PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Aplicación de redes neuronales para estimar la emisión de gases contaminantes con relación al consumo de gasolina de vehículos livianos circulando en Lima Metropolitana en el 2022

Tesis para optar el grado académico de Maestra en Energia que presenta:

Alessandra Meza Lázaro

Asesor:

Julio César Cuisano Egúsquiza

Lima, 2024

Informe de Similitud

Yo, Julio César Cuisano Egúsquiza, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis "Aplicación de redes neuronales para estimar la emisión de gases contaminantes con relación al consumo de gasolina de vehículos livianos circulando en Lima Metropolitana en el 2022", de la autora Alessandra Meza Lázaro dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 17%. Así lo consigna el reporte de similitud emitido por el software Turnitin el 08/10/2024.
- He revisado con detalle dicho reporte y la Tesis, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:
San Miguel, 09 de octubre de 2024.

<table>
<tr><td colspan="2">Cuisano Egúsquiza, Julio César</td></tr>
<tr><td>DNI: 10744493</td><td rowspan="2">Firma:
  </td></tr>
<tr><td>ORCID:0000-0002-2175-3656</td></tr>
</table>

AGRADECIMIENTOS

Primero a mis padres por motivarme cada día a cumplir con mis metas profesionales y a mi asesor, el Dr. Julio César Cuisano Egúsquiza por su apoyo y confianza en la elaboración del presente trabajo de investigación.

Al Dr. Juan Varga Machuca Bueno por su constante guía en el tema y consejos para sacar adelante el trabajo presentado.

DEDICATORIA

A mis padres.

RESUMEN

El presente trabajo de investigación se centró en la predicción de contaminantes de vehículos livianos, siendo los contaminantes medidos el HC, NOx, CO y CO2 en Lima Metropolitana durante el 2022. El objetivo del trabajo es estimar la emisión de gases contaminantes con relación al consumo de gasolina mediante una muestra representativa del parque automotor de la ciudad.

El análisis se realizó mediante el entrenamiento de múltiples redes neuronales, escogiendo el mejor esquema de red y contrastándolo con los resultados del modelo de aprendizaje supervisado llamado "random forest". Este enfoque permitió evaluar el desempeño de las redes neuronales y comparar dos metodologías diferentes, destacando así alternativas que no requieran el uso de redes neuronales. En particular, las redes neuronales resultaron ser una alternativa apropiada al tener resultados de un coeficiente de correlación de 0.85 para el caso del CO_2. Loa valores de precisión obtenidos mediante redes neuronales cumplen con lo requerido para brindar confiabilidad en los resultados, ya que se encuentran en el mismo rango que la investigación de predicción por redes neuronales de Seo J y otros. A diferencia del método de "random forest", se tuvo un mejor control del sobreajuste de los datos pese a tener 0.2 menos de precisión para las predicciones de HC, NOx y CO. Esto principalmente por la naturaleza aleatoria de las mediciones en un contexto real, donde múltiples variables afectan la conducción.

Además, el estudio incluyó la evaluación del desempeño de las metodologías de predicción por tipo de carrocería, permitiendo un análisis más detallado y específico de las emisiones vehiculares. Esto contribuye a una comprensión más profunda de las fuentes de contaminación y las posibles medidas de mitigación.

En base a ello, el estudio busca sentar las bases para posteriores investigaciones sobre la predicción de inventarios de emisiones en más ciudades del Perú. Asimismo, enfatiza la importancia del análisis de los inventarios de emisiones de los vehículos utilizando datos reales de campo, con el objetivo de plantear medidas de reducción de contaminantes y proponer alternativas que mejoren la calidad del aire en centros urbanos.

Índice

LISTA DE SÍMBOLOS

EF	Factor de emisión
ε	Error en una función multivariables
β_i	Variables de ingreso en una función. i indica el número de la variable
f_{rf}^{B}	Función de regresión en árboles de decisión
B	Número de árboles
T_b	Árbol procesado
σ	Función de activación
e	Función exponencial
$max(0, z)$	Máximo valor entre 0 y un valor z
LAT	Latitud (º)
LON	Longitud (º)
H	Altura sobre el nivel del mar (m)
c_{HC}	Concentración volumétrica de HC no quemados (ppm)
c_{CO}	Concentración volumétrica de CO no quemados (%)
c_{CO_2}	Concentración volumétrica de CO_2 no quemados (%)
c_{O_2}	Concentración volumétrica de O_2 no quemados (%)
λ	Factor lambda
T_{bs}	Temperatura de bulbo seco (º)
v	Velocidad del vehículo (km/h)
RPM	Revoluciones por minuto
a	Aceleración del vehículo (m/s^2)
τ	Torque (N.m)
$R_{A/C}$	Dosado estequiométrico
$\dot{m}_{CO_2}$	Gasto másico de emisiones de CO_2 (g/s)
$\dot{m}_{eCO_2}$	Gasto másico específico de emisiones de CO_2 (g/s)
$C_{combustible}$	Consumo volumétrico de combustible (l/h)
P_{ad}	Presión múltiple de admisión (kPa)
T_{ad}	Temperatura múltiple de admisión (ºC)

T_{ext}	Temperatura exterior (ºC)
R^2	Coeficiente de correlación
MSE	Error cuadrático medio
$RMSE$	Error de raíz cuadrada media
MAE	Error absoluto medio
A_{abcd}	Consumo de combustible por distancia recorrida (km), ***a*** responde al tipo de combustible, ***b*** a tipo de vehículo, ***c*** al control de la emisión y ***d*** al tipo de ruta o velocidad del vehículo.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

En América Latina, según Isbell y Alvaréz (2017: 91-120), el aumento de la demanda de petróleo y emisión de gases de efecto invernadero tiene origen en el crecimiento económico de los países, donde gran porcentaje de la contaminación es debido al sector de transporte urbano. El poder adquisitivo de la clase media implica el acceso a nuevas flotas de vehículos debido a la urbanización de múltiples ciudades. Asimismo, la ausencia de organización del transporte público en distintas regiones y los obstáculos para integrar apropiadas alternativas amigables al medio ambiente, no permiten un adecuado control de las emisiones de contaminantes y una transición a un transporte sostenible.

En base a las características de la región latinoamericana, relacionadas a la tasa de crecimiento de la urbanización de ciudades y la ausencia de políticas que establezcan un consumo eficiente de combustibles, existen múltiples variables para identificar la contaminación presente en las ciudades principales de cada país (Isbell y otros, 2017: 91-120). En el caso de Lima, la presencia de contaminantes categoriza la ciudad entre un rango de 51 a 150 de índice de calidad del aire (ICA), teniendo una calidad entre moderada e insalubre para grupos sensibles, según lo establece el Servicio Nacional de Meteorología e Hidrología del Perú – SENAHMI (SENAMHI, 2021). Por otro lado, la demanda de vehículos livianos para la ciudad pronostica un alto porcentaje de emisiones contaminantes, una razón para ello es la combustión de los motores, proceso en el cual las emisiones de gases de efecto invernadero se presentan en altas concentraciones.

El propósito del presente trabajo de investigación es predecir la emisión de contaminantes mediante el análisis de los parámetros vehículos y el consumo de combustible. Para ello, el trabajo se divide en cuatro capítulos principales de: Planteamiento del estudio, marco teórico, metodología, resultados y discusión. Los capítulos buscan esclarecer la relación para la estimación de emisiones y proponer una metodología apropiada para la estructura de la red neuronal que podrá ser usada dependiendo del tipo de contaminante que se quiera predecir. Con ello, se busca cuantificar la magnitud del impacto para los ciudadanos y establecer indicadores para la búsqueda de medidas de mitigación, incluyendo la implementación de políticas públicas para la limitación de emisiones en Lima.

CAPÍTULO I - PLANTEAMIENTO DEL ESTUDIO

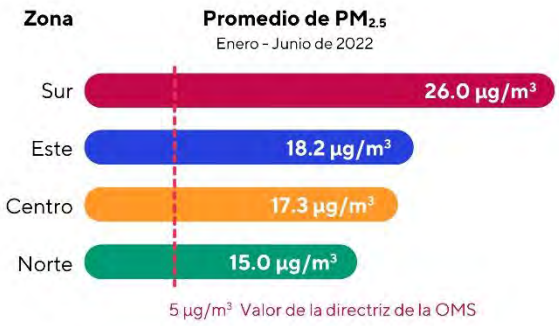
1.1 Planteamiento del problema

Los efectos de las emisiones contaminantes de los motores se dan por la condición de una mezcla rica o pobre de combustible, debido principalmente al exceso o ausencia de aire y otras variables del proceso de combustión (Sher, 1998). En base a ello, a partir de esas condiciones, el manejo de vehículos y la emisión de contaminantes está relacionada con el tipo de motor y, según la Dirección General de Electricidad (DGE) del Ministerio de Energía y Minas (MINEM), el conocimiento del uso del vehículo (2008).

La relación del manejo del vehículo puede ser presentada por ciclos de conducción ya determinados por países específicos. No obstante, múltiples regiones no poseen esta información, como el caso de Perú. Es por ello, que no se cuenta con una estimación o verificación del consumo de combustible y la emisión de contaminantes para Lima Metropolitana en distintos tipos de vehículos para evaluar los inventarios de emisiones. Los ciclos de conducción urbana difieren según la infraestructura vial, las condiciones ambientales, y el tipo y antigüedad de la flota vehicular presente en las ciudades (Amirjamshidi, 2015). La Figura 1, muestra los altos niveles contaminación en Lima Metropolitana, donde los valores sobrepasan el valor directriz brindado por la Organización Mundial de la Salud (Davila, E; 2024).

Figura 1

Ranking de contaminación por aire PM2.5 por zonas de Lima Metropolitana. (Davila, E; 2024).



Nota. Recuperado de "Lima tiene la mayor contaminación de aire por alta densidad de vehículos"

1.2 Formulación del problema

Analizando la estimación de inventarios de emisiones de contaminantes de vehículos livianos, según el consumo de combustible, se puede tener una proyección del impacto en la calidad del aire. Por otro lado, se podría identificar los indicadores de eficiencia energética de acuerdo a las características de los vehículos y la dinámica del transporte en Lima Metropolitana, siendo esta una ciudad representativa de América Latina. En este sentido, surge la siguiente pregunta de investigación, ¿Cómo se relaciona la emisión de gases contaminantes con respecto al consumo de gasolina de vehículos livianos en Lima Metropolitana en el 2022, utilizando redes neuronales para la predicción con datos reales del parque automotor?

A partir de la pregunta de investigación se determinó las siguientes preguntas específicas:

- ¿Qué variables o parámetros son determinantes en el consumo de combustible de los vehículos livianos conducidos en Lima Metropolitana?
- ¿Qué variables o parámetros son determinantes en la emisión de CO_2, CO, HC y NO_x de los vehículos conducidos en Lima Metropolitana?
- ¿Qué relación existe entre las emisiones de CO_2, CO, HC y NO_x con respecto al consumo de gasolina de determinados vehículos livianos conducidos en Lima Metropolitana?

1.3 Objetivos

1.3.1. Objetivo general

Estimar la emisión de gases contaminantes con relación al consumo de gasolina de 15 vehículos livianos circulando en Lima Metropolitana en el 2022, mediante la aplicación de redes neuronales y contrastando los resultados con la técnica de aprendizaje *random forest*.

1.3.2. Objetivos específicos

- Identificar qué variables o parámetros son determinantes en el consumo de gasolina de 15 vehículos livianos conducidos en Lima Metropolitana.
- Identificar qué variables o parámetros obtenidos son determinantes en la emisión de CO_2, CO, HC y NO_x de vehículos conducidos en Lima Metropolitana.
- Relacionar la emisión de CO_2, CO, HC y NOx con respecto al consumo de gasolina de los 15 vehículos livianos conducidos en Lima Metropolitana.

1.4 Justificación e importancia

El impacto del medio ambiente, por el parque automotor de vehículos livianos, presenta un porcentaje elevado en la calidad del aire de las ciudades. Es por ello, que se requiere analizar los inventarios de emisiones de los vehículos a partir de valores de variables bajo un contexto real para la búsqueda de medidas de reducción y alternativas que contribuyan a mejorar la calidad del aire en centros urbanos. Uno de los métodos para realizar estimaciones acerca del consumo de combustible y las emisiones de contaminantes, es a partir de los ciclos de conducción vehicular. No obstante, acorde sugiere el informe presentado por el SENAHMI basado en una "Estimación de emisiones vehiculares en Lima Metropolitana", es importante contar con datos de emisiones que se ajusten a la realidad del parque automotor peruano y de combustibles del país para realizar estudios consistentes y que brinden proyecciones como herramientas de decisión respaldadas (Dawidowski y otros, 2014). Esto principalmente por las características que difieren con otras regiones que cuentan con los indicadores en base a su flota vehicular e infraestructura vial, como el FTP – 75, NEDC, etc (Barlow y otros, 2009).

En este contexto, el presente trabajo de tesis propone una metodología basada en la implementación de redes neuronales y contrastar los resultados con la técnica de random forest para estimar la relación entre los gases contaminantes y el consumo de gasolina de vehículos livianos que circulan en Lima Metropolitana, utilizando datos experimentales.

CAPÍTULO II - MARCO TEÓRICO

En este capítulo, como primera parte se presenta el contexto del estudio y las distintas investigaciones nacionales e internacionales sobre la evaluación de emisiones contaminantes vehiculares y modelos de predicción empleados para distintas ciudades, así como metodologías para la captura y procesamiento de datos de los factores de emisión. A partir de ello, se realiza una evaluación de los estudios implementados para realizar comparativas con los modelos empleados en este trabajo de investigación. En la segunda parte del capítulo, se abordarán los fundamentos teóricos empleados en la elaboración de la red neural para la predicción de emisiones y las características de los factores de emisión a considerar en el proceso de implementación del modelo de predicción.

2.1 Antecedentes del problema

En el 2018 se llevó a cabo la "Cumbre Sudamericana sobre el control de emisiones vehiculares", evento en el cual se trató un plan regional para la adopción de medidas para la contaminación ambiental de vehículos. El objetivo del evento fue brindar información a los reguladores sudamericanos sobre las mejoras del cumplimiento de normas de emisiones y programas especiales para un transporte eficiente. Entre las medidas adoptadas para el monitoreo de emisiones como parte de herramientas innovadoras en el evento, destacaron el monitoreo remoto y el uso de PEMS ("*Portable Emissions Measurement Systems"*). Ambas herramientas se plantearon como recursos potenciales para brindar información real de la emisión de contaminantes para vehículos livianos con fin de demostrar transparencia relacionados a estándares europeos (ICCT, 2018). En base a la información provista, el uso de nuevas metodologías para la obtención de medidas que representen el impacto a la calidad del aire del transporte urbano es esencial para el cumplimiento del plan de acción de mejoras provisto por el Ministerio del Ambiente en Perú. A partir del marco normativo formulado, se implementaron medidas para la mejora y modificación de los Límites Máximos Permisibles (LMP), Estándares de Calidad Ambiental (ECA) y entre otros planes (MINAM, 2017).

2.1.1 Estudios nacionales

En el Perú, se cuenta con estudios de calidad de aire realizados de manera anual por el SENAHMI, como parte de la vigilancia de parámetros contaminantes. Para los estudios utilizan información que proviene de las estaciones meteorológicas y de otros centros internacionales para el análisis de las variaciones de cada variable. Asimismo, como parte de la relación entre el transporte urbano como factor en la contaminación, SENAHMI emplea una herramienta digital para el pronóstico del impacto de las emisiones. Como parte de las conclusiones del estudio se especifica que las emisiones se mantienen menores a los estándares ECA (SENAMHI, 2021). No obstante, los datos no son obtenidos de manera directa del parque automotor, sino por medio de proyecciones y correlacionando valores. A partir de ello, es importante realizar estudios y validaciones con nuevas fuentes de datos que permitan realizar un contraste con las emisiones a tiempo real y aplicando nuevas tecnologías para la predicción.

El trabajo de investigación titulado "Propuesta metodológica para el cálculo de factores de emisión para vehículos livianos a gasolina circulando en Lima Metropolitana" elaborado por Mendoza, J., J., en el 2022, plantea la obtención de información de mediciones en tiempo real de vehículos livianos. La metodología propuesta se centra en la estimación de factores de emisión de los vehículos bajo condiciones de manejo real en la ciudad de Lima. Los vehículos utilizados corresponden, principalmente, a la categoría M1 del tipo de carrocerías Sedán, SUV y Hatchback. Estos fueron analizados en base a rutas aleatorias seguidas por un mismo conductor, con lo cual se obtuvo de conclusiones que las emisiones de contaminantes detectadas presentaban una buena precisión, analizada desde una curva de velocidad media, en donde el promedio de mayores cantidades de emisiones se daba en velocidad menores a 20 km/h (Mendoza, 2022). El presente trabajo de tesis utiliza los datos del autor para el entrenamiento de los modelos. Estos datos fueron seleccionados en base a la metodología propuesta en el Capítulo III. Parte de la importancia de la investigación de Mendoza, es que los datos fueron adquiridos en un contexto real bajo parámetros que evidencian una muestra adecuada del parque automotor de Lima Metropolitana.

La aplicación de modelos de predicción para la evaluación de factores de emisión de los vehículos livianos en la ciudad de Lima, requieren de la validación de modelos utilizados de aprendizaje

supervisado y no supervisado. En el 2020, Vargas, presentó el trabajo titulado "Implementación de un modelo algorítmico para la estimación del nivel de concentración de contaminante PM2.5 en zonas urbanas", el cual tiene como objetivo estudiar diferentes modelos de aprendizaje de máquina para estimar el nivel de contaminación de aire en zonas urbanas en base a los datos de contaminantes y otras variables meteorológicas de la ciudad de Beijing. Los métodos utilizados para la investigación incluyen el modelo de regresión lineal, LightGBM y XGboost. A partir de estos, se precisa que el modelo de regresión lineal tiene un mejor desempeño predictivo entre todos, pero una menor robustez. De esta manera, el autor explica las técnicas del procesamiento inicial de los datos utilizada para la evaluación de los modelos, destacando la normalización de los mismos por medio de la adaptación de "*Inverse Distance Weighting*", lo que condujo a efectos diferentes para la evaluación de cada modelo de aprendizaje implementado. En base a ello, como trabajo futuro se plantea la adaptación de los modelos para la investigación de zonas urbanas en Lima (Vargas, 2020). La implementación de los métodos mencionados, corresponden al uso de técnicas de aprendizaje supervisado, el presente trabajo contrastó los resultados de los modelos seleccionados para evaluar su desempeño en contraste con el algoritmo XGboost debido a su implementación para modelos de concentraciones de contaminantes en regiones como China (Lin y otros, 2022)

En base a los estudios de técnicas de aprendizaje no supervisado, en el caso de la región, se investigaron trabajos aplicados a predicciones de emisiones de contaminantes mediante redes neuronales. El trabajo de investigación titulado "Predicción de la contaminación atmosférica generada por las emisiones del CO2 en el Perú, utilizando los métodos de ARIMA y Redes Neuronales", tuvo como objetivo encontrar un modelo adecuado para la predicción de dióxido de carbono en el Perú a partir de la comparativa del modelo econométrico ARIMA y la implementación de Redes Neuronales (Rosales, 2022). Para ello, utilizó los datos del Banco Mundial acerca de las emisiones anuales de CO_2 desde 1960 hasta el 2018, obteniendo como resultado que el uso de redes neuronales es más efectivo debido a su aplicación a un conjunto de datos no lineales y a su mayor convergencia. A partir de ello, el autor recomienda el acceso a datos de mediciones de CO_2 de manera periódica para futuras investigaciones, con el fin de obtener modelos más eficientes. El presente trabajo de investigación utilizará como referencia específicos parámetros de configuración para la sintonización de los modelos de redes neuronales. Se busca

implementar datos más estandarizados que eviten la no linealidad para la optimización de resultados.

La tesis de maestría titulada "Redes Neuronales para la predicción de contaminación del aire en Carabayllo-Lima", desarrolla un modelo de predicción por etapas para la evaluación del comportamiento de contaminantes medioambientales como el material particulado PM2.5 y otros químicos como el CO, SO_2 y NO capturados por la estación meteorológica de calidad de aire del distrito de Carabayllo en Lima, Perú (Herrera, 2019). El trabajo utiliza una red optimizada para un grupo de datos para la obtención de un modelo óptimo de predicción, en este caso emplea 4 modelos del tipo perceptrón multicapa: NARX, LM, BR y SCG. Entre estos, el algoritmo con mejor desempeño fue el SCG, debido a su bajo nivel de error y adecuada aproximación de pronóstico. Entre las mejoras a futuro, el autor recomienda la implementación de otras técnicas de entrenamiento no solo basado en retropropagación y, además, para una mejor precisión del entrenamiento sugiere a utilizar factores como datos meteorológicos, estacionales y con una mayor amplitud de años. En el caso del estudio mencionado y la investigación desarrollada, se busca evaluar la aplicación de un modelo de red mediante alternativas al uso de retropropagación para examinar el impacto de resultados al parque automotor de Lima Metropolitana. Asimismo, se plantea analizar los resultados en relación al CO y NO del trabajo de Herrera y del trabajo de Mendoza, J. del 2022., con el fin de contrastar las emisiones de Lima Metropolitana con Carabayllo.

2.1.2 Estudios internacionales

En diferentes regiones se ha realizado la implementación de *"machine learning"* para la predicción de emisión de contaminantes de vehículos y estimación de modelos por ciudades. Existen estudios con aplicaciones de múltiples técnicas de procesamiento, sean supervisadas o no supervisadas, orientadas a la contaminación por vehículos urbanos. Estos estudios plantean metodologías para la obtención de información, como características independientes examinadas por cada zona, para la evaluación de factores de emisión. En esta sección se expondrán estudios relevantes que serán utilizados como referencia para la selección de algoritmos y variables a examinar en la predicción de contaminantes y consumo de combustible en vehículos livianos para la ciudad de Lima.

En el 2012, Fonseca, en un estudio llamado "Aspectos de la medición dinámica instantánea de emisiones de motores. Aplicación al desarrollo de un equipo portátil y una metodología para estudios de contaminación de vehículos en tráfico real", planteó como objetivo evaluar los problemas asociados con la medición a tiempo real de equipos relacionados a variables medioambientales, para proponer el desarrollo de un equipo y una metodología de medición de contaminantes y consumo de combustible de vehículos de turismo en tráfico real. A partir de ello, se plantó un aparato de medición basado en concentraciones de emisiones contaminantes y caudales del gas de escape de los vehículos, incluyendo subsistemas de análisis y toma de muestras de vehículos en movimiento. El sistema incorporado de manera interna y que es utilizado en el proceso metodológico para la obtención de datos es el sistema PEMS ("*Portable Emissions Measurement Systems*") adaptado a los vehículos seleccionados para el muestreo. En base al sistema utilizado es posible plantear el análisis de las variables considerando los criterios de muestreo; es por ello, que la autora recomienda realizar ensayos en diferentes vehículos para mantener un estándar, dado que a pesar de que el conductor de los vehículos mantenga un mismo estilo de conducción habrá variaciones en el consumo de los vehículos. Por otro lado, el estudio toma en consideración la duración de cada prueba, donde tiempos prolongados pueden dar pasos variables no apropiados para los resultados (Fonseca, 2012). En base al estudio realizado por Fonseca, se tomará en consideración los resultados de las pruebas experimentales para el análisis de la base de datos que se tiene para el entrenamiento de los modelos a implementar, principalmente para tener referencia de las características específicas de los factores de emisión y consideraciones en el procesamiento de variables.

En el caso de los modelos de consumo de combustible y el análisis de emisiones, el estudio realizado por Tarun en el 2019 titulado "*Artificial Neural Networks for fuel consumption and emisiones modeling in light duty vehicles*" busca medir y realizar un modelo de consumo de combustible en ruta y medición de contaminantes de vehículos livianos de gasolina y diesel en Fort Collins, Colorado. El estudio se centra en la aplicación de modelos lineales multivariables y redes neuronales artificiales para la predicción del consumo de combustible y emisiones de un grupo determinado de vehículos. La medición se realizó mediante el uso de PEMS para la captura de valores de los gases contaminantes del tubo de escape. En base a múltiples configuraciones de

los modelos se realizó una evaluación de la predicción para cada tipo de vehículo. Los modelos de redes neuronales aplicados a los vehículos de gasolina fueron los menos sintonizados, debido a que la captura de datos se realizó en la fase fría de manejo del vehículo, obteniéndose modelos de datos altamente no lineales. A partir de ello, el bajo rendimiento también se atribuyó a la captura de datos por debajo del límite de detección del PEMS, llegando a correlaciones bajas entre valores, lo cual ocurrió de igual manera en el trabajo de Mendoza. Los datos capturados fueron realizados en zonas suburbanas, siendo el modelo diferente a carreteras o zonas altamente transitadas de Colorado (Tarun, 2019). La aplicación de redes neuronales a una base de datos captada permitirá correlacionar resultados con el caso de la ciudad de Colorado. Asimismo, dado que la obtención de datos se centró en suburbios, el análisis de datos del incremento de tendencias de emisiones brindará información sustancial para el caso de Lima en cuanto consumo de combustible.

El proceso de análisis de datos para corroborar las tendencias de emisión en cuanto al consumo de combustible depende de las zonas a evaluar para considerar la dispersión de contaminantes. En ese sentido, el estudio titulado "*Modeling of CO Emissions from Traffic Vehicles Using Artificial Neural Networks*" (Azees y otros, 2019) propuso una metodología para el análisis de emisiones en secciones determinadas de un mapa. La propuesta incluyó 3 pasos principales para la obtención de un modelo que se adapte a la emisión de monóxido de carbono de *New Klang Valley Express Way* en Malasia. Parte de este modelo se centró en la obtención de mapas a microescala de las áreas urbanas analizadas, para lo cual estableció como primer paso la selección de un modelo de predicción de correlación CFS ("*Correlation Based Feature Selection*") seguido de la aplicación de una red neuronal multicapa de perceptrón. Finalmente, se aplicó la creación de los mapas microescala para la predicción. Los autores concluyeron que las emisiones de CO en condiciones de tráfico se encuentran en un rango de 35 ppm en áreas céntricas, en contraste con áreas alejadas donde se presenta un valor de 0 ppm. Asimismo, identificaron que el uso de los mapas desarrollados son un recurso eficaz para la identificación de soluciones de mitigación de tráfico. Las mejoras propuestas por los autores incluyen alternativas de algoritmos que incrementen la precisión de resultados y técnicas para identificar los factores de emisión más importante durante el proceso de recolección de datos, para la reducción de tiempo de procesamiento computacional. A partir de ello, el trabajo de investigación, plantea el modelamiento de la predicción de contaminantes y el consumo de combustible utilizando como variable adicional la georeferencia

de los vehículos testeados para realizar una validación visual mediante mapas de predicción de emisiones.

En base a los estudios mencionados, se adoptará un panorama para el desarrollo del presente trabajo de investigación, tomando en consideración los modelos implementados por Azees y Tarun al igual que las técnicas de filtrado y limpieza de datos de la investigación experimental desarrollada por Mendoza, J.. Asimismo, considerando los estudios emisiones desarrollados por Vargas y Rosales sobre Lima, al igual que estudios de SENAHMI sobre la calidad del aire, se plantea contrastar en términos generales los resultados para evaluar la coherencia de los modelos entrenados al igual que el impacto de la cantidad de vehículos analizados como parte del proyecto de investigación. A diferencia de los estudios internacionales previos de emisiones, se realizará el enfoque del estudio a la ciudad de Lima, una de las ciudades más contaminadas de América Latina (IQAIR, 2021) y utilizando datos reales de vehículos representativos del parque automotor de la ciudad. Asimismo, en contraste con los modelos de redes neuronales implementados, se plantea una mejora en los parámetros de ajuste y procesamiento de filtrado inicial en una red neuronal multicapa.

2.2 Fundamento teórico

2.2.1 Sistemas de medición de emisiones

Los sistemas de medición de emisiones de vehículos han ido evolucionando en el tiempo. En base a ello, se hará referencia a los sistemas utilizados y protocolos correspondientes para el análisis de información en el trabajo de investigación.

Entre los sistemas de medición nuevos, recomendados como iniciativa innovadora para la lectura de variables de emisiones está el PEMS (ICCT, 2018). Este sistema es una alternativa para el monitoreo y recopilación de datos de contaminantes en base a concentraciones de emisiones. La medición de las variables por medio de estos equipos se centra en la portabilidad y acoplamiento externo en el mismo. El sistema cuenta con una serie de sensores capaces de registrar el flujo de escape del vehículo y consumo de combustible, realizando un seguimiento de CO, CO_2, NO, NO2

y partículas adicionales que salgan del vehículo (MAHLE, 2016). Los sistemas PEMS pueden ser adquiridos por compañías que se dedican al diseño y manufactura, por lo cual los sistemas pueden diferir en cuanto a componentes adicionales o mejoras en la captación de datos dependiendo de los proveedores. De esta manera, existen inventos como el desarrollado por Vojtisek-Lom, llamado "Portable On - Board System for Measurement Vehicle Exaust Particulate Emissions" que plantean el uso de fotómetros para un sistema que sensa la temperatura del motor, la temperatura del ambiente, detección de partículas del motor de combustión por medio de fotómetros, fuentes de luz y detectores de determinadas partículas, entre otros componentes. En base a ello, se incluyen sistemas de recolección y procesamiento de datos para la lectura de las variables medidas (Vojtisek-Lom, 2002).

En la Figura 2, se puede observar la instalación de un sistema PEMS donde se precisa que consta de componentes externos acoplados al vehículo y este es portable, de manera que permite el monitoreo en ruta a tiempo real de los parámetros deseados (Valverde y Bonnel, 2017).

Figura 2

Instalación de unidades principales PEMS.

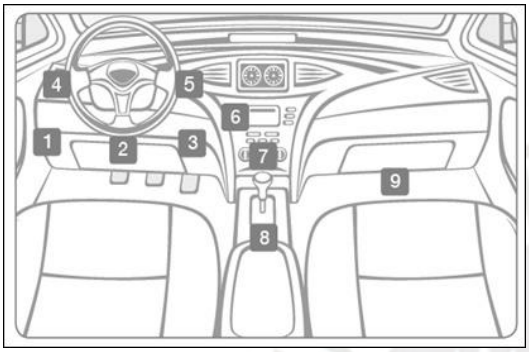
Nota. Tomado de "On-road testing with Portable Emissions Measurement Systems (PEMS)" por Valverde, V. & Bonnel, P. (2017). European Union: Joint Research Centre.

Un sistema adicional, utilizado por normativas para el diagnóstico del vehículo es el OBD ("*On Board Diagnostics*"), el cual está orientado al control y monitoreo del motor y otros dispositivos internos. El puerto ODB – II fue creado como estándar en la manufactura de vehículos para monitorear las emisiones por medio de una interfaz y líneas de mensajes. El sistema se encuentra ubicado en diferentes zonas dependiendo de la empresa que manufactura el vehículo. En la Figura 3, se puede observar los distintos puntos donde puede encontrar el puerto, el cual cuenta con 16

pines, de los cuales solo 9 son utilizados por el estándar SAE. En la Tabla 1 se puede visualizar la descripción de los pines a utilizar por el estándar para la lectura de la información (Klinedinst y King, 2016).

Figura 3

Posiciones del puerto OBD - II.



Nota. Tomado de Verizon (2022). Encuentra el puerto OBD - II. (https://espanol.verizon.com/support/knowledge-base-210135/)

Tabla 1

Diagrama de pines OBD – II utilizado por el estándar SAE.

Pin	Señal	Descripción
2	J1850 Bus +	—
4	CGND	GND
5	SGND	GND
6	CAN High	J-2284
7	ISO 9141.2 K-LINE	Tx/Rx
10	J1850 Bus -	—
14	CAN Low	J-2284
15	ISO 9141-2 L-LINE	Tx/Rx
16	+12 V	Batería

Nota. Adaptado de "SAE J1979" por SAE, 2002. Washington: SAE. (https://law.resource.org/pub/us/cfr/ibr/005/sae.j1979.2002.pdf).

El sistema OBD-II funciona para hacer un seguimiento del vehículo iniciado su proceso de marcha, en donde monitorea parámetros de fallas del encendido, el sistema de combustible, códigos de diagnóstico y otros (DEC, 2000). La verificación de distintos componentes del motor se realiza a

específicas condiciones, para lo cual se debe realizar un monitoreo no continuo de parámetros como: el sistema EGR, sensores de O2, catalizador, sistema evaporativo, sistemas de aire, entre otros (ICCT, 2016). El sistema cuenta con soporte a 5 distintos protocolos de comunicación que dependen de las normativas correspondientes a cada país y manufactureras de vehículos. En el caso de vehículos con una velocidad de transmisión serial de 10.4 kbaud, el protocolo será el ISO 9141-2, donde los mensajes y salida de datos son correspondientes a 12 bytes y protocolo UART. Los vehículos correspondientes a los datos brindados, corresponden a manufacturas europeas y asiáticas. No obstante, para este mismo mercado se encuentra la norma ISO 14230, donde marcas como Audi, Hyundai, Kia, Honda, entre otras, tienen una velocidad de transmisión de datos entre 1.2 y 10.4 kbaud, donde los mensajes llegan a tener 255 bytes de envío; otros protocolos son el ISO 15765 y SAE 1850 (VPW y PWM) (ESAT, 2013). A partir de los datos de las normas correspondientes, el proceso de análisis de información depende del protocolo de cada vehículo.

2.2.2 Proceso y evaluación de mediciones

2.2.2.1 Emisiones de contaminantes

Las emisiones se dan por medio del tubo de escape del vehículo debido a la quema de combustible, donde se dispersan contaminantes como: monóxido de carbono (CO), dióxido de carbono (CO2), óxidos de nitrógeno, total de hidrocarburos (THC) y partículas (Heywood, 2018).

El origen de las emisiones de contaminantes en el caso de los hidrocarburos se da por la quema incompleta de combustible en el motor (Heywood, 2018). Una de las causas es el apagado de la llama cuando entra en contacto con la cámara de combustión. La formación de hidrocarburos se da en la corona del pistón y la pared del cilindro del motor. A partir de ello, la formación de hidrocarburos se da por un arranque abrupto, lo que afecta la combustión del combustible y se mezcla con otros gases como NO_x y CO (Fonseca, 2012). La emisión de CO, se da por la combustión incompleta, en donde el combustible se oxida de manera parcial (Heywood, 2018). El desequilibrio químico generado ocasiona una emisión alta por los tubos de escape, debido al lento proceso de oxidación a una alta temperatura. En el caso del óxido de nitrógeno (NO_x), se da por la reacción entre los átomos de nitrógeno y oxígeno debido a elevadas temperaturas del motor, y un exceso de oxígeno en mezclas homogéneas (Fonseca, 2012).

En los motores de combustión, el proceso de control de emisiones depende de las condiciones del vehículo bajo específicos modos de manejo, en donde múltiples factores se involucran en el proceso de análisis de cada contaminante. Asimismo, la medición de contaminantes de vehículos en distintos escenarios de manejo influye a la razón de cambio de las emisiones debido a las variaciones en cuanto los estados del motor (Kelly y otros, 2017). Por un lado, las emisiones evaporativas corresponden a las emisiones diurnas, emisiones en circulación, emisión del vehículo recién apagado, en reposo con el motor en frío y durante el proceso de descarga. Por otro lado, en las emisiones por el tubo de escape, diferencia de las evaporativas, destacan los hidrocarburos (SEMARNAT, 2009).

En la Figura 4, se evidencia un resumen de las emisiones del vehículo. A partir de esta distribución se puede proyectar el uso de determinados equipos descritos en la sección anterior para la medición de contaminantes. Asimismo, las condiciones del vehículo son esenciales para el análisis de profundidad en la tasa de emisión (SEMARNAT, 2009).

Figura 4

Proceso de emisión de contaminantes en vehículos automotores.

Nota. Tomado de "Guía metodológica para la estimación de emisiones vehiculares en ciudades mexicanas" por SEMARNAT, 2009. México: Secretaría de Medio Ambiente y Recursos Naturales.

2.2.2.2 Factores de emisión

Los factores emisión son valores que asocian las emisiones de un gas con respecto al consumo de energía del mismo, en este caso al consumo de combustible (OECD, 2001). El cálculo de la emisión de estas variables se puede realizar aplicando distintas metodologías, entre ellas se destaca los procesos TIER 1, TIER 2 y TIER 3 desarrollados por El Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC) (IPCC, 2003). En el caso del transporte vehicular, se toman en consideración variables como el consumo de combustible del vehículo y el gas o contaminante disipado, del cual se haya una relación para expresar las emisiones (Kg CO_2/T.Km) en el ambiente:

$$Emisiones = Combustible\ consumido\ (TJ)\ x\ Factor\ de\ emisión\ (Kg\ CO_2/TJ), \quad (1)$$

En la Ecuación 1, el factor de emisión responde a ser la masa de un contaminante con relación a la energía requerida para su formación. Asimismo, las metodologías mencionadas, centradas en el análisis de emisiones, tienen variaciones de variables en el modelamiento de consumo de combustible en términos de eficiencia energética. Esto se debe a que múltiples factores intervienen en las condiciones del vehículo como: eficiencia del motor, carga, condiciones ambientales, año de antigüedad, condiciones del vías y más (India GHG, 2015).

En relación a las metodologías aplicadas también se puede incluir variables cómo el arranque en frío del vehículo y las pérdidas por evaporación de las emisiones a analizar (véase Ecuación 2):

$$Emisiones = \sum_{abcd}(EF_{abcd} x A_{abcd}) + \sum_b C_b + \sum_b E_b\ , \quad (2)$$

EF corresponde al factor de emisión del contaminante, siendo la masa por unidad de tasa de actividad (consumo de combustible), “A” es el consumo de combustible o distancia recorrida, “C” son las emisiones adicionales por arranque en frío y “E” las emisiones por evaporación del vehículo. Además, la variable “a” responde al tipo de combustible, “b” al tipo de vehículo, “c” al control de la emisión y “d” al tipo de ruta o velocidad del vehículo (IPCC, 2001).

En el Perú, en el año 2017 por medio del decreto supremo Nº 010-2017-MINAM se establecieron

los límites máximos permisibles para los vehículos categorías L3 y L5 (Tabla 2) y categorías M y N (Tabla 3) (El Peruano, 2017).

Tabla 2

Límites permisibles de vehículos categorías L3 y L5 del 2017 en adelante.

I.1. Vehículos de categorías L3 a L5 con motores de encendido por chispa de dos tiempos que usan mezclas de gasolina - aceite como combustible y de cuatro tiempos que usan gasolina, GLP o GNV como combustibles								
Año aplicación(*)	Categoría	Norma	Directiva	Ciclo	N° de ruedas	CO [g/Km]	HC [g/Km]	NOx [g/Km]
2017 en adelante	< 150 cc	EURO III o de mayor exigencia	2002/51/EC(B)(1)	ECE R40(2)	2	2,0	0,8	0,15
2017 en adelante	≥ 150 cc	EURO III o de mayor exigencia	2002/51/EC(B)(1)	ECE R40(3)	2	2,0	0,3	0,15
2017 en adelante	Todos	EURO II o de mayor exigencia	2002/51/EC(A)(1)	ECE R40	3(4)	7,0	1,5	0,4
2017 en adelante	vmáx < 130 km/h	EURO III o de mayor exigencia	2006/72/EC(C)(1)	WMTC	2	2,62	0,75	0,17
2017 en adelante	vmáx ≥ 130 km/h	EURO III o de mayor exigencia	2006/72/EC(C)(1)	WMTC	2	2,62	0,33	0,22

Nota. Tomado de "DECRETO SUPREMO N° 010-2017-MINAM". Diario Oficial El Peruano. MINAM. (2017).

Tabla 3

Límites permisibles de vehículos categorías M y N por rangos de fecha.

II.4. Vehículos de categorías M y N con motor de encendido por chispa a gasolina, GLP o GNV como combustible u otros combustibles alternos				
Año de fabricación(*)	Altitud [msnm]	CO [% - v/v]	HC [ppm]	CO + CO2 [% - v/v] mínimo
Hasta 1995	0 a 1800 > 1800	3,0 3,0	400 450	10 [8(1)] 8
1996 a 2002	0 a 1800 > 1800	2,5 2,5	300 350	10 [8(1)] 8
2003 en adelante	A cualquier altitud	0,5	100	12[8(1)]

Nota. Tomado de "DECRETO SUPREMO N° 010-2017-MINAM". Diario Oficial El Peruano. MINAM. (2017).

2.2.3 Modelos de Machine Learning

La visualización de datos y el análisis de los mismos, al igual que la interpretación, corresponden a una meta de cuantificación de datos, donde la metodología corresponde al "*statical learning*". Este término es referido a la interpretación y cuantificación de datos, mientras que las técnicas enfocadas a la predicción de datos utilizando grandes cantidades de información corresponden al término de "*machine learning*" o aprendizaje de máquina (Kroese y otros, 2022). El aprendizaje

de máquina o "*machine learning*" aborda el cómo los sistemas pueden ser entrenados a partir de datos para el proceso de aprendizaje de los mismos, con el objetivo de predicción o toma de decisiones de ese continuo análisis. A partir de un razonamiento lógico, a medida que se incrementa la cantidad de información, se modifican las decisiones o resultados de los datos, debido a variaciones de casos provistos. De ese modo, se puede tener mejores resultados o en múltiples ocasiones debido a la variabilidad y dispersión de la información se puede llegar a no tener decisiones precisas (Ranjan, 2016).

La metodología empleada para la toma de decisiones o aprendizaje, que parte de la inteligencia, requiere de una serie de procesos por etapa para ser definido de manera precisa. Múltiples técnicas de "*machine learning*" vienen del proceso de aprendizaje de humanos y animales por medio de la búsqueda de modelos computacionales que se adapten a dichos procesos (Nilsson, 2018). Por lo tanto, los modelos parten del aprendizaje biológico para conseguir un programa estructurado de datos que permitan realizar tareas asociadas, de alguna manera, a la inteligencia artificial. El proceso de aplicación incluye una serie de etapas como: reconocimiento, análisis, planificación, predicción, entre otras (Nield, 2022). En base a lo escrito por Nilsson y otros, la aplicación de "*machine learning*", pese a ser trivial el aprendizaje de una máquina, implica ventaja en el procesamiento de grandes cantidades de datos, con lo cual se puede llegar a relacionar variables entre sí. Por medio de ello, el cerebro humano, a pesar de estar evolucionado no cuenta con la capacidad de manejo de tanta información (Nilsson, 2018: 1-9). En este trabajo de investigación, se busca relacionar una gran cantidad de datos y dos variables independientes para evaluar en términos de eficiencia energética y contaminación ambiental, las emisiones de contaminantes y consumo de combustibles de vehículos en Lima Metropolitana, a partir de lo cual el uso de técnicas de "*machine learning*" son una herramienta esencial para lograr los objetivos planteados del trabajo.

En base a lo descrito anteriormente, es importante diferenciar las técnicas empleadas en el proceso de aplicación de modelos para realizar un juicio adecuado de los resultados y una implementación metodológica eficaz de lo que requiere el presente estudio. A partir de ello, se introducirá los métodos de aprendizaje de máquina supervisados y no supervisados orientados a la predicción e importancia en el área de estudio a realizar. Por otro lado, se dará énfasis a las bases teóricas de la

aplicación de redes neuronales como técnica no supervisada para el análisis y predicción de emisiones de contaminantes y las etapas a implementar que requiere el modelo.

2.2.3.1 Aprendizaje supervisado

El aprendizaje supervisado es un proceso por el cual el sistema está en la capacidad de producir decisiones basadas en datos de entrada que son filtrados y procesados para ser contrastados con los datos de salida, con lo cual se tiene un aprendizaje basado en la relación entre ambos (Ranjan, 2016). Un ejemplo de aplicación es el uso de hipótesis para hallar una relación entre las variables de una función para una cantidad de muestras, donde la aplicación de aprendizaje supervisado se centra en la relación de cada punto que pueda tenerse para el ajuste de las muestras a la función planteada (Nilsson, 1998). El aprendizaje supervisado abarca la asignación de etiquetas a cada punto de entrada y salida, es decir que se procesa dicha relación para aprender de la misma. Una vez obtenido el modelo de esos datos, se puede predecir el comportamiento de futuros puntos en base a determinados parámetros de entrada o salida, dependiendo de la aplicación. No obstante, una dificultad de ello, es la obtención de datos etiquetados, que requieren de interpretación humana y que esta esté correcta para una adecuado aprendizaje y predicción (Lindholm y otros, 2019).

Los métodos de aprendizaje supervisados más conocidos y de los cuales parten variaciones son regresión, clasificación, árboles de decisión y redes bayesianas (Lindholm y otros, 2019). En el caso del método de regresión se tiene la aplicación de la relación entre variables de cuantitativas y cualitativas de entrada con una variable de salida para la aplicación de modelos en la forma de:

$$y = f(x) + \varepsilon\,. \tag{3}$$

Acorde a la Ecuación 3 se tendrá un parámetro de error el cual se considera como ruido para aquellos valores que no coincidan con el modelo aplicado. A partir de este modelo general, se tienen variaciones conceptuales de la aplicación de regresiones, entre ellas se incluye la regresión lineal, donde el modelo es una combinación de variables x_i escalares para una salida "y". En base a este modelo, se incluyen coeficientes o parámetros β_i correspondientes a cada variable de ingreso (x_i), esto puede verse en la Ecuación 4 (Lindholm y otros, 2019),

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i + \varepsilon \,. \quad (4)$$

correspondiente a la ecuación brindada se puede plantear el uso del modelo para la aplicación en el caso de la predicción de emisiones de contaminantes y consumo de combustible. Por lo tanto, se puede utilizar de base las variables medidas por el sistema PEMS y OBD - II de los datos adquiridos por Mendoza, J., para relacionar los efectos en las variables de salida mencionadas. Asimismo, se puede aplicar una relación multivariable, dado que entre los tipos de regresión aplicada se tiene la oportunidad de establecer modelo de múltiples salidas (Ranjan, 2016), así como la aplicación de modelos de regresión logística (Nield, 2022). No obstante, parte de la teoría básica de la aplicación de los modelos de regresiones, es que se asume una tendencia lineal, en el caso de los modelos de emisiones y consumo de combustible en determinadas características del vehículo, esto puede diferir debido al impacto de las zonas de manejo, la presencia de tráfico o consideraciones de conducción (paradas, inicio, frenado) (Tarun, 2019). El comportamiento no lineal fue un error visto en el trabajo de Tarun. En base a ello, se evaluará la relación entre ambas variables, emisiones y consumo de combustible, para determinar el comportamiento del modelo. En caso de presentar linealidad, bastaría la implementación de modelos de regresión para predecir futuros escenarios de eficiencia energética y contaminación ambiental por parte de los vehículos livianos analizados.

En base a la aplicación de clasificación como técnica de aprendizaje supervisado, se tiene como objetivo la descripción de una relación entre variables, considerando los parámetros descritos en la Ecuación 4. Los métodos de clasificación buscan evaluar la significancia de los coeficientes β_i, donde se evalúa la correlación por medio de un proceso de razonamiento basado en hipótesis que será como fuente para la aplicación de los modelos de regresión lineal en la predicción con futuros datos (Lindholm y otros, 2019). Por otro lado, la clasificación puede partir de la regresión logística como estrategia de predicción para la representación de las variables deseadas. No obstante, para la naturaleza de los valores de emisiones, el recurso no puede ser considerado debido al uso de valores discretos o binarios (Nield, 2022).

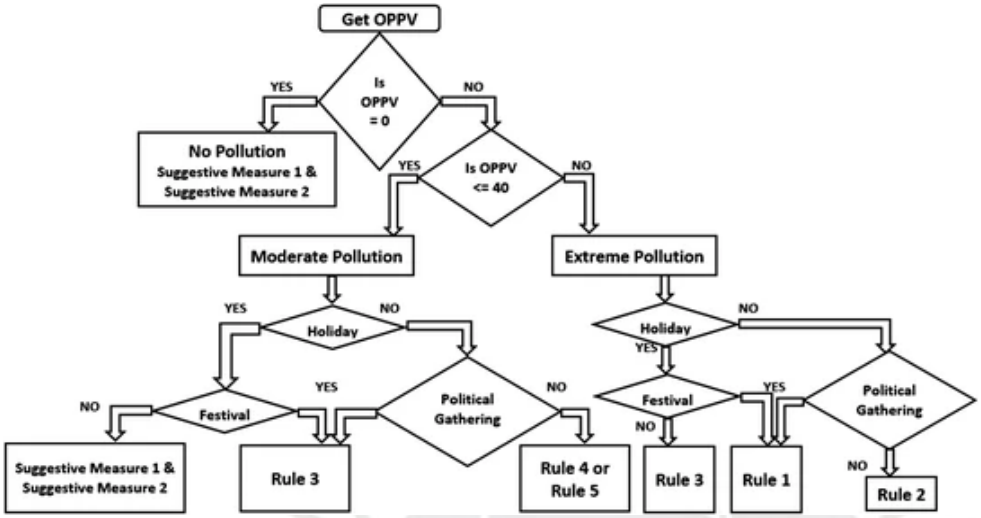
2.2.3.1.1 Bosque aleatorio

Los bosques aleatorios, a diferencia de otros métodos como los árboles de decisión, brindan una menor varianza y mejora sobre el *bagging* aplicado a los árboles de decisión. Mediante ello se reduce la correlación de cada uno de los árboles que se involucran en el proceso (Epifanto, 2021). Para entender los bosques aleatorios es necesario comprender el método de árboles de decisión. En donde, los datos de entrada o atributos a ser analizados son de tipo nominal o numérico para obtención de un modelo de clase objeto. A partir de estos datos, se presenta una división de dos conjuntos en un hiperplano visualizable en una interfaz gráfica, una ventaja que ofrece a diferencia de otros métodos de clasificación. Posteriormente a la segmentación, los modelos de árboles se basan en determinadas condiciones con lo cual cada tramo derivado o hijo tenga una asignación de pesos. Los nodos de los árboles se utilizan para evaluar dicha ponderación y continúan con la sucesión de tramos aplicando específicos algoritmos dependiendo de la aplicación del problema y tipos de datos (Flores, 2021). La implementación de árboles de decisión en el análisis de emisiones de contaminantes puede implementarse siguiendo un esquema de clasificación de regiones geográficas, fechas conmemorativas y contexto político para predecir futuras decisiones acorde a los pesos y segmentación de cada valor (Desarkar y Das, 2017).

En la Figura 5, se puede observar un esquema de clasificación en forma de árbol de decisión, en donde la variable de ingreso llamada OPPV ("*Overall Pollution Percent Violation*") que denota una medida de violación del porcentaje de contaminación general que es clasificada por medio de un estándar de valores establecidos por el autor, para categorizar si se presenta una contaminación moderada o extrema y atribuir contexto a cada valor según festividades o eventos políticos. En base a esta información, se aplican modelos de regresión y clasificación para la obtención de futuras predicciones (Desarkar y Das, 2017). En relación a la implementación de los modelos, se puede evidenciar el uso de distintas técnicas combinadas de aprendizaje supervisado para la predicción de valores tomando en cuenta múltiples estrategias; es por ello, que la selección de modelos depende de lo requerido por el contexto del problema (Kroese y otros, 2022).

Figura 5

Árbol de decisión para la predicción de niveles de contaminación.



Nota. Tomado de "Implementing Decision Tree in Air Pollution Reduction Framework. Smart Computing and Informatics" por Desarkar, A., & Das, A. (2018). Smart Innovation, Systems and Technologies, 77, 104.

La idea del *bagging* mencionado anteriormente busca de reducir el ruido de los árboles de decisión que se puede dar con modelos no bias y para reducir la varianza y captar una completa estructura de los datos de una manera más profunda. Para ello, se utiliza la siguiente secuencia de procesamiento para un problema de regresión (McGill University, 2022):

- Paso 1: Inicialización de datos
  Define el número de árbol con determinadas características del modelo.
- Paso 2: Aplicación de siguiente paso a cada árbol.
  - Muestreo *Bootstrap*: Consiste en la extracción de una muestra aleatorio de un tamaño N de los datos de entrenamiento. Esto se hace mediante el muestreo con reemplazo, lo que significa que algunas observaciones pueden aparecer más de una vez en la muestra, mientras que otras pueden no aparecer en absoluto.
  - Crecer el árbol: Crecer el árbol a los datos de muestreo haciendo que se repita de manera consecutiva el siguiente proceso de nodos por terminal de cada árbol, hasta

que el mínimo de nodos se haya obtenido (parámetro de entrada). El primer paso es seleccionar un número aleatorio de variables de entrada y dividir el nodo en dos nodos hijos.

- Ensamblar el árbol para que estos crezcan y sean un "bosque".

El modelo matemático para el caso de los problemas de regresión sigue la siguiente ecuación:

$$f_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x). \tag{5}$$

En la ecuación (5), el término "B" es el número de árboles seleccionados y "T" corresponde al árbol procesado. Los parámetros que se configuran para este algoritmo son los siguientes (Scikitlearn, 2024):

- Número de estimadores: equivale al número de árboles del bosque
- Máxima profundidad: evita el sobreajuste de los árboles.
- Mínimas muestras de nodo: indica el número mínimo de muestras requerido por árbol, para dividir un nodo.
- Máximas características: es el número para considerar como máximo en una óptima división para que haya diversidad en los árboles.

2.2.3.1.2 Redes neuronales

Las redes neuronales poseen capas internas descritas como "L+1" (dependiendo del tipo), un vector de entrada de "n" valores y una función de salida. Las capas internas están ocultas y cada una contiene un número de nodos específicos y asociados a otras variables (Kroese y otros, 2022). En base a esta moción se pueden implementar modelos de redes ya determinados que constan de variaciones en las capas, entrada de datos, algoritmos de entrenamiento y otros para la aplicación en determinados contextos. En la sección 2.2.4 se aborda a profundidad la aplicación de redes neurales y el esquema de estructura de implementación.

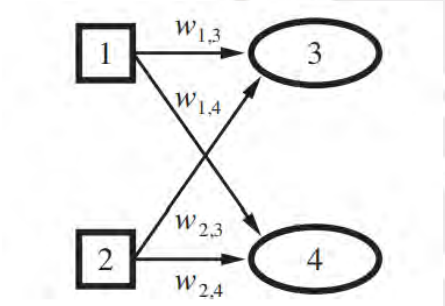
La aplicación de redes neuronales fuera de las técnicas de regresión y clasificación, permiten su implementación a sistemas con comportamientos no lineales. Es por ello, que son una estrategia para el uso de análisis multivariable tanto de entradas como de salidas, para la predicción de futuros resultados (Nield, 2022).

Existen distintas técnicas de implementación que dependen de los modelos aplicados en base a las capas de cada red. Entre los tipos más representativos de redes se encuentran:

- Perceptrón: es una red de una sola capa, donde las entradas están conectadas directamente a las salidas y sigue la moción de que "m" salidas implican "m" redes separadas, como cada peso afecta únicamente a una de las salidas. En la Figura 6, se visualiza la estructura de este tipo de redes con cada peso w_i correspondiente (Russel y Norvig, 2010).

Figura 6

Perceptrón de dos entradas y dos salidas.



Nota. Tomado de "Data Science and Machine Learning: Mathematical and Statistical Method" por Kroese, D., Botev, Z., Taimre, I., & Vaisman, R. (2022). Boca Ratón: Chapman and Hall/CRC.

- "*Feed-forward neural network*" o red neuronal de avance: es una red que tienen una cantidad de capas equivalente a "L + 1", un vector de entrada y salida. Cada capa intermedia contiene un número de nodos determinado (Kroese y otros, 2022). Un ejemplo de este tipo de red es el observado en la Figura 4.
- Perceptrón multicapa: en el caso de un perceptrón, se tienen limitaciones en cuanto a los patrones que pueden generarse, mientras que un perceptrón multicapa contiene más capas ocultas intermedias y es entrenado por medio de un algoritmo de retropropagación (back

propagation) (Larraga y otros, 2007).

- Red neuronal convolucional: una red neuronal convolucional involucra la aplicación del concepto de convolución de dos señales continuas unidimensionales, utilizando sistemas invariantes en el tiempo, en respuesta de una función de entrada. A diferencia de otras redes, se extraen filtros que contienen características de las entradas y su estructura es compleja ya que utiliza múltiples capas entre sí, hasta conseguir una clasificación de la información provista (Bonilla, 2020).

Las técnicas existentes son aplicadas a partir de distintos enfoques metodológicos. No obstante, para la implementación de redes neuronales se requiere de la toma de decisiones en la etapa de diseño a partir de consideraciones técnicas. Es por ello, que como buena práctica se sugiere considerar las siguientes etapas (Bodnovich, 2000):

- Identificación de problema para selección de una red: evaluar la necesidad de implementación de una red al problema planteado y según la disponibilidad de datos.
- Identificación de parámetros de entrada relevantes: incorporar técnicas de regresión para evaluar la importancia de los parámetros, utilizando algoritmos de preprocesamiento o de regresión no paramétrica.
- Recolección y limpieza de datos: implementar técnicas estadísticas para reducir el número de valores atípicos y errores en la base de datos disponible.
- División de datos en datos de entrenamiento, prueba y validación: Los datos de una red se dividen en tres categorías dado que los pesos de la red serán ajustados en base al aprendizaje de la relación de los datos. Es por ello, que se debe evaluar la eficacia de la red con valores no antes entrenados, es decir, datos nuevos.
- Selección de un paradigma de red: el paradigma hace referencia a la estructura de red seleccionada, a la función de transferencia y la regla de aprendizaje. La estructura depende de la selección de capas en base a los tipos de red descritos, la función de transferencia se relaciona con la salida de datos y la regla de aprendizaje depende de cómo los pesos de cada arista son actualizados conforme evoluciona el aprendizaje (Bodnovich, 2000).

En base a lo mencionado anteriormente, se debe considerar la función de activación, con

la cual se dará la probabilidad de ejecutar o no un determinado nodo, considerando su posibilidad de ocurrencia en la red. Las funciones para ello son las siguientes (Permuter, 2022):

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{6}$$

$$\sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}. \tag{7}$$

$$\sigma(z) = \frac{z}{|z|} \tag{8}$$

$$\sigma(z) = max(0, z), \tag{9}$$

donde "z" es la entrada a cada nodo de la red.

- Diseño de la red: determinación de capas y nodos ocultos de la red.
- Selección del espacio de implementación: entorno de programación en donde se desarrollará el programa.
- Entrenamiento de la red: etapa de implementación de la red con los datos separados de entrenamiento. Es una etapa iterativa para la evaluación de los parámetros de configuración y las funciones que regirán la red.
- Prueba y validación de la red: validación con los grupos de datos restantes para ver su eficacia en datos no entrenados. Esta etapa y la anterior se realizan hasta conseguir un desempeño apropiado de la red.
- Selección de un modelo óptimo: selección final del modelo y realización de mejoras o arreglos del programa.

En base a los procedimientos y tipos de red mencionados, recopilando como base los conceptos de aprendizaje supervisado para específicas etapas del aprendizaje no supervisado, se plantea realizar la implementación de una red neuronal para el presente trabajo de investigación abordando cada punto especificado.

2.3 Definición de términos básicos

En esta sección se plantea la definición de los términos básicos del estudio:

- Arista: hace referencia a la línea de intersección entre dos planos. No obstante, en este trabajo parte del concepto de un grafo, donde una arista es la línea de conexión entre dos nodos. Esta conexión tiene como característica un peso que corresponden a la probabilidad de que suceda o no el evento o nodo al cual está conectado (Deo, 2016).
- Categoría vehicular: código de asignación para la estandarización de registro a los vehículos automotores según especificaciones técnicas y características de los mismos, como: número de ruedas, cilindrada del motor, velocidad, peso, si es utilizado como transporte de carga, capacidad de acoplamiento y más.
- *"Children"* o hijo: hace referencia a un nodo que es descendiente de otro en una estructura de árbol, es decir en una estructura jerárquica. Asimismo, se refiere a un evento derivado de otro para que ese mismo ocurra (Leskovec y otros, 2014).
- CO: el monóxido de carbono es un gas dañino en grandes cantidades y es incoloro e inodoro. Se libera a partir de la combustión. En este estudio nos enfocaremos al CO emitido por la quema de combustible, en donde su producción se debe a una combustión incompleta. Este gas produce un efecto peligroso al sistema respiratorio humano ya que puede adherirse a los glóbulos rojos e impedir la absorción de oxígeno de manera temporal (Fonseca, 2012).
- CO_2: el dióxido de carbono es un compuesto formado por un átomo de carbono unido a cada átomo de oxígeno a partir de un doble enlace. Este es producido por la respiración de múltiples seres vivos como: animales, hongos y otros microorganismos que dependen de las plantas. Los procesos de combustión para la generación de energía son contribuyentes a la formación de CO_2 (Fonseca, 2012).
- Entrenamiento: será referido al proceso por el cual una red neuronal iniciará su proceso de aprendizaje en base a unos datos de entrada iniciales y orientados a ese propósito. El objetivo de la etapa es logra llegar a parámetros altos de la red para poder probarla con otro tipo de datos nuevos, diferentes a los entrenados, para ver la eficiencia del modelo de red obtenido (Leskovec y otros, 2014).

- Función de costo: hará referencia netamente a las funciones de costo en una red neuronal, por lo cual es importante no confundir con un parámetro económico de impacto energético o social. En el estudio, se utilizará para referir a la función que se implementará con el fin de evaluar el error entre los valores estimados y reales de salida. Esto para ver técnicas de optimización que permitan la reducción del error (Deo, 2016).
- Función de transferencia: la función de transferencia o de activación, a diferencia del concepto utilizado en instrumentación industrial, se refiere al cálculo de la probabilidad de suceso de preceder a un nodo para el aprendizaje de la red, es decir al evento. De esta manera, se validan los sucesos y su consecuencia cuando se entrena el modelo (Nield, 2022).
- Factor de Emisión: es la relación entre la emisión del gas contaminante con relación al consumo de energía requerido para ello. En este estudio será la cantidad de gramos del contaminante y se utilizará como cantidad de energía el combustible consumido (IPCC, 2001).
- HC: los hidrocarburos son compuestos orgánicos formados por carbono e hidrógeno y se encuentran de fuentes como el petróleo, como en el estado condensado, líquido o gaseoso de gas natural (Fonseca, 2012).
- Nodos: son componentes de los grafos o de una red que se utilizan para la representación de objetos que pueden contener eventos determinados. Estos están conectados mediante aristas entre sí. Acorde al tipo de grafo o estructura de red puede tener múltiples conexiones o solo una (Leskovec y otros, 2014).
- NO_x: los óxidos de nitrógeno son grupos de gases combinados por óxido nítrico y dióxido de nitrógeno que se forman durante la combustión de vehículos y son emitidos por los tubos de escape de los mismos. Estos gases son tóxicos y dañinos, especialmente para el sistema respiratorio (Fonseca, 2012).
- *"Parent"* o padre: en términos de programación, un nodo padre es donde se encuentra el puntero al cual y del cual descienden otros nodos que se llaman hijos. También puede decirse en términos de eventos que es un evento inicial que debe suceder para ser predecesor de otros (Leskovec y otros, 2014).
- Red neuronal: a lo largo de trabajo es referido a una red artificial, siguiendo los principios de ciencia de la computación, más no a una red neuronal biológica. Una red neuronal es un

conjunto de nodos conectados mediante aristas, la estructura se basa en la transmisión de señales para el proceso de aprendizaje. El sistema aprende de manera no supervisada y considerando ponderaciones para que los datos de entrada viajen por medio de la red para la obtención de datos coherentes de salida (Nield, 2022).

2.4 Hipótesis y descripción de variables

2.4.1 Hipótesis general

El consumo de combustible en determinados escenarios del estado del vehículo en ruta influye directamente en la emisión de gases contaminantes, presentando un comportamiento no lineal entre ambas variables.

2.4.2 Hipótesis específicas

Las hipótesis específicas formuladas para el trabajo de investigación son las siguientes:

- La velocidad, aceleración y condiciones de las vías de manejo del vehículo son determinantes en las variaciones del consumo de combustible de los vehículos livianos conducidos en Lima Metropolitana.
- Las condiciones del motor, las variables atmosféricas y las condiciones de las vías de manejo son determinantes en la emisión de CO_2, CO, HC y NO_x de los vehículos conducidos en Lima Metropolitana.
- Las emisiones de CO_2, CO, HC y NO_x con respecto al consumo de gasolina presentan un comportamiento no lineal debido a la cantidad de variables que influyen en su estimación.

2.4.3 Descripción de variables

Las variables evaluadas en este trabajo de investigación son las variables dependientes de gases emitidos y el consumo de combustible de vehículos. Ambas variables son analógicas y cuantitativas nominales y serán analizadas en base a cada vehículo evaluado de los datos de

Mendoza, J.. Asimismo, el entrenamiento del modelo de red neuronal requerirá del análisis de variables independientes de entrada que serán: la humedad, temperatura del ambiente, la posición geográfica, velocidad, aceleración, condiciones del motor durante la conducción, temperatura, torque, RPM, tipo de motor, distancia recorrida, tiempo de marcha, presión de los fluidos del motor.

CAPÍTULO III - METODOLOGÍA

3.1 Método, enfoque y alcance de la investigación

El desarrollo del presente trabajo de investigación se centrará en la evaluación de un modelo que relacione las variables determinantes en la emisión de gases contaminantes de vehículos livianos del parque automotor de Lima. Para ello, se plantea el uso de datos obtenidos por un diseño de investigación no experimental con datos longitudinales. Los datos fueron directamente provistos por la investigación de Mendoza, J., en donde se realizan distintos ensayos de medición de múltiples vehículos de Lima Metropolitana en el 2022. El alcance de la presente investigación se limita al número de vehículos investigados por Mendoza, J. y a las zonas a las cuales fueron transitados en la captura de datos. En base a ello, se tiene una cantidad de 29 muestras de 15 vehículos de marcas representativas del parque automotor; sin embargo, la diversidad de los autos en la ciudad y el año de fabricación de los mismos limitan a que el modelo pueda ser confiable en su totalidad para el total de autos de la ciudad.

3.2 Diseño de la investigación

El modelo planteado para el proyecto actual, se centra en el desarrollo de una comparativa del análisis de variables mediante métodos de aprendizaje supervisado y no supervisado para la obtención de modelos que correlacionen el efecto de distintas variables de un vehículo con los gases contaminantes que se emiten. A partir de ello, el método de aprendizaje seleccionado será redes neuronales, la viabilidad del método seleccionado e importancia en comparación con métodos de aprendizaje supervisado se determinará en base a su porcentaje de confiabilidad. De esta manera se plantea un modelo adecuado que estime el comportamiento de contaminantes en el parque automotor de Lima Metropolitana.

La metodología a aplicar se puede visualizar en la Figura 7, en donde la primera etapa del proceso centrada en la recolección de datos, fuera de la captura de los mismos que fue realizada por Mendoza, J., se refiere a la recolección de las bases de datos de los distintos autos y la clasificación de los ensayos de los vehículos para el posterior pre procesamiento. La recolección constituye una

etapa de entendimiento del origen de cada variable para visibilizar su importancia en el trabajo de investigación. Por otro lado, la etapa de pre procesamiento, se centra en la adecuación de las variables y la limpieza de los datos para la adecuada implementación de los modelos. La tercera etapa, centrada en el análisis de variables, se refiere a un análisis estadístico descriptivo de los datos pre procesados para tener un panorama general del comportamiento entre cada variable con las emisiones de contaminantes. En base a ello, se plantea la selección se parámetros que se correlacionan con las emisiones y se descartan aquellos que no. La cuarta etapa es la evaluación de modelos para la aplicación del aprendizaje supervisado y no supervisado, con lo cual se determinan los métodos a utilizar y los componentes a considerar para el diseño de la arquitectura de las redes neuronales. El diseño de la arquitectura es esencial para realizar el entrenamiento, así como el ajuste constante de parámetros que se utilizarán en la red neuronal para que esta sea óptima y tenga una tasa de confiabilidad elevada. Posterior a la obtención de una red adecuada se procedió a comparar con el método de aprendizaje supervisado para evaluar el grado rendimiento entre ambos. La última etapa es el despliegue del modelo de red obtenido para predecir valores con todos los datos que se tienen y obtener resultados de las predicciones para el caso de Lima Metropolitana.

Figura 7

Metodología a desarrollar.

Nota. Diseño propio

3.3 Población y muestra

Las muestras obtenidas por cada vehículo corresponden a un trabajo de investigación para el cálculo de factores de emisión de vehículos livianos (Mendoza, 2022). En base a los estos se procedió a realizar un estudio de los datos para determinar la muestra más correcta que entrará en el análisis del modelo de entrenamiento a utilizar. La población objeto de este estudio consiste en los automóviles más comunes en la ciudad de Lima, Perú, que han transitado por zonas altamente recurridas.

Dada la diversidad de vehículos presentes en la capital peruana, se optó por centrar la atención en aquellos automóviles que son más representativos y frecuentes en el entorno urbano. Según la investigación de Mendoza, son 6 las características principales a tomar en consideración, correspondientes a 786 mil vehículos de 950 mil que corresponden a los autos nuevos en Lima registrados del 2010 al 2017 por la Asociación Automotriz del Perú (AAP). Estas 6 características son:

- Categoría: el 87% de los vehículos registrados en Lima corresponden a la categoría M1 que refiere a vehículos para el transporte de personas (MTC, 2023).
- Forma de propulsión: en el país, el 83% de vehículos utiliza gasolina y el 17% diesel a pesar de las entradas de nuevas tecnologías como los vehículos eléctricos o híbridos.
- Carrocería: existen múltiples tipos de carrocería y las consideradas con mayor frecuencia en el parque automotor son los Sedán (40%), SUV (17%) y Hatchback (18%), siendo estos el 75% de representación del parque automotor del país (MTC, 2023).
- Tracción: la tracción de los vehículos livianos es de dos tipos, de 2 ruedas (2WD) o 4 ruedas (4WD), esto dependiendo del modelo del vehículo.
- Transmisión: el tipo de transmisión con mayor representación en el mercado corresponde a una frecuencia del 77% en comparación con los otros tipos de transmisión como automática o secuencial.
- Cilindrada: para el tipo de carrocería de vehículos, según la AAP, corresponde a unos rangos típicos de 1.10 a 1,80 litros para el caso de vehículos sedán, entre 1.20 a 2 para los hatchbacks y entre 1.48 a 2.7 para los SUV.

En base a la información de las características de los vehículos livianos en Lima, se concluyó que la muestra de datos tomada por Mendoza representa el parque automotor actual de la ciudad. La relación en resumen del tipo de vehículos se puede ver en la Tabla 4 que contiene los datos diferentes vehículos considerados para el trabajo de tesis, donde los datos fueron tomados en agosto y setiembre del 2022. Asimismo, las muestras de datos se captaron considerando calles y avenidas con fluidez y aquellas de mayor tránsito para captar información de características típicas de conducción en Lima, con un muestreo aleatorio. En este sentido, la muestra se respalda en las rutas con mayor afluencia de personas basado en el plan maestro de transporte urbano (Yachiyo Engineering Co, 2005) del 2005 al 2025, siendo de esta manera las muestras representativas para el análisis (Anexo 2).

Tabla 4

Resumen de vehículos analizados.

Marca	Modelo	Carrocería	Cilindrada	Fórmula Rodante
Chevrolet	Sail	Hatchback	1.5L	2WD
Honda	Pilot	SUV	3.5L	4WD
Hyundai	Accent	Sedan	1.6L	2WD
Hyundai	Elantra	Sedan	2.0L	2WD
Hyundai	Santa Fe	SUV	2.4L	4WD
Kia	Cerato	Sedan	2.0L	2WD
Kia	Rio	Sedan	1.6L	2WD
Mazda	CX-5	SUV	2.5L	4WD
Mazda	Mazda 3	Hatchback	2.0L	2WD
Nissan	Kicks	SUV	1.6L	2WD
Toyota	Agya	Hatchback	1.0L	2WD
Toyota	RAV4	SUV	2.5L	4WD
Toyota	Yaris	Hatchback	1.5L	2WD
Volkswagen	Golf	Hatchback	1.4L	2WD
Volkswagen	Tiguan	SUV	2.0L	4WD

Nota. Basado en el trabajo de Mendoza, J. (2023). *Propuesta metodológica para el cálculo de factores de emisión para vehículos livianos a gasolina circulando en Lima Metropolitana*.

Posterior al análisis de la población y muestra de Mendoza, J., se procedió a realizar un contraste con el informe estadístico automotor de Lima Metropolitana realizado por la Asociación Automotriz del Perú en los meses que se llevaron a cabo las pruebas (Mendoza, 2022). La inserción

de vehículos livianos en el mes de agosto significó un incremento aproximado del 20% en la marca Toyota, 10% Hyundai, 8.7% Kia, 7.4% Chevrolet, 4% Nissan y 3.7% Susuki. En base a estos datos, los vehículos mencionados corresponden al top 10 de Lima. A partir de ello, se reconoce la segmentación de la población de vehículos en base a la marca y la carrocería. Otras características resaltantes del estudio de Mendoza, J. en cuanto a los datos provistos fueron las fuentes de donde se obtuvieron, siendo estas el OBD y el sistema para medición de los contamiantes del gas de escape. Las variables capturadas fueron las siguientes:

Tabla 5

Resumen de variables medidas por sistema.

Sistema	Símbolo	Variables Medidas	Rango de medida y precisión	Instrumento/Equipo de medición
Sistema del analizador	LAT	Latitud (º)	$\pm 3m$	BT-Q1000XT
	LON	Longitud (º)	$\pm 3m$	BT-Q1000XT
	H	Altura sobre el nivel del mar (m.s.n.m)	$\pm 1\%$	Sensor de presión atmosférica 8121
	c_{HC}	Emisiones volumétricas de Hidrocarburo no quemados (ppm)	0 – 9999; $\pm 1\%$	Analizador FGA 4500
	c_{CO}	Emisiones volumétricas de CO (%)	0 – 10; $\pm 1\%$	Analizador FGA 4500
	c_{CO_2}	Emisiones volumétricas de CO_2 (%)	0 – 20; $\pm 1\%$	Analizador FGA 4500
	c_{O_2}	Emisiones volumétricas de O_2 (%)	0 – 25; $\pm 1\%$	Analizador FGA 4500
	Hum	Humedad relativa del aire exterior (%)	0 – 100; ± 2	Sensor HX94
	T_{ext}	Temperatura exterior (°C)	0 – 100; ± 0.6	Sensor HX94
	λ	Factor Lambda medido por el analizador	0 – 5; $\pm 1\%$	Analizador FGA 4500
Sistema del OBD	$R_{A/C}$	Relación aire combustible medido por el analizador	0 – 100; ± 1	OBDII-ELM327
	v	Velocidad del vehículo (km/h)	0 – 250; ± 1	OBDII-ELM327
	RPM	Revoluciones por minuto del	0 – 8000; $\pm 2\%$	OBDII-ELM327

		cigüeñal del motor (RPM)		
	A_g	Consumo de aire calculado del motor (g/s)	0 – 100; ±5%	OBDII-ELM327
	a	Aceleración (m/s^2)	-10 – 10; ±0.1	OBDII-ELM327
	τ	Torque del motor (N.m)	0 – 500; ±5	OBDII-ELM327
	$C_{combustible}$	Consumo volumétrico de combustible (l/h)	0 – 50; ±5%	OBDII-ELM327
	P_{ad}	Presión del aire en el múltiple de admisión del motor (hPa)	500 – 1100; ±1	OBDII-ELM327
	T_{ad}	Temperatura del aire en el múltiple de admisión del motor (°C)	0 – 150; ±1	OBDII-ELM327
	T_{ext}	Temperatura exterior (°C)	0 – 150; ±1	OBDII-ELM327

Nota. Basado en el trabajo de Mendoza, J. (2022). *Propuesta metodológica para el cálculo de factores de emisión para vehículos livianos a gasolina circulando en Lima Metropolitana*.

En el caso del OBD y los datos del sistema de medición por tubo de escape tienen tiempo distantes de captura de los datos, debido a que son sistemas independientes. Por tal razón, se tuvieron consideraciones adicionales en el proceso de pre procesamiento de los datos para la selección del método de muestreo con el cual se implementarían los modelos de aprendizaje. Posteriormente, se procedió con la etapa de preprocesamiento de datos para la inclusión de los métodos de muestreo necesarios en el modelo de aprendizaje no supervisado y supervisado.

3.3.1 Preprocesamiento de datos

La información recopilada de las muestras pasó por una etapa de preprocesamiento de datos utilizando técnicas comunes para su limpieza. Entre los métodos empleados se encuentran la detección y eliminación de valores atípicos, así como la imputación de datos faltantes. Estos procedimientos garantizaron la calidad y confiabilidad de la información utilizada en el análisis desarrollado. Los métodos empleados, en orden, fueron:

1. Detección y eliminación de duplicados:

Primero, se identificaron y eliminaron las entradas duplicadas en el conjunto de datos, considerando como datos duplicados aquella información que tenía una misma hora de registro.

2. Manejo de valores nulos:

Posteriormente, se abordaron los valores nulos o faltantes. Estos fueron tratados mediante diversas técnicas de imputación, como la basada en la mediana entre el valor previo y el posterior.

3. Detección y eliminación de valores atípicos:

Luego, se procedió a la identificación y eliminación de valores atípicos, que son datos que se desvían significativamente del resto del conjunto. La presencia de valores atípicos puede afectar la performance de los modelos y llevar a interpretaciones incorrectas. Es por ello, que se utilizó el análisis de la desviación estándar y el rango intercuartílico (IQR) para filtrar los datos.

4. Normalización de datos:

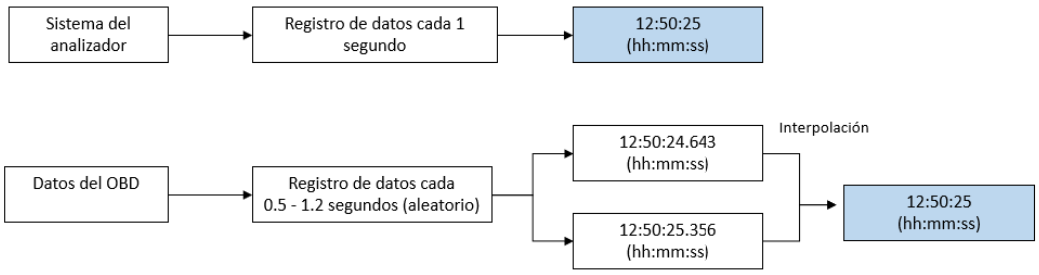
Finalmente, los datos fueron normalizados. Este proceso ajusta los valores a una escala común, mejorando la eficiencia y precisión de los algoritmos de aprendizaje automático. Se utilizó la técnica de normalización min-max para capturar los rangos diferentes de cada variable y ponerlas en un rango específico de 0 a 1.

El preprocesamiento de datos se realizó mediante el uso de la librería de pandas de Python para la filtración de valores vacíos y la eliminación de valores atípicos que se evidenciaron en ciertos vehículos, debido principalmente a los valores de longitud y latitud que en ciertos vehículos tenían el valor de cero. Este hallazgo significó que la herramienta de medición no estaba capturando las variables deseadas debido a posible ruido en el sistema empleado. Una vez realizado el filtro de valores atípicos se procedió a contrastar los datos entre los dos sistemas de medición. Hubo una diferencia en el reloj de registro de datos de cada uno; el sistema de medición de gases (analizador) realizaba un registro cada segundo; mientras que el OBD tenía un registro automático de un

segundo más distintos milisegundos como se muestra en la Figura 8. La variación del registro de datos del OBD se uniformizó con los datos medidos por el analizador, mediante la interpolación de valores para las horas de registro del sistema de medición. De esta manera, los datos del sistema de medición con los del OBD cada segundo fueron los que se utilizaron para los modelos de entrenamiento.

Figura 8

Esquema de interpolación de datos para contrastar OBD con Sistema de Medición.



Nota. Diseño propio

En base al análisis de los datos, debido a la naturaleza de la captura de los mismos, hay presencia de un error inerte de los propios sensores, debido a ello, el filtrado y limpieza de datos fue una parte fundamental para la obtención de los resultados del presente trabajo de investigación.

3.4 Técnicas e instrumentos de análisis de datos

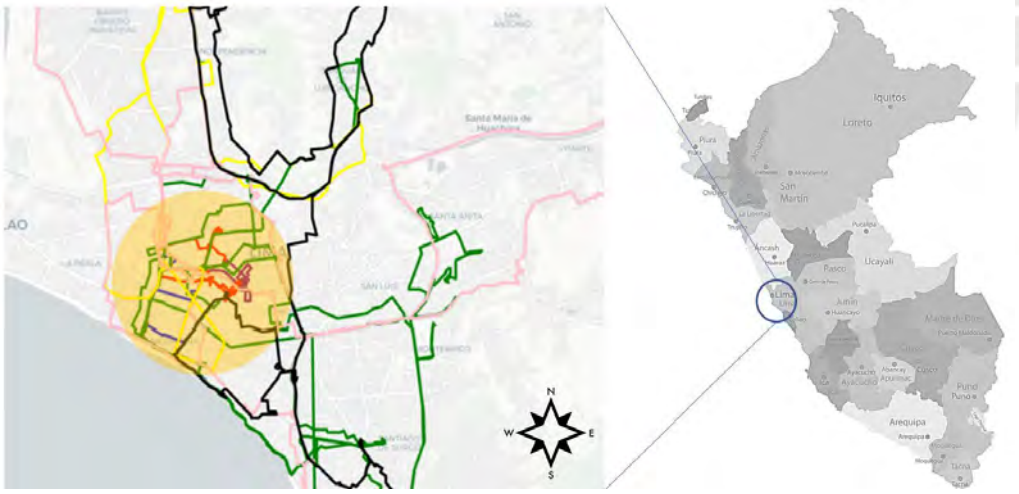
Las técnicas de análisis utilizadas para evaluar el consumo de combustible en determinados puntos de la ciudad se centran en dos modelos de aprendizaje supervisados basados en árboles de decisión y redes neuronales para evaluar el contraste de resultados. En base a ellos, el análisis de las variables previo al entrenamiento de los modelos es clave para comprender los parámetros a utilizar para la estructura de la red.

3.4.1 Análisis de variables

El análisis de los datos se basó en la evaluación de las variables con relación a su trayectoria en cuanto a las rutas recorridas por el conductor del 2022. Conocidas las posiciones de latitud y longitud de cada vehículo muestreado se tuvo la gráfica de la Figura 9, en donde se visualiza las rutas recorridas por tramos en el mes de agosto y setiembre. Entre las rutas que se destacan en esta imagen, 5 de ellas coinciden con las mayores rutas recorridas en la ciudad de Lima. Estas son la Av. Javier Prado, Av. Panamericana Norte, Av. Panamericana Sur, Av. Faucett y Circuito de Playas. Además, se puede ver que los tramos con mayor repetición por los vehículos analizados se centran en la zona circular marcada de color amarillo, que corresponde al centro de Lima. Esta zona es una de la más recurrentes de la ciudad por concentrar comercios, hospitales, colegios, institutos y universidades.

Figura 9

Mapa de distribución de rutas de cada vehículo.



Asimismo, el análisis preliminar de relación de las variables de los datos examinados se realizó mediante una matriz de correlación para ver el peso de las variables y su relación entre sí. El modelo utilizado se llama “Feature Selection”. A partir de la matriz de correlación obtenida (véase Anexo 1) se tuvo como resultado la Tabla 6, que presenta la relación con mayor porcentaje de correlación positiva y/o negativa entre variables. Los pesos indicados corresponden a las variables

con mayor relación entre sí, donde se evidencia que las emisiones tienen una relación directa con la temperatura de admisión T_{ad} y la humedad del mismo, lo cual influye en el proceso combustión. Asimismo, considerando las variables con mayor correlación para cada contaminante, se evidenció que la aceleración y velocidad de los vehículos examinados tuvieron valores altos en el caso del HC y CO_2, por lo cual y en efecto las revoluciones por minuto del cigüeñal del motor también fueron considerado dada su correlación de -0.21 y 0.26.

Tabla 6

Variables con mayor correlación al HC, CO, CO_2 y NO_x

Contaminantes	Variables con mayor peso
Hidrocarburos (HC)	Temperatura de exterior: 0.37 Humedad relativa del aire exterior: -0.32 Temperatura del aire en el múltiple de admisión del motor: 0.29 RPM: -0.21
CO	Velocidad del vehículo: 0.35 Temperatura del aire en el múltiple de admisión del motor: 0.32 Temperatura exterior: 0.3 Humedad relativa del aire exterior: -0.25
CO_2	Relación aire combustible medido del analizador: -0.92 Consumo de aire calculado: 0.7 Presión del aire en el múltiple de admisión del motor: 0.51 Velocidad del vehículo: 0.40 Temperatura exterior: 0.29 RPM: 0.26
NO_x	Temperatura exterior: 0.37 Humedad relativa del aire exterior: -0.34 Temperatura del aire en el múltiple de admisión del motor: 0.21

El análisis de las variables permitió esquematizar los principales datos a evaluar para el entrenamiento de los modelos de predicción. En base a los resultados, se procedió a realizar la estructura de los modelos y sus parámetros de selección para la predicción.

3.4.2 Métodos de aprendizaje

3.4.2.1 Random Forest

Los métodos de aprendizaje utilizados para la elaboración del presente trabajo de investigación se basaron en los utilizados en los trabajos de Yuvaraj N. y otros (2023) al igual que en el trabajo Hai Y (2022). En donde ambos realizan algoritmos de estimación basados en *Machine Learning* para la predicción de emisiones en vehículos livianos. La diferencia principal es que uno utiliza modelos basados en *Random Forest* y el otro un esquema de red neuronal. En el presente trabajo, la metodología utilizada para el modelo de *Random Forest* se centró en el diagrama de la Figura 10. En la metodología descrita se inicia, posterior al procesamiento de los datos, con la división de estos en datos de prueba y de entrenamiento mediante la técnica de muestreo *bagging*. El muestreo realizó la subcategorización de datos en múltiples grupos de manera aleatoria para tener datos representativos en ambos grupos (prueba y entrenamiento). Posterior a ello, los datos de entrenamiento se utilizaron para construir el modelo del Random Forest, que optimizado en múltiples iteraciones. El modelo es validado mediante el error cuadrático medio (MSE) siguiendo lo siguiente:

$$MSE(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \quad (10)$$

donde:

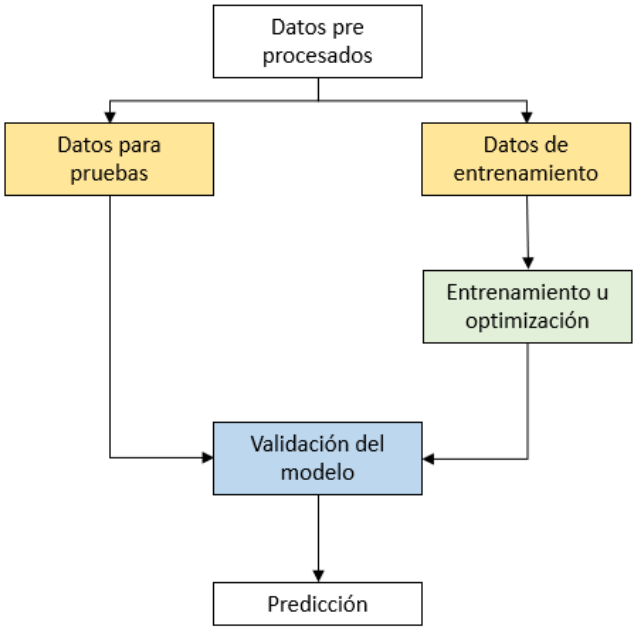
- N es el número de muestras en el conjunto de datos de prueba;
- x_i son los valores reales de la variable objetivo para la muestra i;
- y_i son las predicciones del modelo para la muestra i.

El valor MSE se seleccionó considerando su aplicación para problemas de regresión y mediante el cual se pudo determinar la eficacia de los parámetros selectos. Asimismo, se siguió lo

recomendado en los trabajos de Das, K. y otros (2004).

Figura 10

Esquema de modelo utilizado para Random Forest.



Es importante resaltar que se implementaron modelos separados de entrenamiento con el objetivo de validar el rendimiento del modelo de *Machine Learning* para cada contaminante. Mediante el uso de la librería Optuna se logró obtener las mejores configuraciones para cada modelo, mediante iteraciones de entrenamiento. Los parámetros optimizados fueron los siguientes (Koehrsen, 2018):

- Estimadores: este parámetro representa el número de árboles en el bosque aleatorio (Random Forest). Un mayor número de estimadores generalmente mejora la capacidad de generalización del modelo al reducir el sobreajuste. Sin embargo, aumentar demasiado este valor puede llevar a un mayor costo computacional.
- Máxima profundidad: indica la profundidad máxima de cada árbol en el bosque. Una mayor profundidad permite que los árboles se ajusten más a los datos de entrenamiento, pero también puede aumentar el riesgo de sobreajuste.

- Mínimo número de muestras: es el número mínimo de muestras requerido para dividir un nodo interno en un árbol. Este parámetro controla la cantidad de muestras necesarias para que se realice una división en el árbol. Un valor más alto puede ayudar a evitar divisiones que conduzcan a un sobreajuste.
- Mínimo número de muestras por nodo: representa el número mínimo de muestras requeridas en un nodo hoja. Este parámetro influye en la cantidad de muestras necesarias para formar una hoja en el árbol. Valores más altos pueden ayudar a prevenir divisiones que generen nodos hoja con muy pocas muestras, lo que podría conducir a un sobreajuste.

3.4.2.2 Red Neuronal

El presente trabajo de investigación contemplo la estructura de 5 redes neuronales basadas en la predicción de las emisiones y el consumo de combustible. Se decidió dividir las redes neuronales para cada contaminante debido a la particularidad encontrada en los datos cuando estos fueron pre procesados, a causa que habían determinados escenarios en donde la medición de un contaminante era mejor que en otros periodos de tiempo. A partir de esta consideración, se procedió a optimizar los hiperparámetros de la red mediante el uso de Optuna. Como entrada de datos para el inicio del proceso se tomaron de referencia parámetros de los modelos de N. Abdulakreem y A. Abdulazeez (2021). Se realizaron distintas iteraciones para ver que parámetros se ajustaban mejor a los datos, siguiendo el esquema presentado en la Tabla 7, en donde cada iteración constó de 100 pruebas hasta limitar el rango de parámetros procesados.

Tabla 7

Rango de parámetros para optimización de modelos de emisiones

Iteraciones	Iteración 1	Iteración 2	Iteración 3
Número de capas ocultas	1 - 20	1 - 10	1 - 6
Número de neuronas por cada capa	2 - 200	2 – 100	2 – 50
Función de activación	Rectificación lineal (ReLU), lineal, sigmoid, tangente hiperbólica, ReLU paramétrica y Exponencial lineal (ELU)	Rectificación lineal (ReLU), lineal, ReLU paramétrica y Exponencial lineal (ELU)	Rectificación lineal (ReLU) y lineal
Optimizador	Adam, SGD y RMSprop	Adam y RMSprop	Adam
Tasa de aprendizaje	0.0000001 a 1	0.00001 a 1	0.0001 a 1
Parámetros de regularización	L1: 0.0001 a 0.01 L2: 0.0001 a 0.01	L1: 0.0001 a 0.01 L2: 0.0001 a 0.01	L1: 0.0001 a 0.01 L2: 0.0001 a 0.01
Épocas	0-20	0-20	0-20

A partir de los resultados, se decidió utilizar la Iteración 3, debido a la acotación de rango de hiperparámetros a ser seleccionados, considerando la salida de resultado de las previas iteraciones. Los parámetros mejorados incluyen:

- El número de capas ocultas se seleccionó de forma dinámica entre 1 y 6 capas, permitiendo una mayor flexibilidad en la arquitectura de la red.
- El número de neuronas por capa se ajustó entre 2 y 50 neuronas, influyendo en la capacidad de representación y aprendizaje de la red.
- La función de activación se eligió entre “ReLu” y “lineal”, determinando cómo se propagan

las señales a través de las capas.

- El optimizador se fijó en “Adam”, una opción comúnmente utilizada por su eficacia en el entrenamiento de redes neuronales.
- La tasa de aprendizaje se ajustó de manera logarítmica, entre 1e-4 y 1, optimizando la velocidad y la estabilidad del entrenamiento.
- Los parámetros de regularización, como el coeficiente L1 y el coeficiente L2 se optimizaron entre rangos específicos, utilizando una escala logarítmica para controlar el sobreajuste y mejorar la generalización del modelo.
- Épocas es el número de iteraciones que dará el algoritmo al conjunto de datos de entrada en el entrenamiento.

CAPÍTULO IV - RESULTADOS Y DISCUSIÓN

4.1 Resultados y discusión

Los resultados del presente trabajo de investigación tuvieron como técnica de evaluación de la precisión de la predicción del modelo el valor del error cuadrático medio (RMSE) y el error absoluto medio (MAE). En el caso del RMSE, este es utilizado para tomar en consideración la penalización del error en la predicción debido, especialmente a la variabilidad de los datos. Por otro lado, el MAE será utilizado para brindar robustez al modelo ante los valores atípicos.

4.1.1 Modelo de Random Forest

En base a la metodología aplicada en el apartado 3.4.2.1 se realizaron pruebas por cada contaminante para determinar los valores finales de entrenamiento y precisión de los modelos, en base a varias iteraciones. Asimismo, se añadió una comparación con validación cruzada con el método de k-fold para la evaluación del rendimiento del modelo, en donde se dividen los data para iterar el modelo en datos divididos por pliegues.

A continuación, se presentan lo resultados obtenidos para cada tipo de contaminante:

4.1.1.1 Emisiones de Hidrocarburos (HC)

Los resultados de las emisiones de los contaminantes se realizaron segmentando el tipo de carrocería y de manera general para evaluar el comportamiento de los modelos aplicados. A continuación, se presentan los resultados realizados para el análisis según el tipo de carrocería, siendo vehículos del tipo sedán, hatchback y SUV.

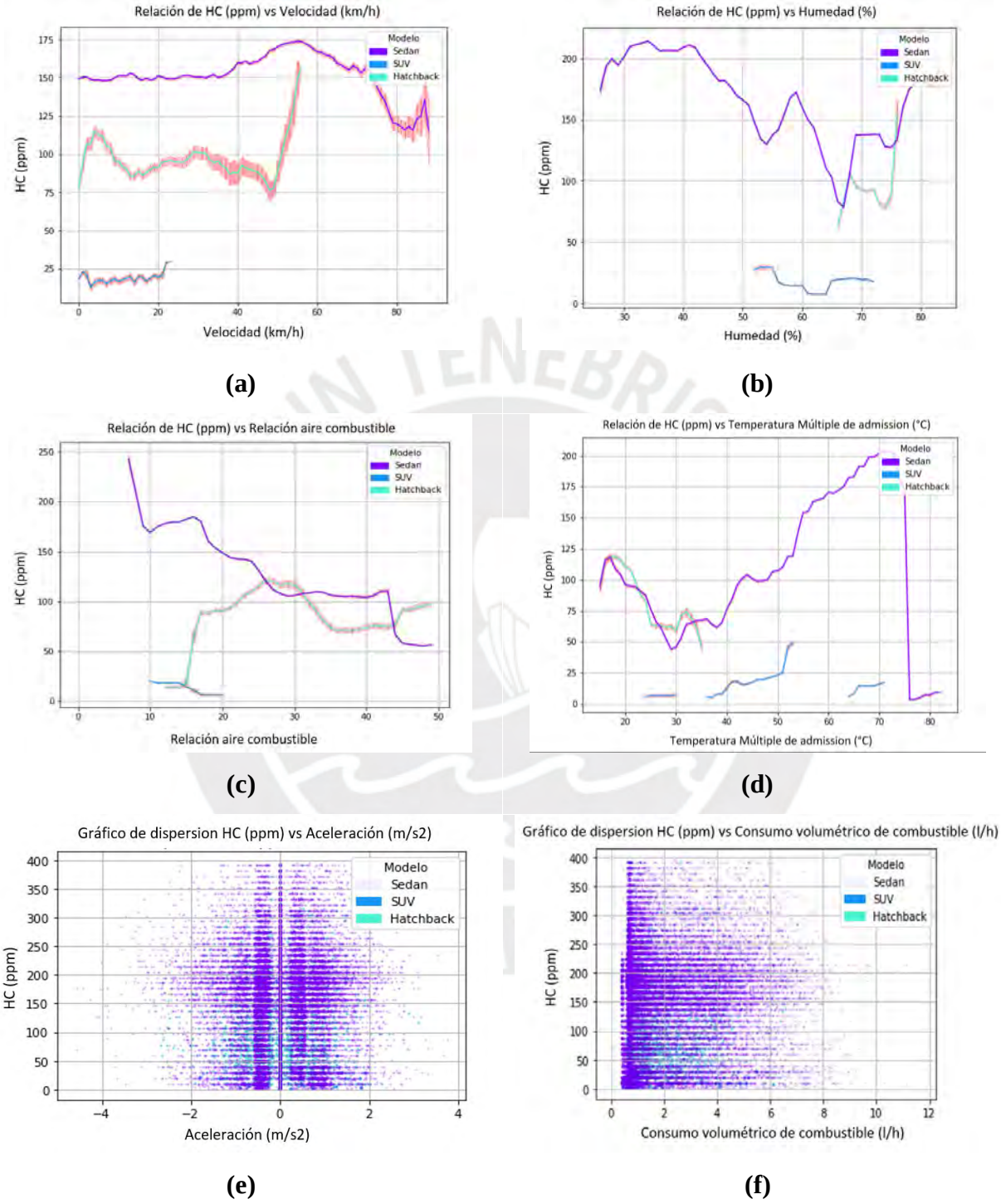
4.1.1.1 Análisis por carrocería

El análisis por tipo de carrocería se realizó mediante la evaluación de las variables medidas por el sistema de medición y el OBD y su impacto en la predicción. En la Figura 11, se precisa el

comportamiento de la media según cada variable seleccionada para el entrenamiento y el NOx. En el caso de la velocidad de los vehículos, se precisó un incremento de emisiones de HC con el aumento de la velocidad (hasta 60 km/h), especialmente evaluando los vehículos de categoría sedán que corresponden a un gran porcentaje de los datos segmentados. Asimismo, en las gráficas realizadas el mayor porcentaje de valores (70% de los datos) de la relación de aire combustible se encontraron entre 10 a 20, ello evidencia el incremento del error a medida que los datos van alejándose del rango. Por otro lado, en el caso de los vehículos del tipo SUV a pesar de tener muestras en rangos de velocidades similares, se vio que estos tenían un menor rango de emisiones de NOx. Esto no solo se manifiesta en la variable de velocidad, sino también en las demás, indicando una mayor contaminación por parte de esta específica carrocería. Para justificar ello, los resultados que corresponden a la relación de aire combustible indican que a mayor dosaje habrá menores emisiones para todas las carrocerías.

Figura 11

Análisis de comportamiento de variables con emisiones de HC según tipo de carrocería



El entrenamiento utilizando *Random Forest* se realizó, para cada tipo de carrocería, mediante el uso de Optuna, obteniendo los parámetros optimizados de cada modelo. En la Tabla 8 se puede observar los resultados finales del entrenamiento realizado.

Tabla 8

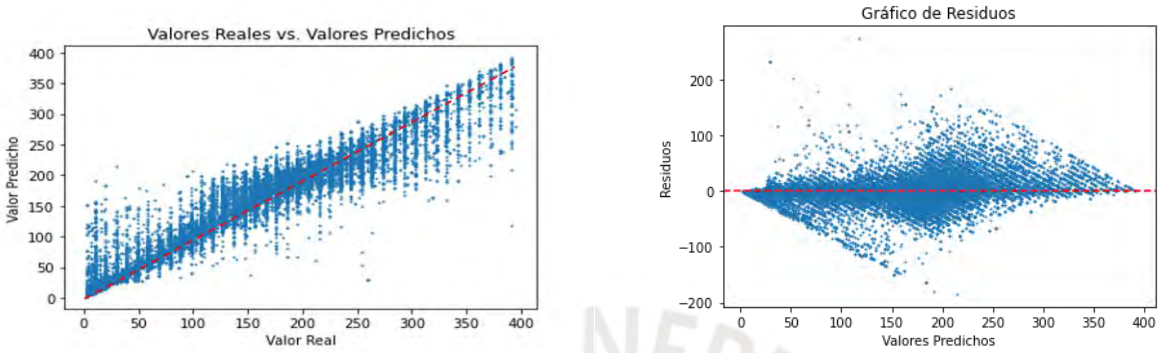
Parámetros de rendimiento de los HC con Random Forest según tipo de carrocería.

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.91*	*24.4*	*15.25*
Hatchback	*0.41*	*50.6*	*35.6*
SUV	*0.85*	*4.2*	*1.4*

Según la Tabla 8, se puede notar que el segundo mejor rendimiento del modelo se dio en los vehículos SUV, pese a la limitada cantidad de datos en comparación con el sedán. No obstante, una característica resaltante, fue que en el caso de los vehículos sedan se tuvo mayor el rendimiento del modelo en cuanto al parámetro de R^2 pero un mayor error absoluto en comparación con los SUV. Esto resalta que, a pesar de pasar por múltiples iteraciones para la optimización del modelo, la aplicación del Random Forest para los vehículos sedan presenciaron un sobreajuste de los datos que se manifiesta en el parámetro del MAE. En la Figura 12, se puede destacar este comportamiento al tenerse una visualización de la tendencia de los valores reales de los residuos con relación a los valores predichos. Asimismo, en el caso de los otros tipos de carrocería se tuvo un menor error en la predicción debido a la uniformidad de los datos. No obstante, a pesar de los resultados del error, el rendimiento nos indica que fue bajo debido a la limitada cantidad de datos capturados para dichos contaminantes.

Figura 12

Resultados del aprendizaje de entrenamiento y los datos reales de los HC para los vehículos sedán.



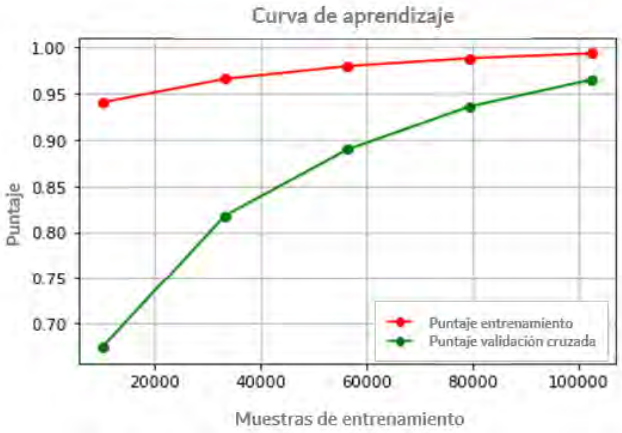
(a) (b)

4.1.1.2 Análisis de datos en conjunto

El análisis del conjunto de datos, excluyendo tipo de carrocería, se realizó optimizando los parámetros mediante Optuna y evaluando el sobreajuste de datos visto en el análisis anterior. A partir de ello, se obtuvo la curva de aprendizaje de la Figura 13. A medida que aumenta el tamaño del conjunto de datos de entrenamiento, el puntaje de validación cruzada tiene un comportamiento que simula el acercamiento a los datos de entrenamiento. Si bien la curva de aprendizaje se va acercado a la curva de validación y reduciendo el error RMSE, se puede precisar que por el comportamiento de los datos hay un sobreajuste debido al ruido interno de los valores.

Figura 13

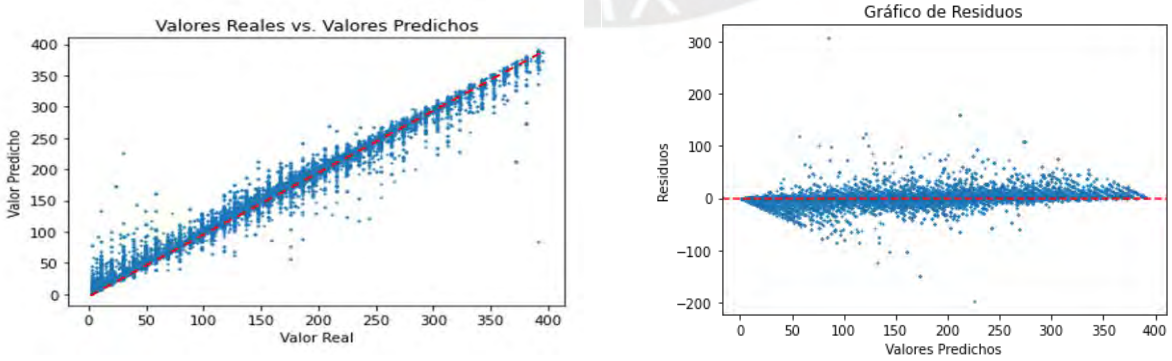
Curva de aprendizaje de entrenamiento y validación de datos de los HC.



A partir de la diferencia de los valores reales con los predichos se puede identificar su rango de residuos y validar el comportamiento del modelo. En la Figura 14, se tiene un adecuado modelo que refleja una linealidad de relación entre lo que fue real con lo que se predijo, esta linealidad indica un adecuado entrenamiento del modelo, especialmente con la Figura 14 (b) donde se ve que el rango de residuos esta entre -100 y -100. No obstante, a juzgar por el comportamiento de la data y la gráfica de predicción de un rango determinado de valores se puede precisar un acercamiento similar, indicando presencia de sobreajuste en el modelo de predicción.

Figura 14

Resultados del aprendizaje de entrenamiento y los datos reales de los HC.



(a) (b)

Considerando los resultados obtenidos en la Tabla 9, el modelo a pesar de tener un RMSE que va disminuyendo, el entrenamiento de los valores no garantiza un adecuado modelo final, especialmente porque no toma en consideración otros contaminantes y categorías de vehículos. No obstante, a pesar de ello, se tuvo unos adecuados valores de predicción que deben tomar en cuenta el sobreajuste de datos mediante modelos adicionales, como el de redes para validar el adecuado rendimiento.

Tabla 9

Parámetros de rendimiento de los HC con Random Forest

R^2	0.97
RMSE	10.3
MAE	4.8

4.1.1.2 Emisiones de NO_x

Al igual que las emisiones de HC, los resultados de las emisiones de NOx se realizó segmentando el tipo de carrocería y de manera general para evaluar el comportamiento de los modelos aplicados.

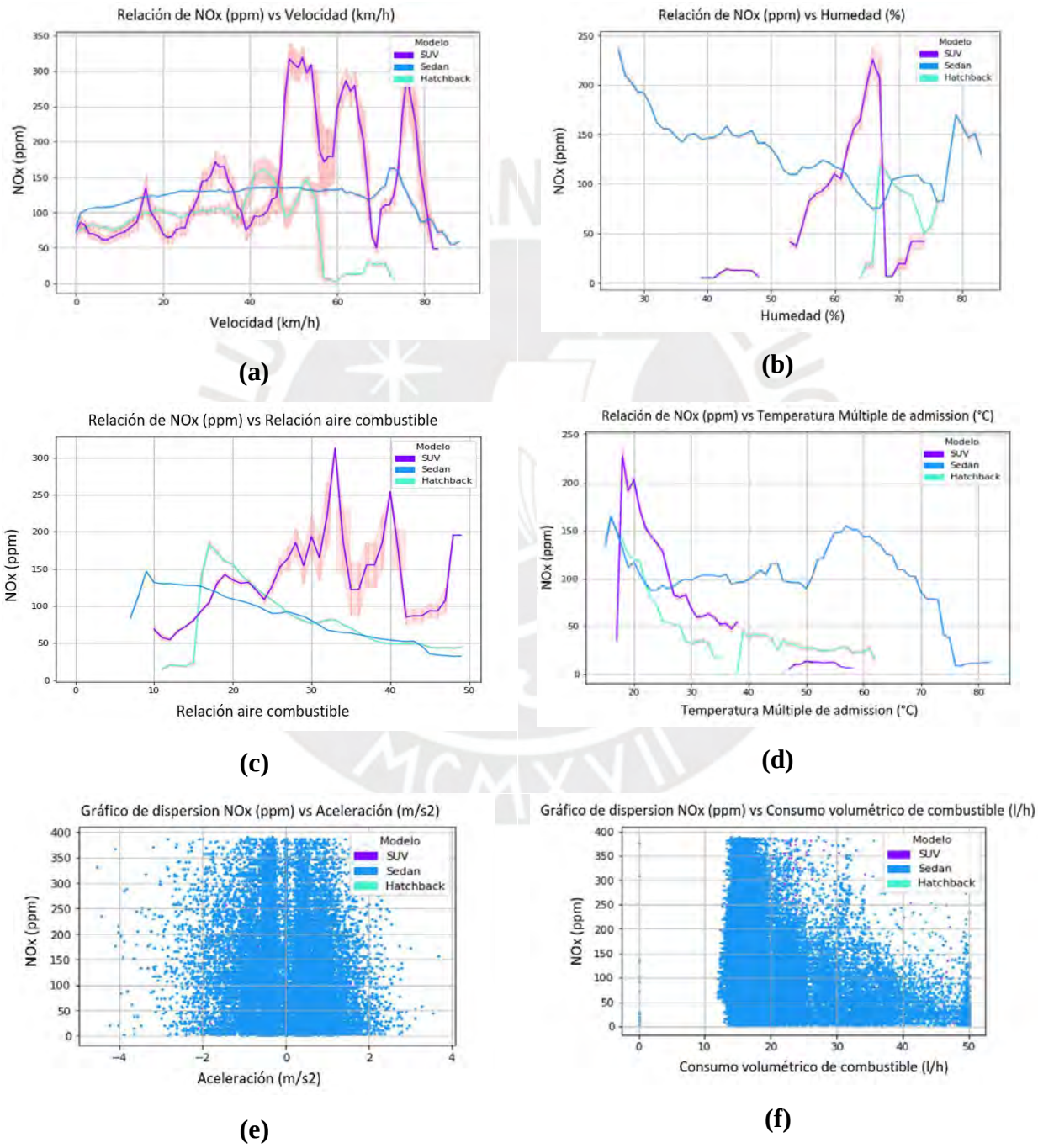
4.1.1.2.1 Análisis por carrocería

Similar a los hidrocarburos, el modelo utilizado para el entrenamiento de NOx en ppm destacó principalmente los modelos que se muestran la Figura 15. En su mayoría resaltaron los datos del tipo Sedan. El comportamiento de los NOx en similar al de los hidrocarburos principalmente porque ambos poseen una relación con el proceso de combustión, en donde a la presentación de hidrocarburos por una combustión incompleta dará menor probabilidad a la formación de NOx. El NOx se forma, principalmente por la presencia de oxígeno en el proceso de combustión a cierta temperatura (Mei, Hui y otros, 2021). A menor presencia de este según nos indica la relación de combustible consumido en la Figura 15 (e) se tendrá una disminución de las emisiones dado el incremento de temperatura que permite una combustión más uniforme. No obstante, según lo que

se visualiza en la Figura 15 (a), esto puede darse a distintas velocidades con la excepción de que a menores velocidad de 60 km/h la emisión será mayor en comparación a mayores velocidades.

Figura 15

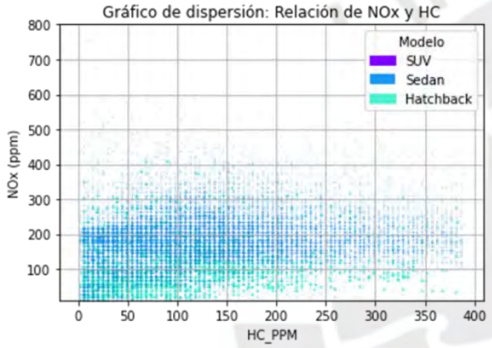
Análisis de comportamiento de variables con emisiones de NOx según tipo de carrocería



En la Figura 16, se muestra un comportamiento entre los hidrocarburos y el NOx, donde a medida que los valores de NOx disminuyen se incrementan los valores de HC hasta un límite, esto evidencia que la formación no solo depende a las variables independientes del vehículo sino de los contaminantes entre sí para tener un mejor acercamiento a su predicción. Las emisiones de HC y NOx están relacionadas entre sí y tienen como variable de mayor correlación en sus modelos predictivos a la relación de aire y combustible. En ese sentido, tanto el incremento u disminución de velocidad y arranques del vehículo determinarán el comportamiento de ambos contaminantes debido al proceso de combustión.

Figura 16

Relación de NOx y HC.



En base a la optimización de parámetros por medio de Optuna, se obtuvo los resultados de rendimiento para cada tipo de carrocería, siendo la del tipo sedán la que obtuvo mejor rendimiento y menor sobre ajuste del modelo según el valor del MAE. En caso de los otros tipos de carrocería, la limitada de cantidad de datos no permite un rendimiento mayor debido a la aleatoriedad que se puede ver en las gráficas de variables, así como la cantidad limitada de datos.

Tabla 10

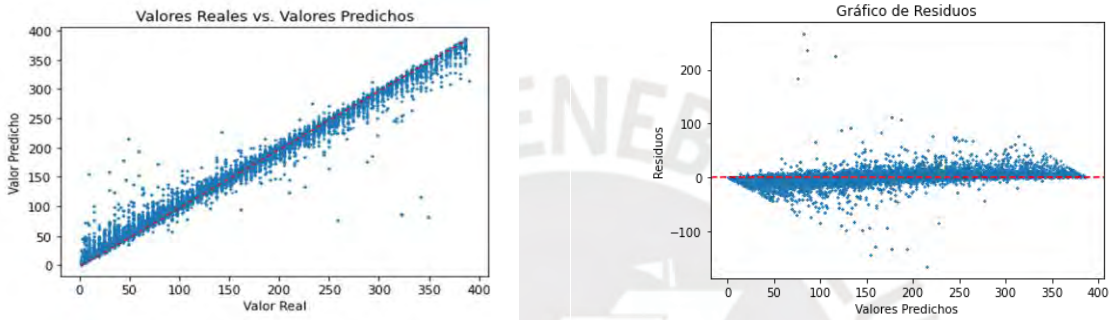
Parámetros de rendimiento de los NOx con Random Forest según tipo de carrocería.

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.91*	*13.4*	*8.2*
Hatchback	*0.49*	*28.5*	*15.6*
SUV	*0.56*	*30.4*	*18.2*

En la Figura 17, se puede precisar los valores reales y predichos para el caso del entrenamiento de los vehículos del tipo sedán, donde se destaca la baja cantidad de diferencia de los residuos y linealidad de la Figura 17 (a) que representa una adecuada predicción de los datos.

Figura 17

Resultados del aprendizaje de entrenamiento y los datos reales del NOx para la carrocería tipo sedán.



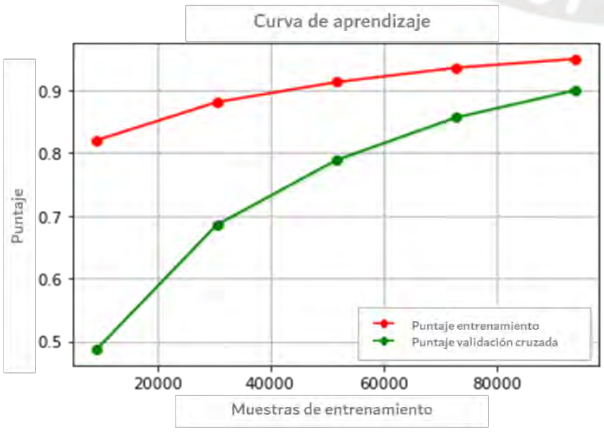
(a) (b)

4.1.1.2.2 Análisis de datos en conjunto

En el análisis de la predicción de los NO_x excluyendo tipo de carrocería, se obtuvo la curva de aprendizaje de la Figura 18, en donde se observó de la misma manera que en el caso anterior un sobre ajuste en los datos, indicando que se está prediciendo el ruido de las señales de entrada.

Figura 18

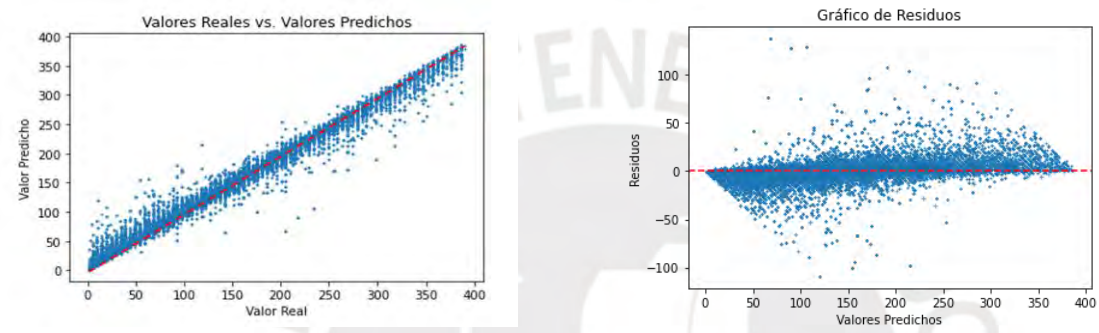
Curva de aprendizaje de entrenamiento y validación de datos de los NOx.



En la Figura 19 se tiene un modelo apropiado de predicción, donde se llegó a un valor de R^2 de 0.92 a diferencia de la evaluación por estadística descriptiva de Mendoza, J. donde se obtuvo un 0.81 de correlación con la velocidad. Asimismo, según los puntos graficados se precisa un error menor a lo esperado contrastándolo con el resultado de los HC y esto puede verse en el valor del MAE.

Figura 19

Resultados del aprendizaje de entrenamiento y los datos reales NOx.



(a) (b)

En la Tabla 11, se puede ver los valores finales de rendimiento obtenidos para el caso de los NOx, en donde los resultados muestran un adecuado nivel de predicción considerando el incremento de variables condicionales de entrada y la distribución de las variables de velocidad, que influyen en las emisiones del vehículo de manera directa.

Tabla 11

Parámetros de rendimiento de los NOx con Random Forest

R^2	0.92
RMSE	13.2
MAE	7.52

4.1.1.3 Emisiones de CO

Las emisiones de CO fueron de los datos mejor capturados para el entrenamiento con lo cual se tuvo mayor retención de datos posterior a la limpieza de los mismos para los tipos de carrocería evaluadas.

4.1.1.3.1 Análisis por carrocería

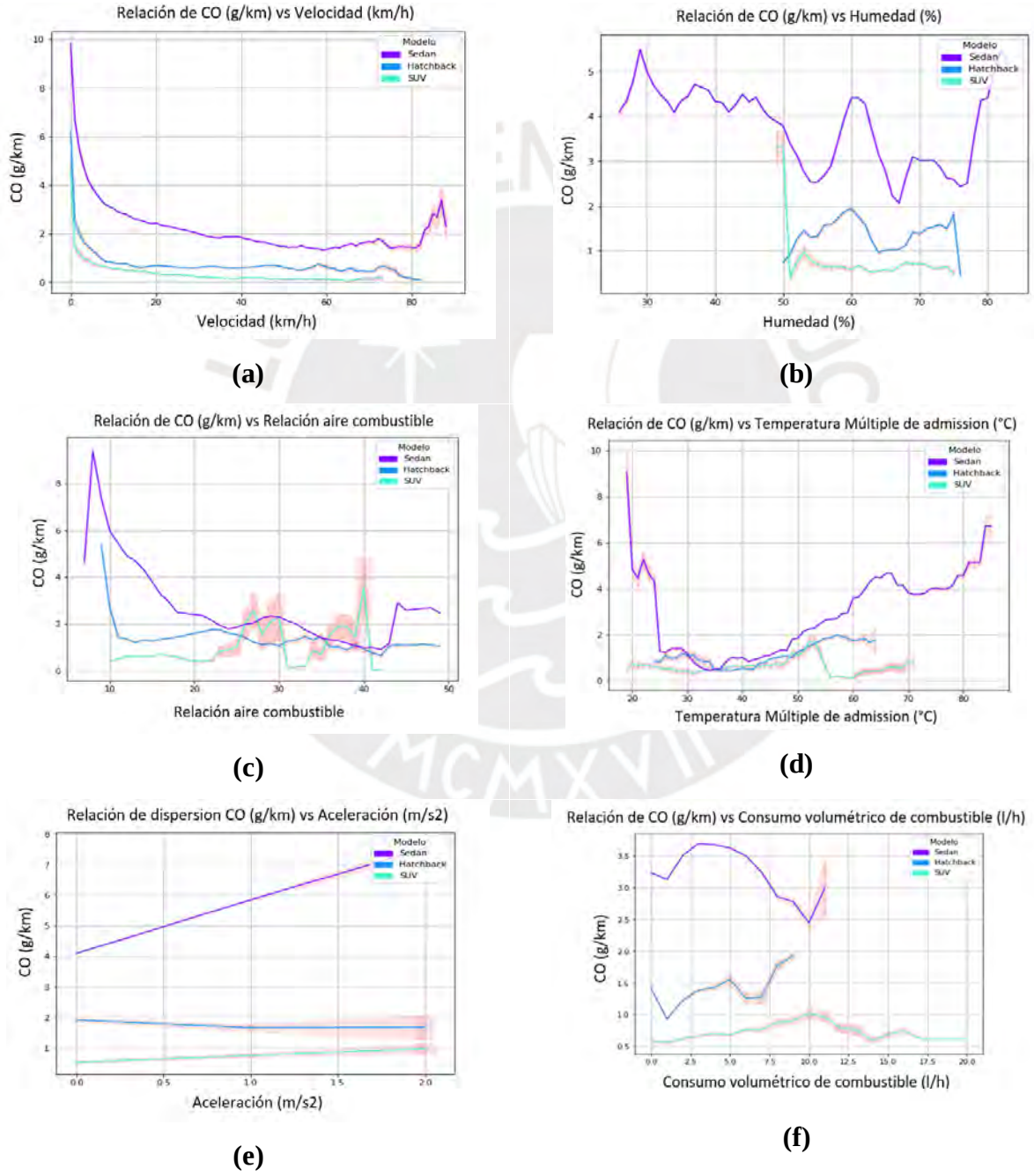
En el caso del CO se tuvieron valores menores a 100 g/km, a diferencia de los otros contaminantes se evidenció una relación más clara y directa con la velocidad reflejando que a medida del incremento de velocidad, las emisiones iban disminuyendo. En el caso de las carrocerías analizadas, se precisó menor emisión de contaminantes en un rango de 0 a 25 g/km como máximo considerando la relación con los parámetros de aceleración. Por otro lado, en relación con las otras variables, en el análisis de importante la temperatura y el consumo de combustible tuvieron pesos considerables de relación como el consumo de combustible. Según el análisis de emisiones y factores de contaminación realizado por Huang, Wang y otros (2016); se menciona que a velocidades medias las emisiones de contaminantes suelen estar entre un 40% a 80% menos que a bajas velocidades; asimismo indican que dependiendo del tipo de vehículo y sus condiciones de temperatura del mismo provocan una variación en el consumo de combustible (Huang, Wang y otros, 2016).

En el caso de la Figura 20, se precisa una caída de la media de valores de humedad en el caso de la carrocería tipo Hatchback, esto principalmente porque los valores oscilan en su mayoría en un porcentaje mayor a 50% a excepción de ciertos modelos atípicos que tiene un error estándar mayor en comparación con otros modelos. En la Figura 20 (c), se puede precisar que en el proceso de mayor dosado, en donde habrá aire adicional o exceso de oxígeno durante la combustión que terminará reduciendo las emisiones de CO. Al añadir más oxígeno del necesario para la combustión completa del combustible, se asegura que todo el carbono se oxide a dióxido de carbono (CO_2) en lugar de formarse CO. Esto resulta en una mezcla más rica en oxígeno, reduciendo así las emisiones de CO y mejorando la eficiencia de combustión. Por otro lado, el caso inverso se precisa en la Figura 20 c, donde a mayor temperatura de admisión, la densidad del

aire es reducida traduciéndose a una quema incompleta de combustible favoreciendo la formación de CO debido que en cada ciclo de admisión el motor estará recibiendo mayores moléculas de oxígeno desfavoreciendo la combustión del motor.

Figura 20

Análisis de comportamiento de variables con emisiones de CO según tipo de carrocería



En base al análisis realizado, en el caso de los resultados de la Tabla 12, se tuvo un RMSE menor, dada la presencia de una mayor cantidad de datos. No obstante, solo en el caso de la carrocería sedan se llegó a unos valores de rendimiento de R^2 óptimos en comparación con las otras carrocerías pese a tener valores de error menores en contraste con el entrenamiento de los otros contaminantes. Esto se debe principalmente a la limitación de la aplicación de Random Forest a los datos obtenidos, pese a las optimizaciones contempladas.

Tabla 12

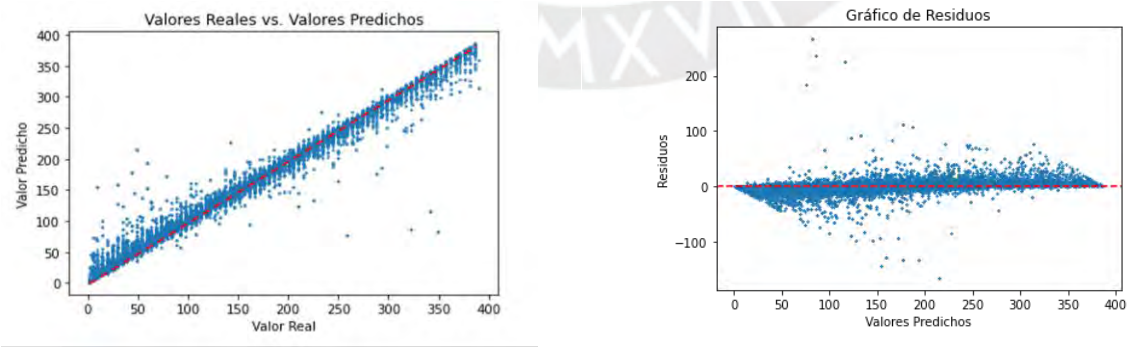
Parámetros de rendimiento del CO con Random Forest por tipo de carrocería.

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.92*	*1.82*	*0.64*
Hatchback	*0.56*	*2.14*	*0.97*
SUV	*0.67*	*1.31*	*0.56*

En la Figura 21, se ve un gráfico de residuos con mayor concentración de valores en contraste con los contaminantes anteriores y una linealidad óptima de la gráfica de valores reales con los predichos.

Figura 21

Resultados del aprendizaje de entrenamiento con los datos reales del CO para carrocería tipo sedán



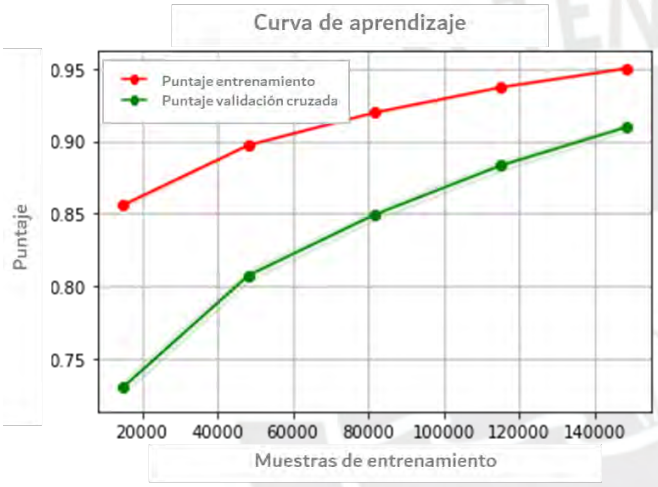
(a) (b)

4.1.1.3.2 Análisis de datos en conjunto

En el análisis de la predicción de los CO se obtuvo la curva de aprendizaje de la Figura 22, en donde se observó un comportamiento similar a los casos de los otros contaminantes, en donde la diferencia entre el entrenamiento y la validación resulta en un error bastante amplio a pesar de las iteraciones de datos realizados.

Figura 22

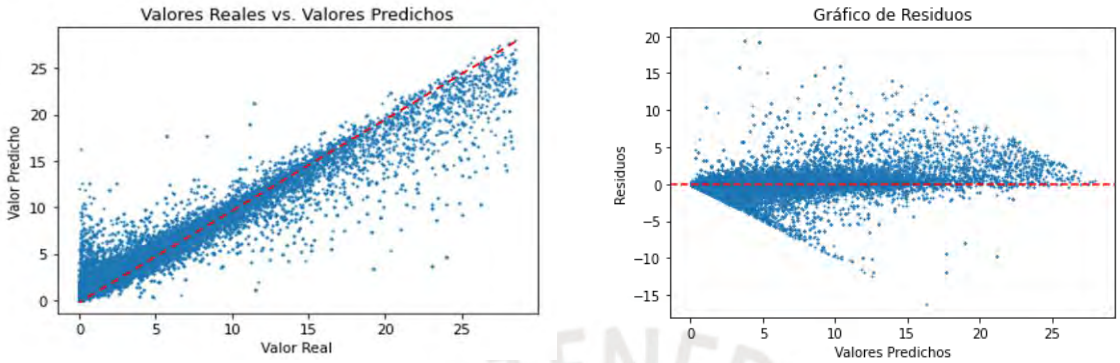
Curva de aprendizaje de entrenamiento y validación de datos de CO.



A partir de la diferencia de los valores reales con los predichos se puede identificar su rango de residuos y validar el comportamiento del modelo. En la Figura 23 se precisa una dispersión mayor que en el caso de los otros contaminantes y mayor rango de residuos por la naturaleza del comportamiento de las variables, lo que implica que no se está calculando de manera apropiada la relación entre las variables. En este caso, debido al gran número de datos, se puede descartar casos de sesgo en el modelo, siendo el sobreajuste un tema relacionado a la varianza, en donde el principal problema es la dispersión de datos que se tiene y la sensibilidad a valores nuevos.

Figura 23

Resultados del aprendizaje de entrenamiento y los datos reales del CO.



(a) (b)

A pesar de los resultados obtenidos, se puede precisar que el R^2 permite tener un panorama general de la emisión de CO en Lima Metropolitana. En la Tabla 13 se observan los resultados obtenidos finales.

Tabla 13

Parámetros de rendimiento del CO con Random Forest.

R^2	0.92
MSE	2.46
MAE	1.11

4.1.1.4 Emisiones de CO_2

En el caso del CO_2 se tuvo un comportamiento similar al CO, los rangos de valores analizados fueron de 0 a 1750 g/km del acorde a su porcentaje volumétrico. En base a los filtros de datos y procesamiento de los mismos.

4.1.1.4.1 Análisis de datos por carrocería

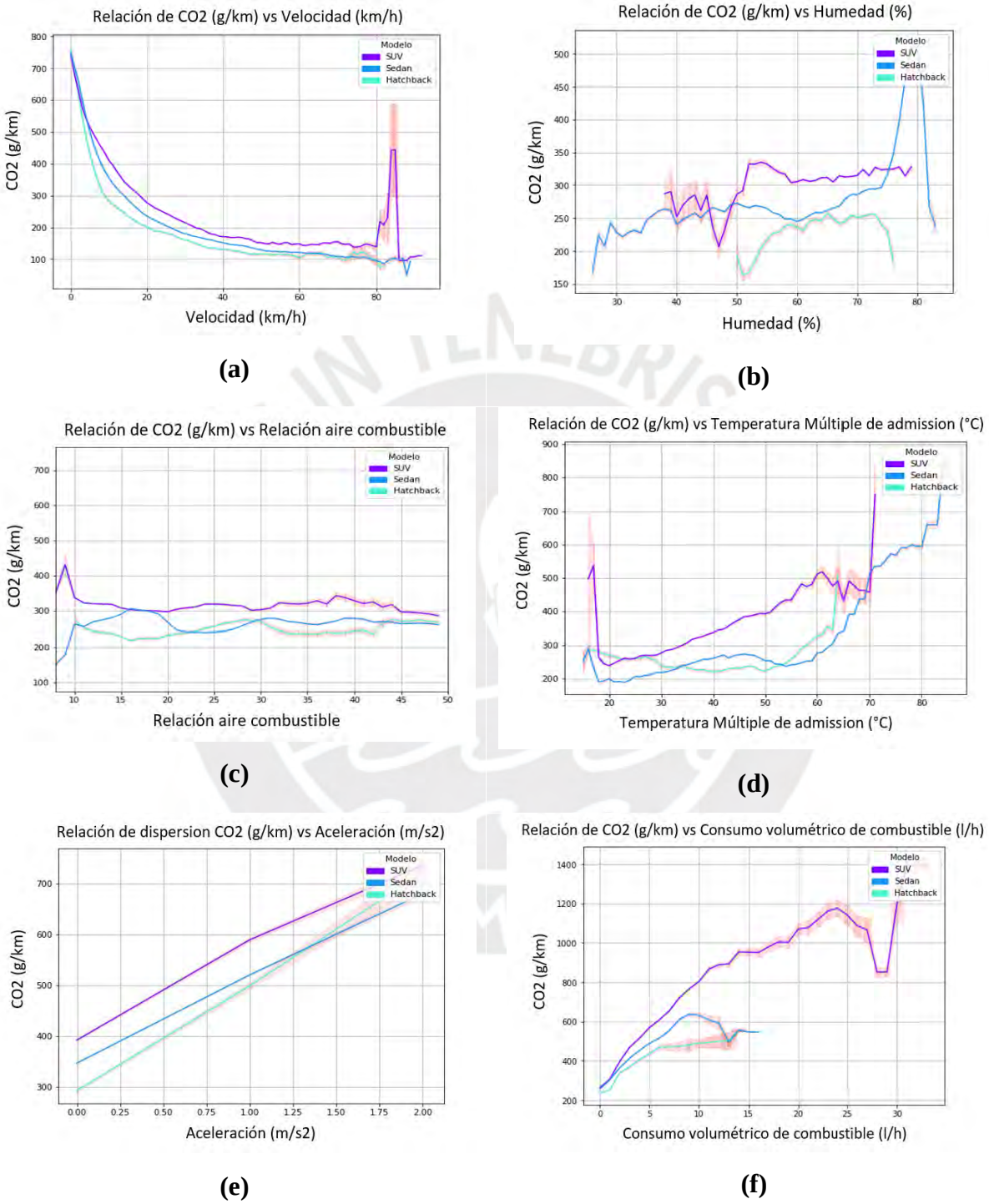
En el caso del CO_2 (Figura 24) es importante resaltar su dispersión con la variable de relación al aire combustible, a medida que este incrementa las emisiones disminuyen por la naturaleza del

consumo lo que se visibiliza en la velocidad. Según R. Kwamoto y otros (2019) durante un análisis de combustión interna tuvieron de conclusión que el manejo de largas distancia disminuía las emisiones de CO_2 de los vehículos analizados en Australia. En este caso, esto se debe a ciclo de vida de las emisiones de dióxido de carbono y a los sistemas de control de emisiones de los vehículos (Kawamoto y otros, 2019). Las emisiones totales de CO2 de un vehículo no solo dependen de las emisiones directas durante la conducción, sino también de las emisiones asociadas con la producción del vehículo y el combustible utilizado, el uso prolongado del vehículo puede amortizar estas emisiones.

En el caso de la Figura 24, se precisa que a medida que hay un aumento de temperatura, habrá un incremento de las emisiones de CO_2 a temperaturas elevadas, sin embargo, en la Figura 24 (d), no se tiene un comportamiento evidente a temperaturas menores de 50ºC. En el caso del CO_2, el consumo de combustible en el proceso de combustión y la relación de aire combustible indicará como resultado si se realizó una combustión completa. Por lo cual, se ve una constancia en la Figura 24 (c), en donde la relación de aire combustible no sufre de múltiples variaciones a lo largo de los datos captados.

Figura 24

Análisis de comportamiento de variables con emisiones de CO_2 según tipo de carrocería.



En la Tabla 14 se precisa el rendimiento de los parámetros optimizados para el caso de cada tipo de carrocería, en donde todos tuvieron resultados positivos en cuando al rendimiento evaluado por el R^2. No obstante, en el caso del error absoluto, solo los vehículos de tipo sedán y SUV tienen datos confiables de predicción en caso de que el modelo sea implementado en la ciudad de Lima.

Tabla 14

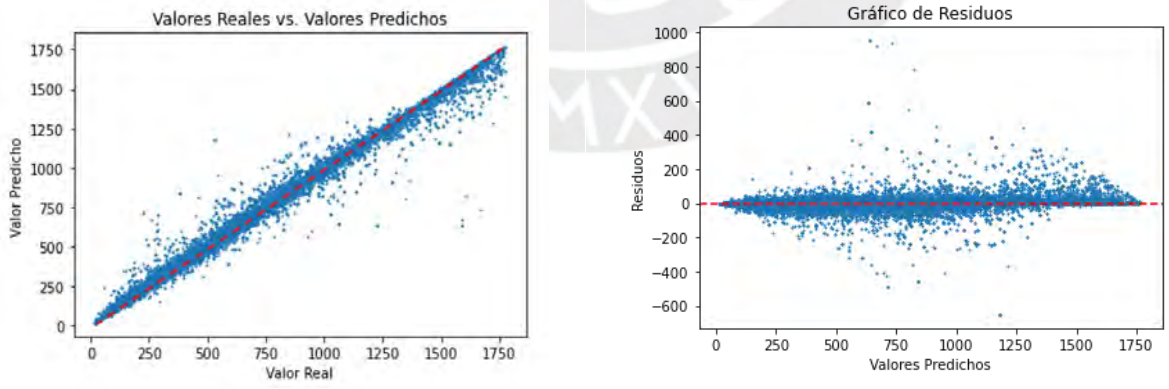
Parámetros de rendimiento del CO_2 *con Random Forest por tipo de carrocería.*

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.96*	*21.3*	*5.76*
Hatchback	*0.88*	*86.9*	*24.5*
SUV	*0.94*	*32.4*	*11.3*

En la Figura 25, se puede apreciar la predicción de los valores y su contraste con los reales, en este caso ocurre una linealidad apropiada con lo cual los datos del tipo sedán son más confiables y representativos de Lima Metropolitana para que puedan ser considerados en escenarios reales de estimaciones los contaminantes del parque automotor.

Figura 25

Resultados del aprendizaje de entrenamiento y los datos reales del CO_2 *para carrocería tipo sedán*



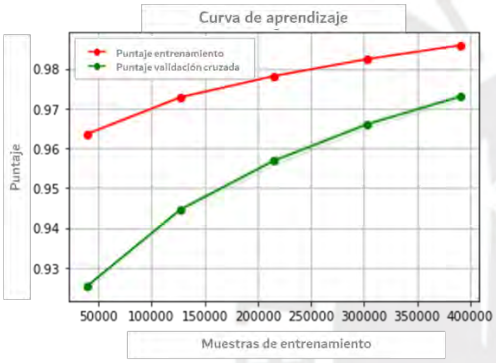
(a) (b)

4.1.1.4.2 Análisis de datos en conjunto

En el análisis de la predicción del CO_2, la curva de aprendizaje de la Figura 26 tuvo un comportamiento similar como en el caso de los otros contaminantes, pero con una tendencia de error mayor en comparación con las otras curvas analizadas. Esto puede deberse principalmente a la dependencia de la variable a la velocidad, en donde lo hace más sensible a sobreajustes en el entrenamiento.

Figura 26

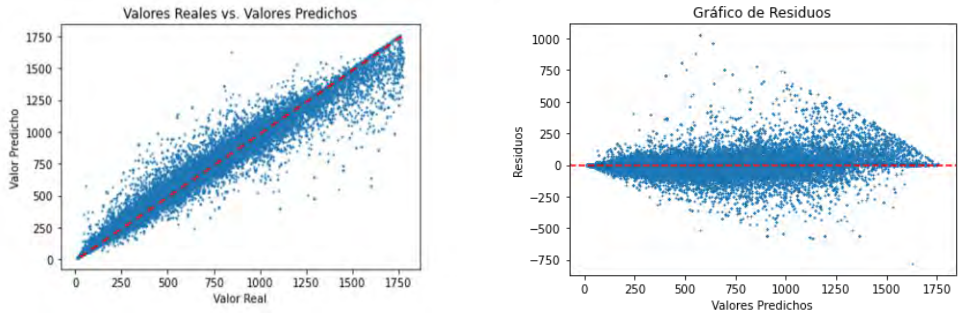
Curva de aprendizaje de entrenamiento y validación de datos del CO_2.



En la Figura 27, se puede observar el comportamiento de la predicción con los valores reales en donde se visualiza una fuerte correlación y un gráfico de residuos concentrado en valores por debajo de un error de 250.

Figura 27

Resultados del aprendizaje de entrenamiento y los datos reales del CO_2.



(a) (b)

Los parámetros de rendimiento obtenidos se encuentran en la Tabla 15, donde el valor del R^2 se puede ver que es mayor en comparación con el CO, el cual está relacionado a este contaminante. Es importante resaltar la diferencia entre el RMSE obtenido en la predicción en comparación con el de la tabla de resultados de rendimiento, en donde la disminución se atribuye a los datos aleatorios de entrenamiento utilizados por el modelo cada vez que se realiza una prueba. El conjunto de datos mantuvo el mejor valor de rendimiento de la carrocería sedán pero incremento los parámetros de error, esto principalmente por la dispersión de datos según el tipo de carrocería, la aleatoriedad del comportamiento y el sobreajuste del modelo pese a la optimización,

Tabla 15

Parámetros de rendimiento del CO_2 con Random Forest.

R^2	0.92
RMSE	42.3
MAE	9.4

4.1.2 Modelo de Redes Neuronales

En el desarrollo del entrenamiento por medio de redes neuronales se buscó reducir el sobreajuste presentado en los modelos de *Random Forest* para contrastarlos con datos de predicciones obtenidos por múltiples autores como Seo J, Abdullah H., Song H. y otros que desarrollaron metodología de aprendizaje para distintos vehículos.

4.1.2.1 Emisiones de HC

El método empleado para las emisiones de hidrocarburos fue similar al de *Random Forest*, donde se realizó un análisis por tipo de carrocería y uno general para ver el desempeño del modelo.

Los resultados en cuanto al modelo de redes dieron una mejora para los valores de rendimiento del R^2 siendo el de la carrocería tipo sedán y SUV superior al obtenido por el método de *Random Forest*. En el caso de los otros dos parámetros también se tuvieron mejoras en la disminución del

error, pero esta no fue significativa.

Tabla 16

Parámetros de entrenamiento de los HC con redes neuronales por tipo de carrocería.

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.78*	*38.2*	*27.6*
Hatchback	*0.44*	*48.5*	*34.4*
SUV	*0.85*	*13.2*	*3.6*

Las emisiones de los hidrocarburos con los datos generales se realizaron con la misma metodología que para los arboles de decisión en la cual se utilizó Optuna y ajusta manual de los parámetros de entrenamiento para obtener el menor error posible en la predicción. A diferencia del ajuste de los parámetros anteriormente mencionados, para las redes neuronales se optó por utilizar el método de "Dropout" para regular el modelo y controlar un porcentaje optimizado de neuronas para evitar el sobreajuste, los parámetros finales para el caso de los hidrocarburos de 100 iteraciones fueron los presentados en la Tabla 17.

Tabla 17

Parámetros de entrenamiento de los HC con redes neuronales.

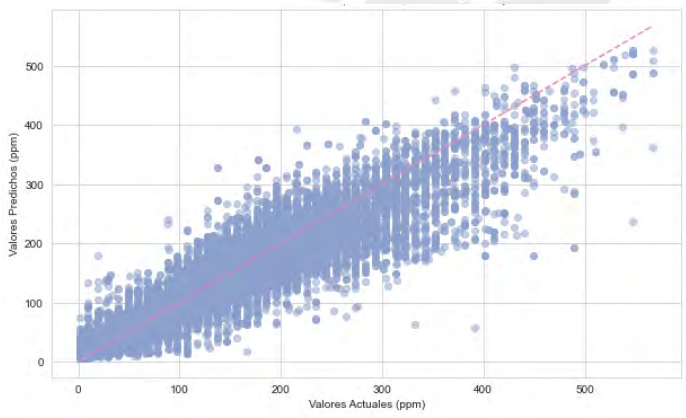
Número de capas ocultas	Número de neuronas por cada capa	Función de activación	Optimizador	Tasa de aprendizaje	Parámetros de regularización
8	87	Relu	Adam	0.0003	L_1: 0.02300 L_2: 0.00046

Los parámetros finales obtenidos evidencian la naturaleza de la complejidad del modelo, al ser una red con variables numéricas no requiere de funciones de activación sinuosidades o tangentes a diferencia de problemas de clasificación. Inicialmente, se planteó orientarlo según el tipo de vehículo. No obstante, las predicciones fueron similares al omitir su importancia en las redes. A

partir de lo descrito, la Figura 28 muestra el comportamiento de la predicción en donde se obtuvo un valor de R^2 igual a 0.64, siendo este por debajo a lo visto en los modelos de *Random Forest* que dieron de resultado un valor de 0.86. Según la investigación de Seo J y otros (2021), centrado en la predicción instantánea de emisiones de vehículos livianos, los valores obtenidos muestran tener una relación con los cuatro modelos redes neuronales de doble capa y 128 neuronas probadas por el autor, en donde sus valores de rendimiento R^2 para HC estuvieron entre 0.04 y 0.70 como valor máximo para los vehículos de prueba. Esto se debió principalmente a sus variables de entrada, donde la relación de aire combustible, aumentó su rendimiento. Asimismo, las características de su red se basaban en una estructura similar a las del presente trabajo, donde se destaca el uso de la misma función de activación y optimizador.

Figura 28

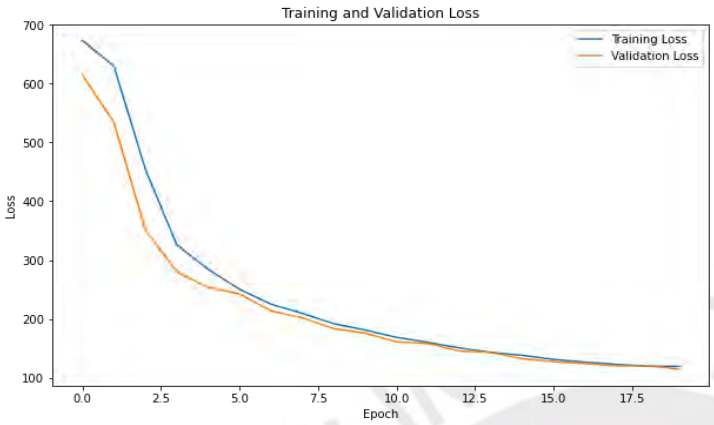
Resultados del aprendizaje de entrenamiento y los datos reales de HC.



La curva de la Figura 29 muestra que el modelo de aprendizaje automático mejora con el tiempo, tanto en el conjunto de entrenamiento como en el conjunto de validación. Esto sugiere que el modelo no está sobreajustando los datos de entrenamiento a pesar de tener un error menor al método por *Random Forest*, ya que la pérdida del conjunto de validación no aumenta significativamente después de un cierto punto.

Figura 29

Curva de pérdidas de entrenamiento y validación para HC.



Los parámetros de rendimiento de la red de los HC se presentan en la Tabla 18, se enfatiza el error superior al obtenido en la Tabla 9, al cual se tiene casi el doble de error. No obstante, en el análisis se destaca el sobreajuste que fue mejorado con el uso de redes neuronales.

Tabla 18

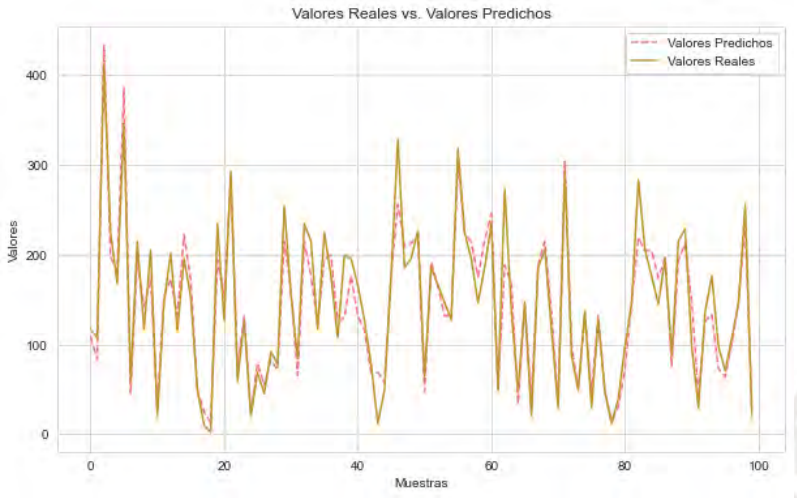
Parámetros de rendimiento de los HC con redes neuronales.

R^2	0.85
RMSE	13.4
MAE	9.8

Por otro lado, en la Figura 30, se evidencia el comportamiento de las predicciones, las cuales se van observando al largo de 100. En estas se precisa como la detección de los picos de valores medidos son los que bajan el rendimiento al modelo. Estos picos se dan de manera aleatoria en el contexto real de conducción de cualquier vehículo.

Figura 30

Valores predichos y reales con 100 muestras de prueba para los HC.



4.1.2.2 Emisiones de NOx

En base al análisis realizado según tipo de carrocería, se tuvo una mejora en los parámetros de rendimiento obtenidos y hubo una disminución del error, los datos se presentan en la Tabla 19. En base a ello, se percibe que la categoría SUV y sedán siguen siendo los modelos con mayor rendimiento y mejor predicción, no obstante, en términos de error en el caso de la SUV se tiene una menor dispersión de valores que puede deberse a la menor cantidad de datos filtrados.

Tabla 19

Parámetros de entrenamiento de los NOx con redes neuronales por tipo de carrocería

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.82*	*34.3*	*24.3*
Hatchback	*0.59*	*43.2*	*29.3*
SUV	*0.89*	*10.3*	*2.4*

Las emisiones de los contaminantes NOx en cuanto al conjunto de datos, al igual que los otros contaminantes tuvieron que calcularse mediante el apagamiento de neuronas con la técnica

Dropout. Los parámetros finales calculados con Optuna y mediante un ajuste manual son presentados en la Tabla 20, donde muestra la reducción del número de capas y una disminución de la tasa de aprendizaje. Asimismo, en este caso la penalización por medio de los parámetros de regularización en el caso del L_1 y L_2, se tuvo valores iguales, indicando que tanto la modificación de los pesos de ciertas variables tanto de manera aditiva como multiplicativa se dio de manera equitativa. Esto quiere decir que las variables de entrada fueron penalizadas de manera similar para evitar el sobreajuste (Velo, E., 2020).

Tabla 20

Parámetros de entrenamiento de los NOx con redes neuronales.

Número de capas ocultas	Número de neuronas por cada capa	Función de activación	Optimizador	Tasa de aprendizaje	Parámetros de regularización
4	81	Relu	Adam	0.0001	L_1: 0.0080 L_2: 0.0013

En la Figura 31, se precisa el rendimiento de la red en donde se obtuvo un valor de R^2 equivalente a 0.54 valor que es inferior a los resultados por el modelo de *Random Forest* por el mismo atributo precisado de sobreajuste por la alta presencia de picos en el comportamiento de las emisiones. A partir de ello, se contrastaron los valores con los obtenidos por Seo J y otros (2021), donde la predicción llegó a estar hasta 0.71 en condiciones diferentes de manejo debido a las pendientes que evalúa el estudio.

La Figura 32, muestra la curva de pérdidas de entrenamiento y validación en donde se ve un error en la convergencia de ambas tendencias debido al sobreajuste de los valores del contaminante. A pesar de presentarse este comportamiento, se tiene una convergencia apropiada para estimar con este método los valores de emisión.

Figura 31

Resultados del aprendizaje de entrenamiento y los datos reales de los NOx.

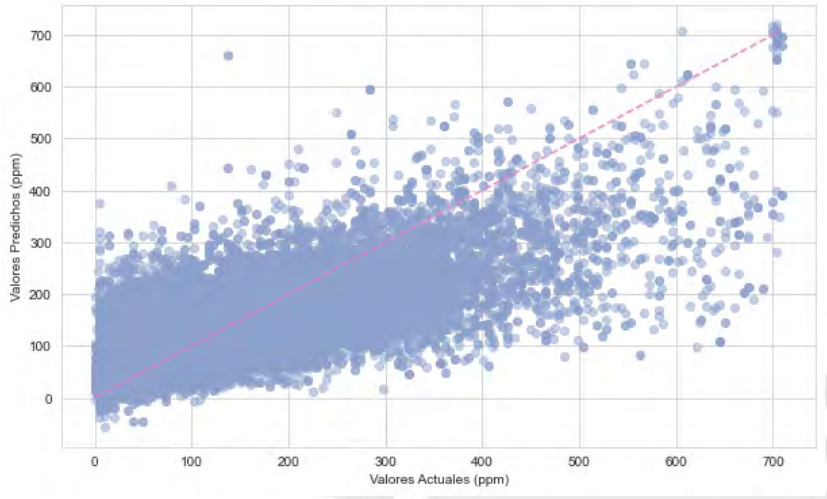
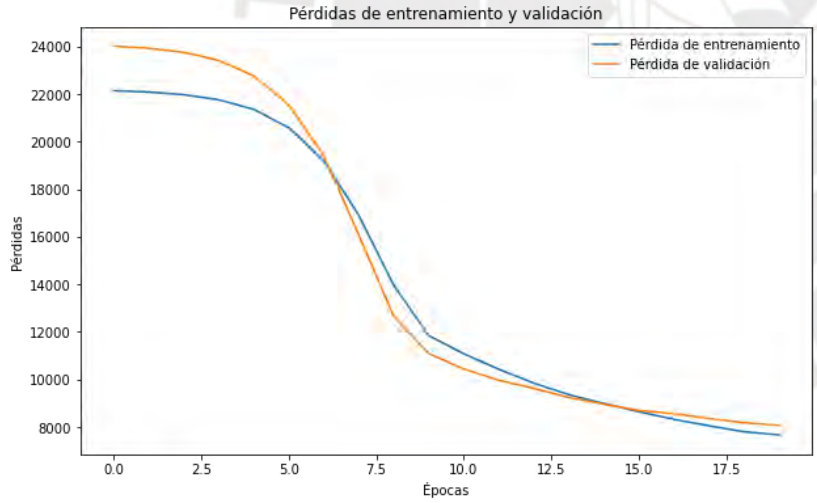


Figura 32

Curva de pérdidas de entrenamiento y validación para NOx



En la Tabla 21, se encuentran los valores de rendimiento obtenidos, donde el error continúa con un valor mayor al detectado por el método de *Random Forest* esto debido al sobreajuste del modelo en la implementación del método anterior.

Tabla 21

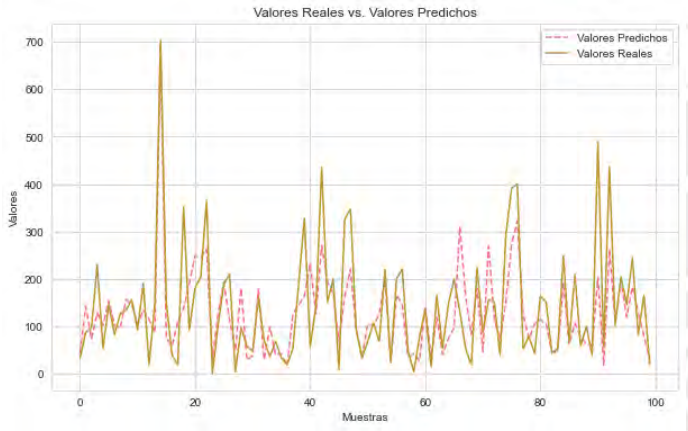
Parámetros de rendimiento de los NOx con redes neuronales

R^2	0.54
RMSE	24.5
MAE	32.3

En la Figura 33, se observa el comportamiento de las muestras graficadas para evidenciar cómo se comporta la predicción a lo largo del tiempo. En base a lo observado, se puede identificar un retraso en la estimación de los parámetros y una omisión de los picos de NOx, lo cual es natural dada la incertidumbre y aleatoriedad de las muestras tomadas y a un retraso por la medición debido los componentes utilizados en el OBD.

Figura 33

Valores predichos y reales con 100 muestras de prueba para los NOx.



4.1.2.3 Emisiones de CO

El análisis de las emisiones de CO, al igual que el método anterior, se realizó mediante el análisis por carrocería (Tabla 22), resultando con valores mejores de rendimiento por R^2 y con un aumento del error RMSE en el caso de los Hatchback y los SUV. Esto se debe a la dispersión de los datos de ambos tipos de vehículo y a la disminución del sobreajuste de los datos en comparación con el método de *Random Forest*.

Tabla 22

Parámetros de entrenamiento del CO con redes neuronales por tipo de carrocería

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.94*	*1.70*	*0.72*
Hatchback	*0.64*	*2.34*	*0.89*
SUV	*0.71*	*1.12*	*0.65*

Las emisiones de CO del conjunto de datos fueron predichas mediante una red construida con los parámetros de la Tabla 23, en donde se requirió un menor número de capas ocultas pero una mayor tasa de aprendizaje. En este caso la regularización y disminución del sobreajuste se realizó únicamente con los parámetros de regularización y se descartó el uso del "Dropout", principalmente, por los parámetros bajos de rendimiento en el entrenamiento y dado que por el método de *Random Forest* se tuvieron valores bajos de rendimiento y con alto sobreajuste.

Tabla 23

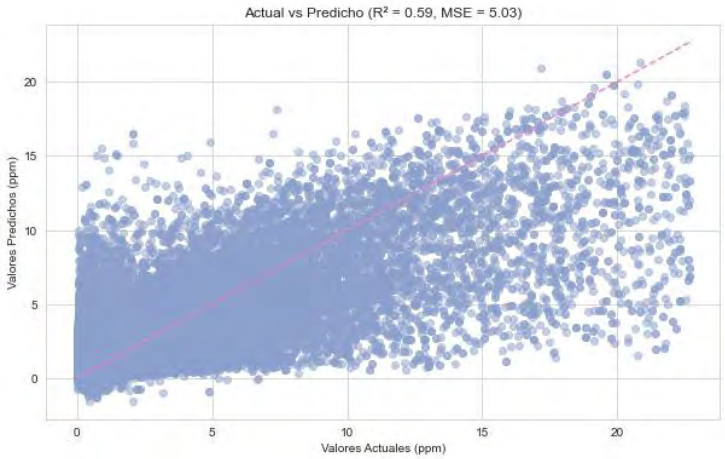
Parámetros de entrenamiento del CO con redes neuronales

Número de capas ocultas	Número de neuronas por cada capa	Función de activación	Optimizador	Tasa de aprendizaje	Parámetros de regularización
3	60	Relu	Adam	0.0005	L_1: 0.00110 L_2: 0.00005

Los resultados del aprendizaje se pueden observar en la Figura 34, en donde se obtuvo un rendimiento de 0.59, el cual fue obtenido posterior a una evaluación de 100 muestras de redes pasadas.

Figura 34

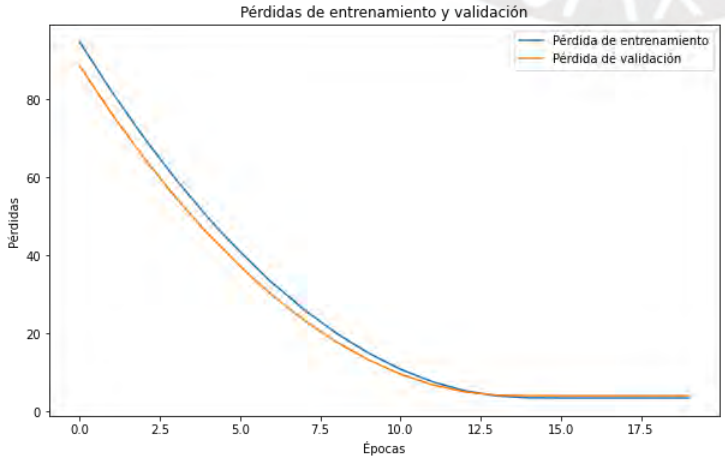
Resultados del aprendizaje de entrenamiento y los datos reales del CO.



En la Figura 35, la pérdida de entrenamiento es consistentemente menor que la pérdida de validación, a diferencia de los otros casos se debe a que el modelo aprende a disminuir los errores en datos nuevos. La pérdida de validación comienza a nivelarse después de la época 12, lo que sugiere que el modelo está empezando a sobreajustarse a los datos de entrenamiento debido a un aprendizaje de los errores de los datos. En el caso de las emisiones de CO, el modelo es apropiado para generalizar los datos a un contexto de emisiones en Lima, debido a la variabilidad de los datos de entrada vistos anteriormente.

Figura 35

Curva de pérdidas de entrenamiento y validación para CO



En la Tabla 24, se presenta el detalle de todas las variables de rendimiento. En el caso del CO, lo valores obtenidos fueron por debajo de los estudios de Seo, J y otros (2021). No obstante, según lo que se puede evidenciar en la Figura 36, se tuvo múltiples picos de CO en los datos muestreados, el error se debe a que al ser unos contaminantes con mayor dependencia a la velocidad su comportamiento tiene a ser más aleatorio en contraste con los otros contaminantes.

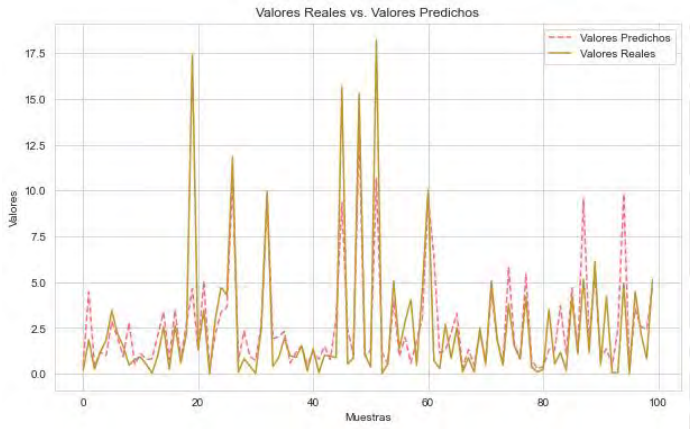
Tabla 24

Parámetros de rendimiento del CO con redes neuronales.

R^2	0.58
MSE	5.03
MAE	1.32

Figura 36

Valores predichos y reales con 100 muestras de prueba para CO.



4.1.2.4 Emisiones de CO_2

El análisis de las emisiones de CO_2, se realizó de la misma manera que los otros contaminantes, teniéndose como resultado (Tabla 25) una disminución del rendimiento del entrenamiento de los datos, debido a la dispersión de los mismos según el tipo de carrocería. No obstante, los valores del error fueron mejores para los casos de los vehículos sedán y SUV, principalmente por la gran cantidad de datos recuperados en el pre procesamiento.

Tabla 25

Parámetros de entrenamiento del CO_2 con redes neuronales por tipo de carrocería

Tipo de carrocería	R^2	RMSE	MAE
Sedan	*0.82*	*17.2*	*4.98*
Hatchback	*0.81*	*85.2*	*21.2*
SUV	*0.87*	*30.2*	*9.8*

Los datos muestreados por el CO_2 comparado con los otros contaminantes fue el que tuvo mejores correlaciones con las variables del vehículo si consideramos que se evitó el sobreajuste de datos. A partir de ello se resaltan los valores de la Tabla 26, donde no se tuvo una red compuesta de 4 capas con valores bajos de ajustes de regularización, pero con el uso de la técnica de "Dropout" debido a que presentaba un rendimiento elevado y una curva de aprendizaje con menor sobreajuste en contraste con los otros contaminantes.

Tabla 26

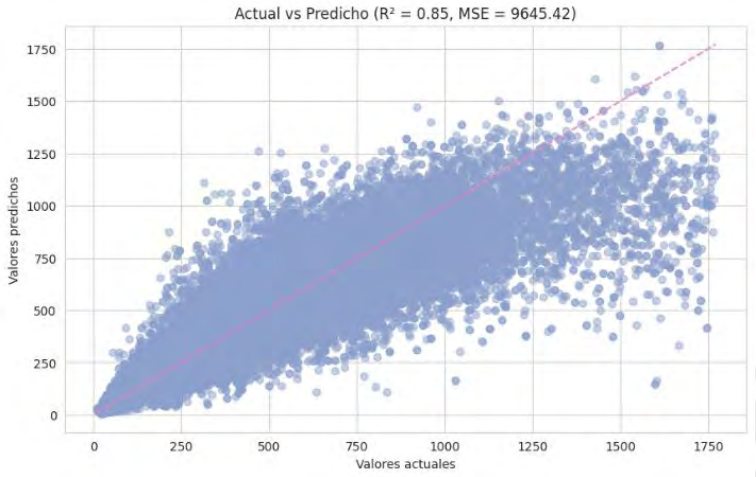
Parámetros de entrenamiento del CO_2 con redes neuronales.

Número de capas ocultas	Número de neuronas por cada capa	Función de activación	Optimizador	Tasa de aprendizaje	Parámetros de regularización
4	40	Relu	Adam	0.009	L_1: 0.01800 L_2: 0.00166

En la Figura 37, se puede observar el rendimiento de 0.85 que es superior a los valores obtenidos por Seo J y Abdullah, donde los resultados que tuvieron en entornos controlados fueron de 0.80 (Seo, J. y otros, 2021) con vehículos de características similares y de 0.76 (Abdullah, H.,2023).

Figura 37

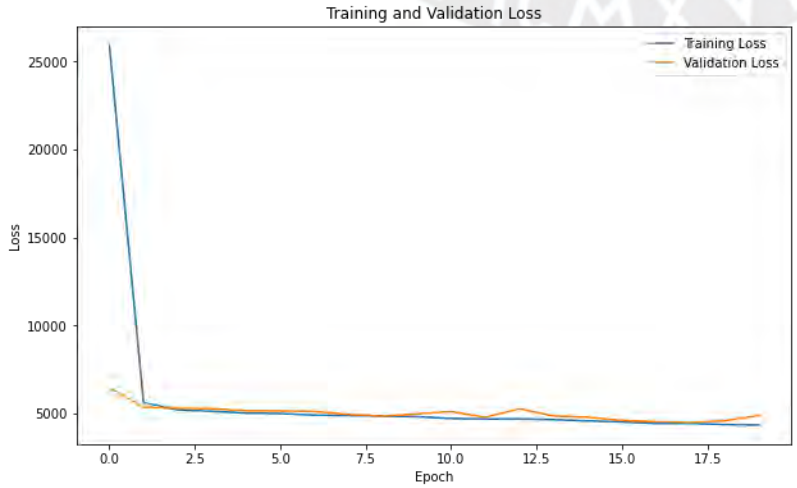
Resultados del aprendizaje de entrenamiento y los datos reales de CO_2.



En la Figura 38, se observa que la pérdida del entrenamiento disminuye rápidamente durante las primeras épocas del entrenamiento, mientras que la pérdida de la validación disminuye a un ritmo más lento. Esto sugiere que el modelo está aprendiendo el conjunto de entrenamiento bastante bien, pero no está generalizando los datos no vistos. Al finalizar, el modelo sufre de un sobreajuste de los datos similar a los presentado por Abdullah, lo que sugiere la complejidad de la aleatoriedad para la predicción de los datos de CO_2.

Figura 38

Curva de pérdidas de entrenamiento y validación para CO_2



El detalle de los parámetros de rendimiento obtenidos se muestra en la Tabla 27, donde se precisa un error menor a los valores del modelo por *Random Forest.*

Tabla 27

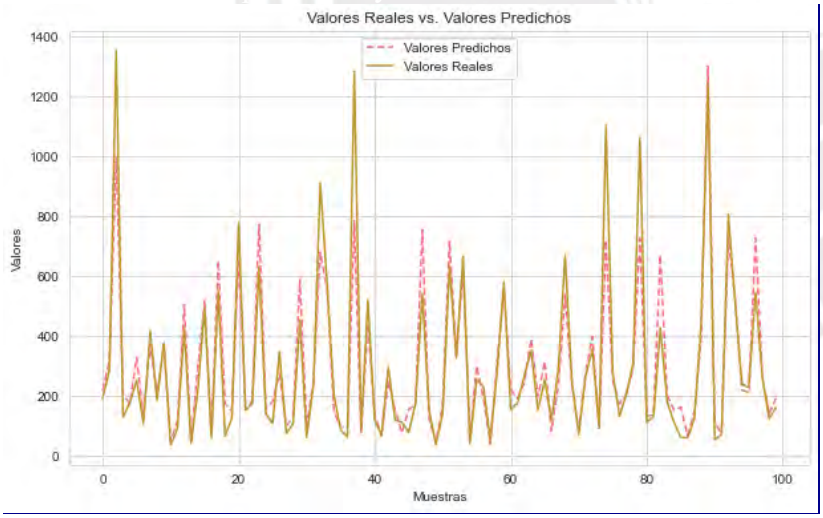
Parámetros de rendimiento del CO_2 con redes neuronales.

R^2	0.85
MSE	6734.8
MAE	20.4

Por otro lado, en la Figura 39, se detalla un entrenamiento apropiado al tener una tendencia que permite calcular determinados picos de la variable de CO_2. En este caso, se debe a que los datos tuvieron mejor correlación con las variables de manejo del vehículo.

Figura 39

Valores predichos y reales con 100 muestras de prueba para CO_2.



4.3. Mapas de predicción de emisiones

En base a los resultados obtenidos para cada contaminante, se realizaron mapas de emisiones con las rutas de prueba consideradas en los modelos de redes neuronales implementados. Los resultados se muestran en las Figuras 40, 41, 42 y 43 en donde z equivale a un rango normalizado de las emisiones de cada contaminante. En las figuras se destaca que en cada caso de contaminante la mayor parte de las emisiones se dieron en calles con alto tránsito y en avenidas céntricas de la ciudad. No obstante, en zonas más lejanas los valores de las emisiones iban disminuyendo principalmente por la disminución de tránsito de otros vehículos que permite viajes a velocidades constantes. Por otro lado, los HC y CO_2, tuvieron mayor presencia de magnitud en los recorridas analizados para la predicción, indicando que ambos valores son los que más dispersos se encontraron en toda la ciudad.

Figura 40

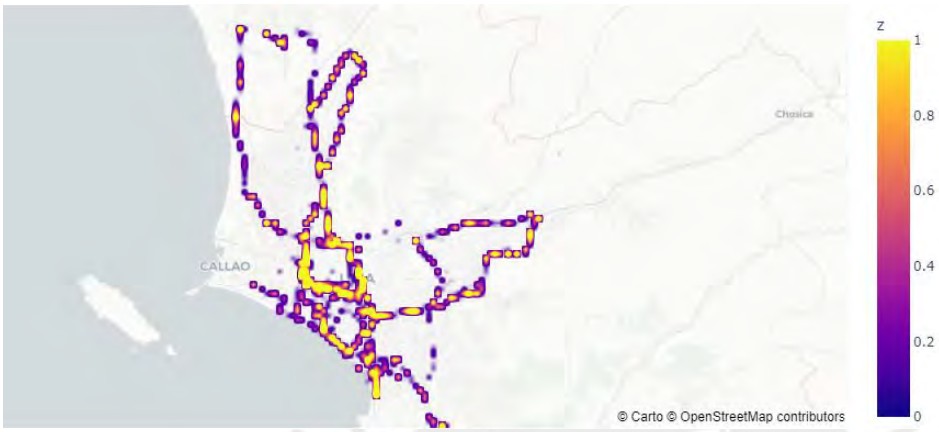
Mapa de predicción de HC en Lima

En el caso del HC y NOx, los rangos de valores de emisiones fue de 0 a 175 ppm y 0 a 400 ppm, su presencia se percibe mayor en la Av. Elmer Faucett y la Panamericana Norte. Acorde al Informe Nº359-2012-OEFA/DE en un estudio realizado a las emisiones en las Av. Elmer Faucett y Av. Argentina, se indicó que la mayoría de las emisiones de autos no pasaba los límites permisibles por el MTC de HC (100 ppm como máximo) y NOx (300 ppm como máximo), siendo esta una referencia de la magnitud de impacto de las emisiones en la Av. Elmer Faucett según los resultados

del presente trabajo.

Figura 41

Mapa de predicción de NOx en Lima



En el caso de la Figura 38 y 39, las emisiones de CO y CO_2 más elevadas se centraron en los distritos de San Miguel, Lince y La Victoria; esto debido a que la mayor cantidad de datos de muestreo correspondieron a las zonas mencionadas. No obstante, es importante resaltar que a pesar de la mayor presencia de CO_2 en zonas específicas, estás no indicaron altos valor de CO.

Figura 42

Mapa de predicción de CO en Lima

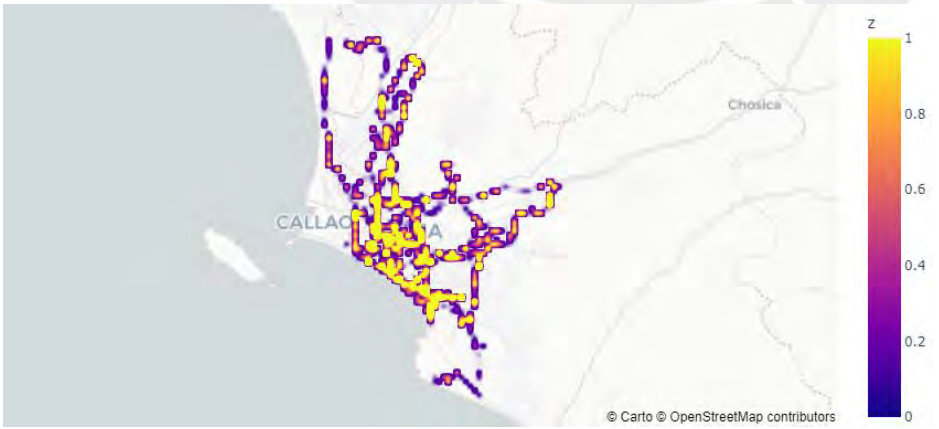
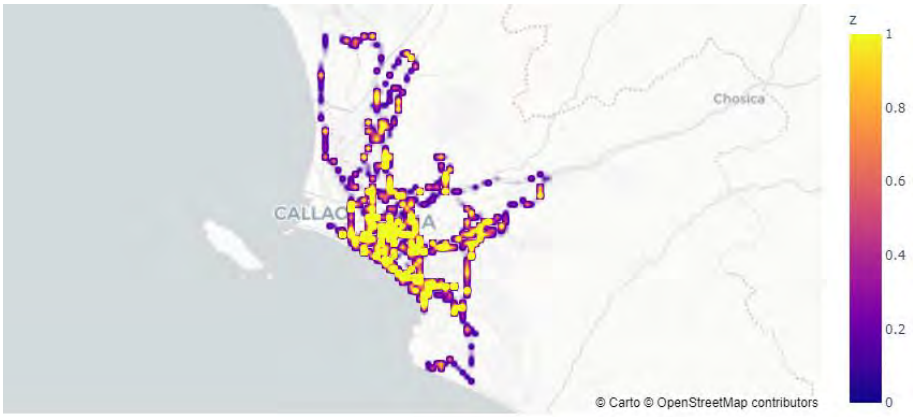


Figura 43

Mapa de predicción de CO_2 en Lima



CONCLUSIONES

El impacto ambiental del parque automotor de vehículos livianos tiene una importante contribución a la calidad del aire. Por ello, este trabajo de investigación se centró en analizar las emisiones de una flota de autos y predecirlas mediante técnicas de *Machine Learning* en Lima Metropolitana, una de las ciudades más representativas de América Latina.

- Los parámetros determinantes que afectan directamente el consumo de combustible en los vehículos livianos, en base al análisis de los datos, son el torque, la velocidad y la entrada de aire al sistema que tiene una importante correlación con la relación de aire combustible. Esto se vio mediante el análisis de "feature importance" realizado para cada contaminante.

- Las emisiones de contaminantes se predijeron utilizando redes neuronales y se validaron con el método de "Random Forest", en ambos casos las variables más importantes identificadas como entradas a los modelos mediante el método de "feature importance" fueron la velocidad del vehículo, la temperatura exterior, la humedad relativa, la relación de aire combustible, la aceleración, el consumo de combustible, la temperatura y presión del aire en el múltiple de admisión.

- Las predicciones de CO_2 por el método de redes neuronal se adecua a resultados de otros ya que se encuentran en el mismo rango que la investigación de predicción por redes neuronales de Seo J y otros (2021) con coeficiente de correlación alto (R2=0.85). En el caso de "Random Forest", las predicciones de HC, NOx y CO fueron superiores en comparación con el método de redes neuronales, mostrando una diferencia de 0.2 en la precisión. Las redes neuronales son una estrategia adecuada para evitar el sobreajuste de datos dada la aleatoriedad de comportamiento que tienen las emisiones de contaminantes.

- Finalmente, las variables analizadas dependían de distintos factores, destacándose la velocidad, para el caso del CO y CO2, y el consumo de combustible para todas las variables. Se concluyó así la importancia de los parámetros de conducción en la emisión de contaminantes de vehículos, así como del entorno geográfico y la flota vehicular

característica del parque automotor de la ciudad. A partir de ello, se precisó que el consumo de combustible y variables de manejo de los vehículos mencionadas son determinantes para obtener valores altos de entrenamiento que permiten una adecuada predicción y brindan un que se adapta a las emisiones reales de las calles analizadas en lima metropolitana.

RECOMENDACIONES

El presente trabajo de investigación tuvo resultados acorde al alcance inicial contemplado. Los resultados dieron oportunidades de mejora para poder ampliar el estudio a nivel nacional con el objetivo de poder contrastar el comportamiento de las emisiones en el parque automotor de más ciudades tomando en consideración características geográficas (altura, temperatura, humedad, entre otros). Con el propósito de cumplir esta meta se propone las siguientes recomendaciones:

- Mejorar la precisión y confiabilidad de la captura de los datos para evitar toma de valores nulos y errados que afecten el rendimiento del modelo de red neuronal.
- Implementar alternativas de una red neuronal basado en redes recurrentes para evaluar el efecto entre sí de contaminantes.
- Categorizar los vehículos y zonas distritales de muestreo para utilizar como variable de entrada estos valores y ampliar el alcance de las predicciones.

REFERENCIAS BIBLIOGRÁFICAS

Adbulkareem, A., Pradhan, B., Chakraborty, S. utores. (2021). *An Optimized Deep Neural Network Approach for Vehicular Traffic Noise Trend Modeling.* ResearchGate. https://www.researchgate.net/publication/353542513_An_Optimized_Deep_Neural_Network_Approach_for_Vehicular_Traffic_Noise_Trend_Modeling

Aguirre, D.. (2012). Evaluación de emisiones vehiculares en el Callao. (Nº359-2012-OEFA/DE).

Amirjamshidi, G. (2015). *Assessment of Commercial Vehicle Emissions and Vehicle Routing of Fleets using Simulated Driving Cycles.* Toronto: University of Toronto.

(2012). *Aspectos de la medición dinámica instantánea de emisiones de motores. Aplicación al desarrollo de un equipo portátil y una metodología para estudios de contaminación de vehículos en tráfico real.* Tesis doctoral. Bogotá: Universidad de los Andes.

Barlow, T., Latham, S., McCrae, I., & Boulter, P. (2009). *A reference book of driving cycles for use in the measurement of road vehicle emissions.* TRL.

Berry, M., Mohamed, A., & Wah, B. (2020). *Supervised and Unsupervised Learning for Data Science.*

Bodnovich, T. (2000). Neural Networks: Proposed Methodology for Applications Development.

Bonilla, S. (2020). *Redes Convolucionales.* Universidad de Sevilla. Recuperado el 18 de Noviembre de 2022, de https://idus.us.es/bitstream/handle/11441/115221/TFG%20DGMyE%20Bonilla%20Carri%C3%B3n%2C%20Carmelo.pdf?sequence=1&isAllowed=y

Cruzado, A., & Valdez, B. (2013). *Guía metodológica para la estimación de emisiones de fuentes fijas.* México: Secretaría de Medio Ambiente y Recursos Naturales. Recuperado el 17 de Noviembre de 2022, de https://www2.congreso.gob.pe/sicr/cendocbib/con4_uibd.nsf/2F1AECB7E100DA9705257D4D005632DA/$FILE/Gu%C3%ADaMetodol%C3%B3gicaParaEstimaci%C3%B3nDeEmisiones.pdf

Das, K., Jiang, J., & Rao, J. N. K. (2004). Mean squared error of empirical predictor. Annals of Statistics, 32. https://doi.org/10.1214/009053604000000201

Dawidowski, L., Sánchez - Ccoyllo, O., & Alarcón, N. (2014). *Estimación de emisiones vehiculares en Lima Metropolitana.* Lima, Perú: SENAHMI.

DEC. (2015). *On Board Diagnostics (OBD II)*. Vermont: Deparment of Environmental Conservation. Recuperado el 18 de Noviembre de 2022, de https://dec.vermont.gov/sites/dec/files/documents/OBD_20150311.pdf

Deo, N. (2016). *Graph Theory.* New Jersey: Dover Publications.

Desarkar, A., & Das, A. (2018). Implementing Decision Tree in Air Pollution Reduction Framework. *Smart Computing and Informatics . Smart Innovation, Systems and Technologies, 77*, 101-113. doi:https://doi.org/10.1007/978-981-10-5544-7_11

Dridi, S. (2021). *Unsupervised Learning - A System Literature Review.* doi:https://doi.org/10.31219/osf.io/qxng6

Eggleston, S., & M, W. (2001). Emissions: Energy, Road Transport. En IPCC, *Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories* (págs. 55 - 70). Montreal: Intergovernmental Panel on Climate Change.

EPA. (2016). *Light-Duty Automotive Technology, Carbon Dioxide Emissions, and Fuel Economy Trends: 1975 Through 2016.* EPA (United States Environmental Protection Agency).

ESAT. (2013). *OBD II Sepecifications and Connections.* Euro Systems Automotive Training Inc. Recuperado el 22 de Noviembre de 2022, de http://www.esatinc.ca/page011_news_letters.html

GHG, I. (2015). *India Specific Road Transport Emission Factors.* India: India GHG Program. Recuperado el 13 de Noviembre de 2022, de https://shaktifoundation.in/wp-content/uploads/2017/06/WRI-2015-India-Specific-Road-Transport-Emission-Factors.pdf

Hernández, R., Fernández, C., & Baptista, M. (2014). *Metodología de la investigación* (6 ed.). (McGraw-HILL, & Interamericana editores S.A., Edits.) México.

Herrera, J. (2019). *Redes Neuronales para la predicción de contaminación del aire en Carabayllo -Lima.* Tesis para optar por el grado académico de maestro en ingeniería de sistemas. Lima: Universidad Nacional Federico Villareal.

Huang, X., Wang, Y., Xing, Z., & Du, K. (2016). Emission factors of air pollutants from CNG-gasoline bi-fuel vehicles: Part II. CO, HC and NO x. Science of The Total Environment, 565, 698–705. doi:10.1016/j.scitotenv.2016.05.069

ICCT. (2016). *On-Board Diagnostic (OBD) Checks for Inspection and Maintenance in India.* ICCT. Recuperado el 18 de Noviembre de 2022, de https://theicct.org/wp-

content/uploads/2021/06/ICCT_IM-OBD-India_201606.pdf

ICCT. (2018). *Anual report.* Washinton: Internation Council on Clean Transportation.

IPCC. (2019). *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories.* Intergovernmental Panel on Climate Change. Recuperado el 10 de Noviembre de 2022, de https://www.ipcc.ch/site/assets/uploads/2019/12/19R_V0_01_Overview.pdf

IQAir. (2021). *World Air Quality Report.* IQAir. Obtenido de https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2021-en.pdf

Kawamoto, R., Mochizuki, H., Moriguchi, Y., Nakano, T., Motohashi, M., Sakai, Y., & Inaba, A. (2019). Estimation of CO2 Emissions of Internal Combustion Engine Vehicle and Battery Electric Vehicle Using LCA. Sustainability, 11, 2690. https://doi.org/10.3390/su11092690

Klinedinst, D., & King, C. (2016). *On Board Diagnostics: Risks and Vulnerabilities of the Connected Vehicle.* EEUU: Carnegie Mellon University.

Kroese, D., Botev, Z., Taimre, I., & Vaisman, R. (2022). *Data Science and Machine Learning: Mathematical and Statistical Method.* Boca Ratón: Chapman and Hall/CRC.

Larrañaga, P., Inza, I., & Moujahid, A. (2007). *Tema 8. Redes Neuronales.* País Vasco: Departamento de Ciencias de la Computación e Inteligencia Artificial - Universidad del País Vasco. Recuperado el 14 de Noviembre de 2022, de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t8neuronales.pdf

Leskovec, J., Rajaraman, A., & Ullman, J. (2020). Clustering. En J. Leskovec, A. Rajaraman, & J. Ullman, *Mining of Massive Datasets* (págs. 240 - 280). EEUU: Cambridge University Press. Recuperado el 8 de Noviembre de 2022, de http://www.mmds.org/

Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., & Ashraf, T. (2022). Estimating PM2.5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China. *Remote Sensing*.

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. (2019). *Supervised Machine Learning.* Department of Information Technology, Uppsala University. Recuperado el 10 de Noviembre de 2022, de https://mwns.co/blog/wp-content/uploads/2020/01/Supervised-Machine-Learning.pdf

MAHLE. (2016). *MAHLE Portable Emissions Measurement System (PEMS).* United Kingdom: MAHLE.

Mendoza, J., J. (2022). *Propuesta metodológica para el cálculo de factores de emisión para*

vehículos livianos a gasolina circulando en Lima Metropolitana. Tesis para obtener el grado académico de Magíster en Energía. , Lima: Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería.

MINAM. (2017). *DECRETO SUPREMO N° 010-2017-MINAM.* Diario Oficial El Peruano.

Nickischer, A. (2020). Environmental Impacts of Internal Combustion Engines and Electric Battery Vehicles. *D. U. Quark*.

Nield, T. (2022). *Essential Math for Data Science.* EEUU: O'Reilly Media, Inc.

Nilsson, N. (1998). *Introduction to Machine Learning.* California: Stanford University. Recuperado el 15 de Noviembre de 2022, de https://ai.stanford.edu/~nilsson/MLBOOK.pdf

Omer, S., Biswajeet, P., Helmi, Z., Nagesh, S., Chang-Wook, L., & Hossein, M. (2019). Modeling of CO Emissions from Traffic Vehicles Using Artificial Neural Networks. *Applied Science*.

Permuter, H. (2022). *Lecture 4 - Neural Networks.* Negev: Ben-Gurion University of the Negev. Recuperado el 18 de Noviembre de 2022, de https://www.ee.bgu.ac.il/~haimp/ml/lectures.html

Pinho, J. (2010). *Métodos de clasificación basados en asociación aplicados a sistemas de recomendación.* Tesis Doctoral. Salamanca, España: Universidad de Salamanza. Departamento de Informática y Automática.

Rosales, J. (2022). Predicción de la contaminación atmosférica generada por las emisiones del CO2 en el Perú utilizando los métodos ARIMA y Redes Neuronales. *TecnoHumanismo. Revista Científica*, 114-125.

Russel, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach.* New Jersey: Pearson Education, Inc.

SAE. (2002). *SAE J1979.* Washington: SAE. Obtenido de https://law.resource.org/pub/us/cfr/ibr/005/sae.j1979.2002.pdf

SEMARNAT. (2009). Guía metodológica para la estimación de emisiones vehiculares en ciudades mexicanas. México: Secretaría de Medio Ambiente y Recursos Naturales.

SENAMHI. (2021). *Informe: Vigilancia de la calidad del aire.* Área Metropolitana de Lima y Callao: SENAMHI.

Sher, E. (1998). *Handbook of Air Pollution from Internal Combustion Engines.* San Diego, USA: Academic Press.

Tarun, S. (2019). *Artificial neural networks for fuel consumption and emissions modeling in light duty vehicles.* Master of Science. Colorado: Colorado State University, Department od Mechanical Engineering.

Toward Data Science. (s.f.). Hyperparameter Tuning the Random Forest in Python using Scikit-Learn. Toward Data Science. Recuperado de https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Valverde, V., & Bonnel, P. (2017). *On-road testing with Portable Emissions Measurement Systems (PEMS).* European Union: Joint Research Centre.

Vargas, I. (2020). *Implementación de un modelo algorítmico para la estimación del nivel de concentración de contaminante PM2,5 en zonas urbana.* Pontificia Universidad Católica del Perú, Tesis para optar por el grado académico de Magistar en Informática con mención en Ciencias de la Computación. Lima: Facultadad de Ciencia e Ingeniería.

Verizon. (s.f.). Hum by Verizon - Encuentra el puerto OBD - II. Recuperado el 20 de Noviembre de 2022, de https://espanol.verizon.com/support/knowledge-base-210135/

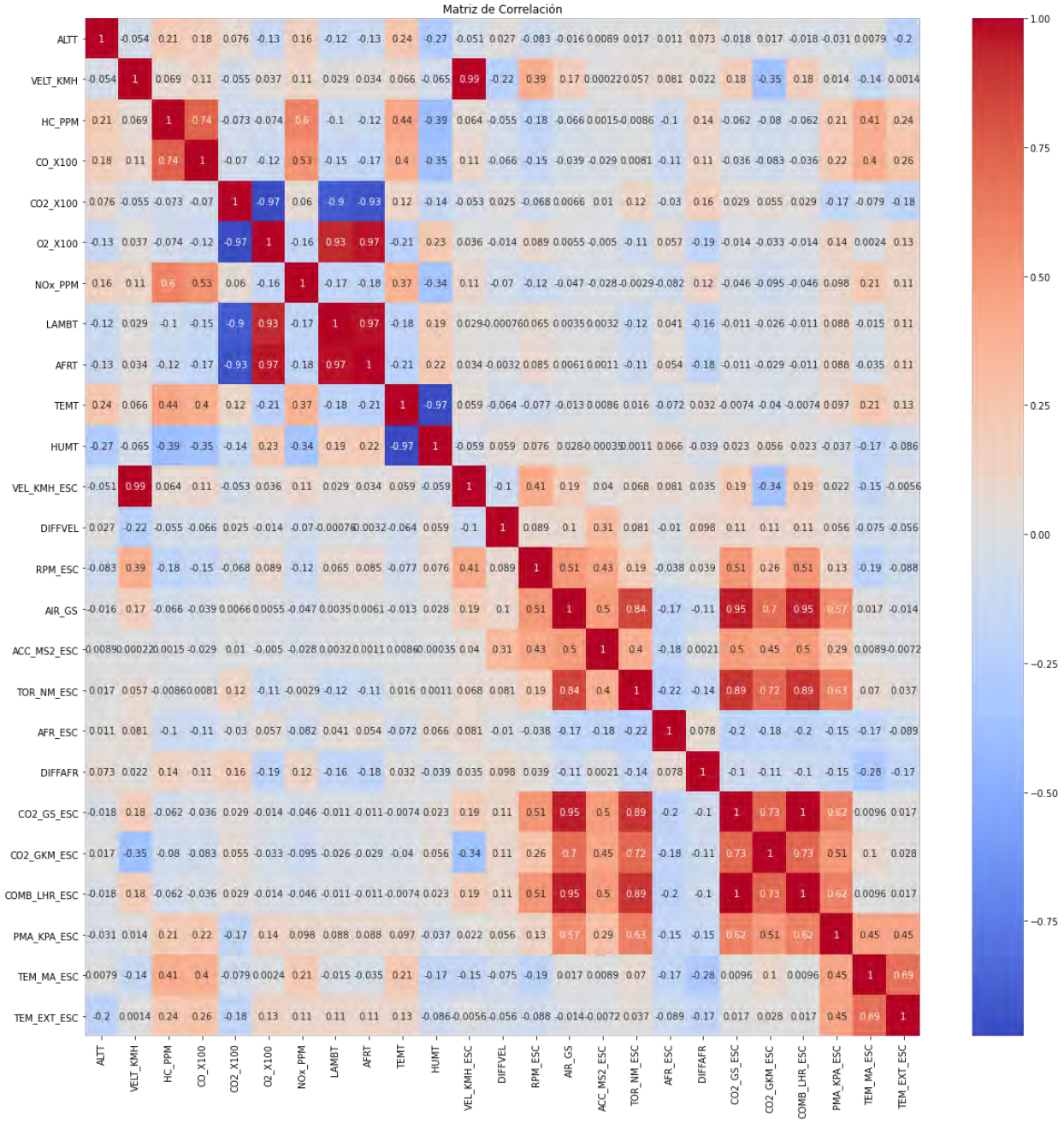
Viscidi, L., & O' Connor, R. (2017). The Energy of Transportation: A Focus on Latin American Urban Transportation. En P. Isbell, & E. Alvárez, *Energy and Transportation in the Atlantic Basin* (págs. 91 - 127). Washinton: Center for Transatlantic Relations.

Vojtisek-Lom, M. (2000). *EE.UU Patente nº 6435019.*

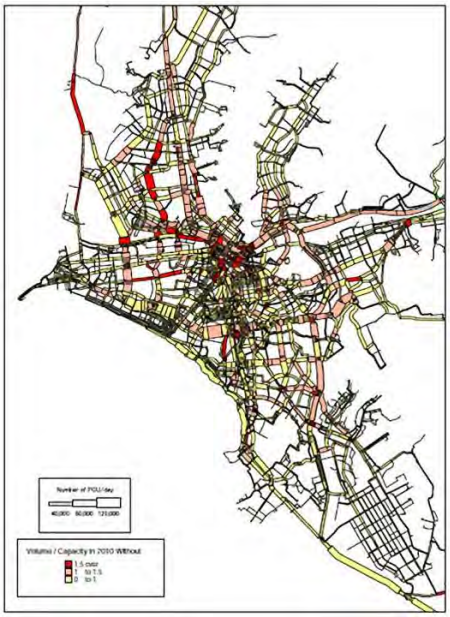
Wallington, T. J., Kaiser, E. W., & Farrell, J. T. (2006). Automotive fuels and internal combustion engines: a chemical perspective. Chemical Society Reviews, 35(4), 335. https://doi.org/10.1039/b410469m

ANEXOS

Anexo 1. Matriz de correlación de variables de utilizando el método de feature importante.



Anexo 2. Rutas de demanda de transporte de Lima. Fuente: Yachiyo Engineering Co, 2005.



En base a ello, las avenidas tomadas en consideración que corresponden a las más recorridas (MML, 1999) son:

- Av Grau
- Av. Venezuela
- Av. Brasil
- Av. Panamericana Sur
- Av. Panamericana Norte
- Av. Universitaria –Av. Tomas Valle
- Av. Angamos
- Av. Néstor Gambeta
- Av. La Molina
- Av. Canevaro
- Av. Faucett
- Circuito de Playas
- Av. Javier Prado