si la matrícula del vehículo está activa, inhabilitada, cancelada o trasladada.

DE_COM: Variable que identifica el tipo de combustible utilizado por los vehículos registrados en Bogotá (por ejemplo, gasolina, diésel, eléctrico).

A partir del cruce de las variables “clase de combustible” y “estado del vehículo”, se obtuvieron los siguientes resultados:

La clase de combustible predominante es la gasolina en los vehículos con matrícula activa, representando el 98.3% del conjunto de datos, este porcentaje es alto comparándolos con los porcentajes de otros tipos de combustibles, esto se traduce en la emisión de aproximadamente 7.474.696 facturas, lo que indica que el mayor recaudo por concepto del impuesto proviene de vehículos que utilizan gasolina.

En relación con otras alternativas de combustible presentes en el conjunto de datos, tales como: Diesel- ACPM, Gas-Gasolina, Gaso Elec. y Eléctrico, se identificó que la frecuencia absoluta acumulada para los combustibles mencionados es de 536.256 vehículos.

Este primer análisis evidencia una diferencia de 7.067.707 vehículos activos que utilizan

gasolina, siendo superior en comparación con el total de vehículos activos que emplean alguno de los combustibles mencionados. Este dato revela que la emisión de facturas está estrechamente relacionada con los vehículos que funcionan con gasolina.

Un dato importante que se debe considerar respecto a la variable “estados”, se relaciona con los vehículos de combustibles diesel-ACPM, dado que este tipo de vehículo genera el mayor porcentaje de cancelaciones y traslados, esta situación le representa a Bogotá 22.011 vehículos menos a facturar y por ende menos recaudo. Esta misma situación en menor medida, se presenta con los vehículos de gas – gasolina, que entre cancelaciones y traslados representan 4.222 vehículos menos a facturar. Esta situación.

Debe tenerse presente que este análisis es inicial, por tanto, se continúa el estudio con las tablas cruzadas para combinar la clase de combustible y nuevas variables, de modo que se tenga un panorama explícito sobre los patrones de comportamiento asociados con el uso del combustible.

Tabla 8

Acumulado Relativo Entre Estado y Clase Combustible Vigencia 2022-2025

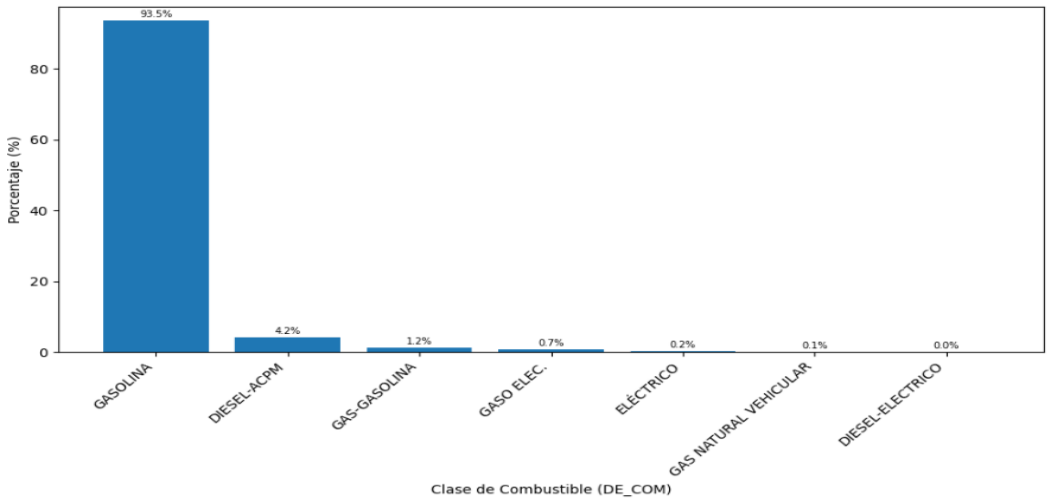
CLASE DE COMBUSTIBLE	ACTIVO	CANCELACIÓ	INHABILITAD	TRASLADAD	TOTAL
GASOLINA	98,3	0,9	0	0,8	7603963
DIESEL-ACPM	93,9	4	0	2,1	360841
GAS-GASOLINA	95,9	3,2	0	0,9	102977
GASO ELEC.	98,6	1,2	0	0,2	55966
ELÉCTRICO	99,6	0,3	0	0,1	16472
GAS NATURAL VEHICULAR	99,5	0,3	0	0,1	6960
DIESEL-ELECTRICO	99,9	0,1	0	0	2294
BIODISEL	99,8	0	0	0,2	559
TRACCIÓN ANIMAL	99,2	0,8	0	0	240
HIDROGENO	100	0	0	0	13

CLASE DE COMBUSTIBLE	ACTIVO	CANCELACIÓ	INHABILITAD	TRASLADAD	TOTAL
ETANOL	87,5	12,5	0	0	8
GLP	100	0	0	0	2

Para coadyuvar con el análisis se genera el top 7 de los combustibles con mayor presencia en el parque automotor de la ciudad. La siguiente imagen refuerza la idea que la emisión de facturas está ligada a los vehículos con combustible a gasolina, su representación asciende al 93.5% del total, en segundo lugar, con una participación muy distante, se encuentran los vehículos a diesel-ACPM con un 4.2%, el tercer puesto corresponde para los vehículos gas-gasolina con 1.2% de representación, los siguientes puestos están por debajo del 1% mostrando que son poco o nada representativos para el parque automotor de la ciudad.

Figura 13

Top 7 Tipos Combustible vs Vehículos Activos a 2025



Nota. Tomado de Jupyter Notebook.

Clase de Combustible y Año de Matrícula

En este apartado, se analiza la relación entre el tipo de combustible utilizado por los vehículos y su año de matrícula, con el objetivo de identificar tendencias en el uso de gasolina frente a otros combustibles (como diésel o eléctrico) a lo largo del tiempo. Para ello, se cruzan las variables "clase de combustible" y "año de matrícula.

El año de matrícula corresponde a la fecha de registro del vehículo en Bogotá, extraída de la variable FE_REMATRI. Los datos se agruparon en intervalos de 15 años para facilitar su análisis y estandarización. A continuación, se presentan los resultados

Gasolina: Los vehículos que usan gasolina han mantenido un predominio significativo a lo largo del tiempo, aunque con ligeros descensos. En el intervalo 2011-2025, representaron el 92.4% de las matrículas, lo que configuró una reducción del 1.6% respecto al intervalo anterior.

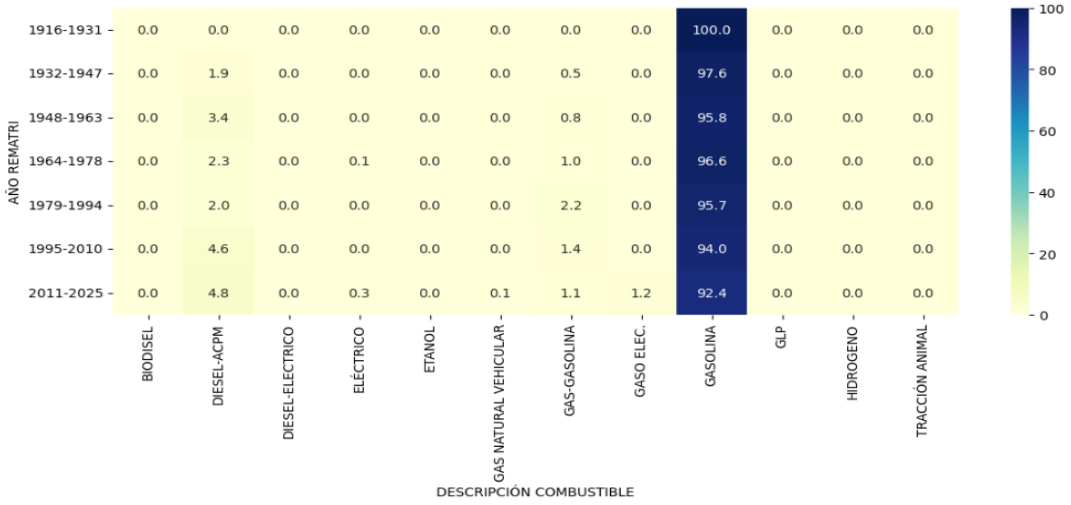
Diésel – ACPM: Este combustible que ocupa el segundo lugar con mayor cantidad de vehículo activos en Bogotá, en el intervalo 2011-2025, se registraron 222,817 vehículos, equivalentes al 4.8% del total, marcando su mayor crecimiento en último período.

Gas-gasolina: Caso contrario ocurrió con el tipo de combustible gas- gasolina que pese a estar ocupando el tercer lugar con mayor cantidad de vehículos activos en la ciudad, tuvo en el período de los últimos 15 años (2011-2025) un descenso de 0.3% y con un crecimiento acumulado de 1.1% que representa 51.062 vehículos.

Híbrido (gasolina-eléctrico): El año 2018 es el inicio de las primeras 284 matrículas en la ciudad, para el período de los últimos 15 años (2011-2025) la participación de este tipo de vehículo creció a 1.2% representando para la ciudad 55.704 vehículos.

Figura 14

Acumulado Relativo Clase de Combustible y Año de Matrícula



Nota. Tomado de Jupyter Notebook.

Top 3 de matrículas por tipo de combustible (2011-2025): Este análisis refuerza el predominio de los vehículos de gasolina en el parque automotor de Bogotá durante el período 2011-2025. Cada matrícula activa genera una factura emitida por la entidad, lo que hace que el comportamiento de las matrículas sea un indicador clave para la gestión administrativa y fiscal de la entidad. A continuación, se presenta el ranking de los combustibles con más matrículas activas:

Gasolina: Representa el 92.4% del total de matrículas, equivalente a 4,289,222 vehículos registrados.

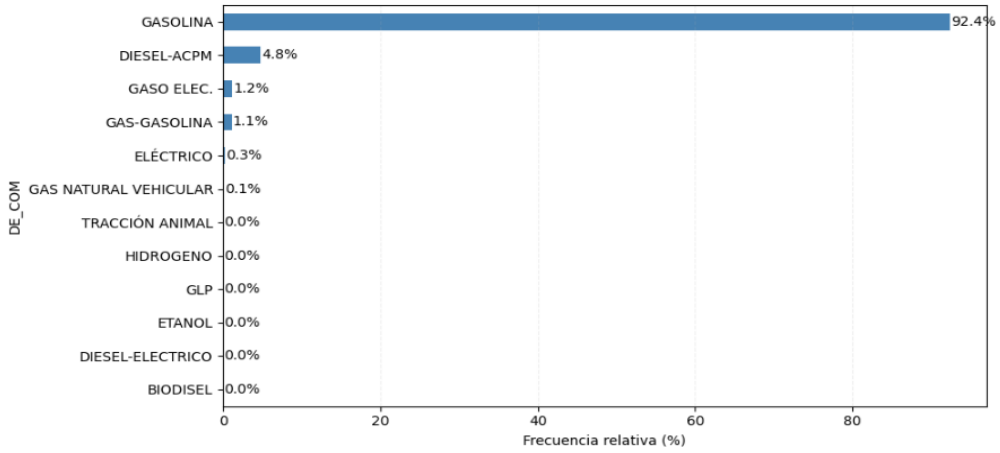
Diésel – ACPM: Ocupa el segundo lugar con el 4.8%, correspondiente a 222,817 matrículas.

Híbrido (gasolina-eléctrico): Se posiciona en tercer lugar con el 1.2%, equivalente a 55,704 matrículas.

La siguiente gráfica ilustra la distribución de las matrículas por tipo de combustible en el período 2011-2025, destacando el predominio de la gasolina frente a otros combustibles.

Figura 15

Comportamiento de las matrículas de vehículos en Bogotá, período 2011-2025



Nota. Tomado de Jupyter Notebook.

Clase de Combustible y Tipo de Vehículo (2011-2025)

Para complementar el análisis de las clases de combustible, se cruzó la variable "clase de combustible" con "tipo de vehículo matriculado en Bogotá" durante el período 2011-2025. En los análisis previos se ha destacado el predominio de la gasolina en las matrículas activas, una tendencia que se mantiene constante. En este análisis se identifica los tipos de vehículos donde la gasolina es más común, sin embargo, se examina el uso del diésel – ACPM. Los resultados se presentan a continuación:

Automóviles a gasolina: Del total de vehículos, le corresponde el 99,2% (3.997.451).

Motos y motocarros a gasolina: Con el 99.9% del total de este tipo de vehículos, lo que corresponde a 1,603,347, este dato refleja un dominio casi total para las motos y motocarros.

Camperos y camionetas a gasolina: Del total de vehículos matriculados en Bogotá, los camperos y camionetas ocupan el segundo lugar, y del total de camperos, el 91.6% son a gasolina (1,527,320)

El combustible diésel – ACPM tiene una presencia moderada, que se ve reflejada así:

Camionetas doble cabina: Son el tipo de vehículo con mayor uso de diésel – ACPM, representando el 44.5% de las matrículas de esta categoría (117,243 vehículos).

Vehículos de carga: El 26.5% de estos vehículos utilizan diésel – ACPM, equivalente a 66,522 matrículas.

Vehículos de pasajeros: Ocupan el tercer lugar, con un 19% de las matrículas (48,193 vehículos). De otro lado, se resalta la superioridad del combustible tipo gasolina y el tipo de vehículo tipo automóvil, esta combinación de tipo de vehículo y tipo de combustible tiene predominio en el parque automotor de Bogotá. A continuación, se presenta la tabla con la distribución de las matrículas según el tipo de vehículo y la clase de combustible, expresada en términos de frecuencia relativa (porcentaje respecto al total de cada tipo de vehículo).

Tabla 9

Frecuencia Relativa Entre Clase Combustible y Tipo Vehículo

DE_TVH	BIODI	DIES ACPM	DIESE ELEC	ELÉCT	ETANO	GAS NAT VEH	GAS GASOLI	GASO ELE	GASOLI	GLP	HIDROG	TRACCI ANIM	TOTA
AUTOMÓVILE	0	0,2	0	0	0	0	0,6	0	99,2	0	0	0	4029689
CAMPEROS Y CAMIONETAS	0	7,1	0	0	0	0	1,2	0,1	91,6	0	0	0	1667380
MOTOS Y MOTOCARROS	0	0	0	0,1	0	0	0	0	99,9	0	0	0	1604952
CAMIONETAS DOBLE CABINA	0	44,5	0,1	0,2	0	0	2,9	0,5	51,8	0	0	0	263468
PASAJEROS	0	19	0,4	0,1	0	1,7	13,1	0	65,8	0	0	0	253646
CARGA	0	26,5	0	0	0	0,1	6,5	0	66,8	0	0	0,1	251027
HIBRIDOS	1	0,1	2	0	0	0	0,1	90,9	6	0	0	0	57071
ELÉCTRICOS	0	0	0	99,8	0	0	0	0	0,1	0	0	0	14747

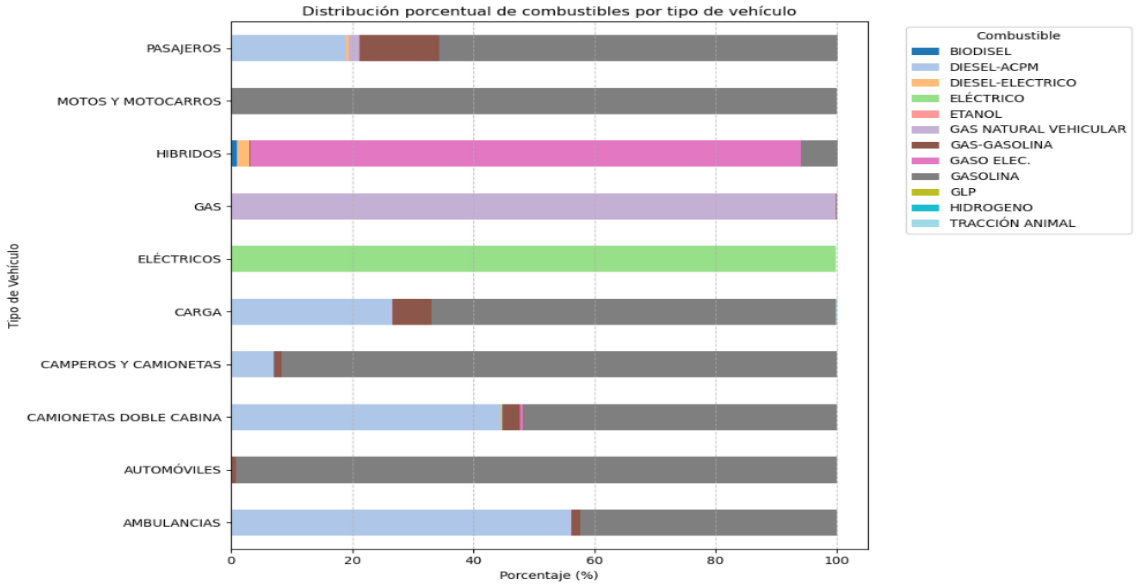
DE_TVH	BIODI	DIES ACPM	DIESE ELEC	ELÉCT	ETANO	GAS NAT VEH	GAS GASOLI	GASO ELE	GASOLI	GLP	HIDROG	TRACCI ANIM	TOTA
AMBULANCIA	0	56,1	0	0	0	0,1	1,5	0	42,3	0	0	0	6415
GAS	0	0	0	0	0	99,8	0,2	0	0	0	0	0	1900
Total												**8150295**	

Nota. Tomado de Jupyter Notebook.

Además, se presenta el gráfico que complementa el análisis y la *Tabla 9*

Figura 16

Distribución Porcentual de Combustible por Tipo de Vehículo



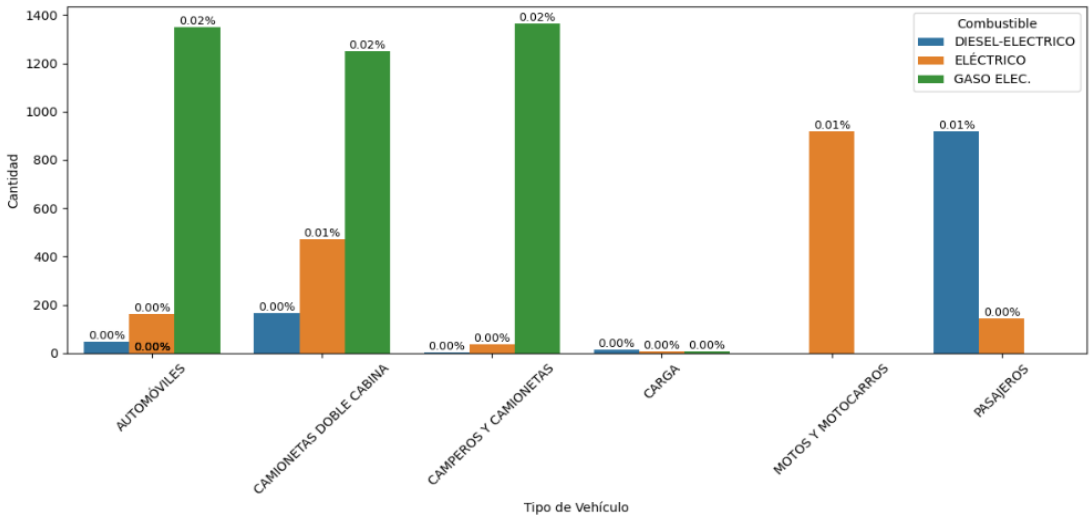
Nota. Tomado de Jupyter Notebook.

En los anteriores análisis se ha confirmó la presencia dominante de los vehículos a gasolina, en especial los automóviles que funcionan con este tipo de combustible. Sin embargo, es importante examinar el comportamiento de los combustibles alternativos, como los eléctricos, gasoeléctricos y diésel-eléctricos. A partir de la Tabla 8, se observa que el total de vehículos

activos matriculados en Bogotá asciende a 7.993.904 vehículos (corte 2025). Dentro de este conjunto, se calcula la participación porcentual de los vehículos que utilizan combustibles alternativos. La siguiente gráfica de barras presenta el porcentaje de participación correspondiente a los cinco tipos de vehículos activos que operan con estos combustibles.

Figura 17

Distribución Tipo de Vehículos Activos con Combustible Alternativos



Frecuencias Relativas Cruzadas con la Variable Estado

Estado y Año de Matrícula (2011-2025)

Este análisis examina el comportamiento del parque automotor de Bogotá en función del estado de las matrículas (activas, inhabilitadas, canceladas o trasladadas) y su relación con el año de matrícula durante el período 2011-2025. El objetivo es determinar si existe variabilidad significativa en el estado de las matrículas o si hay tendencia hacia la estabilidad. Para ello, se cruzaron las variables "estado de la matrícula" y "año de matrícula", obteniendo los siguientes resultados.

Tabla 10

Frecuencia Absoluta Cruzada Entre Estado y Año de Matrícula

ESTAD	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	TOTA
ACTIV	681639	604459	511790	620852	499768	367556	273900	279006	258438	179179	147075	78454	42681	24265	39	4569101
CANCELAD	5345	4472	3607	4120	3325	2510	1794	1771	1322	759	5239	119	19	9	0	34411
INHABILITA	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
TRASLADA	5457	5399	4321	4670	3721	3280	2870	2850	2574	1488	1479	268	87	38	0	38502
Total	692441	614330	519718	629642	506814	373346	278564	283627	262334	181426	153793	78841	42787	2313	39	4642015

Nota. Tomado de Jupyter Notebook.

Tendencia general: Tendencia de matrículas nuevas: Entre 2011 y 2025, se registraron 4,642,015 nuevas matrículas, de las cuales 4,569,101 (98.4%) son activas, de acuerdo con la Tabla 10. Se muestra una disminución sostenida en las matrículas nuevas activas, especialmente en los últimos años, este proceso va en descenso sin punto de pare en el 2024, el 2025 aún no se tiene un consolidado porque está vigente. Por tanto, se procede con el comparativo con el año inmediatamente anterior a partir del 2020, para identificar las tendencias de acuerdo con los diferentes estados.

Matriculas activas: Los porcentajes de tendencia decreciente son los siguientes:

2020: 30,67%

2021: 17,92%

2022: 46,66%

2023: 45,60%

2024: 43,15%

La caída de 2020 se vio influenciada por la emergencia sanitaria global por COVID-19, sin embargo, la tendencia de decrecimiento ya era evidente desde 2015, situación que impacta directamente los ingresos tributarios de la ciudad, configurándose en un aspecto muy importante para la administración y para el recaudo de impuesto vehicular de la Bogotá.

Matrículas inhabilitadas: Este estado es poco frecuente, con solo 1 matrícula inhabilitada registrada en 2024.

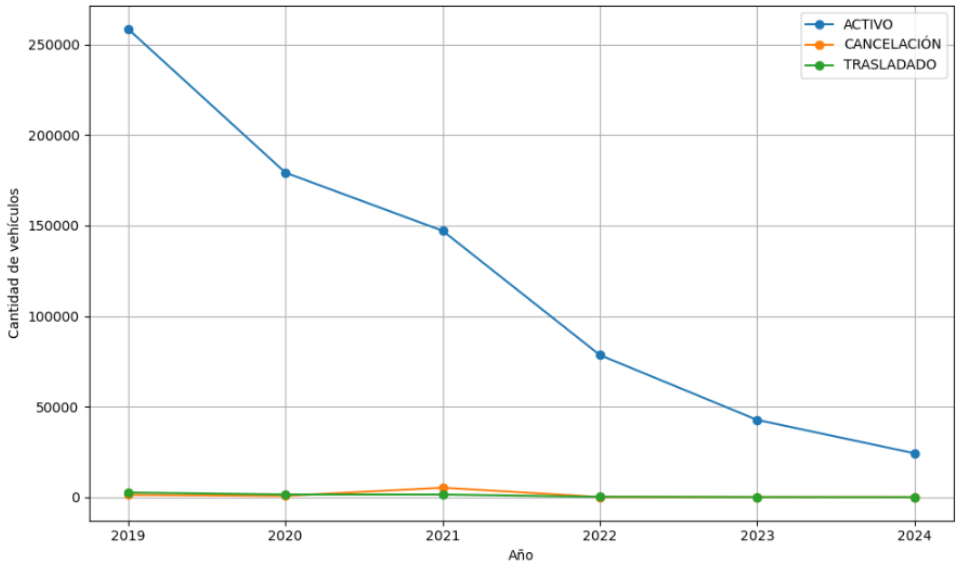
Matrículas canceladas: Totalizan 34,410 entre 2011 y 2025. Desde 2020, las cancelaciones muestran variabilidad: aumentaron un 590.25% en 2021 (de 759 en 2020 a 5,239), pero luego disminuyeron un 97.73% en 2022 (a 119), un 84.03% en 2023 (a 19), y un 52.63% en 2024 (a 9). Esta reducción reciente es favorable para mantener más matrículas activas.

Matrículas trasladadas: Suman 38,502 en el período. En 2020, disminuyeron un 42.54% respecto a 2019 (de 2,574 a 1,479). En 2021, se mantuvieron estables (1,479), pero luego cayeron un 81.88% en 2022 (a 268), un 67.54% en 2023 (a 87), y un 56.32% en 2024 (a 38).

Teniendo este contexto, se evidencia una variabilidad significativa en las matrículas nuevas, con una tendencia decreciente desde 2015, exacerbada en 2020 por la pandemia, pero persistente hasta 2024. Esta disminución impacta los ingresos tributarios, afectando las finanzas de la ciudad. La reducción en cancelaciones y traslados en años recientes es positiva, pero no compensa la caída general de nuevas matrículas. La siguiente gráfica muestra la evolución de las matrículas nuevas por estado, sin embargo, se redujo el periodo desde 2019 al 2024.

Figura 18

Estado de las Matrículas de Vehículos vs Año de Matrículas 2019-2024 Bogotá



Nota. Tomado de Jupyter Notebook.

Estado y Tipo de Vehículo

Continuando el análisis referente al estado de las matrículas de vehículos en la ciudad, para validar la tendencia de decrecimiento observada en los resultados de la variable fecha de matrícula, ahora se combina con los tipos de vehículos, para conocer en qué tipo de vehículos se evidencia el crecimiento o decrecimiento, en este sentido, se encontró lo siguientes resultados:

En relación con los automóviles, ya se tenía identificada la tendencia mayoritaria en la cantidad de las matrículas, adicionalmente, se reafirma que el 98.2% están activas, correspondiendo a un total de 3,957,154, y que del 99.2% corresponde al combustible tipo gasolina, todo este conjunto de datos se consolida como información relevante para la toma de decisiones en el proceso de emisión de facturas.

Los camperos y camionetas en el top 2 de mayor cantidad de matrículas y que de las mismas, el 98.1% están activas, significa esto que hay 1,635,699.

Finalmente, las motos y motocarros completan el top 3 de los vehículos con más número de matrículas, y de dicho total el 98.9% están activas, esto representa 1,587,297 de motos y motocarros con matrícula activa para Bogotá.

De otro lado, se hace énfasis en los traslados que afectan directamente los ingresos de la ciudad, aunado a ello, se identificó que las matrículas vienen en descenso, motivo por el cual se resalta el hecho que las camionetas doble cabina son los vehículos que más traslados presentan con un 2.4%, esto es 6,323, le siguen los vehículos de carga con un 1.5 % que representan 3,765 y el top 3 con mayor los traslados corresponde a los camperos y camionetas con un 1.1% que 18,341, sumados este top 3 de traslados, observamos que se asciende a 28,429 traslados, aunque se debe hacer un estudio frente al valor monetario que estos traslados afectan y cuanto representa el aporte en impuesto por tipo de vehículo, es valioso hacer seguimiento a este tipo de novedades que al final afectan en cierta medida las finanzas de la ciudad.

Y con una representación pequeña pero valiosa, se encuentran las cancelaciones de matrículas, para ello se observa en la tabla que los vehículos de pasajeros son los que sobresalen con las cancelaciones de matrícula en Bogotá, con un 3.8% que representa 9,639 cancelaciones, nuevamente los vehículos de carga hacen presencia en el top, con seguido de los vehículos de carga con un 3.4% cancelaciones que significa 8,535 cancelaciones, se finaliza el top 3 con las ambulancias con un 1.9% que representa 122 cancelaciones, sumados este top de cancelaciones, se observa que asciende a 18,296 cancelaciones.

Tabla 11

Relación Entre Estado y Tipo de Vehículo

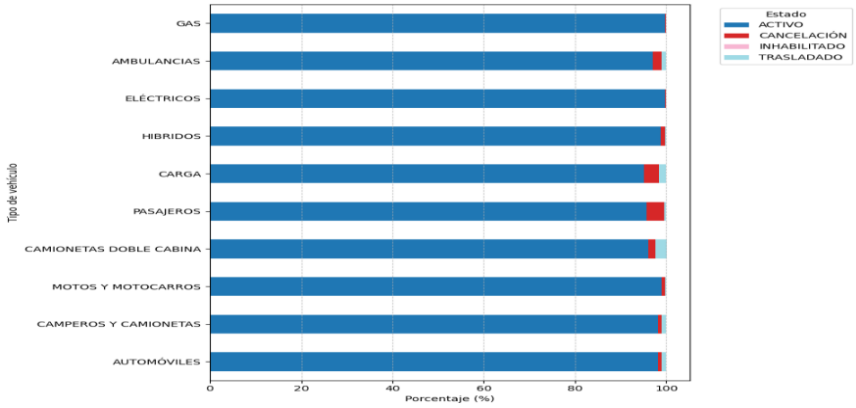
TIPO DE VEHÍCULO	ACTIVO	CANCELACIÓN	INHABILITADO	TRASLADADO	TOTAL
AUTOMÓVILES	98,2	0,8	0	0,9	4029689
CAMPEROS Y CAMIONETAS	98,1	0,8	0	1,1	1667380
MOTOS Y MOTOCARROS	98,9	0,8	0	0,3	1604952
CAMIONETAS DOBLE CABINA	96,1	1,6	0	2,4	263468
PASAJEROS	95,7	3,8	0	0,5	253646
CARGA	95	3,4	0	1,5	251027
HIBRIDOS	98,8	1	0	0,2	57071
ELÉCTRICOS	99,7	0,2	0	0,1	14747
AMBULANCIAS	97	1,9	0	1	6415
GAS	99,7	0,3	0	0	1900
Total			8150295		

Nota. Tomado de Jupyter Notebook.

Para complementar visualmente la información, se presenta el gráfico con el comportamiento de los estados frente a los tipos de vehículos.

Figura 19

Distribución Porcentual del Estado por Tipo de Vehículo



Nota. Tomado de Jupyter Notebook

Estado y Vigencia

Para completar el análisis relacionado con el estado de los vehículos frente a la matrícula, se pretende vislumbrar el comportamiento de los estados de matrículas a partir de las vigencias del 2022 al 2025, a continuación, se relaciona la información relevante:

Los vehículos activos presentan una estabilidad sin sobresaltos en las cantidades, como punto de partida del presente análisis, se observa que la vigencia 2022 tiene 1,027,217 vehículos activos, seguido le devino un crecimiento considerable en el 2023 con 1,128,330, se debe aclarar que dicho crecimiento pudo deberse a posibles traslados, dado que en el análisis de la Tabla 10, evidenció que las matrículas van en descenso desde el 2015, en relación a la vigencia 2024 se continua levemente en ascenso, esto es creció en 311,315 vehículos activos, finalmente para la vigencia 2025 hay un descenso leve, representado en 121,449 vehículos activos, como se observó el estado activo en los últimos 4 años muestran estabilización, solo se presenta un pico alto favorable entre el 2022 y 2023, pero seguido se ve fluctuaciones moderadas.

Situación diferente se presenta con el comportamiento de las cancelaciones, en contraposición con el leve crecimiento de vehículos activos para la vigencia 2024, las cancelaciones tuvo un crecimiento importante, esto es para el 2024 ascendió a 37,599 que porcentualmente es 37,6%, si la comparamos con el crecimiento que tuvo los activos al 2024 y que fue de 30.8%, significa claramente que la frecuencia de las cancelaciones superó las activaciones para esa vigencia, sin embargo, las novedades no se pueden controlar pero si sería importante hacer un debido seguimiento.

De otro lado, los traslados son igualmente importantes y deben ser objeto de seguimiento por parte de la entidad, nuevamente la vigencia del 2024 resalta porque se generaron el mayor número de traslados, ascendiendo a 34,316

Cabe resaltar que la vigencia 2024 fue un año especial para cancelar y trasladar vehículos, significando pérdidas importantes para el recaudo de la ciudad, ya que en esta vigencia entre cancelaciones, traslados e inhabilitaciones ascendió a 71,923, los traslados de la vigencia 2023 también fuero representativas en 29,181 traslados de vehículos.

Tabla 12

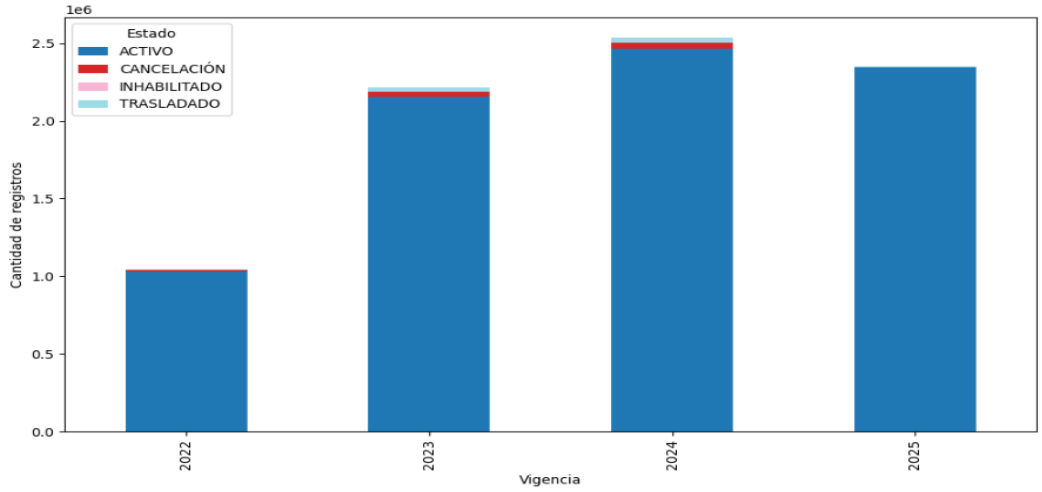
Acumulado Estado de Matrículas Entre las Vigencias 2022-2025

VIGEN	ACTIVO	CANCELACIÓN	INHABILITADO	TRASLADADO	TOTAL
2022	1027217	12312	0	3850	1043379
2023	2155547	31547	3	29181	2216278
2024	2466862	37599	8	34316	2538785
2025	2345413	1687	25	4728	2351853
		Total		8150295	

Nuevamente se acompaña el análisis de la presentación visual a través de la siguiente gráfica.

Figura 20

Comportamiento del Estado de las Matrículas Entre las Vigencias 2022 al 2025



Frecuencias Relativas Cruzadas con la Variable Tipo de Vehículo

Tipo de Vehículo y Año de Matrícula

Finalizando todo el estudio analítico descriptivo, mediante las tablas cruzadas se pretende explorar el comportamiento de las matrículas frente a los tipos de vehículos en el lapso de los últimos seis años, a continuación, se detalla la siguiente información:

Respeto del comportamiento de los automóviles, se había identificado previamente que este tipo de vehículo domina en cantidad el registro de matrículas del parque automotor en estado activo, aunado a esto, se identificó que el pico alto de matrícula se produjo en la vigencia 2019 con 123.086 matrículas.

Es de recordar que en el análisis de la *Tabla 10* se evidenció que la última vigencia con registros altos de matrícula se produjo en la vigencia 2019, desde ahí, hay tendencia del decrecimiento de matrículas incluida la vigencia en curso.

Los camperos y camionetas son los que ocupan el segundo lugar con alta cantidad de vehículos matriculados, con pico de crecimiento en el 2019 igual que los automóviles, motos y motocarros, camioneta doble cabina, pasajeros y carga.

En relación con los tipos de vehículos híbridos, el crecimiento de matrículas se dio en los años 2021al 2023, sumando 40.769 que representa un 72.5%.

Los vehículos eléctricos han presentado un comportamiento estable, sin altos ni bajos picos, el máximo número de matrícula se dio en el 2022 con 3.332 matrículas, en promedio se infiere que llega a 1.500 matriculadas contabilizadas desde el 2019.

Finalmente, los vehículos de gas solo han presentado en el 2020 un comportamiento significativo en matrículas ha tenido un solo año significativo de matrículas con 1.569.

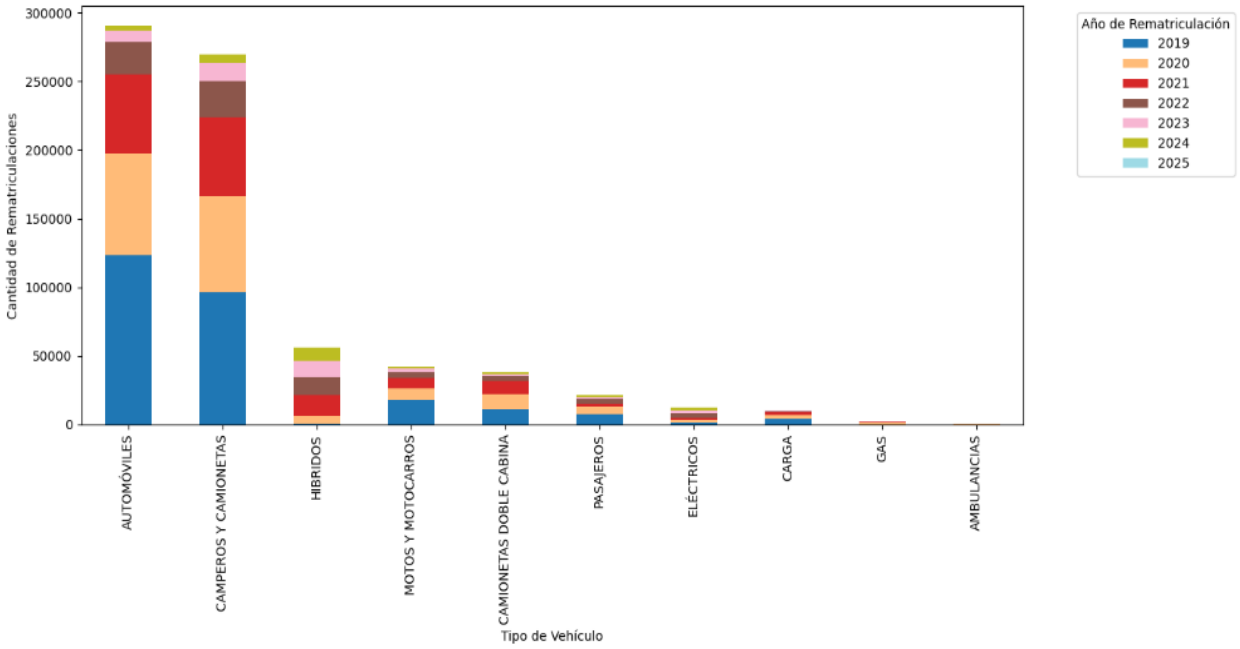
Tabla 13

Relación Entre Tipo de Vehículo y Año de Matrícula

TIPO VEHÍCULOS	2019	2020	2021	2022	2023	2024	2025	Total
AUTOMÓVILES	123086	74228	57721	23728	8416	3292	12	290483
CAMPEROS Y CAMIONETAS	96263	69913	57891	26137	13505	5789	18	269516
HIBRIDOS	1009	4981	15191	13522	12056	9511	4	56274
MOTOS Y MOTOCARROS	18017	8412	7586	4297	2564	1309	1	42186
CAMIONETAS DOBLE CABINA	10747	11375	9474	3438	2000	951	3	37988
PASAJEROS	7329	5580	2437	3280	1644	1200	0	21470
ELÉCTRICOS	1263	2469	1354	3332	2059	2136	1	12614
CARGA	4427	2646	1668	984	526	120	0	10371
GAS	0	1569	284	0	0	0	0	1853
AMBULANCIAS	193	253	187	123	17	5	0	778
Total	262334	181426	153793	78841	42787	24313	39	743533

Figura 21

Distribución Tipo de Vehículo por Año de Matrícula 2019- 2025



Modelo Red Neuronal Recurrente Predicción Facturas 2026-2028

Entrenar y Evaluar Modelo 1

Se implementa el modelo de redes neuronales recurrentes (RNN) con LSTM autorregresivo para hacer predicción, dentro del modelo se incorporó la evaluación con validación cruzada hacia atrás en series de tiempo. El código fuente correspondiente puede consultarse en el **Anexo A**.

Etapa de Exploración y Modelado Inicial

El objetivo modelo RNN es predecir la emisión de facturas del impuesto vehicular durante los años 2026 al 2028 a partir de los datos históricos de las vigencias 2022 al 2025.

Variables para el modelo RNN: Se tiene el dataframe consolidado con las siguientes variables:

VIGEN: Es la variable candidata a ser la variable objetivo porque permite el entrenamiento mediante el uso de serie de tiempo, es viable utilizarla para predicciones en secuencia en el tiempo, la limitante representa anualidad y solo se cuenta para registros que van de 2022 al 2025.

FE_REMATRI: También es candidata a ser la variable objetivo, está compuesta por año, mes y día, apta para el estudio el aprendizaje en secuencia de tiempo o se puede utilizar como una característica histórica al modelo.

ID_COM: Esta variable representa una característica valiosa que aporta detalle al objetivo del modelo, dado puede determinar el tipo de combustible relevante por año.

ID_TVH: Variable que representa una característica valiosa y aporta al detalle del objetivo del modelo, complementando información, ejemplo: el tipo de vehículo más común por

año, ID_EST: Es fundamental dado que determina si un vehículo se le genera factura conforme a su condición.

Proceso de Entrenamiento del Modelo

Los pasos que se surtieron para la construcción y entrenamiento del modelo son:

Preparación de datos

Crear secuencias (con ventanas deslizantes)

Validación cruzada inversa – Entrenar con años recientes, validar en años pasados

Predicción facturas 2026-2027 y 2028.

Fuente de Información

El alistamiento se hace con las facturas históricas de las vigencias 2022 al 2025, para la creación de las series del modelo, bajo el supuesto de prevenir que se cumplan uno o varios de los escenarios identificados como limitantes en un modelo de predicción. Por ello, se organiza la fuente de datos en la siguiente tabla.

Tabla 14

Combinación de Variables en una Sola Clave

CLAVE	Descripción de la clave	2022	2023	2024	2025
0101	Gasolina Automóvil	498649	1058253	1233170	1135470
0102	Gasolina Campero	179597	393591	470072	457220
0103	Gasolina Camioneta	16476	38072	40556	36625
0105	Gasolina Moto y motocarro	232807	440202	475820	437458
0106	Gasolina Pasajero	18025	47921	39138	56635
0107	Gasolina Carga	18025	47921	39138	56635
0201	Diesel ACPM Automóvil	263	3799	1032	868
0202	Diesel ACPM Campero	13507	32145	35540	33954
0203	Diesel ACPM Camioneta	13546	30760	35229	32474

CLAVE	Descripción de la clave	2022	2023	2024	2025
0205	Diesel ACPM Moto y motocarro	38	133	58	69
0206	Diesel ACPM Pasajero	6090	12573	13075	11961
0207	Diesel ACPM Carga	7254	15857	18343	16831
301	Gas Natural Automóvil	30	51	78	75
302	Gas Natural Campero	28	46	56	64
303	Gas Natural Camioneta	13	18	19	21
305	Gas Natural Moto y motocarro	6	8	8	3
306	Gas Natural Pasajero	64	1159	1173	1941
0307	Gas Natural Carga	3	41	59	66
0401	Gas Gasolina Automóvil	6713	8080	5951	3532
0402	Gas Gasolina Campero	3046	4484	5410	6205
0403	Gas Gasolina Camioneta	906	1620	3011	1830
0405	Gas Gasolina Moto y motocarro	1	13	19	1
0406	Gas Gasolina Pasajero	2796	7544	10647	10931
0407	Gas Gasolina Carga	2101	3902	4996	4880
0601	Eléctrico Automóvil	96	48	15	1
0602	Eléctrico Campero	5	11	8	12
0603	Eléctrico Camioneta	144	314	13	0
0605	Eléctrico Moto y motocarro	179	393	344	0
0606	Eléctrico Pasajero	12	11	120	0
0607	Eléctrico Carga	3	2	2	0
1001	Gasolina Eléctrico Automóvil	244	775	278	54
1002	Gasolina Eléctrico Campero	237	516	510	103
1003	Gasolina Eléctrico Camioneta	26	1207	16	1
1007	Gasolina Eléctrico Carga	0	5	0	0

Nota. Con esta tabla se ejecuta el modelo.

Limitantes en el Entrenamiento del Modelo RNN

Escases de los datos: La variable VIGEN es la indicada para hacer el proceso de entrenamiento del modelo de red neuronal recurrente (RNN), sin embargo, tiene solo 4 registros para entrenar el modelo, esta situación afecta el resultado de la predicción.

Registros con sparsidad: Corresponde a las filas que contienen valores repetidos y/o cero y/o valores mínimos, en búsqueda de robustecer el modelo, se combinan las siguientes variables en un campo llamado “Clave” que se compone del ID_COM y el ID_TVH. Estas combinaciones se visualizan en la *Tabla 14*. Sin embargo, en esa misma tabla se visualizan registros con con sparsidad. Ver comportamiento de la clave= 1007.

Vigencia 2022= 0

Vigencia 2023= 5

Vigencia 2024= 0

Vigencia 2025= 0

La sparsidad, obstaculiza el entrenamiento, porque hay valores repetitivos con ceros o nulos, ocasiona que los modelos aprendan muy poco y se desoriente al momento de entrenar.

Desbalance extremo entre categorías: Esta situación también se presenta en la *Tabla 15*, en donde se identifican registros con valores muy altos y bajos, se evidencia claramente en las claves 0101 tienen cantidades de facturas así:

Vigencia 2022 = 498,649

Vigencia 2023 = 1,058,253

Vigencia 2024= 1,233,170

Vigencia 2025= 1,135,470

Clave 305, de la misma tabla, tiene los siguientes:

Vigencia 2022= 6

Vigencia 2023= 8

Vigencia 2024= 8

Vigencia 2025= 3

Esta situación genera un desbalance de clases, generando predicciones dominadas por la categoría mayoritaria en consecuencia las categorías pequeñas quedan mal modeladas.

Outliers y fluctuaciones interanuales: Se evidencia en el dataframe valores muy bajos en comparación a valores muy altos. Ejemplo: Clave 306 tiene los siguientes datos:

Vigencia 2022= 64 facturas

Vigencia 2023= 1,159 facturas

Vigencia 2024= 1,173 facturas

Este es un claro ejemplo de picos o fluctuación con la vigencia 2022 y vigencia 2023. Los cambios bruscos impiden que el modelo como RNN aprendan patrones de forma consistente y por el contrario se está incorporando ruido, inestabilidad y dificultad para que el modelo aprenda patrones generalizables y por el contrario genere patrones erráticos o poco generalizables.

Resultado Predicción Emisión de Facturas 2026 al 2028

Tabla 15

Resultado Vigencias 2026 al 2028

CLAVE	2026	2027	2028
101	1272946	1294403	1350233
102	522183	546142	585626
103	16476	16476	16476
105	474915	477254	492184

CLAVE	2026	2027	2028
106	56459	65614	70201
107	60704	63556	69566
201	263	263	263
202	37343	38258	39723
203	36199	36715	38063
205	56	56	54
206	12779	12774	12987
207	18722	18780	19289
301	93	101	114
302	73	82	91
303	23	25	28
305	3	3	3
306	2245	2885	3526
307	3	3	3
401	4175	4049	4087
402	6478	6870	7125
403	906	906	906
405	0	0	0
406	2796	2796	2796
407	5511	50791	6197
601	1	1	1
602	5	5	5
603	0	0	0
605	103	100	106
606	48	63	46
607	1	0	0
1001	54	54	54

CLAVE	2026	2027	2028
1002	216	205	211
1003	1	1	1
1007	0	0	0

Informe Predicción de Emisión de Facturas para 2026 al 2028

En este análisis, se evidencia que, a pesar del proceso de consolidado realizado durante el prealistamiento del dataframe, el comportamiento histórico de la emisión de facturas entre 2022 y 2025 presentó las siguientes características, afectando el entrenamiento del modelo de red neuronal recurrente.

Registros sin facturas en más de una vigencia o fluctuaciones significativas entre las distintas combinaciones: Esta situación se justifica, ya que se presentó sin emisión de factura en más de una vigencia y para las diferentes claves (combinación entre clase de combustible y tipo de vehículo), o al consultar la cantidad de facturas emitidas en una vigencia para las combinaciones de gasolina y tipo de vehículo, existe una alta diferencia en la cantidad de emisión de facturas entre las distintas combinaciones. En el caso de las facturas de vehículos con combustible tipo gasolina, se observó durante el análisis descriptivo que esta es una variable predominante en el parque automotor matriculado y activo en Bogotá. Lo mismo ocurre con la cantidad de automóviles, donde los vehículos a gasolina representan la mayor parte del parque automotor, por encima de otros tipos de vehículos y combustibles.

Por lo tanto, las combinaciones de "gasolina-automóvil", "gasolina-campero" y "gasolina-moto" son las que más se destacan, mostrando siempre una alta cantidad de emisión de facturas en comparación con otras combinaciones, como "diésel-ACPM-moto", que tiene una emisión de facturas mucho más baja. Otros casos particulares incluyen las combinaciones

"eléctrico-carga" y "gas-eléctrico-carga", donde la emisión de facturas es aún menor. Al consolidar estos datos en una sola fuente, las combinaciones menos representativas, especialmente las que involucran vehículos eléctricos o híbridos, generan picos de ruido debido a su comportamiento atípico, lo que afecta negativamente el entrenamiento del modelo.

Análisis de Crecimiento y Decrecimiento Datos Históricos Facturas 2022 al 2025

En el análisis de los datos históricos para los años 2022 al 2025, ver *Tabla 14*, se infiere la siguiente información para la clave 0101:

En 2023, respecto a 2022, hubo un crecimiento de 559.604 facturas.

En 2024, respecto a 2023, también hubo un crecimiento, aunque mucho menor, con 174.917 facturas adicionales.

En 2025 respecto a 2024, se registró una disminución de 97.700 facturas.

Es de resaltar que la clave 0101 es una de las más estables en comparación con otras claves. Observando los datos de la tabla 14 se evidencia la presencia de: Escases de los datos, registros con sparsidad, desbalance extremo entre categorías, outliers y fluctuaciones interanuales, todas estas situaciones se deben a varios factores:

Variabilidad en la emisión de facturas: La emisión de facturas no sigue un patrón uniforme ni completamente predecible. Aunque se realiza un proceso de prealistamiento, no todas las facturas se generan automáticamente para todos los contribuyentes. Factores como cambios en los datos del vehículo, modificación de titularidad, información incompleta o trámites administrativos pueden alterar el proceso. Además, a lo largo del año, los contribuyentes tienen la posibilidad de generar su factura, Recibo de Orden de Pago (ROP) o declaración a través de la oficina virtual, lo que genera una distribución temporal irregular que puede extenderse incluso por varios años.

Calidad de los datos: La calidad de los datos es un factor crítico, ya que existen registros faltantes, incompletos o inconsistentes, lo que genera que las facturas no se emitan en el año correspondiente.

Omisos: Muchos contribuyentes, aunque tienen la obligación de pagar el impuesto vehicular, no realizan el pago en el año correspondiente. Estos omisos no reciben la factura o el ROP hasta que se acerquen a un proceso administrativo o vendan el vehículo. Esto genera fluctuaciones en la emisión de facturas de un año a otro.

Evaluación del Modelo

La evaluación incorporada al modelo fue la validación cruzada hacia atrás (Reverse Time Series CV), se utilizó dos años de validación 2023 y 2022 y se hizo lo siguiente:

Usar años futuros como entrenamiento

Predecir el año de validación para cada fila (clave)

Entrenar un modelo LSTM para cada fila (clave)

Calcular la métrica incorporada:

mse = mean_squared_error(val_trues, val_preds)

mae = mean_absolute_error(val_trues, val_preds)

Error Cuadrático Medio (MSE): Mide el promedio de los cuadrados de los errores entre las predicciones y los valores reales. Se obtuvo el siguiente valor: 1717893581.5911443, se puede considerar un valor grande, lo que se puede considerar que se presentó errores en las predicciones

Error Absoluto Medio (MAE): Mide el promedio de los valores absolutos entre las predicciones y los valores reales. Se obtuvo el siguiente valor: 13174.881584416737, se considera que las predicciones están desviadas en 13.17

Entrenar y Evaluar Modelo 2

Etapa de Exploración y Modelado Inicial

El objetivo modelo RNN es predecir la cantidad de matrículas nuevas en las vigencias 2025 al 2028 a partir de los datos históricos de las matrículas del 2010 al 2024. El código fuente correspondiente puede consultarse en el Anexo B.

AÑO_REMATRI: Es la variable objetivo, se derivó de la variable FE_REMATRI, solo tiene el año de matrícula de los vehículos en Bogotá.

Proceso de Entrenamiento del Modelo

Los pasos que se surtieron para la construcción y entrenamiento del modelo son:

Cargar y preparar los datos

Crear ventanas deslizantes

Parámetros

Modelo RNN simplificado

Predicción futura auto regresiva con suavizado

Fuente de Información

El alistamiento se hace con las cantidades de matrículas entre las vigencias 2010 al 2024, para este modelo se utilizó la variable ID_TVH que identifica el tipo de vehículo. Los datos son:

Tabla 16

Cantidad de Matrículas Vigencias 2010 – 2024

ID TVH	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
1	182073	306507	259317	203624	264337	222927	189695	140023	138410	121350	73415	55309	23581	8383	3282
2	54474	83250	97977	103919	148409	116044	103945	88099	98766	94822	69055	55268	25976	13472	5766
3	22887	22008	17770	13692	16850	12106	12490	11018	11221	10440	11108	8983	3407	1991	947

ID TVH	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
5	122074	238746	204420	164721	155637	127479	47322	23945	18650	17739	8242	7149	4277	2555	1309
6	12976	19970	16542	15482	25481	14609	9136	6579	5225	7286	5550	2368	3277	1643	1200
7	6393	10691	8009	9551	9265	5814	4312	3851	5684	4353	2587	1604	976	525	119

Resultado Predicción Cantidad Matrículas 2010 al 2024

Los resultados obtenidos después de entrenar el modelo son los siguientes.

Tabla 17

Predicciones Matrículas Vehículos 2010-2024

ID_TVH	2025	2026	2027	2028
1	7378	6836	6762	7102
2	6237	5947	5787	5788
3	1013	973	950	950
5	2013	1899	1858	1916
6	1707	1630	1636	1680
7	134	126	120	120

Comparativo Datos Históricos vs Predicciones

A continuación, se visualiza el consolidado entre los datos históricos de matrículas desde la vigencia 2010 hasta la vigencia 2024.

Tabla 18

Comparativo Datos Históricos Matrículas vs Predicciones Matrículas 2025-2028

ID TVH	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
1	182073	306507	259317	203624	264337	222927	189695	140023	138410	121350	73415	55309	23581	8383	3282	7378	6836	6762	7102
2	54474	83250	97977	103919	148409	116044	103945	88099	98766	94822	69055	55268	25976	13472	5766	6237	5947	5787	5788
3	22887	22008	17770	13692	16850	12106	12490	11018	11221	10440	11108	8983	3407	1991	947	1013	973	950	950
5	122074	238746	204420	164721	155637	127479	47322	23945	18650	17739	8242	7149	4277	2555	1309	2013	1899	1858	1916
6	12976	19970	16542	15482	25481	14609	9136	6579	5225	7286	5550	2368	3277	1643	1200	1707	1630	1636	1680
7	6393	10691	8009	9551	9265	5814	4312	3851	5684	4353	2587	1604	976	525	119	134	126	120	120

ID_TVH= 1: Corresponde a la información de los Automóviles, se puede observar en este primer registro que los datos históricos están completos, la variable objetivo se presume con una secuencia importante, no hay registros con novedad de sparsidad, si hay fluctuación interanual, la tendencia es descendente, la predicción de la vigencia 2025 estima un crecimiento muy alto en relación al dato histórico que registra para la vigencia 2024, por tanto, la predicción es un patrón anormal que no están dentro de los rangos que se infieren en la fluctuación para este tipo de vehículo.

ID_TVH= 2: Corresponde a la información de los camperos, no hay registros con novedad de sparsidad, si hay fluctuación interanual, la tendencia es descendente que se mantiene desde el 2015, sin embargo, la predicción de la vigencia 2025 estima un crecimiento leve, pero no va en concordancia con la tendencia, aunque no es un patrón del todo anormal.

ID_TVH= 3: Corresponde a la información de la camioneta, sucede exactamente igual que la predicción de los camperos, la tendencia es descendente, sin embargo, la predicción de la vigencia 2025 estima un crecimiento leve, por tanto, la estimación no está por fuera del rango del histórico de matrículas para este tipo de vehículo.

ID_TVH= 5: Corresponde a la información de las motos, pese a que la tendencia muestra que hay un decrecimiento alto desde el 2012, la predicción del 2025 es muy positiva conforme a la evidente tendencia de los datos históricos.

ID_TVH= 6: Corresponde a la información de los vehículos de pasajeros, la tendencia de los datos históricos muestra fluctuaciones interanuales tanto de crecimiento como de decrecimiento, así pues, la predicción no se puede demarcar dentro o fuera de un rango establecido porque no está plenamente definido.

ID_TVH= 7: Corresponde a la información de los vehículos de carga, se observa en los datos históricos que la tendencia está marcada al decrecimiento desde el 2014, sin embargo, la estimación del 2025 es prudente con un crecimiento de 15 unidades, que no se puede demarcar como fuera de un rango posible.

Evaluación del Modelo2

La evaluación incorporada al modelo, son:

MSE (Mean Squared Error): 15,357,834.80, el error cuadrático promedio es alto, por tanto, las predicciones se desvían significativamente de los valores reales.

RMSE (Root Mean Squared Error): 3,918.91 el promedio en escala de número de matrículas es un valor elevado comparado con los datos históricos.

MAPE (Mean Absolute Percentage Error): 85.63% este resultado indica que el error porcentual está muy lejos de los valores reales, es decir, la predicción no cumple.

Conclusiones

A corte de 2025, el uso de combustibles alternativos, como los híbridos y eléctricos, presenta una baja representatividad dentro del parque automotor matriculado en Bogotá. La tendencia es evidente: según el análisis cruzado entre la clase de combustible y el estado del vehículo, se identificó una diferencia de 7.302.060 unidades a favor de los vehículos a gasolina, lo que confirma que este tipo de combustible continúa siendo el predominante en la ciudad. Esta afirmación se respalda en la Tabla 8, así como en las Figuras 13 y 14.

Entre 2011 y 2025, las matrículas de vehículos a gasolina en Bogotá disminuyeron un 1,6 % en comparación con el período 1995-2010, lo que representa una reducción de 121.676 vehículos. No obstante, esta disminución resulta poco significativa en el contexto de una transición hacia tecnologías más limpias, como los vehículos eléctricos o híbridos. Pese a la tendencia global hacia la movilidad sostenible y a los incentivos fiscales ofrecidos en la ciudad como descuentos del 60 % en el impuesto vehicular para vehículos eléctricos y del 40 % para híbridos, la adopción de estas tecnologías sigue siendo limitada. Esto significaría que posiblemente hay influencia de otros factores como el costo de adquisición, la disponibilidad de infraestructura o las preferencias del mercado local. Dichos factores podrían explorarse con mayor profundidad en futuros análisis complementarios. Ver gráfico figura 15.

Continuando con el análisis del tipo de combustible, a partir de los datos presentados en la Tabla 8 se identifica que, a corte de 2025, el parque vehicular activo en Bogotá asciende a 7.993.904 vehículos. De este total, la participación de los vehículos que utilizan combustibles alternativos limpios es la siguiente:

Totalmente eléctricos: 16.406

Gasolina-eléctrico (híbridos): 55.410

Diésel-eléctrico: 2.292

En conjunto, estos vehículos suman 73.108 unidades, lo que representa apenas el 0,91% del total del parque automotor activo. Esto indica que, a pesar de los incentivos económicos existentes y de la tendencia global hacia la movilidad sostenible, la adopción de tecnologías limpias en Bogotá continúa siendo limitada. Al detallar los tipos de vehículos que emplean combustibles alternativos en 2025, se encuentran los siguientes resultados:

Gasolina-eléctrico (híbridos):

Camperos y camionetas: 1.366

Automóviles: 1.351

Camionetas doble cabina: 1.250

Vehículos de carga: 5

Eléctricos puros:

Motos y motocarros: 916

Camionetas doble cabina: 471

Automóviles: 160

Vehículos de pasajeros: 143

Camperos y camionetas: 36

Vehículos de carga: 7

Diésel-eléctrico:

Camionetas doble cabina: 164

Automóviles: 46

Vehículos de carga: 12

Camperos y camionetas: 2

Aunque estos vehículos aún tienen una baja participación en el total de matriculados, se destaca que el tipo de combustible alternativo más utilizado es el gasolina-eléctrico, siendo los camperos y camionetas los más representativos dentro de este grupo.

Este dato es relevante, ya que el valor comercial de este tipo de vehículos en el mercado colombiano se encuentra entre los $100.000.000 y $200.000.000 COP, dependiendo de la marca y modelo. Esto sugiere que la adopción de tecnologías limpias está, en gran medida, concentrada en propietarios con un nivel adquisitivo alto.

Otro análisis muy importante tanto en el estadístico de tablas cruzadas con la variable Estado y año de matrícula (2011-2025), está relacionado con la tendencia a las matrículas nuevas en el periodo 2011 al 2025, la pandemia del COVID-19 declarada en el año 2020, estuvo marcada por una situación atípica que afectó muchos ámbitos de la vida, y en Bogotá se visibilizó en el comportamiento de las matrículas vehiculares.

No obstante, más allá de la situación del año 2020, se resalta conforme a la Tabla 10, que el decrecimiento en las matrículas nuevas va en descenso sin punto de pare en el 2024.esta situación es preocupante para la finanza de la ciudad ya que las matrículas y los vehículos activos representan la población objetivo para emitir las facturas y en consecuencia depende el recaudo del impuesto vehicular en Bogotá. A continuación, se muestra el porcentaje en la variabilidad negativa, mostrando un decrecimiento alto de un año a otro desde el año 2020.

2020: 30,67%

2021: 17,92%

2022: 46,66%

2023: 45,60%

2024: 43,15%

Frente a la evaluación del modelo 1 de predicción de la emisión de facturas entre 2026 y 2028, se identificó que los resultados presentan un bajo nivel de confiabilidad. Las métricas obtenidas fueron:

MSE (Error Cuadrático Medio): 1.717.893.581,59

MAE (Error Absoluto Medio): 13.174,88

Estos valores evidencian errores significativos en las predicciones, especialmente considerando que el MAE implica una desviación promedio de más de 13 mil unidades en las estimaciones. Las posibles causas de este bajo desempeño incluyen:

Calidad deficiente de los datos

Fluctuaciones interanuales atribuibles a omisos o comportamientos atípicos

Naturaleza no lineal y no uniforme del proceso de emisión de facturas, que se realiza durante todo el año.

Existencia múltiples documentos asociados al proceso, tales como. Factura, Recibo Oficial de Pago (ROP) y actos administrativos.

Escasez de datos en la variable objetivo

Además, se evidenció que variables clave como ORIGEN no pudieron ser consideradas en el análisis debido a la alta proporción de valores nulos, lo cual obstaculizó su inclusión en el modelo, por tanto, se debe realizar una evaluación previa y sistemática del estado de las variables a estudiar antes del desarrollo de modelos predictivos.

Finalmente, el resultado de la predicción de matrículas nuevas (2025-2028), a través del modelo RNN (Modelo No. 2) se entrenó utilizando datos históricos de matrículas entre 2011 y 2024. El objetivo era proyectar la evolución de las matrículas, dado que cada matrícula activa

genera una factura asociada al impuesto vehicular. Sin embargo, los resultados obtenidos no fueron favorables, como se detalla a continuación:

Resultados del modelo, las métricas de evaluación reflejan un rendimiento deficiente, tales como se evidencia con las siguientes:

MSE (Mean Squared Error): El resultado muestra desviaciones significativas entre las predicciones y los valores reales.

MAPE (Mean Absolute Percentage Error): Evidencia que las predicciones se desvían, en promedio, un 85.63% de los valores reales, un error inaceptablemente alto.

R^2 (Coeficiente de Determinación): Demuestra que el modelo no captura la tendencia de los datos. Factores que afectan el rendimiento: El bajo rendimiento del modelo puede atribuirse a la alta variabilidad de los datos históricos, como se observó en la Tabla 10. Entre 2011 y 2024, las matrículas nuevas muestran fluctuaciones significativas, como la caída del 30.67% en 2020 (de 258,438 a 179,179) y del 50.05% en 2024 (de 85,454 a 42,681), además de un decrecimiento sostenido desde 2015. Esta variabilidad, junto con el volumen limitado de datos (14 años), probablemente distorsionó el entrenamiento del modelo, dificultando su capacidad para capturar patrones consistentes.

Recomendaciones

Uno de los mayores retos durante el desarrollo de este trabajo consistió en la preparación de los datos, esto a causa de gran cantidad de registros con valores nulos. Fue necesario imputar estos datos para evitar la pérdida de información útil tanto en el análisis estadístico como en la construcción de los modelos de predicción. En este sentido, se recomienda a la Secretaría Distrital de Hacienda implementar mecanismos de validación de la completitud y calidad de los datos desde su origen. Al provenir de bases de datos estructuradas, se esperaba mayor consistencia en el registro de la información. Una base de datos completa y precisa es fundamental para asegurar resultados confiables y útiles para la toma de decisiones.

Durante el análisis estadístico de las vigencias 2022 a 2025, se identificó un incremento significativo en los estados de cancelación y traslado de matrículas, así como una tendencia a la disminución en el número de vehículos matriculados. Por tanto, se sugiere realizar un estudio más profundo que evalúe los factores internos y externos que estén influyendo en esta situación, dado que estos eventos tienen un impacto directo en el recaudo del impuesto vehicular y, por ende, en los ingresos para la ciudad.

Otro aspecto relevante identificado es que los vehículos eléctricos, híbridos y combinados, están siendo adoptados principalmente por propietarios con una alta capacidad económica. Esto sugiere que los beneficios tributarios actuales, como los descuentos del 60% para vehículos eléctricos y 40% para híbridos, están siendo aprovechados por una población que posiblemente ya cuenta con los recursos para cumplir con sus obligaciones tributarias, en consecuencia, se recomienda revisar y evaluar el diseño de estos incentivos, con el fin de asegurar un mayor equilibrio entre el beneficio ambiental y el impacto económico, que conlleve a la promoción de política más equitativa y eficiente para la ciudad de Bogotá.

Teniendo en cuenta todo el proceso desarrollado para la construcción del modelo predictivo de emisión de facturas para las vigencias 2026 a 2028, se resalta la necesidad de que, desde la entidad, se adelanten acciones orientadas a:

Consolidar procedimientos que garanticen la calidad de los datos desde su origen en el sistema ERP.

Implementar políticas de gobernanza del dato que aseguren la completitud, consistencia y calidad de la información.

Referencias Bibliográficas

Alarcon Garcia, R. E. (2021). *Sistema Analítico Basado en un Modelo Predictivo de Procesamiento de Datos en la Big Data en la Educación Superior*. Obtenido de https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/228973?page=7

ANDI y Fenalco. (marzo de 2025). *En Febrero de 2025 se Registraron 17.103 Vehículos Nuevos*. Obtenido de https://www.andi.com.co/Home/Noticia/17808-andi-y-fenalco-en-febrero-de-2025-se-re

Bogotá. (10 de febrero de 2023). *Medios para Obtener las Facturas de los Impuestos Predial y de Vehículos.* Obtenido de https://bogota.gov.co/mi-ciudad/hacienda/facturas-de-impuesto-predial-y-de-vehiculos-llegaran-de-3-maneras

Bogotá. (18 de abril de 2025). *Inició Entrega de 2 Millones de Facturas de Impuesto Vehículos 2025 en Bogotá.* Obtenido de https://bogota.gov.co/mi-ciudad/hacienda/pago-impuesto-vehiculos-bogota-2025-inicio-entrega-2-millones-recibos

Cano, D. (2024). *Proyecto TFM: Predicción de Series Temporales con Modelos de IA.* Obtenido de https://blog.structuralia.com/proyecto-tfm-predicci%C3%B3n-series-temporales-modelos-ia

Casas Roma, J. -N.-J. (2019). *Big Data: Análisis de Datos en Entornos Masivos*. Obtenido de https://elibro-net.bibliotecavirtual.unad.edu.co/es/lc/unad/titulos/117744?as_all=Big__data:__an%C3%A1lisis__de__datos__en__entornos__masivos&as_all_op=unaccent__icontains&prev=as

Cepal. (Octubre de 2021). *Ciudades inclusivas, sostenibles e inteligentes en el marco de la Agenda 2030 para el desarrollo en América Latina y El Caribe.* Obtenido de

https://www.cepal.org/es/proyectos/ciudades-inclusivas-sostenibles-inteligentes-marco-la-agenda-2030-desarrollo-america

Colegio de Estudios Superiores de Administración – CESA. (2019). *Trabajo de Investigación "Evaluación para Financiamiento de Vehículos en Colombia".* Obtenido de https://repository.cesa.edu.co/bitstream/handle/10726/2517/MBA_80171349_2020_1.pdf?sequence=4

DaimlerChrysler Research & Technology FT3/KL. (2025). *CRISP-DM: Towards a Standard Process Model for Data Mining.* Obtenido de https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Data Science PM. (2024). *What is CRISP DM?* Obtenido de https://www.datascience-pm.com/crisp-dm-2/

Goodman, J. (Octubre de 2022). *As Electric Vehicle Growth Squeezes Gas Tax Revenues, Data Helps States Prepare*. Obtenido de https://www.pewtrusts.org/en/research-and-analysis/articles/2022/10/03/as-electric-vehicle-growth-squeezes-gas-tax-revenues-data-helps-states-prepare

Herrera Vargas, S. H. (2023). *Diseño de un Modelo de Analítica de Datos para la Medición del Desempeño de Proyectos.* Obtenido de https://repository.universidadean.edu.co/entities/publication/c871ba8c-3103-4b60-9802-7c9e8eaaa3b7

Ira Ronzheimer, J. D. (2022). *Towards the Measurement of Electromobility in International Trade.* (P. a. Nations, Ed.) Obtenido de https://www.cepal.org/en/publications/48007-towards-measurement-electromobility-international-trade

Normas APA. (Marzo de 2024). *La Guía Definitiva para Presentar Trabajos Escritos*. Obtenido de https://normasapa.in/

Rodríguez Montequín, M. T., Álvarez Cabal, J. V., & Mesa Fernández. (s.f.). *Metodologías para la Realización de Proyecto de Datamining*. Obtenido de https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Sandoval, E. E. (Diciembre de 2022). *Situación Financiera de la Movilidad Urbana en Bogotá*. Obtenido de https://www.cepal.org/es/publicaciones/48554-situacion-financiera-la-movilidad-urbana-bogota

Scala Learning. (2000). *Metodología Desarrollo Proyectos Minería de Datos RS4*. Obtenido de https://gc.scalahed.com/recursos/files/r161r/w25341w/MetodologiaDesarrollo_proyectos_Mineria_de_Datos_RS4.pdf

Secretaria Distrital de Hacienda. (Septiembre de 2024). *Cumplimiento del Impuesto de Vehículos Automotores por Vigencia, Clase y Uso. Bogotá D.C*. Obtenido de https://datosabiertos.bogota.gov.co/dataset/cumplimiento-del-impuesto-de-vehiculos-automotores-bogota-d-c

Unidad de Planeación Minero Energética – UPME. (2020). *Realizar un Estudio que Permita Identificar las Clases de vehículos y Modalidades de Transporte Susceptibles de Realizar el Ascenso Tecnológico hacia Tecnologías de Cero y Bajas Emisiones a Nivel Nacional.* Obtenido de https://www1.upme.gov.co/DemandayEficiencia/Documents/Informe_final_Ascenso_tecnologico.pdf

Universidad de La Salle. (2025). *Modelo de Predicción de Ventas del (2022-2023) para la Empresa Metal Acero El Trébol.* Obtenido de

https://ciencia.lasalle.edu.co/server/api/core/bitstreams/7229fa61-be18-43b7-b275-3ac4ac75d5a4/content

Universidad Militar Nueva Granada. (2013). *Propuesta de Estrategias para el Mejoramiento del Recaudo del Impuesto de Vehículos Departamento de Cundinamarca.* Obtenido de https://repository.umng.edu.co/server/api/core/bitstreams/a33df386-06cd-4ac5-9a37-e72a552e5625/content

Universidad Tecnológica de Pereira. (2015). *Modelo de Predicción de Fallos para Proyectos de Software de la Universidad Tecnológica de Pereira Utilizando Redes Neuronales.* Obtenido de https://www.iadb.org/es/proyecto/CO-T1784

Apéndices

Apéndice A

Código Modelo 1 RNN

```
 !pip install thefuzz[speedup]
 pip install tensorflow
 import pandas as pd
 import numpy as np
 import seaborn as sns
 sns.set_style('darkgrid')  # Reemplaza plt.style.use('seaborn-darkgrid')
 import pandas as pd
 import numpy as np
 import matplotlib.pyplot as plt
 from sklearn.preprocessing import MinMaxScaler
 from tensorflow.keras.models import Sequential
 from tensorflow.keras.layers import SimpleRNN, Dense
 import seaborn as sns
 from keras.layers import Input
 from keras.models import Sequential
 from keras.layers import LSTM, Dense
 from sklearn.metrics import mean_squared_error, mean_absolute_error
 sns.set_style('darkgrid')
 print("Todas las librerías están instaladas correctamente.")

#Cargar dataframe
dfinalmodelo=pd.read_csv("dfinalmodelo.csv")
dfinalmodelo

#Verificar nulos
print(dfinalmodelo.isnull().sum())

#Información básica del dataframe
dfinalmodelo.info()

# Paso 1: Preparacón de los datos para la secuencia temporal
# Filtrar solo registros con estado activo (ID_EST == 1)
df_activos = dfinalmodelo[dfinalmodelo['ID_EST'] == 1]

# Agrupar por VIGEN, ID_COM, ID_TVH y contar las facturas (cada fila es una
factura)
df_grouped = df_activos.groupby(['VIGEN', 'ID_COM',
'ID_TVH']).size().reset_index(name='CANTIDAD_FACTURAS')

# Eliminar los registros con ID_COM igual a 1998, 3, 5, 7, 8, 9  11 y 12,
quedan 1, 2, 3, 4, 6, 10
df_grouped = df_grouped[~df_grouped['ID_COM'].isin([1998,5,7,8,9,11,12])]

# Eliminar los registros con ID_TVH igual a 4, 7, 8, 9 , 10 quedan 1, 2, 3,
5, 6, 7
df_grouped = df_grouped[~df_grouped['ID_TVH'].isin([4,8,9,10])]
```

```
# Crear la columna CLAVE como ID_COM + ID_TVH,
df_grouped['CLAVE'] = (df_grouped['ID_COM'].astype(str).str.zfill(2) +
df_grouped['ID_TVH'].astype(str).str.zfill(2))

# Pivotear para crear la serie temporal (CLAVE as index, VIGEN as columns)
df_series = df_grouped.pivot(index='CLAVE', columns='VIGEN',
values='CANTIDAD_FACTURAS').fillna(0)

# Verificar el resultado
print("df_grouped:")
print(df_grouped.head())
print("\ndf_series:")
print(df_series.head())
df_series.to_excel("facturasultima.xlsx")

#Paso 2: Función para crear secuencias
def create_sequences(data, n_steps):
    X, y = [], []
    for i in range(len(data[0]) - n_steps):
        X.append(data[0][i:i + n_steps])
        y.append(data[0][i + n_steps])
    return np.array(X), np.array(y)

# Paso 3: Validación cruzada inversa – Entrenar con años recientes, validar
en años pasados
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Input

# Parámetros
n_timesteps = 2
n_features = 1
years = [2022, 2023, 2024, 2025]
future_years = [2026, 2027, 2028]
n_future = len(future_years)

# Cargar los datos
df = pd.read_excel("facturasultima.xlsx")
df = df.set_index('CLAVE').astype(float)

# Resultados de validación
validation_results = {}

# VALIDACIÓN: predecir 2022 y 2023 usando años futuros
for val_year in [2025, 2024]:
    train_years = years[:years.index(val_year)]  # años anteriores al año de
validación
    val_preds, val_trues = [], []
    for i in range(df.shape[0]):
        row = df.iloc[i]
        train_vals = row[train_years].values
```

```
        val_val = row[val_year]

        # Filtros de seguridad
        if np.count_nonzero(train_vals) < n_timesteps:
            continue
        if len(train_vals) <= n_timesteps:
            continue
        if np.all(train_vals == train_vals[0]):
            continue

        # Escalar
        scaler = MinMaxScaler()
        try:
            train_vals_scaled = scaler.fit_transform(train_vals.reshape(-1,
1)).flatten()
        except:
            continue

        # Crear secuencias
        X, y = create_sequences(train_vals_scaled.reshape(1, -1),
n_timesteps)
        if len(X) == 0:
            continue
        X = X.reshape((X.shape[0], n_timesteps, n_features))

        # Modelo LSTM
        model = Sequential([
            Input(shape=(n_timesteps, n_features)),
            LSTM(20, activation='relu'),
            Dense(1)
        ])
        model.compile(optimizer='adam', loss='mse')
        model.fit(X, y, epochs=100, verbose=0, batch_size=1)

        # Predicción
        last_seq = train_vals_scaled[-n_timesteps:].reshape((1, n_timesteps,
1))
        pred_scaled = model.predict(last_seq, verbose=0).flatten()[0]
        pred_real =
scaler.inverse_transform(np.array([[pred_scaled]])).flatten()[0]
        val_preds.append(pred_real)
        val_trues.append(val_val)

    # Métricas
    if val_preds:
        mse = mean_squared_error(val_trues, val_preds)
        mae = mean_absolute_error(val_trues, val_preds)
        validation_results[f'Train {train_years} -> Val {val_year}'] =
{'MSE': mse, 'MAE': mae}
    else:
        validation_results[f'Train {train_years} -> Val {val_year}'] = {
            'MSE': None, 'MAE': None, 'Note': 'No valid data for prediction'
        }
```

```
# MOSTRAR VALIDACIONES
print("Resultados de validación:")
print(validation_results)

#  Paso 4: Predicción para 2026–2028 (auto-regresiva)
pred_df = pd.DataFrame(index=df.index, columns=future_years)
for i in range(df.shape[0]):
    row = df.iloc[i].values

    # Filtros
    if np.count_nonzero(row) < n_timesteps:
        continue
    if len(row) <= n_timesteps:
        continue
    if np.all(row == row[0]):
        continue

    # Escalado
    scaler = MinMaxScaler()
    try:
        row_scaled = scaler.fit_transform(row.reshape(-1, 1)).flatten()
    except:
        continue

    # Crear secuencias
    X_train, y_train = create_sequences(row_scaled.reshape(1, -1),
n_timesteps)
    if len(X_train) == 0:
        continue
    X_train = X_train.reshape((X_train.shape[0], n_timesteps, n_features))

    # Modelo
    model = Sequential([
        Input(shape=(n_timesteps, n_features)),
        LSTM(10, activation='relu'),
        Dense(1, activation='relu')  # evitar negativos
    ])
    model.compile(optimizer='adam', loss='mse')
    model.fit(X_train, y_train, epochs=200, verbose=0, batch_size=1)

    # Predicción futura
    current_seq = row_scaled[-n_timesteps:].tolist()
    future_preds_scaled = []
    for _ in range(n_future):
        X_input = np.array(current_seq[-n_timesteps:]).reshape((1,
n_timesteps, 1))
        pred_scaled = model.predict(X_input, verbose=0).flatten()[0]
        future_preds_scaled.append(pred_scaled)
        current_seq.append(pred_scaled)

    # Inversión del escalado
    row_extended = np.concatenate([row_scaled, future_preds_scaled])
    row_inverse = scaler.inverse_transform(row_extended.reshape(-1,
1)).flatten()
```

```
    future_preds = row_inverse[-n_future:]
    pred_df.iloc[i] = np.round(future_preds).astype(int)

# EXPORTAR RESULTADOS
pred_df.to_excel("predicciones_facturas_2026_2028.xlsx")
print("Archivo exportado: predicciones_facturas_2026_2028.xlsx")
import matplotlib.pyplot as plt

# Elige una clave (puedes usar .iloc para la primera o poner directamente un
valor de índice)
clave_ejemplo = df.index[29]  # O por ejemplo: clave_ejemplo = 123456

# Verifica que esté en ambos DataFrames
if clave_ejemplo in df.index and clave_ejemplo in pred_df.index:
    historico = df.loc[clave_ejemplo, [2022, 2023, 2024, 2025]]
    predicho = pred_df.loc[clave_ejemplo, [2026, 2027, 2028]]
        # Combinar en una sola serie
    serie_completa = pd.concat([historico, predicho])
    serie_completa.index = serie_completa.index.astype(int)
    # Graficar
    plt.figure(figsize=(10, 6))
    plt.plot(serie_completa.index, serie_completa.values, marker='o',
label='Valores')
    plt.axvline(x=2025.5, color='gray', linestyle='--', label='Inicio
predicción')
    plt.title(f'Predicción de facturas para la clave {clave_ejemplo}')
    plt.xlabel("Año")
    plt.ylabel("Facturas emitidas")
    plt.grid(True)
    plt.legend()
        plt.tight_layout()
    plt.show()
else:
    print(f"La clave {clave_ejemplo} no se encuentra en los datos.")

# Ver metricas
validation_results
for k, v in validation_results.items():
    print(f"{k}:")
    if isinstance(v, dict):
        for metric, value in v.items():
            print(f"   {metric}: {value}")
    else:
        print(f"   {v}")
```

Apéndice B

Código Modelo 2 RNN

```
 !pip install thefuzz[speedup]
 pip install tensorflow
 !pip install seaborn
 import pandas as pd
 import numpy as np
 import seaborn as sns
 sns.set_style('darkgrid')  # Reemplaza plt.style.use('seaborn-darkgrid')
 import pandas as pd
 import numpy as np
 import matplotlib.pyplot as plt
 from sklearn.preprocessing import MinMaxScaler
 from tensorflow.keras.models import Sequential
 from tensorflow.keras.layers import SimpleRNN, Dense
 import seaborn as sns
 from keras.layers import Input
 from keras.models import Sequential
 from keras.layers import LSTM, Dense
 from sklearn.metrics import mean_squared_error, mean_absolute_error
 from tensorflow.keras import Input
 sns.set_style('darkgrid')
 print("Todas las librerías están instaladas correctamente.")

#Cargar Dataframe
dfinalmodelo=pd.read_csv("dfinalmodelo.csv")
dfinalmodelo
import pandas as pd
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Input
from sklearn.preprocessing import MinMaxScaler

# Cargar y preparar los datos
df = pd.read_excel("cantidad_matriculasvehiculo.xlsx", sheet_name="Sheet1")
df_long = df.melt(id_vars="ID_TVH", var_name="AÑO", value_name="MATRICULAS")
df_long["AÑO"] = df_long["AÑO"].astype(int)

# Función para crear ventanas deslizantes
def create_sequences(data, n_steps):
    X, y = [], []
    for i in range(len(data) - n_steps):
        X.append(data[i:i+n_steps])
        y.append(data[i+n_steps])
    return np.array(X), np.array(y)

# Parámetros
N_STEPS = 5
EPOCHS = 100
FUTURE_YEARS = [2025, 2026, 2027, 2028]
results = {}
```

```
for clave in df["ID_TVH"].unique():
    serie = df_long[df_long["ID_TVH"] ==
clave].sort_values("AÑO")["MATRICULAS"].values.reshape(-1, 1)

    # Verificar si hay suficientes datos
    if len(serie) <= N_STEPS:
        print(f"Clave {clave} no tiene suficientes datos, se omite.")
        continue

    # Verificar si la serie es constante
    if np.all(serie == serie[0]):
        print(f"Clave {clave} tiene valores constantes, se omite.")
        continue

    # Escalado por fila
    scaler = MinMaxScaler()
    serie_scaled = scaler.fit_transform(serie)
        X, y = create_sequences(serie_scaled, N_STEPS)
    X = X.reshape((X.shape[0], X.shape[1], 1))

    # Modelo RNN
    model = Sequential([
        Input(shape=(N_STEPS, 1)),
        LSTM(50, activation='relu'),
        Dense(1, activation='relu'), # Garantizar no negativos
        Dense(1)  # No usar activación aquí
    ])
    model.compile(optimizer='adam', loss='mse')
    model.fit(X, y, epochs=EPOCHS, verbose=0)

    # Predicción futura autoregresiva
    last_window = serie_scaled[-N_STEPS:].reshape(1, N_STEPS, 1)
    future_preds = []
    for _ in FUTURE_YEARS:
        pred = model.predict(last_window, verbose=0)
        pred = pred.reshape(1, 1, 1)
        future_preds.append(pred[0, 0, 0])
        last_window = np.concatenate([last_window[:, 1:, :], pred], axis=1)

    # Invertir escalado
    future_preds_rescaled =
scaler.inverse_transform(np.array(future_preds).reshape(-1, 1)).flatten()
    results[clave] = dict(zip(FUTURE_YEARS, future_preds_rescaled))

# Guardar resultados
df_preds = pd.DataFrame.from_dict(results, orient="index").reset_index()
df_preds = df_preds.rename(columns={"index": "ID_TVH"})
df_preds.to_excel("prediccion_matriculavehiculo.xlsx", index=False)

#print("Predicción guardada como prediccion_matricula.xlsx")
import pandas as pd
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Input, Dropout
```

```
from sklearn.preprocessing import MinMaxScaler
import tensorflow as tf

# Custom activation to ensure non-negative outputs
def non_negative_activation(x):
    return tf.maximum(x, 0.0)

# Custom loss function to penalize flat predictions
def custom_loss(y_true, y_pred):
    mse = tf.reduce_mean(tf.square(y_true - y_pred))  # Reemplazo correcto
    # Penalize predictions that are too similar to each other (encourages
trend learning)
    penalty = tf.reduce_mean(tf.square(y_pred - tf.reduce_mean(y_pred)))
    return mse + 0.05 * penalty # Reduced penalty weight for stability

# Cargar y preparar los datos
df = pd.read_excel("cantidad_matriculasvehiculo.xlsx", sheet_name="Sheet1")
df_long = df.melt(id_vars="ID_TVH", var_name="AÑO", value_name="MATRICULAS")
df_long["AÑO"] = df_long["AÑO"].astype(int)

# Función para crear ventanas deslizantes
def create_sequences(data, n_steps):
    X, y = [], []
    for i in range(len(data) - n_steps):
        X.append(data[i:i+n_steps])
        y.append(data[i+n_steps])
    return np.array(X), np.array(y)

# Parámetros
N_STEPS = 4
EPOCHS = 200
FUTURE_YEARS = [2025, 2026, 2027, 2028]
results = {}
for clave in df["ID_TVH"].unique():
    serie = df_long[df_long["ID_TVH"] ==
clave].sort_values("AÑO")["MATRICULAS"].values.reshape(-1, 1)

    # Verificar si hay suficientes datos
    if len(serie) <= N_STEPS:
        print(f"Clave {clave} no tiene suficientes datos, se omite.")
        continue

    # Apply modified log transformation to stabilize small values
    serie_log = np.log1p(serie + 1)  # Add 1 to avoid log(0) and stabilize
    scaler = MinMaxScaler()
    serie_scaled = scaler.fit_transform(serie_log)
        X, y = create_sequences(serie_scaled, N_STEPS)
    X = X.reshape((X.shape[0], X.shape[1], 1))

    # Modelo RNN simplificado
    model = Sequential([
        Input(shape=(N_STEPS, 1)),
        LSTM(50, activation='relu', return_sequences=False),
        Dropout(0.2),  # Add dropout to prevent overfitting
```

```
        Dense(10, activation='relu'),
        Dense(1, activation=non_negative_activation)
    ])
    model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.001),
loss=custom_loss)

        # Early stopping and learning rate scheduler
    early_stopping = tf.keras.callbacks.EarlyStopping(monitor='val_loss',
patience=30, restore_best_weights=True)
    lr_scheduler = tf.keras.callbacks.ReduceLROnPlateau(monitor='val_loss',
factor=0.5, patience=10, min_lr=1e-6)

        # Train with validation split
    model.fit(X, y, epochs=EPOCHS, validation_split=0.2,
callbacks=[early_stopping, lr_scheduler], verbose=0)

    # Predicción futura autoregresiva con suavizado
    last_window = serie_scaled[-N_STEPS:].reshape(1, N_STEPS, 1)
    future_preds = []
    for _ in FUTURE_YEARS:
        pred = model.predict(last_window, verbose=0)

        # Apply slight smoothing to avoid flat predictions
        pred = 0.9 * pred + 0.1 * np.mean(last_window, axis=1, keepdims=True)
        future_preds.append(pred[0, 0])
        last_window = np.concatenate([last_window[:, 1:, :], pred.reshape(1,
1, 1)], axis=1)

    # Invertir escalado y transformación log
    future_preds_rescaled =
np.expm1(scaler.inverse_transform(np.array(future_preds).reshape(-1,
1)).flatten()) - 1
    future_preds_rescaled = np.maximum(future_preds_rescaled, 0)  # Ensure
non-negative
    results[clave] = dict(zip(FUTURE_YEARS, future_preds_rescaled))

# Evaluar el modelo con métricas
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import numpy as np

# Predecir valores para el conjunto de entrenamiento/validación
y_true = serie_scaled[N_STEPS:]  # Valores reales escalados
y_pred = model.predict(X, verbose=0)
# Invertir escalado para comparación
y_true_rescaled = np.expm1(scaler.inverse_transform(y_true).flatten()) - 1
y_pred_rescaled = np.expm1(scaler.inverse_transform(y_pred).flatten()) - 1

# Calcular métricas
mse = mean_squared_error(y_true_rescaled, y_pred_rescaled)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_true_rescaled, y_pred_rescaled)
mape = np.mean(np.abs((y_true_rescaled - y_pred_rescaled) / y_true_rescaled))
* 100
r2 = r2_score(y_true_rescaled, y_pred_rescaled)
```

```
# Análisis de sesgo
bias = np.mean(y_pred_rescaled - y_true_rescaled)
over_predictions = np.sum(y_pred_rescaled > y_true_rescaled) /
len(y_true_rescaled) * 100

# Imprimir resultados
print(f"Clave {clave} - Métricas:")
print(f"MSE: {mse:.2f}")
print(f"RMSE: {rmse:.2f}")
print(f"MAE: {mae:.2f}")
print(f"MAPE: {mape:.2f}%")
print(f"R²: {r2:.4f}")
print(f"Sesgo promedio: {bias:.2f} (negativo = subestimación, positivo =
sobreestimación)")
print(f"Porcentaje de predicciones sobreestimadas: {over_predictions:.2f}%")

# Guardar métricas junto con las predicciones
results[clave]["metrics"] = {
    "MSE": mse,
    "RMSE": rmse,
    "MAE": mae,
    "MAPE": mape,
    "R²": r2,
    "Bias": bias,
    "Over_Predictions": over_predictions
}

# Actualizar DataFrame para incluir métricas
df_preds = pd.DataFrame.from_dict(results, orient="index").reset_index()
df_preds = df_preds.rename(columns={"index": "ID_TVH"})
df_preds.to_excel("prediccion_matriculavehiculo_con_metricas.xlsx",
index=False)
# Guardar resultados
df_preds = pd.DataFrame.from_dict(results, orient="index").reset_index()
df_preds = df_preds.rename(columns={"index": "ID_TVH"})
df_preds.to_excel("prediccion_matriculavehiculo.xlsx", index=False)

#Grafica
import pandas as pd
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Input, Dropout
from sklearn.preprocessing import MinMaxScaler
import tensorflow as tf

# Visualización
import matplotlib.pyplot as plt
for clave in df["ID_TVH"].unique():
    historical = df_long[df_long["ID_TVH"] == clave].sort_values("AÑO")
    years = historical["AÑO"].values
    matriculas = historical["MATRICULAS"].values
    pred_years = FUTURE_YEARS
    pred_values = [results[clave][year] for year in FUTURE_YEARS]
```

```
        plt.figure(figsize=(10, 6))
    plt.plot(years, matriculas, label="Historical", marker='o')
    plt.plot(pred_years, pred_values, label="Predicted", marker='x',
linestyle='--')
    plt.title(f"Vehicle Registrations for ID_TVH={clave}")
    plt.xlabel("Year")
    plt.ylabel("Registrations")
    plt.legend()
    plt.grid(True)
    plt.savefig(f"plot_id_tvh_{clave}.png")
    plt.close()
```

Apéndice C

Validaciones Previas

```
 !pip install thefuzz[speedup]
 pip install tensorflow
 !pip install seaborn
 import pandas as pd
 import numpy as np
 import seaborn as sns
 sns.set_style('darkgrid')  # Reemplaza plt.style.use('seaborn-darkgrid')
 import pandas as pd
 import numpy as np
 import matplotlib.pyplot as plt
 from sklearn.preprocessing import MinMaxScaler
 from tensorflow.keras.models import Sequential
 from tensorflow.keras.layers import SimpleRNN, Dense
 import seaborn as sns
 sns.set_style('darkgrid')
 print("Todas las librerías están instaladas correctamente.")

#Cargar Dataframe
dfinal=pd.read_csv("dfinalmodelo.csv")

#Correlación de Pearson
# Agrupar por año (VIGEN) y contar facturas
facturas_por_vigencia =
dfinal.groupby('VIGEN').size().reset_index(name='NUM_FACTURAS')

# Variables temporales
dfinal['AÑO_REMATRI'] = pd.to_datetime(dfinal['FE_REMATRI']).dt.year
dfinal['MES_REMATRI'] = pd.to_datetime(dfinal['FE_REMATRI']).dt.month

# Agrupar por VIGEN con resumen de variables relevantes
df_features = dfinal.groupby('VIGEN').agg({
    'ID_COM': 'nunique',
    'DE_COM': lambda x: x.mode()[0],  # más frecuente
    'ID_TVH': 'nunique',
    'DE_TVH': lambda x: x.mode()[0],
    'ID_EST': lambda x: (x == 1).mean(),  # % activas
    'ESTADO': lambda x: x.mode()[0],
    'AÑO_REMATRI': 'median',
    'MES_REMATRI': 'median'
}).reset_index()
df_modelo = df_features.merge(facturas_por_vigencia, on='VIGEN')

# Convertir categóricas a dummies
df_modelo = pd.get_dummies(df_modelo, columns=['DE_COM', 'DE_TVH', 'ESTADO'],
drop_first=True)

# Normalizar si se desea
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
df_modelo_scaled = df_modelo.copy()
df_modelo_scaled.iloc[:, 1:] = scaler.fit_transform(df_modelo_scaled.iloc[:,
1:])
import seaborn as sns
import matplotlib.pyplot as plt

# Matriz de correlación
corr = df_modelo_scaled.corr()

# Visualizar
plt.figure(figsize=(12, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", center=0)
plt.title("Correlación entre variables y número de facturas")
plt.show()
```