Grado en Ingeniería en Tecnologías de Telecomunicación

Trabajo de Fin de Grado

Modelo de Predicción de Emisiones

Autor
Beatriz Ordóñez Becker

Supervisado por
Luis Francisco Sánchez Merchante

Madrid
Junio 2024

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Modelo de Predicción de Emisiones

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2023-24 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.:  Fecha: 02 / 07 / 2024

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.:  Luis Francisco Sánchez Merchante 2024.07.02 18:22:46 +02'00' Fecha: ..... / ..... / .....

Grado en Ingeniería en Tecnologías de Telecomunicación

Trabajo de Fin de Grado

Modelo de Predicción de Emisiones

Autor
Beatriz Ordóñez Becker

Supervisado por
Luis Francisco Sánchez Merchante

Madrid
Junio 2024

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a todas las personas que han hecho posible la realización de este proyecto.

En primer lugar, agradezco a mi director de proyecto, Luis Francisco Sánchez Merchante, por su apoyo continuo y valiosos consejos durante todas las fases de esta investigación.

En segundo lugar, quiero dar un reconocimiento especial a mi familia, quienes han compartido conmigo innumerables horas de estudio y trabajo. Cada uno a vuestra manera habéis hecho que este proyecto fuese posible.

A mis padres por ser un pilar fundamental en mi vida y apoyarme incondicionalmente. Os admiro profundamente y os estaré siempre agradecida. A mi abuela Carmen, que ha hecho que este sueño fuese una realidad. A mi tía María, que me ha tratado siempre como a una hija. A mi tía Marichu por acogerme en su casa todos estos años y, en especial, a mi tío Carlos Becker, por ser una fuente de inspiración constante, un ejemplo a seguir y un apoyo incondicional.

A todos ustedes, muchas gracias.

MODELO DE PREDICCIÓN DE EMISIONES

Autor: Beatriz Ordóñez Becker

Director: Luis Francisco Sánchez Merchante

Entidad Colaboradora: ICAI - Universidad Pontificia Comillas.

Resumen del proyecto

Este proyecto se centra en el desarrollo de un modelo de predicción de la calidad del aire utilizando técnicas de aprendizaje automático. El objetivo principal es identificar las variables que contribuyen a la acumulación de contaminantes en áreas urbanas y predecir los niveles de calidad del aire en tiempo real basándose en estas variables.

El estudio implica la recopilación y el análisis de datos sobre contaminantes atmosféricos, particularmente el óxido nítrico (NO), las partículas PM2.5 y PM10, de varias estaciones de monitoreo en la Comunidad de Madrid. Estos datos se combinan con información sobre el tráfico y datos meteorológicos para crear un conjunto de datos integral que permita entrenar los modelos de predicción. Asimismo, se aplican varias técnicas de regresión y estrategias de refinamiento del modelo, incluyendo la selección de hiperparámetros, la eliminación de valores atípicos y la reducción de dimensiones, para mejorar la precisión de las predicciones.

Los resultados pretenden proporcionar una comprensión detallada de la dinámica de la contaminación del aire urbano y ofrecer una herramienta para la planificación urbana y la toma de decisiones en salud pública, contribuyendo al objetivo más amplio de crear ciudades sostenibles.

Palabras clave: Contaminantes, Calidad del aire, Área urbana, Modelo de predicción, Precisión.

1. Introducción

Hoy en día, uno de los problemas más preocupantes en nuestra sociedad es la contaminación. Aunque no es un fenómeno reciente, su relevancia ha aumentado significativamente en los últimos años. La preocupación por la contaminación incrementa, así como sus efectos sobre el planeta y sobre nuestra salud. Actualmente, un 20 % de la incidencia total de enfermedades puede atribuirse a factores medioambientales [1]. Desde la revolución industrial, que introdujo nuevas formas de producción, hasta la actual revolución tecnológica, han surgido herramientas que, aunque facilitan nuestra vida cotidiana, también generan una gran cantidad

de contaminantes atmosféricos. Dichos contaminantes serán el objeto de estudio de este proyecto.

2. Definición del proyecto

El objetivo principal de este proyecto es desarrollar un modelo de predicción de calidad del aire utilizando técnicas basadas en Machine Learning. Para ello, se han de identificar las variables más influyentes en la acumulación de partículas contaminantes en áreas urbanas y realizar mediciones que, aplicadas a modelos de predicción, permitan predecir los niveles de contaminación en tiempo real. Este proyecto se adentra en el ámbito de la calidad del aire contribuyendo a desarrollar estudios como *"Numerical simulation model for predicting air quality along urban main roads: first report, development of atmospheric diffusion model"* [2], publicado en el año 1998. A continuación se detalla el proceso de modelización empleado en este estudio.

3. Proceso de modelización

Para la modelización de los contaminantes se han seleccionado un total de 14 variables [1] recogidas en la Tabla 1. La tarea de predicción de los tres contaminantes seleccionados sigue en todo momento un mismo patrón, análisis de la presencia de outliers, disminución de la dimensión del dataset original, aplicación de técnicas de selección de variable, y realización de ajustes de hiperparámetros, utilizando en todo momento el coeficiente de determinación r2 como métrica de precisión. El proceso de modelización se describe en la Figura 1.

[1] La variable *Altura* se ha desglosado en 16 medidas distintas (detalle en el capítulo 3), logrando un total de 29 variables para entrenar los modelos

VARIABLES
ID de la estación
Número de carriles
Área que abarca la estación
Hora del día
Día de la semana
Velocidad del viento
Dirección del viento
Temperatura
Humedad
Presión barométrica
Radiación
Precipitación
Altura

Tabla 1: Lista de variables consideradas en el proyecto.

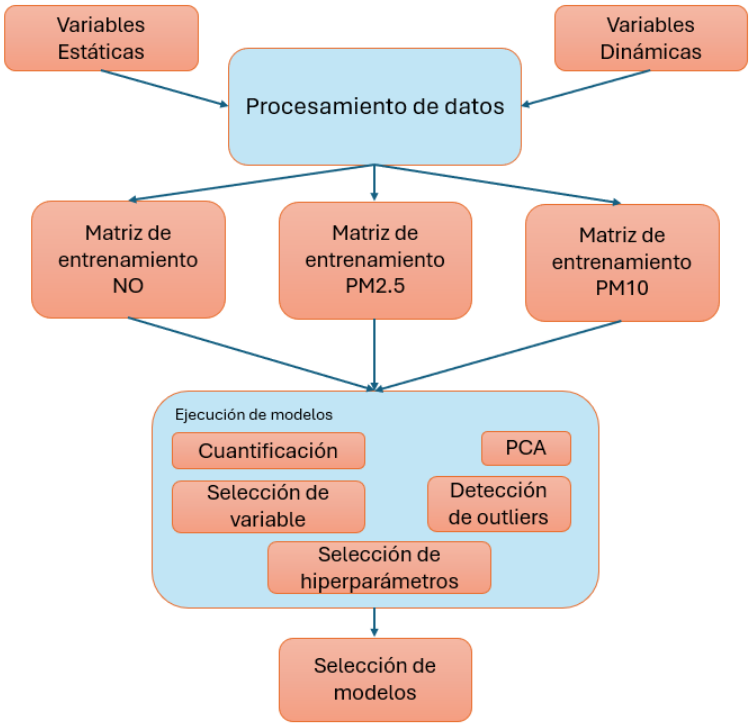

Figura 1: Desarrollo del proyecto

La lista de modelos y sus mejores valores de precisión se presenta en el apartado de resultados.

4. Resultados

Analizando los resultados representados en la Tabla 2, vemos como el modelo *Random Forest* se selecciona como el mas adecuado para la predicción de NO y PM10, mientras que el modelo *Ensamble* para PM2.5. La selección de los modelos se basa no solo en la magnitud del coeficiente de determinación r2 sino que también en su parsimonia y capacidad interpretativa. Cabe destacar que el uso de un número reducido de variables implica una mejor interpretabilidad de los modelos y, por lo tanto, puede predominar frente a otros, siempre y cuando no afecte en gran medida a su precisión.

5. Conclusiones

Los resultados obtenidos destacan la utilidad de modelos predictivos para obtener mediciones de contaminación en tiempo real sin instalar equipos específicos. Estos modelos permiten estimar emisiones a partir de variables disponibles, mejorando la precisión del mapa de calidad del aire en Madrid, actualmente limitado a 38 estaciones.

	NO	PM2.5	PM10
Decision Tree	0.7384	0.6389	0.8332
Redes Neuronales	0.7746	0.6438	0.8293
Random Forest	0.8292	0.7402	0.8959
Nearest Neighbors	0.7139	0.6828	0.8410
Xgboost	0.8428	0.7378	0.8852
Ensamble	0.8392	0.7547	0.8904
Linear Regressor	0.1764	0.3774	0.3639
KRR	0.5986	0.5519	0.6960
Lasso	0.1774	0.3772	0.3639

Tabla 2: Mejores valores de precisión de cada modelo para cada contaminante (r2)

Es importante mencionar que cualquier número de estaciones de medida que no implique desplegarlas en todas las calles de una ciudad, siempre será insuficiente, ya que la contaminación puede variar drásticamente entre calles adyacentes debido al tráfico vehicular. Es por esto que la conclusión de este trabajo cobra una relevancia particular demostrando que se puede predecir la contaminación con una precisión aceptable. Además, estos modelos facilitan la simulación de escenarios urbanos para optimizar la planificación, con el objetivo de reducir la concentra-

ción de contaminantes y sus impactos en la salud pública, mitigando problemas respiratorios y reduciendo costos sanitarios.

6. Trabajos Futuros y Discusión

Los resultados para los modelos seleccionados han superado en todo momento el 75 % de precisión a pesar de disponer de una matriz de entrenamiento limitada, debido a la escasa disponibilidad de estaciones de medida de calidad del aire en vías en las que también se registraba el tráfico. Los resultados sugieren que estos modelos podrían llegar a reemplazar los equipos de medición actuales, reduciendo costos de infraestructura y mantenimiento. Cabe destacar que esta sustitución no puede ser completa, ya que los patrones de comportamiento de los madrileños pueden variar, así como la sustitución de la tecnología de combustión por vehículos eléctricos, lo que obliga a realizar re-entrenamientos con cierta frecuencia. Sin embargo, si que permiten obtener una mayor granularidad de medidas de emisiones. Los resultados de este trabajo muestran que la posibilidad de predecir la contaminación es posible y se puede utilizar para intervenir en la planificación de zonas urbanas.

Después de profundizar en las variables predictoras y en su eficacia para la predicción de métricas de contaminación, creemos que queda lugar para la investigación y mejora de los resultados obtenidos en este trabajo. El uso de modelos mas avanzados, como las redes neuronales recurrentes, para el análisis de series temporales, la ampliación del conjunto de variables explicativas y la consideración del CO2 como contaminante altamente presente en zonas urbanas, son algunas de las medidas que pueden contribuir a un mayor desarrollo de este proyecto.

7. Bibliografía

[1] Francisco Vargas Marcos. *La contaminación ambiental como factor determinante de la salud*. 2005

[2] Yasuo Yoshikawa, Hitoshi Kunimi y Shizuo Ishizawa. "Numerical simulation model for predicting air quality along urban main roads: first report, development of atmospheric diffusion model". En: *Heat Transfer-Japanese Research: Co-sponsored by the Society of Chemical Engineers of Japan and the Heat Transfer Division of ASME* 27.7 (1998), págs. 483-496

EMISSIONS PREDICTION MODEL

Author: Beatriz Ordóñez Becker

Advisor: Luis Francisco Sánchez Merchante

Collaborating Entity: ICAI - Universidad Pontificia Comillas.

Abstract

This project focuses on developing an air quality prediction model using machine learning techniques. The primary objective is to identify the variables that contribute to the accumulation of pollutants in urban areas and predict air quality levels in real-time based on these variables.

The study involves collecting and analyzing data on atmospheric pollutants, particularly nitric oxide (NO), PM2.5, and PM10 particles, from various monitoring stations in the Community of Madrid. This data is combined with traffic information and meteorological data to create a comprehensive dataset that allows to train the prediction models. Various regression techniques and model refinement strategies are applied, including hyperparameter selection, outlier removal, and dimensionality reduction, to improve prediction accuracy.

The results aim to provide a detailed understanding of urban air pollution dynamics and offer a tool for urban planning and public health decision-making, contributing to the broader goal of creating sustainable cities.

Keywords: Pollutants, Air quality, Urban area, Prediction model, Accuracy.

1. Introduction

Today, one of the most concerning problems in our society is pollution. Although it is not a recent phenomenon, its relevance has significantly increased in recent years. The concern about pollution and its effects on the planet and our health is growing. Currently, 20 % of the total incidence of diseases can be attributed to environmental factors [1]. From the industrial revolution, which introduced new forms of production, to the current technological revolution, tools have emerged that, while making our daily lives easier, also generate a significant amount of atmospheric pollutants. These pollutants are the subject of study in this project.

2. Project Definition

The primary objective of this project is to develop an air quality prediction model using machine learning techniques. To this end, the most influential variables in the accumulation of pollutants in urban areas must be identified and measurements taken that, when applied to prediction models, allow real-time pollution level predictions. This project explores the field of air quality, contributing to the development of studies such as *"Numerical simulation model for predicting air quality along urban main roads: first report, development of atmospheric diffusion model"* [2], published in 1998. The modeling process used in this study is detailed below.

3. Modeling Process

A total of 14 variables[2] listed in Table 3 have been selected for pollutant modeling. The task of predicting the three selected pollutants follows a consistent pattern: analysis of outliers, reduction of the original dataset dimension, application of variable selection techniques, and hyperparameter tuning, always using the coefficient of determination r2 as the accuracy metric. The modeling process is described in Figure 2.

VARIABLES
Station ID
Number of lanes
Area covered by the station
Time of day
Day of the week
Wind speed
Wind direction
Temperature
Humidity
Barometric pressure
Radiation
Precipitation
Height

Tabla 3: List of variables considered in the project.

[2]The *Height* variable has been broken down into 16 different measures (detailed in chapter 3), achieving a total of 29 variables for training the models

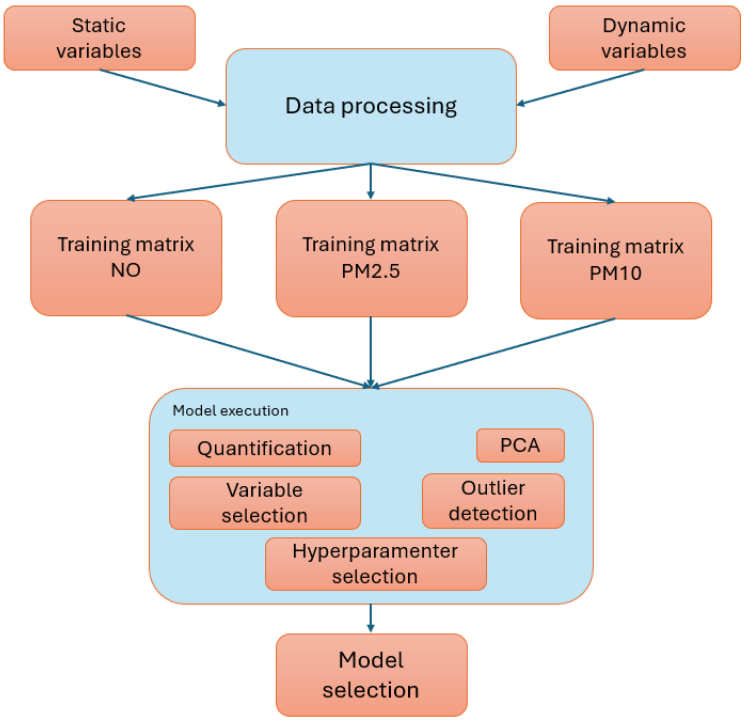

Figura 2: Project development

The list of models and their best accuracy values is presented in the results section.

4. Results

	NO	PM2.5	PM10
Decision Tree	0.7384	0.6389	0.8332
Neural Networks	0.7746	0.6438	0.8293
Random Forest	0.8292	0.7402	0.8959
K-nearest Neighbors	0.7139	0.6828	0.8410
XGBoost	0.8428	0.7378	0.8852
Ensemble	0.8392	0.7547	0.8904
Linear Regression	0.1764	0.3774	0.3639
Kernel Ridge Regression	0.5986	0.5519	0.6960
Lasso	0.1774	0.3772	0.3639

Tabla 4: Precision values of each model for each pollutant (r2)

Analyzing the results presented in Table 4, we see that the *Random Forest* model is selected as the most suitable for predicting NO and PM10, while the *Ensemble* model is used for PM2.5. Model selection is based not only on the magnitude of the coefficient of determination but also on its parsimony and interpretative capacity. It is noteworthy that using a reduced number of variables implies better model interpretability and, therefore, can prevail over others, as long as it does not significantly affect accuracy.

5. Conclusions

The results highlight the usefulness of predictive models for obtaining real-time pollution measurements without installing specific equipment. These models allow estimating emissions from available variables, improving the accuracy of the air quality map in Madrid, which is currently limited to 38 stations.

It is important to mention that any number of measurement stations that does not involve deploying them on every street in a city will always be insufficient, as pollution can vary drastically between adjacent streets due to vehicular traffic. This is why the conclusion of this work is particularly relevant, demonstrating that pollution can be predicted with acceptable accuracy. Additionally, these models facilitate the simulation of urban scenarios to optimize planning, with the aim of reducing pollutant concentrations and their impacts on public health, mitigating respiratory problems, and reducing healthcare costs.

6. Future Work and Discussion

The results for the selected models have consistently exceeded 75 % accuracy despite having a limited training matrix, due to the scarce availability of air quality measurement stations on roads where traffic was also recorded. The results suggest that these models could potentially replace current measurement equipment, reducing infrastructure and maintenance costs. It should be noted that this replacement cannot be complete, as the behavior patterns of Madrid residents may vary, as well as the replacement of combustion technology with electric vehicles, in this case periodic re-trainings are needed. However, they do allow for greater granularity in emission measurements. The results of this work show that it is possible to predict pollution and that this can be used for urban planning interventions.

After delving into the predictive variables and their effectiveness for predicting pollution metrics, we believe there is room for further investigation and improvement of the results obtained in this work. The use of more advanced models, such as recurrent neural networks for time series analysis, the expansion of the set of explanatory variables, and the consideration of CO2 as a highly prevalent

pollutant in urban areas, are some of the measures that could contribute to further development of this project.

7. Bibliography

[1] Francisco Vargas Marcos. *La contaminación ambiental como factor determinante de la salud*. 2005

[2] Yasuo Yoshikawa, Hitoshi Kunimi y Shizuo Ishizawa. "Numerical simulation model for predicting air quality along urban main roads: first report, development of atmospheric diffusion model". En: *Heat Transfer-Japanese Research: Co-sponsored by the Society of Chemical Engineers of Japan and the Heat Transfer Division of ASME* 27.7 (1998), págs. 483-496

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Uno de los múltiples problemas que nos atañen a día de hoy es la contaminación. Aunque su origen no es reciente, su importancia se podría decir que si lo es. Actualmente la preocupación acerca de la contaminación, sus efectos sobre nuestro planeta y, consecuentemente, sobre todos nosotros como habitantes del mismo, no ha hecho más que aumentar. Desde una revolución industrial que introduce nuevas formas de trabajo, hasta la revolución tecnológica que vivimos hoy en día, han aparecido en nuestras vida nuevas herramientas que sin duda facilitan la vida cotidiana, pero que también tienen un gran impacto en la emisión de contaminantes a la atmósfera.

1.1. Justificación

Dado el contexto de preocupación, este proyecto nace de la necesidad de automatizar la identificación de zonas urbanas con un alto nivel de contaminación. A lo largo de los años se han intentado definir las fuentes de partículas contaminantes con el objetivo de poder eliminarlas. Sin embargo, es de vital importancia definir aquellas variables que hacen que dichas partículas se acumulen en zonas urbanas. La apertura de la calle, la altura de los edificios o la cantidad de zonas verdes son solo algunas de las que se nos vienen a la mente. En caso de ser capaces determinar que dichas variables influyen en los niveles de contaminación, entonces no solo nos centraríamos en eliminar las fuentes de contaminación, tarea generalmente complicada, sino que en modificar aquellos detalles que contribuyen a su acumulación.

1.2. Objetivos

Acorde a lo mencionado anteriormente, el principal objetivo de este proyecto es la creación de un modelo de predicción de calidad del aire a través de Machine Learning. Es decir, ser capaces de determinar las variables mas influyentes en la acumulación de partículas contaminantes en zonas urbanas, y realizar mediciones para poder predecir lo niveles de contaminación en tiempo real. Este modelo combina la programación, mediante la recopilación y análisis de datos, con el medio ambiente y el desarrollo de ciudades sostenibles.

El Parlamento Europeo ha aprobado una ley europea del clima que tiene como objetivo principal reducir en un 55 % las emisiones netas de gases de efecto invernadero para el año 2030 [3]. Este proyecto busca contribuir a dicho de desarrollo, a la construcción de una Europa de países sostenibles. Asimismo, esta iniciativa no solo contribuye con las autoridades, sino que también con todo ciudadano. La calidad del aire se ha convertido en un problema para aquellas personas con problemas respiratorios. Este proyecto busca también proporcionar una herramienta sencilla para el acceso de todo ciudadano a los datos de calidad del aire. A día de hoy la mayoría de iniciativas es este campo se centran principalmente en la recopilación de datos en tiempo real. Sin embargo, este proyecto va un paso más allá, el objetivo es utilizar dichos datos para entrenar un modelo que sea capaz de predecir valores de calidad del aire en un entorno nuevo.

Como se ha mencionado, la información obtenida será de utilidad no solo para las autoridades encargadas de imponer medidas para la reducción de los contaminantes, sino que también para aquellos ciudadanos cuya salud se vea altamente perjudicada ante la presencia de dichos contaminantes.

Capítulo 2

Estado del arte

La calidad del aire es un problema que atañe a nivel mundial. En los últimos años el nivel de contaminación ha crecido exponencialmente y con él, el riesgo que supone para todos nosotros respirar dicho aire. Dado este contexto, unido al desarrollo de la inteligencia artificial y las herramientas para el aprendizaje automático conocido como *Machine Learning*, múltiples asociaciones han desarrollado mecanismos de detección de altos niveles de contaminación en las calles. En este estado del arte se exponen, a modo de ejemplo, una serie de iniciativas relacionadas con el tema descrito, así como un compendio de artículos relevantes que detallan la evolución de los modelos de predicción de calidad del aire a lo largo de los años. Se pretende dar una visión global sobre el tema que permita entender el contexto de este proyecto.

En el ámbito de la calidad del aire, se han empleado una gran variedad de técnicas para predecir la calidad en entornos urbanos. En el año 1999 investigadores japoneses desarrollaron un modelo centrado en vías urbanas, cuyo objetivo era predecir el efecto de diferentes características de estas como edificios, tráfico y reacciones químicas en la calidad del aire. En este ensayo se presenta la creación de un modelo centrado en el coeficiente de difusión de concentración. Coeficiente que relaciona la tasa de cambio de concentración de una sustancia en un medio con el gradiente de concentración presente en dicho medio. Este modelo resulta en un aumento de la capacidad predictiva de la calidad el aire. [2]

Mas tarde, año 2011, se publica un artículo en la Universidad Aristóteles de Tesalónica relacionado con el tema. Este artículo habla sobre la relación existente entre los contaminantes generados por el tráfico y los problemas respiratorias o incluso las muertes prematuras. Su objetivo principal es desarrollar un modelo que permita realizar previsiones de los niveles de Benceno en zonas urbanas. El resultante modelo de regresión demuestra ser capaz de captura las tendencias de concentración del benceno y permite observar una fuerte relación entre este y el

monóxido de carbono (CO) [4].

Así, en un periodo en el cual el cuidado del medio ambiente se convierte en un tema altamente relevante, se plantea la dificultad de realizar mediciones de calidad del aire de manera precisa. Un artículo publicado en el año 2013, "A method for targeting air samplers for facility monitoring in an urban environment" [5] plantea una nueva metodología para escoger los puntos de medición óptimos para la obtención de medidas más precisas en un entorno complejo y desafiante. Resalta el hecho de que múltiples herramientas de medida requieren de unas determinadas condiciones climáticas combinadas con un modelado y simulación precisas y detalladas, lo cual resulta ser difícil de lograr e influye en la fiabilidad de las mediciones. La dirección del viento generada por los edificios cercanos o la cantidad de zonas verdes son solo algunos de los muchos factores influyentes. En resumen, esta nueva metodología permite una evaluación más precisa del transporte y la dispersión de contaminantes atmosféricos en entornos urbanos al considerar una amplia gama de condiciones meteorológicas.

A medida que pasa el tiempo, se empiezan a crear múltiples y variados modelos de predicción de calidad del aire. La mayoría de ellos a través del aprendizaje supervisado. El artículo "Ambient Air Quality Estimation using Supervised Learning Techniques" [6] publicado en el 2019, presenta una serie de modelos basados en técnicas de clasificación, regresión y ensamble. En él se establecen los métodos de árboles de decisión, regresión de vectores de soporte y "Stacking Ensemble", como los más eficaces dentro de los estudiados.

En el 2020, el artículo "Prediction of air quality based on Gradient Boosting Machine Method" [7] particulariza en el uso de Light Gradient Boosting Machine y eXtreme Gradient Boosting para predecir la calidad del aire en Beijing. Tras medir la eficacia de ambos modelos se determina que Light GBM posee mayor precisión y eficiencia para la medición del PM2.5.

Ese mismo año se publica el artículo "A Machine Learning Approach to Predict Air Quality in California" [8] que utiliza también Regresión de vectores de soporte (SVR) para predecir las concentraciones de contaminantes por hora en California. Este estudio resulta obtener mejores resultados considerando todo el conjunto de variables propuestas en vez de realizando un análisis de componentes principales. Como se puede ver cada estudio centra sus esfuerzos en determinar el modelo de aprendizaje supervisado que mejor se ajuste a su conjunto de datos. Dichos modelos determinan la relevancia de las variables propuestas como influyentes en la calidad del aire.

Asimismo, el auge de las técnicas basadas en redes neuronales ha potenciado que muchas de las publicaciones más recientes se hayan centrado en su aplicación al ámbito de la predicción de la calidad del aire. Esto se ve reflejado en artículos como "Artificial neural network model for ozone concentration estimation and Monte Carlo analysis"[9] o "Air Quality Prediction Using Improved PSO-BP Neural Network" [10]. El primero de ellos se centra en investigar la viabilidad de utilizar redes neuronales, con parámetros meteorológicos como variables de entrada, para predecir las concentraciones de ozono en el área urbana de Jinan, China. El segundo profundiza en el tema desarrollando un modelo de redes neuronales de retropropagación o propagación hacia atrás, basado el algoritmo de optimización de enjambre de partículas. Modelo que posee una rápida convergencia a la solución óptima.

Estos estudios demuestran colectivamente la diversa gama de técnicas utilizadas para predecir la calidad del aire en entornos urbanos, y son solo algunos de los proyectos desarrollados en pro de la creación de ciudades sostenibles. Como se puede ver, este es un tema muy actual y en pleno auge que requiere de la colaboración de todos nosotros como ciudadanos. Estos proyectos no solo contribuyen a la toma de decisiones sobre planificación urbana y transporte sostenible, sino que también contribuyen a la creación de mapas de calidad del aire en tiempo real.

Cabe destacar que los esfuerzos por mitigar la acumulación de contaminantes en zonas urbanas no se limita al área de la investigación. La Fundación Gas Natural Fenosa, en su propósito de difundir y crear conciencia sobre temas relacionados con la energía, el medio ambiente y la sostenibilidad, lleva mas de 15 años realizando publicaciones y seminarios por España y América Latina [11]. Su preocupación e interés en la calidad del aire queda reflejado en artículos como "Calidad del aire urbano, salud y tráfico rodado"[12] del año 2006 o "Mejora de la calidad del aire por cambio de combustible a gas natural en automoción. Aplicación a Madrid y Barcelona"[13] año 2009. Asimismo, en el año 2018, Fundación Gas Natural Fenosa publica el libro "La calidad del aire en las ciudades. Un reto mundial", este libro recalca una vez mas el problema de la calidad del aire, subrayando que "se trata de un problema tan importante como el cambio climático y con efectos adversos más inmediatos, cuya solución requiere de la colaboración de todos para ser mitigado"[11].

Capítulo 3

Fundamentos teóricos

En este capítulo abordaremos los fundamentos teóricos que sustentan este proyecto, cuyo objetivo es predecir las emisiones contaminantes utilizando modelos de inteligencia artificial. Comenzaremos con una introducción a los contaminantes y tipos de partículas que más preocupan en los entornos urbanos, entendiendo su origen, clasificación y los efectos que tienen sobre el medio ambiente y la salud pública. Esta comprensión es esencial para configurar adecuadamente los modelos predictivos y garantizar su relevancia y precisión.

Posteriormente, profundizaremos en la modelización aplicada, explorando las diversas técnicas de regresión que nos permitirán establecer relaciones entre las variables de entrada y las emisiones contaminantes como variable de salida. Dentro de este contexto, describiremos los tipos de modelos utilizados, detallando sus características, ventajas y limitaciones. Además, examinaremos diferentes técnicas para mejorar la precisión de los modelos como la elección de hiperparámetros, o técnicas de reducción de dimensión como la selección de variables o las técnicas de proyección a baja dimensión. También estudiaremos los efectos de la eliminación de outliers y la cuantificación de la variable *target*. Estos pasos son cruciales para afinar nuestros modelos y asegurar predicciones precisas y fiables. Con este marco teórico, sentamos las bases para desarrollar modelos de IA capaces de abordar la complejidad y la importancia crítica de predecir las emisiones contaminantes, contribuyendo así a esfuerzos más amplios para mitigar el impacto ambiental.

3.1. Contaminantes y tipos de Partículas

La emisión de contaminantes a la atmósfera produce lo que hoy en día conocemos como contaminación atmosférica, problema destacado por la Organización Mundial de la Salud como una prioridad. Esta se refiere a la "presencia en la atmósfera de materias, sustancias o formas de energía que impliquen molestia gra-

ve, riesgo o daño para la seguridad o la salud de las personas, el medioambiente y demás bienes de cualquier naturaleza."[14]

La contaminación atmosférica puede tener un origen tanto natural, derivado de procesos naturales, como antropogénico, es decir, resultado de la actividad humana. En esta ocasión, nos centraremos en los contaminantes de origen antropogénico y en los factores naturales que pueden influir en su acumulación a nivel de calle. Estos contaminantes emanan tanto de fuentes móviles como de fuentes fijas de combustión, y adquieren particular importancia a partir de la revolución industrial. En este periodo surgen fábricas y maquinaria que reemplazan métodos de producción artesanales y manuales, así como nuevas tecnologías que aunque facilitan enormemente nuestra vida diaria, tienen un impacto negativo en nuestra salud y la del planeta. En esta sección nos referiremos a estudios que confirman los efectos de la exposición a ellos, afirmando su impacto negativo en especial con mayor severidad en los sistemas respiratorio y cardiocirculatorio. [15]

Los tipos de contaminantes existentes a día de hoy son numerosos. Sin embargo, en este proyecto se recopila información sobre tres de ellos: Monóxido de Nitrógeno (NO), Partículas menores a 2.5 µm (PM2.5) y Partículas menores a 10 µm (PM10).

3.1.1. Óxido Nítrico

El Monóxido de Nitrógeno u Óxido Nítrico pertenece a la familia de contaminantes conocida como óxidos de nitrógeno. Estos gases son emitidos tanto desde fuentes naturales como desde fuentes atropogénicas, siendo la combustión de motores de diésel su principal origen [16]. En la actualidad, los óxidos de nitrógeno son considerados como uno de los principales contaminantes atmosféricos. Así, a día de hoy, se llevan y se han llevado acabo múltiples estudios centrados en sus efectos sobre la salud. Estos estudios permiten observar como los datos registrados oscilan alrededor del límite establecido por la normativa europea y española, y alertan de los riesgos que suponen dichos niveles [14]. La presencia de estas sustancias en el ambiente puede desencadenar una variedad de enfermedades, desde trastornos respiratorios como la reducción de la función pulmonar, el asma, la bronquitis y el cáncer de pulmón, hasta problemas cardiovasculares como la cardiomegalia y el colapso circulatorio. Además, pueden causar irritaciones en los ojos y la piel. En situaciones más graves, esta exposición puede incluso llevar a un desenlace fatal prematuro.[16]

3.1.2. Partículas PM

En lo que respecta a las partículas PM, partículas finas y respirables en suspensión en el aire, comparten el mismo origen que las mencionadas anteriormente. Estas sustancias son liberadas directamente a la atmósfera de forma natural o como resultado de la actividad humana. El tráfico en áreas urbanas constituye la principal fuente de estas emisiones, sin embargo, también hay otros orígenes como pueden ser los procesos de combustión derivados de diversas actividades industriales y de construcción.

Dentro de este conjunto de partículas contaminantes diferenciamos distintos grupos acordes a los diferentes diámetros que poseen. Las partículas PM2.5 tienen un diámetro igual o inferior al 2.5 micrómetros, mientras que las PM10 igual o inferior a 10 micrómetros.

Al igual que los óxidos de nitrógeno, estudios realizados en España permiten observar la influencia de estas partículas en nuestra salud. Concretamente detallan, en el caso las partículas PM10, como ün incremento de 10 µg/m3 en sus niveles atmosféricos se asocia con un aumento de un 0,2 a un 1 % en la mortalidad por todas las causas y un 0,5 a un 2 % en la mortalidad cardiorrespiratoria..^simismo, en cuanto a las PM2.5, demuestran su asociación con la mortalidad por todas las causas, la mortalidad por enfermedades del aparato circulatorio y por cáncer de pulmón.[14]

Como se puede ver, todos estos contaminantes tienen orígenes comunes que serán objeto de estudio en este proyecto. Se intentará predecir la cantidad de estas partículas contaminantes, emitidas por vehículos y factores naturales presentes en las calles. Para ello es necesario tener en cuenta, como se ha mencionado anteriormente, todos aquellos factores meteorológicos y urbanísticos que influyen en su acumulación o dispersión a nivel de calle.

3.2. Modelización aplicada a la Regresión/Técnicas de Regresión

Para poder entender la modelización es necesario abordar el concepto de "Machine Learning". Este campo de la inteligencia artificial, está compuesto por una serie de algoritmos con capacidad de aprendizaje. Algoritmos que, gracias a esta habilidad, son capaces de identificar patrones de comportamiento en conjuntos de datos, lo que a su vez permite crear modelos para abordar problemas más complejos.

En él encontramos dos tipos de aprendizaje, supervisado y no supervisado. El aprendizaje supervisado utiliza conjuntos de datos etiquetados mientras que

el aprendizaje no supervisado utiliza conjuntos de datos no etiquetados. Dada la naturaleza del problema que este proyecto quiere resolver, la existencia de variables definidas que aparentemente influyen en la contaminación atmosférica, nos enfocaremos en el aprendizaje supervisado. El objetivo será determinar el grado de influencia de cada una de ellas en la contaminación atmosférica.

El aprendizaje supervisado es capaz de resolver problemas de clasificación y de regresión. La clasificación intenta asociar los datos de salida a una determinada clase o grupo. Por otro lado, la regresión proporciona valores continuos que tienen significado fuera de una clase o grupo. En este proyecto se quiere predecir un valor de contaminación, es por ello que utilizaremos modelos de regresión para la combinación de todas las variables. La regresión es un proceso estadístico que permite analizar la relación entre dos o más variables, siendo una de ellas dependiente del resto. Nos centraremos entonces en modelos de regresión que nos ayuden a predecir nuestra variable dependiente, la calidad del aire. Modelos que serán descritos haciendo uso de la herramienta Scikit-learn [17].

Antes de adentrarnos en la descripción de los modelos, detallaremos brevemente el proceso llevado acabo para la construcción y aplicación de cada uno de ellos. Los datos recolectados se organizan en una matriz, detallando claramente sus filas y columnas, y este extenso conjunto de datos se divide en dos subconjuntos en una proporción de 80/20, cada uno con un propósito específico: uno para entrenamiento y otro para pruebas. El conjunto de entrenamiento se emplea, como su nombre lo sugiere, para capacitar al modelo. Una vez que el modelo ha sido debidamente entrenado, se procede a evaluar su eficacia con el conjunto de pruebas. Para evaluar dicha eficacia, se recurre al coeficiente de determinación r2 (R cuadrado), medida estadística que indica la proporción de varianza en la variable target que se puede predecir a partir del resto de variables. Este medidor calculará la diferencia entre el valor real y el predicho dentro del conjunto de test.

3.2.1. Tipos de Modelos (Descripción de los usados)

- Decision Tree Regressor model

  Modelo con la capacidad de predecir el valor de una variable dependiente al aprender reglas de decisión simples derivadas de las características presentes en los datos. Estas reglas de decisión se estructuran en forma de árbol, donde cada nodo representa una pregunta sobre una característica específica de los datos. A medida que se desciende por el árbol, se formulan más preguntas y se toman decisiones más detalladas. Esto permite segmentar el espacio de

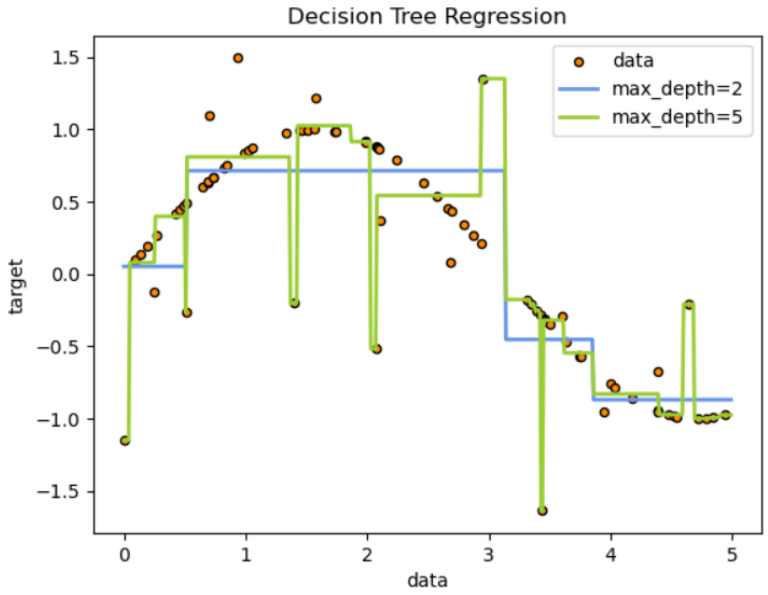

Figura 3.1: Decision Tree Regressor model [17]

características en regiones más pequeñas y definir relaciones más precisas entre las variables de entrada y la variable de salida.

Cuanto más profundo sea el árbol, más complejas serán las reglas de decisión y más ajustado será el modelo a los datos de entrenamiento. Sin embargo, es importante considerar que un árbol muy profundo puede llevar a un sobreajuste, donde el modelo se adapta excesivamente a los datos de entrenamiento y no generaliza bien a nuevos datos.

Para evitar el sobreajuste, es necesario ajustar hiperparámetros del modelo como la profundidad y el número mínimo de muestras, para lograr un equilibrio adecuado. En la Figura 3.1 vemos como se ajusta este modelo a un conjunto de datos acorde a distintas profundidades.

- Random Forest Regressor model

  Este modelo de aprendizaje automático se centra en construir un conjunto de árboles de decisión utilizando muestras aleatorias de datos. Así, cada árbol en el conjunto vota por una predicción, y la predicción final se determina mediante el promedio de todas estas votaciones.

  La ventaja principal de este modelo radica en su habilidad para mejorar la precisión y evitar el sobreajuste al combinar múltiples árboles de decisión, entrenados cada uno de ellos con diferentes subconjuntos de datos.

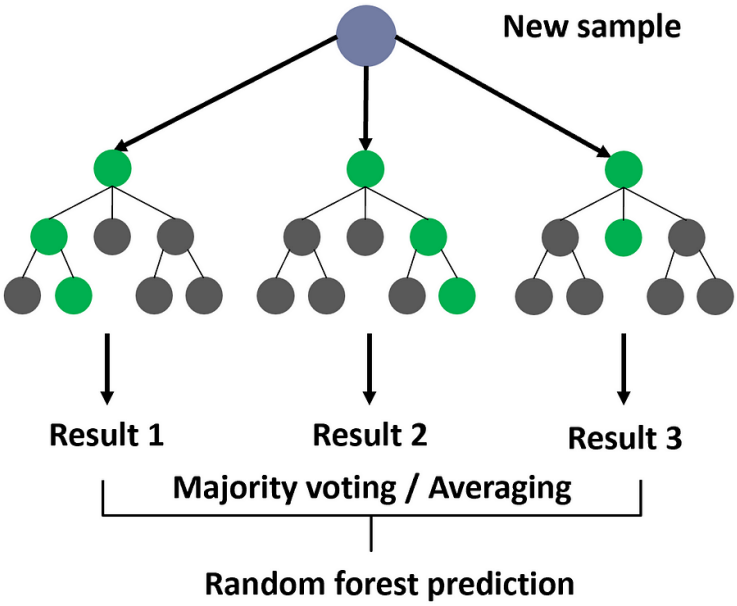

Figura 3.2: Random Forest [18]

De esta forma, el modelo Random Forest Regressor representado en la Figura 3.2, tiende a tener un mejor rendimiento en comparación con un solo árbol de decisión. Al promediar las predicciones de múltiples árboles, este modelo también puede proporcionar estimaciones más estables y robustas.

- Xgboost model

  Gradient Boosting para regresión es un algoritmo sofisticado de aprendizaje automático que construye un modelo predictivo mediante la combinación de múltiples árboles de regresión de manera iterativa. En este proceso, se optimiza una función de pérdida diferenciable en cada paso con el objetivo de mejorar continuamente las predicciones del modelo.

  La característica distintiva de Gradient Boosting es su capacidad para mejorar progresivamente las predicciones mediante la construcción secuencial de árboles de regresión, donde cada nuevo árbol se ajusta para corregir los errores del modelo en etapas anteriores.

  Asimismo, al optimizar la función de pérdida de forma diferenciable, el algoritmo es capaz de aprender de manera eficiente y efectiva, incluso en conjuntos de datos de gran escala y alta dimensionalidad. En la Figura 3.3 vemos una representación del funcionamiento de este modelo.

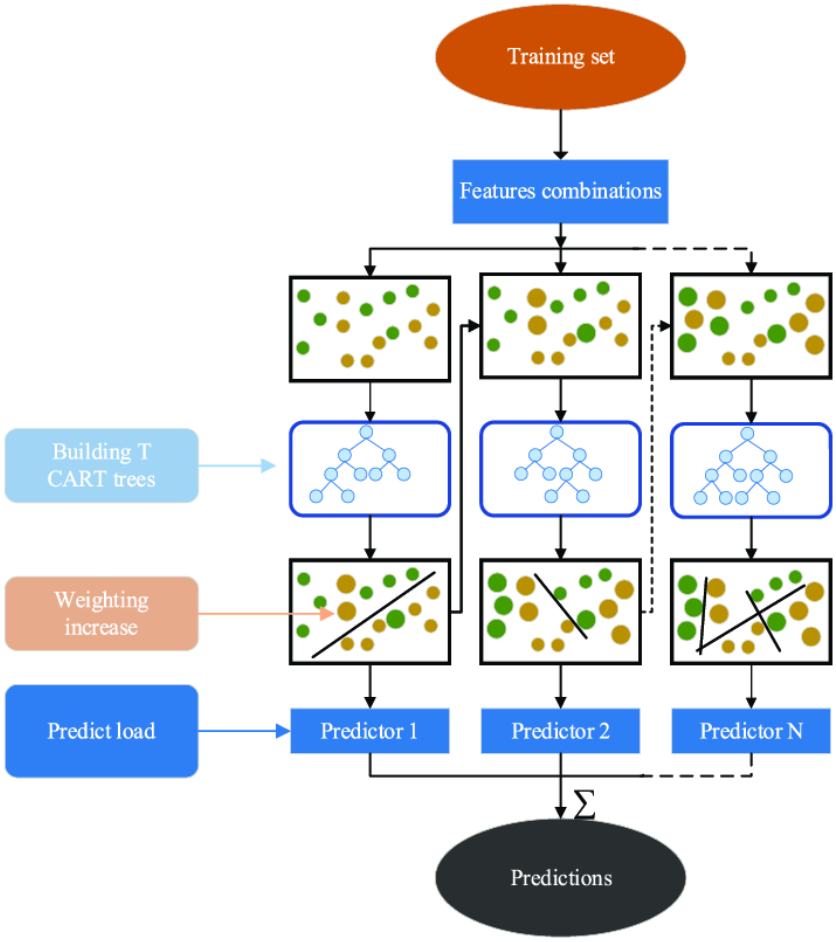

Figura 3.3: Xgboost [19]

- Linear Regressor model

  La regresión lineal es un método estadístico que ajusta un modelo lineal con coeficientes, para minimizar la suma residual de cuadrados entre los valores observados en el conjunto de datos y los valores predichos por la aproximación lineal. Proceso representado en la Figura 3.4. Este enfoque se basa en la suposición de que existe una relación lineal entre la variable dependiente y una o más variables independientes. El objetivo principal de la regresión lineal es estimar los coeficientes de la ecuación lineal que mejor se ajusten a los datos observados.

  La regresión lineal utiliza métodos de optimización para encontrar dichos valores. Se ajustan iterativamente los coeficientes del modelo hasta que se alcanza un punto en el que la suma residual de cuadrados es mínima. Una vez que se ha ajustado el modelo, se pueden utilizar los coeficientes estimados

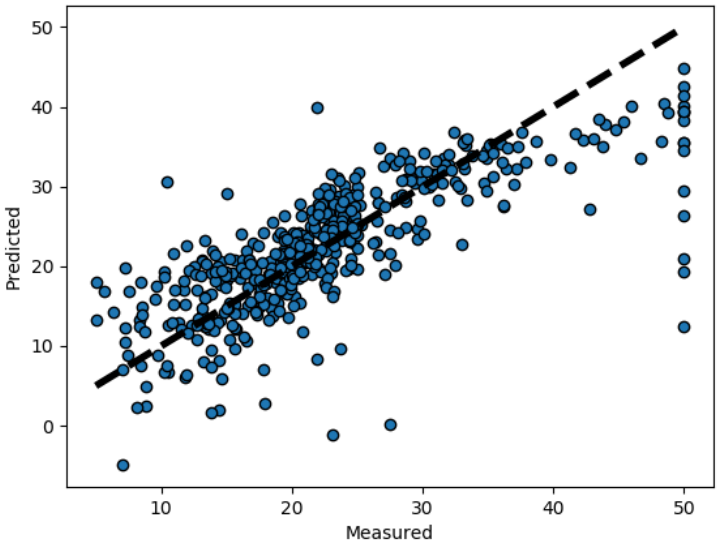

Figura 3.4: Linear Regressor [17]

para predecir los valores de la variable dependiente para nuevas observaciones.

Aunque la regresión lineal es un método simple y fácil de interpretar, es importante tener en cuenta sus limitaciones. Por ejemplo, la regresión lineal asume una relación lineal entre las variables, lo que puede no ser apropiado en todos los casos. Además, la regresión lineal puede ser sensible a valores atípicos en los datos y puede no capturar relaciones no lineales entre las variables. Sin embargo, en muchos casos, la regresión lineal sigue siendo un punto de partida útil para el análisis de datos y la construcción de modelos predictivos.

- Lasso model

  Este modelo se centra en estimar coeficientes dispersos. Matemáticamente es un modelo lineal con una termino de regularización adicional, modelo que resuelve así la minimización de la penalización de mínimos cuadrados. Este término de regularización penaliza la magnitud de los coeficientes de regresión, lo que ayuda a prevenir el sobreajuste y promueve la selección de características importantes al forzar algunos coeficientes a ser exactamente cero. En la Figura 3.5 vemos como varia el comportamiento del modelo al incluir el término de penalización.

  Este enfoque es especialmente útil cuando se trabaja con conjuntos de da-

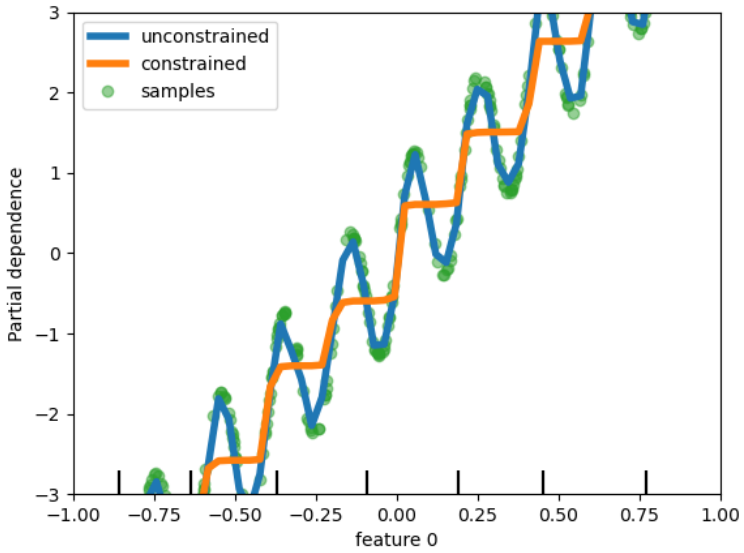

Figura 3.5: Lasso [17]

tos de alta dimensionalidad, donde el número de características es grande en comparación con el número de muestras. Al introducir la regularización, el modelo puede manejar eficazmente la multicolinealidad entre las características y seleccionar automáticamente las más relevantes, lo que conduce a modelos más simples e interpretables. Además, la capacidad de establecer algunos coeficientes en cero permite una selección eficiente de características.

- Modelo de Redes Neuronales

  Modelo que, a diferencia de la regresión lineal, permite capturar relaciones de datos no lineales utilizando arquitecturas de redes neuronales. Un modelo que consta de capas de neuronas conectadas entre sí, donde cada capa combina ponderadamente sus entradas y aplica una función no lineal a dicha combinación. Así se propagan las combinaciones hasta llegar a la final.

  Esta estructura en capas permite que las redes neuronales sean capaces de aprender y representar relaciones complejas entre las variables de entrada y salida. Además, la capacidad de ajustar automáticamente los pesos de conexión durante el entrenamiento permite que las redes neuronales se adapten a una amplia variedad de problemas de aprendizaje. Este modelo de capas queda representado en la Figura 3.6.

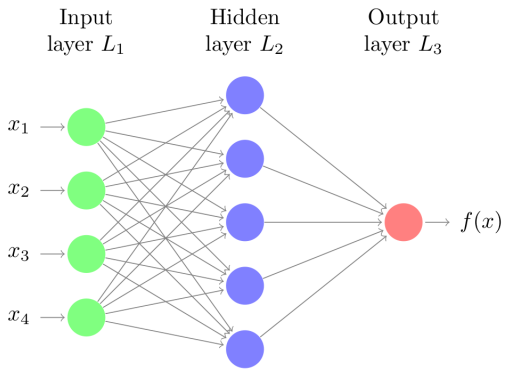

Figura 3.6: Redes Neuronales [20]

- Ensamble model

  Modelo que combina las predicciones de varios estimadores con el objetivo de mejorar los resultados de un único estimador. En el contexto de este proyecto, el modelo ensamble combina las predicciones de modelos como Xgboost y Random Forest, logrando mejorar considerablemente los resultados.

  Al combinar sus predicciones, el modelo de ensamble puede reducir el sesgo y la varianza del modelo final, lo que conduce a una mejor generalización y rendimiento predictivo en datos nuevos. Esta estrategia es ampliamente utilizada en la práctica para mejorar la robustez y la precisión de los modelos de aprendizaje automático.

- Nearest Neighbours model

  El principio tras el método de predicción de K-Nearest Neighbors (K-NN) radica en su enfoque intuitivo de encontrar un conjunto predefinido de muestras de entrenamiento que están cercanas al punto que se desea predecir, como se puede ver en la Figura 3.7. Esta proximidad se determina en función de la distancia entre los puntos en un espacio de características multidimensional. Una vez identificados estos vecinos más cercanos, el modelo realiza la predicción basándose en las etiquetas de estas muestras.

  En el contexto específico de la regresión, donde las etiquetas de los datos son continuas en lugar de discretas, el método K-NN calcula la media de los valores de las etiquetas de los vecinos cercanos para predecir la etiqueta del nuevo punto. Este enfoque se basa en la suposición de que los puntos cercanos en el espacio de características tendrán etiquetas similares, lo que lo hace adecuado para problemas donde existe una relación de proximidad en los datos. Sin embargo, es importante considerar cuidadosamente la elección del parámetro "k", que representa el número de vecinos considerados, ya que

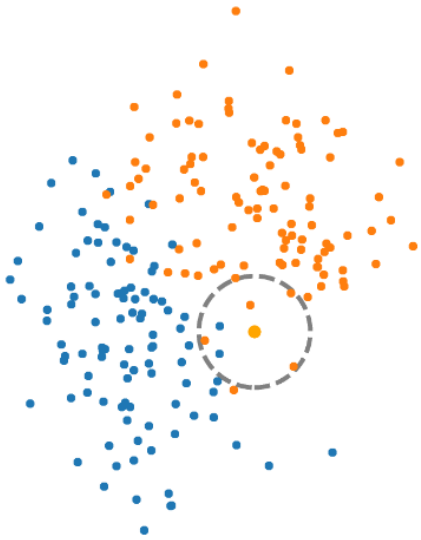

Figura 3.7: Vecinos cercanos [21]

un valor incorrecto puede llevar a un sesgo o a una varianza no deseada en las predicciones.

- Kernel Ridge Regressor model (KRR)

  La regresión de ridge con kernel (KRR) es una técnica que combina la regresión de ridge con el truco del kernel, permitiendo aprender una función que puede ser lineal o no lineal en función de los datos y del kernel utilizado. Esta combinación ofrece flexibilidad en la modelización de relaciones complejas entre las características y la variable objetivo, lo que la hace especialmente útil en conjuntos de datos donde las relaciones son no lineales o de alta dimensionalidad. El uso del truco del kernel permite mapear los datos a un espacio de características de mayor dimensión, donde la relación entre las características y la variable objetivo puede ser más fácilmente capturada por un modelo lineal en ese espacio transformado.

  KRR utiliza la pérdida de error cuadrático como función de pérdida durante el entrenamiento, y puede ajustarse de forma más rápida para conjuntos de datos de tamaño mediano en comparación con otros métodos más intensivos computacionalmente. Esto lo convierte en una opción atractiva cuando se trabaja con datos de tamaño moderado y se requiere un equilibrio entre rendimiento y tiempo de entrenamiento. Sin embargo, es importante tener en cuenta que los modelos KRR tienden a ser más lentos en la fase de predicción. En la Figura 3.8 se pueden ver los tiempos de ejecución de este modelo

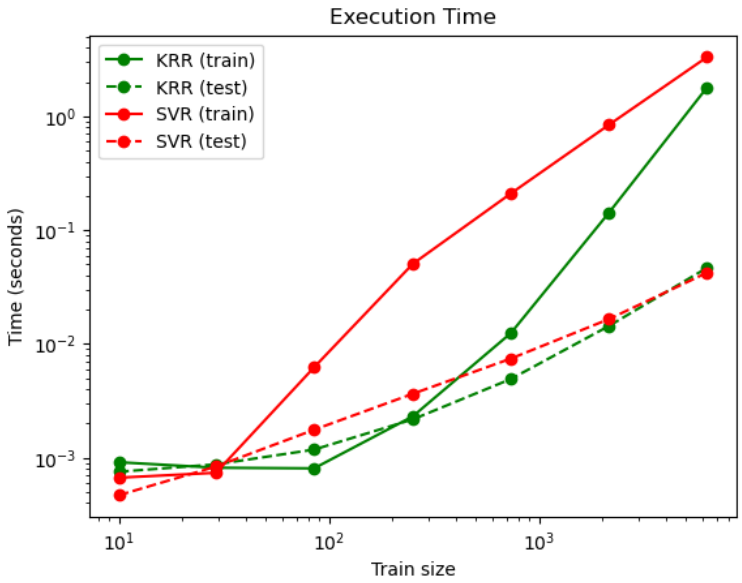

Figura 3.8: KRR [17]

comparado con el modelo SVR (*Support Vector Regression*).

En resumen, la regresión de ridge con kernel ofrece una herramienta versátil para la modelización de relaciones no lineales en conjuntos de datos de tamaño moderado.

Una vez discutidos los modelos empleados, nos adentraremos en el proceso de entrenamiento destinado a predecir la calidad del aire en las calles.

3.3. Refinamientos de Modelos

En la sección dedicada al refinamiento de modelos, nos centraremos en las estrategias clave que incrementan la precisión y la eficiencia de nuestros modelos de inteligencia artificial destinados a la predicción de emisiones. Comenzaremos abordando la eliminación de outliers, una etapa esencial para limpiar nuestros datos y evitar sesgos que podrían desviar las predicciones de la realidad.

La selección de hiperparámetros constituye también un paso fundamental para optimizar el rendimiento de los modelos. Aquí exploraremos cómo ajustar estos parámetros puede significativamente alterar la capacidad de predicción de los modelos, equilibrando entre el overfitting y el underfitting para lograr una generalización óptima. Proseguiremos con la reducción de dimensión, donde discutiremos tanto la selección de variables críticas como las técnicas de proyección, ambas

dirigidas a simplificar los modelos sin sacrificar su capacidad predictiva. Estas estrategias no solo mejoran la velocidad y eficiencia de los modelos, sino que también facilitan la interpretación de los resultados.

Finalmente, abordaremos la cuantificación, un paso que nos permite evaluar y mejorar la precisión de nuestras predicciones. A través de este enfoque multifacético para el refinamiento de modelos, apuntamos a desarrollar herramientas robustas y eficaces en la lucha contra la contaminación ambiental, respaldando nuestros esfuerzos con una base sólida de técnicas avanzadas de procesamiento y análisis de datos.

3.3.1. Selección de hiperparámetros

Los hiperparámetros son configuraciones ajustables que se eligen para entrenar el modelo y que rigen proceso de entrenamiento [22]. Cada uno de los algoritmos de aprendizaje automático o modelos, posee una serie de parámetros que pueden ser ajustados para mejorar su capacidad predictiva y velocidad de entrenamiento. El entendimiento del impacto de cada parámetro en el rendimiento del modelo es vital para seleccionar la mejor configuración y así obtener los mejores resultados que cada modelo pueda ofrecer. Las técnicas de ajuste de hiperparámetros permiten obtener de manera automática las configuraciones que optimizan la exactitud de las estimaciones, estas incluyen técnicas de búsqueda exhaustivas, aleatorias y otras [22]. En este proyecto, la selección de hiperparámetros se lleva acabo comparando los valores de precisión del modelo ante cada combinación de valores de los parámetros. Dichos valores se han establecido cuidadosamente teniendo en cuanta el significado de cada parámetro, siendo el objetivo principal maximizar la precisión de cada uno de los modelos.

3.3.2. Eliminación de Outliers

Otro aspecto a considerar en el refinamiento de los modelos es la eliminación de outliers. Los outliers son datos u observaciones que difieren numéricamente en gran medida del resto del conjunto de datos. Estos valores anómalos pueden sin duda entorpecer el entrenamiento de los modelos introduciendo ruido que distorsione las medidas estadísticas de cada variable. Una de las técnicas más populares de eliminación de outliers, por su sencillez, requiere establecer un rango numérico de valores considerados aceptables, de esta manera, todos los valores que se salgan de ese rango serán considerados outliers. Este rango no esta universalmente definido, depende en gran medida de la cantidad de datos obtenidos, la media y la moda de sus valores. En el contexto de este proyecto se establece la frontera de outliers como una variación de tres veces la desviación estándar de cada variable. En la Figura 3.9 queda representada esta técnica.

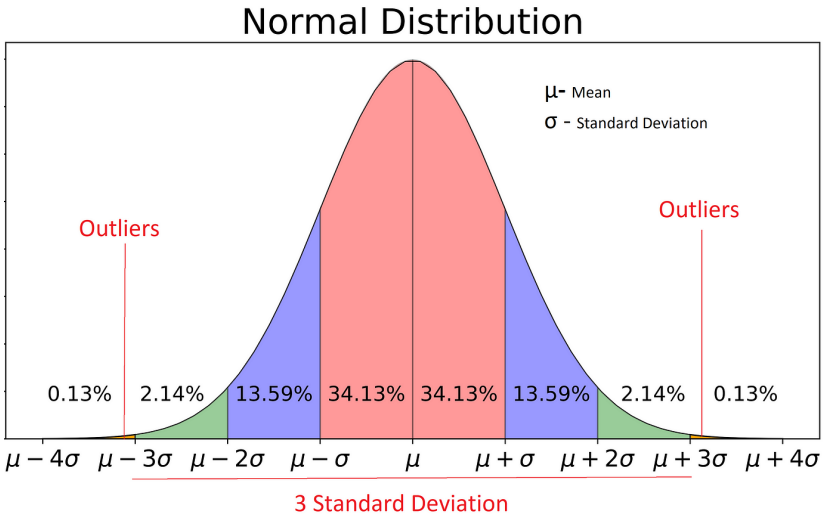

Figura 3.9: Imagen representativa del proceso de detección y eliminación de datos atípicos, conocidos como outliers. [23]

3.3.3. Reducción de dimensión

Selección de variables

Como se mencionará en capítulos posteriores, la selección de variables hecha para la predicción de calidad del aire es extensa. Sin embargo, siendo el objetivo del proyecto determinar cuales de ellas intervienen en la acumulación de partículas contaminantes, es lógico que algunas resulten significativamente menos relevantes que otras. En tal caso, dichas variables entorpecen el rendimiento del modelo y deben ser eliminadas. Esto es lo que conocemos como reducción de variables. Cabe destacar que la importancia de las variables varia en cada modelo, por ello, este análisis se lleva acabo en todos ellos de manera independiente. En nuestro caso, la selección de variables se ha realizado evaluando la variación del coeficiente de determinación (r2) ante la eliminación de variables de manera iterativa. Así, se obtiene la combinación que mas favorece la predicción.

Proyección a baja dimensión

Existen numerosas técnicas de proyección a baja dimensión. La más conocida es el Análisis de Componentes Principales (PCA). La Figura 3.10 muestra un ejemplo. Esta técnica pretende convertir un conjunto de variables correlacionadas en un conjunto de variables no correlacionadas, llamadas componentes principales. Así, la primera componente principal captura la mayor cantidad de variabilidad en los datos, la segunda la siguiente mayor cantidad de variabilidad y así sucesivamente

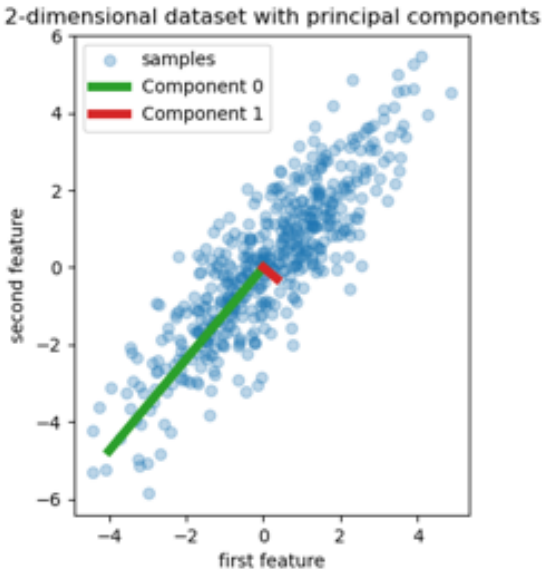

Figura 3.10: Imagen representativa del proceso de proyección a baja dimensión conocido como PCA. [17]

hasta llegar al número de componentes principales deseado. Esta técnica permite visualizar los datos de manera mas clara, intentando mantener la mayor cantidad de información posible, lo cual puede ayudar al rendimiento de los modelos.

3.3.4. Cuantificación

La cuantificación en el contexto de este proyecto se podría definir como el proceso de agrupamiento de valores en intervalos reducidos. Así, cada dato lleva asociado una etiqueta correspondiente al intervalo al que pertenece. A modo de ejemplo, la calidad del aire se podría dividir en tres intervalos: Calidad del aire baja, media o alta. Si definimos, en el caso del NO, que un valor menor o igual 50 µg/m3 representa un bajo nivel de contaminación, entonces nuestro modelo deberá asignar cada valor inferior a 50 µg/m3 dentro de este intervalo. De esta forma, el rendimiento del modelo puede llegar a mejorar notablemente dado que pasamos de intentar predecir un valor exacto de contaminación a simplemente intentar predecir en que intervalo se encuentra. Esta técnica queda representada en la Figura 3.1.

Valores (ug/m3)	Etiqueta
0 - 10	Contaminación insignificante
10 - 20	Contaminación baja
20 - 30	Contaminación media
30 - 40	Contaminación elevada

Tabla 3.1: Valores de cuantificación para el contaminante PM2.5.

Todas estas técnicas se probarán para el conjunto de datos del proyecto y todos los modelos mencionados, con el objetivo de realizar una predicción precisa que represente de manera objetiva el valor de contaminación en una calle.

Capítulo 4

Orígenes de datos

Para lograr entender el origen de los datos es necesario hablar en primer lugar de la naturaleza del modelo que se pretende crear. La calidad del aire en entornos urbanos depende de una serie de variables que podemos dividir en estáticas y dinámicas. Estáticas son todas aquellas variables cuyo valor no depende del instante de tiempo en el que se ha realizado la medición. Por otro lado, las variables dinámicas son aquellas que si dependen de dicho instante de tiempo, por ello su obtención es más compleja.

La forma de recopilar estos dos tipos de variables difiere enormemente. Las estáticas se han recopilado manualmente mientras que las dinámicas a través del portal de datos abierto de la Comunidad de Madrid [24]. Este portal, del que hablaremos en mayor profundidad, proporciona datos en tiempo real de mediciones realizadas por estaciones situadas en múltiples puntos de la comunidad.

Como se ha mencionado, esta investigación se centrará en la Comunidad de Madrid y sus estaciones de medida, estaciones que publican sus datos periódicamente. Para lograr predecir la calidad del aire se han seleccionado dos tipos de variables dinámicas: tráfico y meteorología. Por lo tanto, serán necesarios datos no solo de las estaciones de medida de calidad del aire (datos necesarios para entrenar el modelo) sino que también de las estaciones de tráfico y meteorología. Se seleccionarán entonces todas las ubicaciones que posean información acerca de estas tres variables y posteriormente se recopilarán las variables estáticas asociadas a ellas.

Tras haber definido los tipos de variables, se detallará la relación de cada una de ellas con la variable que se quiere predecir (conocida como variable target), así como los criterios en base a los cuales se han obtenido.

4.1. Variable Objetivo: Calidad del aire

Las estaciones de calidad del aire permiten conocer en cada momento los datos de contaminación atmosférica, proporcionan datos de múltiples contaminantes. Sin embargo, en el contexto de este proyecto, se han seleccionado tres por ser los presentes en un mayor número de estaciones: Monóxido de Nitrógeno (NO), Partículas menores a 2.5 µm (PM2.5) y Partículas menores a 10 µm (PM10).

Los datos obtenidos se actualizan cada hora entre los minutos 20 y 30. Así, se corresponden con la media de los 6 valores diezminutales existentes en una hora. En la Figura 4.1 se puede ver la ubicación de cada una de estas estaciones en el mapa. Nuestro modelo de predicción de calidad del aire se centra en la mediciones a nivel de calle. Es por ello que dado el mapa anterior y la clasificación de la ubicaciones, es necesario descartar todas aquellas estaciones situadas en zonas suburbanas. Estas estaciones no proporcionarán una medición real ya que se encuentran en zonas con intensidad de tráfico nula y no poseen información sobre el resto de variables determinantes. Las estacionadas seleccionadas quedan reflejadas en la Tabla 4.1.

Asimismo, es importante tener en cuenta la estructura de los datos recopilados, Figura 4.2. El campo *punto de muestreo* indica el código completo de la estación (provincia, municipio y estación) así como la magnitud de la medición y la técnica de muestreo. De la misma forma, los campos *HO1* Y *V01* se corresponden con el dato de la 1 de la mañana de ese día y su código de validación respectivamente. Así sucesivamente las 24 horas del día. Únicamente son válidos aquellos datos con código de validación "V".

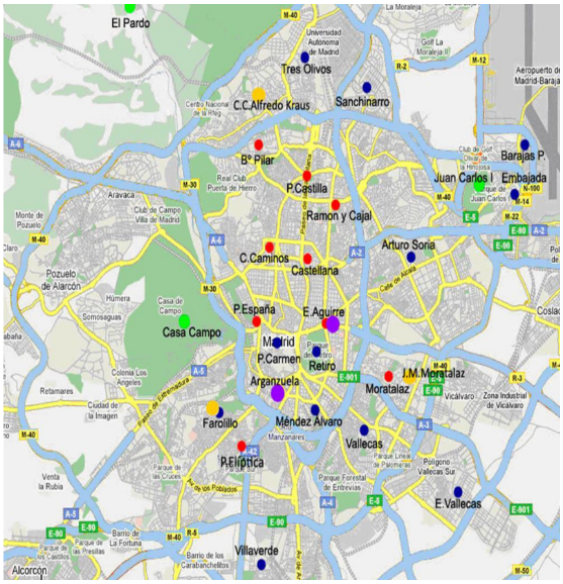

Figura 4.1: Mapa estaciones calidad del aire. [24]

Nombre	ID
Arturo Soria	16
Barrio del Pilar	39
Castellana	48
Cuatro Caminos	38
Escuelas Aguirre	8
Moratalaz	36
Plaza Elíptica	56
Plaza de Castilla	50
Plaza de España	4
Plaza del Carmen	35
Ramón y Cajal	11
Sanchinarro	57

Tabla 4.1: Estaciones de calidad del aire seleccionadas

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	H02	V02
28	79	4	1	28079004_1_38	2019	1	1	23	V	17	V

Figura 4.2: Estructura datos de calidad del aire. [24]

4.2. Variables Predictoras

En este apartado se definen aquellas variables relacionadas con la variable objetivo. Variables que se considera que pueden estar directamente relacionadas con la acumulación de partículas contaminantes a nivel de calle.

4.2.1. Tráfico

Las estaciones de tráfico nos proporcionan datos en tiempo real acerca de la densidad de vehículos cada 5 minutos. Este periodo se establece en base al tiempo mínimo de varios ciclos de semáforo, evitando así que la medición se vea afectada si el semáforo se encuentra abierto o cerrado. Unida a esta información recopilamos la ubicación exacta del medidor para poder cruzar la información de las estaciones de calidad del aire. Una vez representados ambos tipos de estaciones en el mapa, se puede ver como cada estación de calidad del aire se asocia a múltiples estaciones de tráfico. Esto se debe a que podemos encontrar varias de estas últimas en una misma calle dependiendo de el ancho y número de carriles, así como también podemos encontrar varias calles que desembocan en una misma zona de medida, nuevamente

cada una de ellas con múltiples medidores de tráfico. La estructura de estos datos y la definición de cada apartado se pueden ver en la Figura 4.3 y 4.4.

Campo	Descripción
idelem	Identificador del punto de medida. Se corresponde con el campo "idelem" presente en el fichero georreferenciado y que permite su posicionamiento sobre plano e identificación del vial y sentido de la circulación.
descripcion	Denominación del punto de medida.
accesoAsociado	Código de control relacionado con el control semafórico para la modificación de los tiempos.
intensidad	Intensidad de número de vehículos por hora. Un valor negativo implica la ausencia de datos.
ocupacion	Porcentaje de tiempo que está un detector de tráfico ocupado por un vehículo. Por ejemplo, una ocupación del 50% en un periodo de 15 minutos significa que ha habido vehículos situados sobre el detector durante 7 minutos y 30 segundos. Un valor negativo implica la ausencia de datos.
carga	Parámetro de carga del vial. Representa una estimación del grado de congestión, calculado a partir de un algoritmo que usa como variables la intensidad y ocupación, con ciertos factores de corrección. Establece el grado de uso de la vía en un rango de 0 (vacía) a 100 (colapso). Un valor negativo implica la ausencia de datos.
nivelServicio	Parámetro calculado en función de la velocidad y la ocupación. Con ellos se forma una matriz de 4x4 con la que se determina cada uno

Figura 4.3: Estructura datos de trafico parte 1. [24]

	de los niveles de servicio posibles: tráfico fluido (0), tráfico lento (1), retenciones (2) y congestión (3). Los umbrales de velocidad y los de ocupación que determinan dichos niveles de servicio varían en función del punto de medida.
intensidadSat	Intensidad de saturación de la vía en veh/hora y que se corresponde con el máximo número de vehículos que pueden pasar en el acceso a la intersección manteniéndose la fase verde del semáforo.
error	Código de control de la validez de los datos del punto de medida.
subarea	Identificador de la subárea de explotación de tráfico a la que pertenece el punto de medida.
st_x	Coordenada X UTM del centroide que representa al punto de medida en el fichero georreferenciado.
st_y	Coordenada Y UTM del centroide que representa al punto de medida en el fichero georreferenciado.

Figura 4.4: Estructura datos de trafico parte 2. [24]

4.2.2. Información meteorológica

En el caso de estas estaciones, cada una de ellas proporciona información acerca de varias variables meteorológicas. Las seleccionadas para la creación de este modelo son: velocidad y dirección del viento, temperatura, humedad, presión barométrica, radiación y precipitación. La medición del viento es útil para la dispersión de contaminantes y, por lo tanto, la reducción de concentraciones. La temperatura puede tener diversas influencias en la calidad del aire ya que afecta procesos físicos y químicos que pueden tener un impacto en la composición atmosférica. Al igual que la temperatura, la humedad puede alterar en gran medida las propiedades físicas del aire, por ejemplo, su capacidad para dispersar contaminantes o la formación y el tamaño de partículas. En el caso de la presión barométrica, esta puede disminuir en gran medida las concentraciones de O2. Asimismo, refiriéndonos a la radiación, los materiales radioactivos liberados al ambiente pueden causar la contaminación del aire, el agua, los suelos y muchos otros elementos. Por último, las precipitación puede ayudar a asentar los contaminantes diluyendo altas concentraciones trasportadas en el aire. Así, cada una de las variables meteorológicas mencionadas contribuirán a la predicción de la calidad del aire notablemente.

En cuanto a su recopilación, igual que en las estaciones de calidad del aire, sus ficheros se actualizan cada hora entre los minutos 20 y 30. La estructura de los datos queda refleja en la Figura 4.5. En ella vemos como el campo *dato* indica la hora del día en que se ha realizado la medición, junto con su correspondiente código de validación en el campo *Código de validación*. Por último, el campo *magnitud* nos indica la variable meteorológica que se esta midiendo.

Es importante mencionar las dificultades encontradas a la hora de recopilar esta variable. Los datos meteorológicos son reducidos, no poseemos información para todas las estaciones de medida de calidad del aire seleccionadas previamente. Por ello, se ha decidido tomar los valores de una única ubicación. La estación seleccionada se encuentra en la zona centro de Madrid, punto medio entre todas las estaciones de medida de calidad del aire. Así, se ha comprobado que representa adecuadamente los valores de todas las estaciones, los cuales no varían demasiado en la zona de estudio. Todo esto justifica la decisión de tomar las mediciones meteorológicas de una única estación y aplicarlas para el resto de estaciones de calidad del aire.

PROVINCIA	MUNICIPIO	ESTACIÓN	MAGNITUD	TÉCNICA	PERIODO ANÁLISIS	AÑO	MES	DÍA	DATO	CÓDIGO DE VALIDACIÓN
28	079	104	82	98	02	2019	01	01	00023	V

Figura 4.5: Estructura datos meteorológicos. [24]

4.2.3. Altura media de los edificios colindantes

Esta variable estática mide la altura de los edificios que se encuentran dentro del radio de acción de la estación. Existen diferentes formas de plantear la captura de esta variable. En un primer momento se contempló hacer la media de las alturas de los edificios de la región que se había considerado como zona de influencia de la estación de medida de contaminación. Esta decisión se descartó por un motivo muy sencillo, el algoritmo podría confundir una región con un edificio de 15m en mitad de un descampado con una región con 10 edificios de 15 metros en la misma superficie. Por este motivo se decidió dividir el radio de acción en una rejilla de 4x4 elementos, reemplazando la altura como un único valor escalar por un valor vectorial de 16 escalares. Como se podrá apreciar en imágenes posteriores, la rejilla se centra en cada estación de medida. Dicha rejilla está compuesta por 16 celdas de igual dimensión. Siguiendo este criterio se medirá la altura de todos los edificios que se encuentren dentro de alguno de dichos cuadrados. Para poder realizar estas mediciones se ha hecho uso de la aplicación *Google Earth*, por ello, para que el área que abarca la rejilla sea la misma en cada ubicación, se han tomado todas las imágenes a la misma altura de vista y con el norte apuntando a la parte superior del papel. Así, el área en el cual se tienen en cuenta las alturas de los edificios es el mismo en cada caso. En la Figura 4.6 se pueden ver las alturas seleccionadas para una de las estaciones y la rejilla. Cabe destacar que el tamaño de dicha rejilla se ha establecido en base a un estudio exhaustivo de las áreas de apertura y la presencia de edificios en cada estación de medida.

4.2.4. Amplitud de la estación de medida

Esta variable mide el área que rodea a la estación de medida de calidad del aire, la amplitud de la zona. Para que el criterio sea objetivo se utilizará la misma rejilla de 4 x 4, donde la estación estará en el centro. Esta variable medirá el área (superficie) contenida dentro de la rejilla que no tenga ningún edificio.

Como se puede ver en la Figura 4.6, la medición de estas dos primeras variables nos permite hacernos una idea de la superficie disponible para la evacuación de partículas contaminantes. A mayor área despejada y menor altura de los edificios, mayor probabilidad que dichas partículas evacuen la zona y haya un menor nivel de contaminación.

Figura 4.6: Altura, área y amplitud de una estación.

4.2.5. Orientación

Esta variable mide la orientación de la calle y esta directamente relacionada con las condiciones meteorológicas. Si la orientación de la calle coincide con la dirección del viento a lo largo del día, entonces la concentración de contaminantes en la zona será menor dada la evacuación de las partículas por el viento. Lo importante aquí es la dirección de la calle y no el sentido, ya que las partículas se moverán fuera de la zona si el viento lleva la misma dirección, sea cual sea el sentido de la calle.

En la Figura 4.6, con la ayuda de *Google Maps*, se ha posicionado el norte en la parte superior. Así, se puede ver como la orientación de la calle es noreste.

4.2.6. Número de carriles

Esta variable mide el número de carriles en la zona de la estación de medida de calidad del aire que posean medidores de tráfico. Una intensidad de tráfico elevada en un número reducido de carriles genera mayor cantidad de partículas contaminantes que una intensidad de tráfico elevada en una cantidad de carriles mayor. Esta variable está muy ligada a la amplitud de las estaciones de medida, sin embargo, en vez centrarse en el área que rodea a los medidores de calidad del aire, se centra en los medidores de tráfico y el área que abarcan, para relacionarlo con la intensidad de vehículos.

La Figura 4.6 muestra una calle con un total de 2 carriles por los que el tráfico circula dentro del radio de acción de la estación.

4.2.7. Zonas verdes

Esta variable mide la cantidad de árboles en torno a las fuentes de generación de CO2 y, a su vez, la cantidad de zona verde en torno a los medidores de calidad del aire. A mayor número de árboles en torno a los medidores, mayor absorción de contaminantes y mejor calidad del aire. Lo mismo ocurre con las fuentes de CO2, la vegetación a su alrededor actúa como filtro para los contaminantes, atrapando las partículas nocivas en sus hojas y corteza. Ambas cosas son determinantes para el modelo, es por ello que esta variable genera un valor promedio siguiendo los siguientes criterios.

Se han creado tres categorías: Zona verde escasa, zona verde media y zona verde abundante. Cada una de estas categorías está asociada a un número entero (0,1 y 2 respectivamente). Para el cálculo del promedio se sumarán, para cada ubicación, la cantidad de zona verde entorno a cada estación de tráfico y estación de calidad del aire, y se dividirán (a modo de normalización) entre la suma de estos mismo valores en un caso hipotéticamente ideal. Es decir, la suposición de que en torno a todas las estaciones hay zona verde abundante. En la Figura 4.7 se pueden ver dos estaciones, la primera con zona verde media y la segunda con zona verde abundante. Queda también representado el área de acción de cada una de ellas.

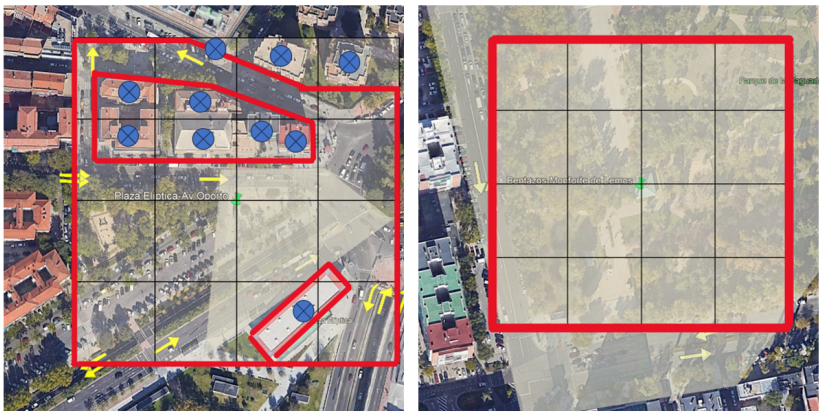

Figura 4.7: Categorías de concentración de zonas verdes.

4.3. Dificultades presentes en la creación de la matriz de entrenamiento

Una vez determinadas todas las variables, tanto estáticas como dinámicas, llega el momento de capturar sus datos en el tiempo. Se quiere crear una matriz de

entrenamiento lo suficientemente grande como para que el modelo sea objetivo. Las variables estáticas asociadas a cada estación han sido recopiladas y ordenadas en una matriz llamada *matriz estática*. El objetivo de esta apartado es capturar periódicamente la variables dinámicas de cada estación y añadirle sus correspondientes estáticas para formar una única linea de información por estación y por hora. Dicha fila será añadida a la matriz de entrenamiento.

Para lograr el objetivo planteado es necesario generar un código que realice llamadas periódicas al API de la Comunidad de Madrid, capture los datos, los ordene y los unifique con la matriz estática. Todo ello para después añadirlo a la matriz de entrenamiento. Sin embargo, hay muchos detalles a tener en cuenta. En primer lugar, la actualización de los ficheros no siempre se realiza en horas fijas sino que en múltiples ocasiones se retrasan. En segundo lugar, se requiere de la verificación de los datos que en ocasiones tarda mas tiempo en actualizarse que el propio dato. Por todo esto se hace a continuación un resumen de las dificultades presentes en la recopilación de cada variable dinámica.

4.3.1. Retos en la captura de datos de tráfico

Capturar los datos de tráfico para cada estación seleccionada no es complejo. Por lo general, los datos se actualizan en la página regularmente y no falta información de ninguna estación. Sin embargo, se ha decidido que es conveniente comprobar que no falte ninguna de las estaciones cada vez que una llamada devuelva información. Así, antes de guardarla se verifica cuidadosamente con la ayuda de un diccionario adicional y se eliminan aquellas horas en las que la información sea deficiente.

4.3.2. Retos en la captura de datos de calidad del aire

Los datos de calidad del aire, como los de tráfico, se actualizan regularmente en el portal y no suelen presentar ningún tipo de problema. Sin embargo, se ha realizado también una comprobación adicional cada hora para verificar la calidad de los datos obtenidos.

4.3.3. Retos en la captura de datos de meteorología

Existen dos grandes problemas entorno a este set de datos. Como se ha mencionado en previos apartados, estos datos poseen un campo llamado *Código de validación* que indica si el dato está o no validado, es decir, si es o no utilizable. Sin embargo, muchas veces encontramos un dato distinto de cero que parece válido pero no tiene código de validación “V”. Puede pasar que dicho dato no este validado en la actualización de esa hora y a la hora siguiente cambie y se valide,

o que simplemente no se valide. Los datos deben de actualizarse entre los minutos 20 y 30 de cada hora, pero por lo general tardan más en actualizarse y no suele ser a intervalos regulares. Esto es un problema, ya que puede ser que el dato de las 12 de la mañana aparezca a las 14 horas. Si estos detalles no se tienen en cuenta a la hora de realizar el código entonces es muy probable que se capture información errónea o que se pierda información.

Capítulo 5

Resultados

En este capítulo se presentan los resultados obtenidos para cada modelo de predicción de contaminantes, así como un análisis exhaustivo de su desempeño. Además, se examinan en detalle los resultados de las distintas técnicas utilizadas para mejorar la precisión y fiabilidad de estos modelos.

Estas técnicas abarcan desde estrategias previas a la modelización, hasta aquellas aplicadas posterior a esta. Todas ellas han sido expuestas en capítulos anteriores y han servido de base para el desarrollo de los modelos. Así, se proporcionará una visión completa y detallada del proceso de construcción y mejora de los modelos de predicción de óxido nítrico, partículas PM2.5 y partículas PM10.

5.1. Análisis descriptivo de las variables

Como se menciona en capítulos anteriores, el análisis de variables previo a la creación de modelos es de vital importancia para el conjunto de datos. Este análisis nos permite encontrar fallos en las mediciones, datos atípicos (outliers) y variables constantes que no aportan información. Asimismo, nos proporciona medidas como el rango de valores, la media y la moda. Toda esta información nos ayuda a interpretar los datos correctamente así como a construir modelos que se ajusten a la realidad con la mayor precisión posible. A continuación se presentan aquellos histogramas considerados mas relevantes.

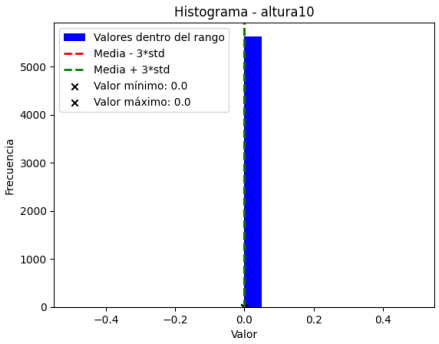

Figura 5.1: Valores de altura obtenidos por las estaciones de la Comunidad de Madrid. Datos muestran que se trata de una variable constante.

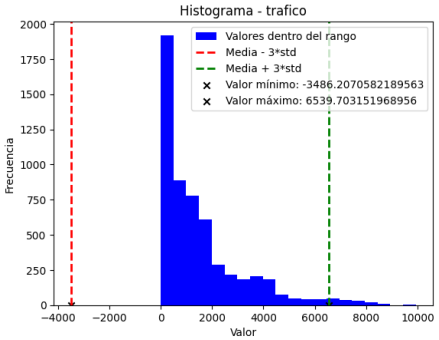

Figura 5.2: Valores del tráfico obtenidos por las estaciones de la Comunidad de Madrid. Se muestran valores que exceden tres veces la desviación estándar.

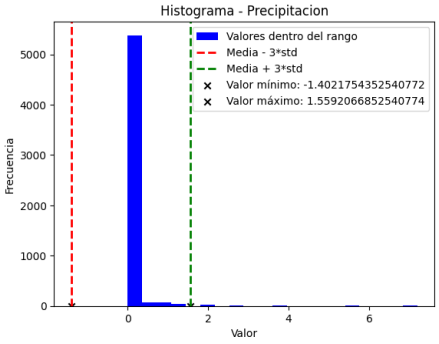

Figura 5.3: Valores de precipitación obtenidos por las estaciones de la Comunidad de Madrid. Se muestran valores que exceden tres veces la desviación estándar.

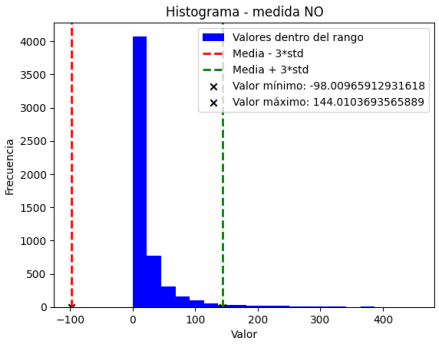

Figura 5.4: Valores de NO obtenidas por las estaciones de la Comunidad de Madrid.

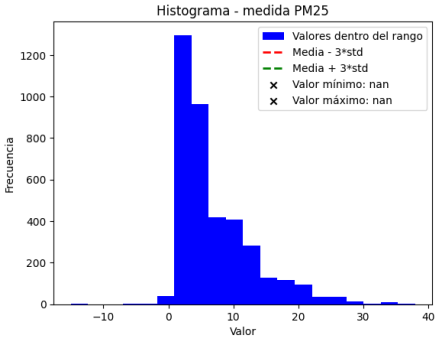

Figura 5.5: Valores de PM2.5 obtenidos por las estaciones de la Comunidad de Madrid.

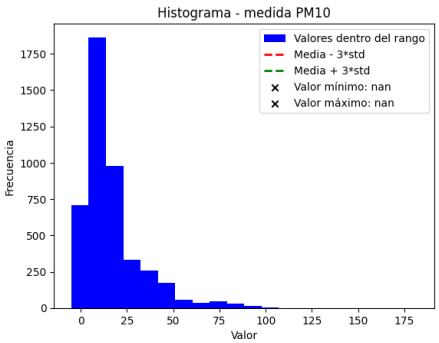

Figura 5.6: Valores de PM10 obtenidos por las estaciones de la comunidad de Madrid.

Tras analizar cada uno de estos histogramas se pueden destacar tres cosas. En primer lugar, nos centramos en la variable *altura 10*, Figura 5.1, la cual pertenece a

una variable más general llamada *altura*. Esta última, como ya se ha mencionado, pretende representar la altura media de los edificios colindantes con la estación y su influencia en la acumulación de las partículas contaminantes. Así, *altura 10* representa la altura de los edificios en una posición determinada dentro de la rejilla de medición. Como se puede ver, esta es nula y constante, no existen edificios en esa posición en ninguno de los casos registrados. Esta variable no aporta entonces ningún tipo de información a nuestro modelos y debe ser eliminada.

En segundo lugar, se puede ver como tanto para la variable *Tráfico* como para la variable *Precipitación* la mayor parte de resultados se agrupan entorno a los mismos valores. Figura 5.2 y 5.3 respectivamente. Valores que representan una clara tendencia de estas variables. Sin embargo, se pueden también observar una serie de outliers, valores que se alejan de esta tendencia. La eliminación de estos outliers, por ser valores atípicos en cantidades insignificantes, puede ayudar a mejorar los resultados de los modelos de predicción.

Por último, nos fijamos en los valores de los tres contaminantes: Óxido Nítrico, PM2.5 y PM10. En el caso del óxido nítrico, Figura 5.4, vemos como la mayor parte de mediciones se concentran en el intervalo de 0 a 150 µg/m3. Los valores superiores a esa cantidad de partículas son insignificantes comparados con los pertenecientes al intervalo mencionado. Por ello, su eliminación podría suponer una mejora de los resultados.

En el caso de PM2.5 Y PM10, Figuras 5.5 y 5.6, se puede ver como ninguno de los valores registrados excede tres veces la desviación estándar. Es decir, no se observan outliers. El rango de valores es aproximadamente de 0 a 40 µg/m3 y de 0 a 100 µg/m3 para las partículas 2.5 y 10 respectivamente.

5.2. Predicción de NO

Una vez detalladas las características de las variables recopiladas para los tres contaminantes, pasamos a centrarnos únicamente en el óxido nítrico. En esta sección se muestra el proceso de modelización asociado a este contaminante y los resultados obtenidos.

5.2.1. Selección de hiperparámetros

Se presentan los valores de los hiperparámetros de cada modelo, así como una tabla que permite visualizar los valores del coeficiente de determinación r2 antes y después de la selección de hiperparámetros.

1. Decision Tree Regressor
    - max features = 20

- max depth = 10

2. Linear Regressor
   - fit intercept = True
   - normalize = False
3. Kernel Ridge Regressor
   - alpha = 1.0
   - degree = 3
   - gamma = 0.1
   - kernel = "poly"
4. Lasso
   - alpha = 0.1
5. Random Foreste regressor
   - max features = 14
   - n estimators = 301
6. Redes neuronales
   - alpha = 0.0001
   - hidden layer sizes = (100,50)
   - learning rate = çonstant"
7. Vecinos cercanos
   - n neighbours = 5
   - p = 1
   - weights = "distance"
8. XGoost
   - colsample bytree = 0.8
   - learning rate = 0.1
   - max depth = 7

- subsample = 0.9

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7066	0.7384
Redes	0.1819	0.7435
Random Forest	0.8292	0.832
Nearest Neighbors	0.6183	0.6222
Xgboost	0.656	0.8363
Ensamble	0.8392	-
Linear Regressor	0.1764	0.1764
KRR	0.1768	0.5987
Lasso	0.1703	0.1774

Tabla 5.1: Selección de hiperparámetros. La tabla muestra los valores del coeficiente de determinación r2 antes (valores por defecto) y después de la selección de hiperparámetros.

En la Figura 5.1 vemos los resultados. Se puede observar claramente como la selección de hiperparámetros mejora el coeficiente de determinación para todos los modelos, a excepción de *Linear Regressor* que mantiene su valor inicial. Asimismo, vemos como algunos modelos aumentan su precisión considerablemente, como por ejemplo el modelo de *Redes Neuronales* o el modelo *KRR*, mientras que otros modelos lo aumenta ligeramente. A modo de conclusión, estos resultados reflejan la importancia de la selección de los hiperparámetros para cada modelo.

5.2.2. Selección de variables

En esta sección nos centramos en la reducción de variables, en la selección de aquellas que son mas influyentes en la creación y acumulación del óxido nítrico y aquellas que solamente entorpecen la precisión del modelo. Se presentará la reducción a 5, 10 y 15 variables, además de la evolución de los valores del coeficiente de determinación en cada caso. Se mencionarán las variables seleccionadas para aquellos casos que muestren varianza en la precisión tras la reducción de variables. Siendo el principal objetivo mejorar el coeficiente de determinación r2, es importante destacar que trabajaremos siempre con la mejor versión de cada modelo lograda hasta el momento. Así, dado que se ha concluido que la selección de hiperparámetros mejora los resultados, aplicaremos la reducción de variables a dicha versión de los modelos.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7384	0.7255
Redes	0.7435	0.7049
Random Forest	0.832	0.8009
Nearest Neighbors	0.6222	0.7114
Xgboost	0.8363	0.8047
Ensamble	0.8392	0.8066
Linear Regressor	0.1764	0.1603
KRR	0.5987	0.4729
Lasso	0.1774	0.1604

Tabla 5.2: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación r2 antes y después de la reducción a 5 variables.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision tree	0.7384	0.7308
Redes	0.7435	0.7527
Random Forest	0.832	0.8209
Nearest Neighbors	0.6222	0.7139
Xgboost	0.8363	0.8428
Ensamble	0.8392	0.8351
Linear Regressor	0.1742	0.1603
KRR	0.5987	0.5852
Lasso	0.1774	0.174

Tabla 5.3: La tabla presenta los valores del coeficiente de determinación r2 antes y después de la reducción a 10 variables.

En la Tabla 5.2 observamos la reducción a 5 variables. De ella podemos concluir que dicha reducción no es en general favorable para la precisión de nuestros modelos. La excepción a esta regla es *Nearest Neighbors*, modelo que aumenta su precisión de un 0.62 a un 0.71 al reducir el espectro de variables. Este determina que la acumulación de NO a nivel de calle depedende en mayor medida del día de la semana, la dirección del viento, la temperatura, la radiación y la altura de los edificios colindantes. Sugiriendo así que el resto de variables no son relevantes a la hora de predecir la calidad del aire en base a este contaminante.

En cuanto a la reducción a 10 variables, Tabla 5.3, aunque se puede ver un ligero incremento en varios modelos, observamos como tampoco supone una mejora sig-

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7384	0.7118
Redes	0.7435	0.7746
Random Forest	0.832	0.8216
Nearest Neighbors	0.6222	0.741
Xgboost	0.8363	0.8294
Ensamble	0.8392	0.8306
Linear Regressor	0.1764	0.1771
KRR	0.5987	0.5871
Lasso	0.1774	0.177

Tabla 5.4: La tabla presenta los valores del coeficiente de determinación r2 antes y después de la reducción a 15 variables.

nificativa. *Xgboost* aumenta su precisión de un 0.836 a un 0.842, *Redes Neuronales* de un 0.743 a un 0.752 y *Nearest Neighbors* de un 0.622 a un 0.713. Las variables día de la semana, hora, humedad y presión barométrica son seleccionas por los tres modelos entre las mas influyentes.

Por último, en el caso de 15 variables, Tabla 5.4, vemos una mejora en los modelos *Linear Regressor*, *Redes Neuronales* y *Nearest Neighbors*. La mejora es nuevamente ligera y prácticamente imperceptible en algunos casos.

A modo de conclusión, se podría decir que la reducción de variables no resulta ser muy significativa en nuestro estudio, aunque puede llegar a mejorar la precisión levemente y por ello ha de ser tenida en cuenta. La mejora mas notable se da para el modelo *Nearest Neighbors*, en el cual la reducción a 15 variables supone un incremento en la precisión de 0.118.

5.2.3. Proyección a baja dimensión

En esta sección se presentan los resultados del Análisis de Componentes Principales conocido como PCA. Analizándose también cuidadosamente sus resultados.

	PCA
Decision Tree	-0.1825
Redes	-0.0008
Random Forest	-0.2422
Nearest Neighbors	-0.1261
Xgboost	-0.0398
Ensamble	-0.0050
Linear Regressor	0.0024
KRR	0.0072
Lasso	0.0023

Tabla 5.5: Evolución de valores tras PCA

En la Tabla 5.5 se puede ver claramente como el análisis de componentes principales otorga para nuestro conjunto de datos unos resultandos bastante deficientes.

5.2.4. Eliminación de outliers

Llegados a este punto podemos observar que el mejor resultado lo obtiene el modelo *Xgboost* con 10 variables. El valor del coeficiente de determinación en este caso es de 0.8428, valor que logra mejorar el inicial en un 0.1868. Dado que queremos encontrar el modelo que mejor se ajuste a los datos, nos enfocaremos ahora en mejorar el valor de precisión de *Xgboost*.

En esta sección nos centramos en la eliminación de outliers encontrados en el inicial análisis descriptivo de las variables. Este proceso se dividirá en dos grupos. En primer lugar, trataremos únicamente con los outliers identificados en la variable altura y los propios datos de NO y, en segundo lugar, añadiremos la precipitación y el tráfico. Antes de comenzar se realizará el filtro de variables y una vez obtenida la matriz reducida se hará la búsqueda de hiperparámetros y selección de características, evaluando el coeficiente de determinación en cada caso.

$Precisión_{Incial}$	$Precisión_{Hyper}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.619	0.7459	0.7097	0.7693	0.7597

Tabla 5.6: Eliminación de outliers [NO y altura]. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y la reducción a 5, 10 y 15 variables de la mejor versión del modelo *Xgboost*.

$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.6722	0.7851	0.7104	0.7877	0.7886

Tabla 5.7: Eliminación de outliers [NO, altura, precipitación y tráfico]. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y la reducción a 5, 10 y 15 variables de la mejor versión del modelo *Xgboost*.

Las Tablas 5.6 y 5.7 muestran como la eliminación de outliers en el contexto de nuestro proyecto no supone una mayor precisión para el modelo *Xgboost* a la hora de predecir los niveles de NO.

5.2.5. Cuantificación

Esta sección se centra en la cuantificación de los valores de NO previa a la aplicación del modelo *Xgboost*. Dado que el valor máximo de NO es de 144.01 µg/m3 se han creado 16 intervalos, donde cada uno de ellos tiene una amplitud 10 µg/m3.

Esta cuantificación se aplica en primer lugar a la matriz completa, realizando búsqueda de hiperparámetros y selección de variables. En segundo a la matriz reducida cuyo modelo ha obtenido mejores resultados, *Xgboost* con 10 variables. Los resultados se muestran a continuación.

$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.6498	0.8231	0.7897	0.8161	0.8151

Tabla 5.8: Cuantificación matriz completa. Tabla presenta los valores del coeficiente de determinación tras las selección de hiperparámetros y reducción a 5, 10 y 15 variables de la matriz original.

$Precisión_{Inicial}$	$Precisión_{Hyper}$
0.8131	0.8179

Tabla 5.9: Cuantificación *Xgboost* 10 variables. Tabla presenta los valores del coeficiente de determinación antes y después de la selección de hiperparámetros con el conjunto de datos reducido a 10 variables.

En las Tablas 5.8 y 5.9 vemos como la cuantificación, aunque otorga buenos resultados, no mejora significativamente los obtenidos previamente. Es decir, la

cuantificación del conjunto de datos de óxido nítrico no mejora el modelo de predicción de este contaminante.

5.2.6. Entrenamiento de modelos por estación

Hasta ahora hemos contemplado un único modelo que capture el comportamiento de las partículas contaminantes a partir de una lista de 14 variables predictoras y que sea de aplicación en cualquier parte de la ciudad. Existe la posibilidad de que los resultados mejoren si realizamos un modelo por estación. Esto podría justificarse con la presencia de algunas variables relevantes para el proceso de acumulación de partículas contaminantes que definen la zona en la que se instala la estación de medida y que no se han tenido en cuenta. Esto resultaría en que entrenar un modelo por estación podría, potencialmente, arrojar mejores resultados que un modelo predictivo generalista.

Esta sección explora la posibilidad de obtener una mayor precisión de predicción al realizar un modelo por estación. Esto quiere decir que la matriz original se divide en 12 matrices reducidas, donde cada una de ella contiene datos de una estación específica. Así, en vez de crear un modelo general aplicable a todas las zonas, se creará un modelo de predicción de calidad del aire para cada ubicación.

Es importante destacar que para llevar acabo esta técnica es necesario eliminar todas aquellas variables que resulten constantes para cada estación. Estas son todas las variables que hemos denominado como estáticas. Así, el número total de variables se reduce a 11 y solo tiene sentido realizar la reducción de variables a 5. Aferrándonos al objetivo de mejorar el modelo *Xgboost* con 10 variables, aplicaremos en primer lugar la modelización por estación únicamente a dicho modelo. Una vez completado, observaremos la estación con la mejor precisión y probaremos para dicha estación el resto de modelos. A continuación se presentan las tablas con los resultados obtenidos.

Como se puede ver en la Tabla 5.10, la precisión del modelo *Xgboost* por estación es baja, el valor más alto se obtiene para la estación número 8.

En la Tabla 5.11 se muestra la precisión de todos los modelos aplicados a la estación número 8, valores que continúan siendo bajos. Podemos así concluir que la creación de modelos por estación no mejora la precisión de un modelo aplicado a todas las estaciones de medida y sus localizaciones.

	$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$
Estación 4	0.5701	0.5923	0.6555
Estación 8	0.4395	0.5389	0.7461
Estación 16	0.2914	0.4311	0.4506
Estación 11	-0.3275	-0.1725	-0.2133
Estación 36	0.4916	0.7014	0.5857
Estación 38	0.4637	0.5668	0.4922
Estación 39	0.2325	0.3723	0.4227
Estación 48	0.1156	-0.1019	0.2025
Estación 50	0.1597	0.2589	0.3720
Estación 56	0.6069	0.7266	0.7341
Estación 57	0.3709	0.4375	0.2263

Tabla 5.10: Modelo *Xgboost* para cada estación. Tabla presenta los resultados del coeficiente de determinación tras la selección de hiperparámetros y reducción a 5 variables para cada estación de medida.

	$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$
Decision Tree	0.2334	0.3922	0.2983
Random Forest	0.5191	0.5693	0.6575
Xgboost	0.4395	0.5389	0.7461
Linear Regressor	0.1883	0.1883	0.1977
Lasso	0.1898	0.1898	0.1888
Redes	0.2212	-	-
Nearest Neighbors	0.7107	0.7366	0.7166
Ensamble	0.557	-	0.7241
KRR	0.1887	0.6501	0.5203

Tabla 5.11: Modelos para la estación 8. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y reducción a 5 variables para cada modelo, aplicado únicamente a la estación número 8.

5.3. Predicción de partículas PM2.5

5.3.1. Selección de hiperparámetros

Se presentan los valores de los hiperparámetros de cada modelo, así como una tabla que permite visualizar los valores del coeficiente de determinación antes y después de la selección de hiperparámetros.

1. Decision Tree Regressor
    - max features = 14
    - max depth = 10
2. Linear Regressor
    - fit intercept = True
    - normalize = False
3. Kernel Ridge Regressor
    - alpha = 1.0
    - degree = 3
    - gamma = 0.1
    - kernel = "poly"
4. Lasso
    - alpha = 0.1
5. Random Forest regressor
    - max features = 14
    - n estimators = 301
6. Redes Neuronales
    - alpha = 0.05
    - hidden layer sizes = (100,50)
    - learning rate = çonstant"
7. Vecinos cercanos
    - n neighbours = 5
    - p = 1
    - weights = "distance"
8. Xgboost
    - colsample bytree = 0.8

- learning rate = 0.1
- max depth = 7
- subsample = 0.8

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.4428	0.5788
Redes	0.3711	0.5777
Random Forest	0.7336	0.7347
Nearest Neighbors	0.5137	0.5549
Xgboost	0.6313	0.7325
Ensamble	0.7410	-
Linear Regressor	0.3774	0.3774
KRR	0.3772	0.4739
Lasso	0.1562	0.3772

Tabla 5.12: Selección de hiperparámetros. La tabla muestra los valores del coeficiente de determinación r2 antes y después de la selección de hiperparámetros.

En la Tabla 5.12 vemos los resultados. Se puede observar claramente como la selección de hiperparámetros mejora o en el peor de los casos mantiene, el coeficiente de determinación para todos los modelos. Asimismo, vemos como algunos modelos aumentan su precisión considerablemente, como por ejemplo el modelo de redes neuronales o el modelo *Decision Tree*, mientras que otros modelos lo aumenta ligeramente. A modo de conclusión, estos resultados reflejan, al igual que para el monóxido de nitrógeno, la importancia de la selección de los hiperparámetros para cada modelo de predicción de partículas PM2.5.

5.3.2. Selección de variable

En esta sección nos centramos en la reducción de variables, en la selección de aquellas que son mas influyentes en la creación y acumulación de partículas PM2.5 y aquellas que solamente entorpecen la precisión del modelo. Se presentará la reducción a 5, 10 y 15 variables, además de la evolución de los valores del coeficiente de determinación en cada caso. Se mencionarán las variables seleccionadas para aquellos casos que muestren varianza en el coeficiente de determinación tras la reducción de variables. Siendo el principal objetivo mejorar el coeficiente de determinación, es importante destacar que trabajaremos en todo momento con la mejor versión de cada modelo lograda hasta el momento. Así, dado que se ha concluido

que la selección de hiperparámetros mejora los resultados, aplicaremos la reducción de variables a dicha versión de los modelos.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.5788	0.6389
Redes	0.5777	0.6058
Random Forest	0.7347	0.7402
Nearest Neighbors	0.5549	0.6503
Xgboost	0.7325	0.7104
Ensamble	0.741	0.7178
Linear Regressor	0.3774	0.3551
KRR	0.4739	0.4691
Lasso	0.3772	0.3551

Tabla 5.13: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 5 variables.

En la Tabla 5.13 observamos la reducción a 5 variables. De ella podemos concluir que dicha reducción es favorable para la precisión de algunos modelos. *Decision Tree*, *Random Forest*, *Redes Neuronales* y *Nearest Neighbors* son los modelos beneficiados. Tras el análisis de las variables seleccionadas por cada uno de ellos, se puede ver como claramente difieren. Sin embargo, se observa una única variable común a todas ellas, la temperatura. Esto determina que la acumulación de PM2.5 a nivel de calle depedende en mayor medida de la temperatura. Sugiriendo así que el resto de variables son menos relevantes a la hora de predecir la calidad del aire en base a este contaminante.

En cuanto a la reducción a 10 variables, Tabla 5.14, se puede también observar como supone un incremento en la mayoría de modelos. *Decision Tree*, *Redes neuronales*, *Nearest Neighbors*, *Ensamble* y *KRR* son los modelos beneficiados por la reducción a 10 variables. Todos ellos poseen múltiples variables en común como el día de la semana, la hora del día, la temperatura, humedad, presión y radiación. Es decir, todos ellos consideran esta serie de variables determinantes en la acumulación de PM2.5 a nivel de calle. Asimismo, nos fijamos en el modelo *Ensamble*, el cual aumenta su coeficiente de determinación de 0.741 a 0.7437 y se convierte en el valor mas alto hasta el momento.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.5788	0.6258
Redes	0.5777	0.6438
Random Forest	0.7347	0.731
Nearest Neighbors	0.5549	0.6828
Xgboost	0.7325	0.725
Ensamble	0.741	0.7437
Linear Regressor	0.3774	0.3754
KRR	0.4739	0.5519
Lasso	0.3772	0.3755

Tabla 5.14: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 10 variables.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.5788	0.6249
Redes	0.5777	0.6408
Random Forest	0.7347	0.7345
Nearest Neighbors	0.5549	0.6693
Xgboost	0.7325	0.7378
Ensamble	0.741	0.7376
Linear Regressor	0.3774	0.3773
KRR	0.4739	0.5368
Lasso	0.3772	0.3767

Tabla 5.15: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 15 variables.

Por último, en el caso de 15 variables, Tabla 5.15, vemos una mejora en los modelos *Decision Tree*, *Xgboost*, *Redes Neuronales* y *KRR*. La mejora es ligera y prácticamente imperceptible en algunos casos.

A modo de conclusión, se podría decir que la reducción de variables resulta ser mas significativa en las partículas PM2.5 que en las partículas de NO. La mejora mas notable se da en los tres casos para el modelo *Nearest Neighbors*, en el cual la reducción a 5, 10 y 15 variables supone un incremento en la precisión del 0.0954, 0.1279 y 0.1144 respectivamente.

5.3.3. Proyección a baja dimensión

En esta sección se presentan los resultados del Análisis de Componentes Principales conocido como PCA.

	PCA
Decision Tree	-0.1720
Redes	0.0074
Random Forest	-0.0126
Nearest Neighbors	-0.1936
Xgboost	-0.0398
Ensamble	-0.0203
Linear Regressor	0.0064
KRR	0.0138
Lasso	0.0064

Tabla 5.16: Evolución de valores tras PCA

En la Tabla 5.16 se puede ver claramente como el análisis de componentes principales otorga para nuestro conjunto de datos unos resultandos bastante deficientes.

5.3.4. Eliminación de outliers

Llegados a este punto podemos observar que el mejor resultado lo obtiene el modelo *Ensamble* con 10 variables, modelo que combina *Xgboost* y *Random Forest*. El valor del coeficiente de determinación en este caso es de 0.7437, valor que logra mejorar el inicial en un 0.0027. *Ensamble* combina los mejores modelos ya con sus hiperparámetros óptimos. Es por ello que en este caso el análisis de hiperparámetros no es necesario. Dado que queremos encontrar el modelo que mejor se ajuste a los datos, nos enfocaremos ahora en mejorar el valor de precisión del modelo *Ensamble*.

En esta sección nos centramos en la eliminación de outliers encontrados en el inicial análisis descriptivo de las variables. De manera contraria a los datos de NO, los datos obtenidos de PM2.5 no poseen outliers. Por ello nos centraremos en los presentes en la precipitación y el tráfico. Antes de comenzar se realizará el filtro de variables y una vez obtenida la matriz reducida se hará la selección de características, evaluando el coeficiente de determinación en cada caso.

$Precisión_{Inicial}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.7547	0.7199	0.7488	0.7501

Tabla 5.17: Eliminación de outliers [precipitación y tráfico]. Tabla presenta los valores del coeficiente de determinación tras la reducción a 5, 10 y 15 variables de la mejor versión del modelo *Ensamble*.

La Tabla 5.17 muestran como la eliminación de outliers en el contexto de nuestro proyecto no supone mayor precisión para el modelo Ensamble a la hora de predecir los niveles de PM2.5.

5.3.5. Cuantificación

Esta sección se centra en la cuantificación de los valores de PM2.5 previa a la aplicación del modelo *Ensamble*. Para este contaminante se ha mantenido un rango de amplitud 10 µg/m3. Sin embargo, al poseer un rango de valores inferior al de las partículas de NO, el número de intervalos se ve reducido también.

Esta cuantificación se aplica en primer lugar a la matriz completa, realizando la selección de variables. En segundo a la matriz reducida cuyo modelo ha obtenido mejores resultados, *Ensamble* con 10 variables. Los resultados se muestran a continuación.

$Precisión_{Inicial}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.5943	0.5725	0.6089	0.6158

Tabla 5.18: Cuantificación matriz completa. Tabla presenta los valores del coeficiente de determinación tras la reducción a 5, 10 y 15 variables de la matriz original.

$Precisión_{Incial}$	$Precisión_{10}$
0.6089	0.6165

Tabla 5.19: Cuantificación *Ensamble* 10 variables. Tabla presenta los valores del coeficiente de determinación tras la cuantificación del conjunto de datos reducido a 10 variables.

Como se puede observar en las Tablas 5.18 y 5.19, la cuantificación efectivamente ayuda a aumentar la precisión. Sin embargo, no a los niveles esperados.

5.3.6. Entrenamiento de modelos por estación

Como ya se ha mencionado, esta sección explora la posibilidad de obtener una mayor precisión de predicción al realizar un modelo por estación. Es importante volver a destacar que para llevar acabo esta técnica es necesario eliminar todas aquellas variables constante para cada estación, estas son todas las variables que hemos denominado como estáticas. Así, el número total de variables se reduce a 11 y solo tiene sentido realizar la reducción de variables a 5. Aferrándonos al objetivo de mejorar el modelo *Ensamble* con 10 variables, aplicaremos en primer lugar la modelización por estación únicamente a dicho modelo. Una vez completado, observaremos la estación con la mejor precisión y probaremos para dicha estación el resto de modelos. A continuación se presentan las tablas con los resultados obtenidos.

	$Precisión_{Inicial}$	$Precisión_5$
Estación 8	0.5102	0.4881
Estación 38	0.5252	0.4601
Estación 48	0.4497	0.4502
Estación 50	0.4903	0.4859
Estación 56	0.4753	0.5472
Estación 57	0.5593	0.4844

Tabla 5.20: Modelo *Ensamble* para cada estación. Tabla presenta los resultados del coeficiente de determinación tras la reducción a 5 variables para cada estación de medida.

	$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$
Decision Tree	0.1602	0.1957	0.2525
Random Forest	0.5476	0.5347	0.5307
Xgboost	0.5435	0.5586	0.4688
Linear Regressor	0.3859	0.3859	0.3837
Lasso	0.1501	0.1501	0.1501
Redes	0.3858	0.4072	0.495
Nearest Neighbors	0.556	0.5877	0.5864
Ensamble	0.5593	0.5593	0.4844
KRR	0.3859	0.554	0.5174

Tabla 5.21: Modelos para la estación 57. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y reducción a 5 variables para cada modelo, aplicado únicamente a la estación número 57.

Como se puede ver en la Tabla 5.20, la precisión del modelo *Ensamble* por estación es baja comparada con los valores de precisión previamente obtenidos. El valor más alto se obtiene para la estación número 57. Por ello, en la Tabla 5.21 se muestra la precisión de todos los modelos aplicados a la estación número 57, valores que continúan siendo bajos.

Podemos así concluir que la creación de modelos por estación no mejora la precisión de un modelo aplicado a todas las estaciones de medida y sus localizaciones.

5.4. Predicción de partículas PM10

5.4.1. Selección de hiperparámetros

Se presentan los valores de los hiperparámetros de cada modelo, así como una tabla que permite visualizar los valores del coeficiente de determinación antes y después de la selección de hiperparámetros.

1. Decision Tree Regressor
   - max features = 14
   - max depth = 10
2. Linear Regressor
   - fit intercept = True
   - normalize = False
3. Kernel Ridge Regressor
   - alpha = 1.0
   - degree = 3
   - gamma = 0.1
   - kernel = "poly"
4. Lasso
   - alpha = 0.01
5. Random Forest regressor
   - max features = 14
   - n estimators = 201

6. Redes neuronales
   - alpha = 0.0001
   - hidden layer sizes = (50,50)
   - learning rate = çonstant"

7. Vecinos cercanos
   - n neighbours = 5
   - p = 1
   - weights = "distance"

8. XGboost
   - colsample bytree = 0.9
   - learning rate = 0.1
   - max depth = 7
   - subsample = 0.9

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7526	0.7982
Redes	0.364	0.7442
Random Forest	0.8848	0.8936
Nearest Neighbors	0.6588	0.6968
Xgboost	0.7485	0.8805
Ensamble	0.8904	-
Linear Regressor	0.3639	0.3639
KRR	0.3635	0.6561
Lasso	0.3262	0.3636

Tabla 5.22: Selección de hiperparámetros. La tabla muestra los valores del coeficiente de determinación r2 antes y después de la selección de hiperparámetros.

En la Tabla 5.22 vemos los resultados. Se puede observar claramente como la selección de hiperparámetros mejora o en el peor de los casos mantiene, el coeficiente de determinación para todos los modelos. Asimismo, vemos como algunos modelos aumentan su precisión considerablemente, como por ejemplo el modelo

Xgboost o el modelo de *Redes Neuronales*. Incrementando este último la precisión en un 0.3802. A modo de conclusión, estos resultados reflejan, al igual que en previos contaminantes, la importancia de la selección de los hiperparámetros para cada modelo de predicción de partículas PM10.

5.4.2. Selección de variable

En esta sección nos centramos en la reducción de variables, en la selección de aquellas que son mas influyentes en la creación y acumulación de partículas PM10 y aquellas que solamente entorpecen la precisión del modelo. Se presentará la reducción a 5, 10 y 15 variables, además de la evolución de los valores del coeficiente de determinación en cada caso. Una vez mas, siendo el principal objetivo mejorar el coeficiente de determinación, es importante destacar que trabajaremos en todo momento con la mejor versión de cada modelo lograda hasta el momento. Así, dado que se ha concluido que la selección de hiperparámetros mejora los resultados, aplicaremos la reducción de variables a dicha versión de los modelos.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision tree	0.7982	0.8142
Redes	0.7442	0.7832
Random Forest	0.8936	0.8679
Nearest Neighbors	0.6968	0.841
Xgboost	0.8805	0.8616
Ensamble	0.8904	0.8738
Linear Regressor	0.3639	0.3471
KRR	0.6561	0.5806
Lasso	0.3636	0.3472

Tabla 5.23: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 5 variables.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7982	0.8332
Redes	0.7442	0.8293
Random Forest	0.8936	0.8810
Nearest Neighbors	0.6968	0.8231
Xgboost	0.8805	0.8806
Ensamble	0.8904	0.8884
Linear Regressor	0.3639	0.3603
KRR	0.6561	0.695
Lasso	0.3636	0.3604

Tabla 5.24: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 10 variables.

	$Precisión_{antes}$	$Precisión_{despues}$
Decision Tree	0.7982	0.8118
Redes	0.7442	0.812
Random Forest	0.8936	0.8868
Nearest Neighbors	0.6968	0.7904
Xgboost	0.8805	0.8852
Ensamble	0.8904	0.8874
Linear Regressor	0.3639	0.3638
KRR	0.6561	0.696
Lasso	0.3636	0.3639

Tabla 5.25: Evolución de valores tras selección de variables. La tabla presenta los valores del coeficiente de determinación antes y después de la reducción a 15 variables.

Al igual que en los contaminantes anteriores, vemos como en cada caso la selección de hiperparámetros parece contribuir al aumento de precisión a la hora de predecir valores de contaminación. Aunque en ciertos casos puede mantenerse o disminuir levemente. En el caso de las partículas PM10, como se puede observar en las tablas anteriores, es nuevamente el modelo *Nearest Neighbors* el que mas se beneficia, aumentando la precisión en un 0.1442, 0.1263 y 0.0936 para 5, 10 y 15 variables respectivamente. Datos reflejado en las Tablas 5.23, 5.24 y 5.25.

5.4.3. Proyección a baja dimensión

En esta sección se presentan los resultados del Análisis de Componentes Principales conocido como PCA.

	PCA
Decision Tree	-0.1264
Redes	-0.0007
Random Forest	0.0150
Nearest Neighbors	-0.0059
Xgboost	0.0263
Ensamble	0.0720
Linear Regressor	-0.0025
KRR	-0.0014
Lasso	-0.0025

Tabla 5.26: Evolución de valores tras PCA

En la Tabla 5.26 se puede ver claramente como el análisis de componentes principales otorga para nuestro conjunto de datos unos resultandos bastante deficientes.

5.4.4. Eliminación de outliers

Llegados a este punto podemos observar que el mejor resultado lo obtiene el modelo *Random Forest* sin selección de variables. El valor del coeficiente de determinación en este caso es de 0.8936, valor que logra mejorar el inicial en un 0.0088. Dado que queremos encontrar el modelo que mejor se ajuste a los datos, nos enfocaremos ahora en mejorar el valor de precisión del modelo *Random Forest*.

En esta sección nos centramos en la eliminación de outliers encontrados en el inicial análisis descriptivo de las variables. Al igual que los datos de PM2.5, los datos obtenidos de PM10 no poseen outliers. Por ello nos centraremos nuevamente en los presentes en la precipitación y el tráfico. Antes de comenzar se realizará el filtro de variables y una vez obtenida la matriz reducida se hará la selección de características, evaluando el coeficiente de determinación en cada caso.

$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.8879	0.8959	0.8866	0.8808	0.8883

Tabla 5.27: Eliminación de outliers [precipitación y tráfico]. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y la reducción a 5, 10 y 15 variables de la mejor versión del modelo *Random Forest*.

La Tabla 5.27 muestra como la eliminación de outliers en el contexto de nuestro proyecto supone una mayor precisión para el modelo *Random forest* a la hora de predecir los niveles de PM10. Esta técnica obtiene un coeficiente de determinación de 0.8959, mejorando el previo valor máximo de 0.8936.

5.4.5. Cuantificación

Esta sección se centra en la cuantificación de los valores de PM10 previa a la aplicación del modelo *Random forest*. Para este contaminante se ha mantenido también un rango de amplitud 10 µg/m3. Esta cuantificación se aplica en primer lugar a la matriz completa, realizando la selección de variables. En segundo a la matriz reducida cuyo modelo ha obtenido mejores resultados, *Random forest* con 10 variables. Los resultados se muestran a continuación.

$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$	$Precisión_{10}$	$Precisión_{15}$
0.8588	0.8643	0.8705	0.8855	0.8845

Tabla 5.28: Cuantificación matriz completa. Tabla presenta los valores del coeficiente de determinación tras la reducción a 5, 10 y 15 variables de la matriz original.

$Precisión_{Inicial}$	$Precisión_{Hyper}$
0.8761	0.877

Tabla 5.29: Cuantificación *Random Forest* 10 variables. Tabla presenta los valores del coeficiente de determinación tras la cuantificación del conjunto de datos reducido a 10 variables.

Como se puede observar, la cuantificación efectivamente ayuda a aumentar la precisión. Sin embargo, al igual que con las partículas de PM2.5, no a los niveles esperados.

5.4.6. Entrenamiento de modelos por estación

Como ya se ha mencionado, esta sección explora la posibilidad de obtener una mayor precisión de predicción al realizar un modelo por estación. Es importante volver a destacar que para llevar acabo esta técnica es necesario eliminar todas aquellas variables constante para cada estación, estas son todas las variables que hemos denominado como estáticas. Así, el número total de variables se reduce a 11 y solo tiene sentido realizar la reducción de variables a 5. Aferrándonos al objetivo de mejorar la predicción del modelo *Random forest*, aplicaremos en primer lugar la modelización por estación únicamente a dicho modelo. Una vez completado, observaremos la estación con la mejor precisión y probaremos para dicha estación el resto de modelos. A continuación se presentan las tablas con los resultados obtenidos.

	$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$
Estación 8	0.5613	0.5781	0.5175
Estación 38	0.6703	0.6789	0.6624
Estación 48	0.725	0.7329	0.7353
Estación 50	0.6400	0.6261	0.6124
Estación 56	0.6256	0.6173	0.5938
Estación 57	0.6659	0.6622	0.6769

Tabla 5.30: Modelo *Random forest* para cada estación. Tabla presenta los resultados del coeficiente de determinación tras la selección de hiperparámetros y la reducción a 5 variables para cada estación de medida.

	$Precisión_{Inicial}$	$Precisión_{Hyper}$	$Precisión_5$
Decision Tree	0.3175	0.3175	0.4263
Random Forest	0.725	0.7329	0.7353
Xgboost	0.6927	0.7043	0.611
Linear Regressor	0.4289	0.4289	0.3923
Lasso	0.4058	0.4331	0.4144
Redes	0.4329	0.5589	0.5944
Nearest Neighbors	0.6372	0.7278	0.7178
Ensamble	0.7239	0.7239	0.699
KRR	0.4298	0.7372	0.6505

Tabla 5.31: Modelos para la estación 48. Tabla presenta los valores del coeficiente de determinación tras la selección de hiperparámetros y reducción a 5 variables para cada modelo, aplicado únicamente a la estación número 48.

Como se puede ver en la Tabla 5.30, la precisión del modelo *Random forest* por estación es baja comparada con los valores de precisión previamente obtenidos, pero considerablemente alta en ciertos casos. El valor más alto se obtiene para la estación número 48. Por ello, en la Tabla 5.31 se muestra la precisión de todos los modelos aplicados a dicha estación, valores en un rango muy amplio de los cuales se puede concluir que la creación de modelos por estación no mejora la precisión de un modelo aplicado a todas las estaciones de medida y sus localizaciones.

Capítulo 6

Conclusiones

Después de analizar los resultados obtenidos para cada uno de los contaminantes, procedemos a elaborar una conclusión final que integre todos los hallazgos del proyecto.

A lo largo de los capítulos anteriores, hemos realizado un estudio sobre la predictibilidad de diferentes medidas de contaminación en relación con un conjunto de variables explicativas. Dicho análisis concluye con la creación de modelos de predicción capaces de estimar la cantidad de partículas contaminantes de una manera muy cercana a la realidad, empleando para ello un total de 14 variables.

El primer bloque de este documento se ha dedicado a la captura y análisis de dichas variables. Parte de ellas han sido cuidadosamente seleccionadas y obtenidas de diversas fuentes públicas, mientras que otras se han construido manualmente tras considerar que podrían tener cierta correlación con los niveles de partículas contaminantes.

Como es sabido, existen numerosas métricas de contaminación, pero desafortunadamente no todas las estaciones de medida con las que hemos trabajado están equipadas con los sensores necesarios para obtener valores de todas ellas. Después de revisar la capacidad de medida de cada una de dichas estaciones, decidimos enfocarnos en las medidas de óxido nítrico (NO), partículas inferiores a 2.5 micrómetros (PM2.5) y partículas inferiores a 10 micrómetros (PM10), que son algunas de las medidas más utilizadas a la hora de discutir el efecto de la contaminación sobre la salud de los seres humanos. Cabe destacar que hubiese sido deseable realizar el mismo análisis con medidas de CO2 dada su relevancia en los estudios sobre movilidad ciudadana y su repercusión en las políticas de restricciones de tráfico. Sin embargo, ninguna de las estaciones desplegadas en la ciudad de Madrid dispone de estas capacidades de medida.

La tarea de predicción de estas tres variables ha seguido en todo momento un mismo patrón, análisis de la presencia de outliers, disminución de la dimensión del dataset original, aplicación de técnicas de selección de variable, y realización de

ajustes de hiperparámetros para los modelos más típicos empleados en este tipo de estudios.

En el caso de la predicción de NO el mejor valor de predicción se obtuvo con el modelo *Xgboost* reducido a 10 variables. Modelo que otorga un coeficiente de determinación r2 de 0.8428. Sin embargo, cabe destacar los resultados obtenidos con el modelo *Random forest*. En este caso, el valor del coeficiente de determinación r2 es de 0.8320 sin reducción de variables. Ante esta situación, nos debatimos entre la precisión del modelo, la interpretabilidad y la necesidad de computación. Dado que la diferencia de precisión entre el *Xgboost* y el *Random forest* no es muy significativa, aun siendo el *Random forest* un modelo con menor necesidad de computación, nos decidimos por el primero dado que un menor número de variables resulta en un modelo de mejor interpretabilidad. Factor de gran importancia para la toma de decisiones.

En el caso de la predicción de PM2.5 el mejor valor de predicción se obtuvo con el modelo *Ensamble* aplicando la reducción del dataset original para la eliminación de outliers. En este caso concreto las variables filtradas son el tráfico y la precipitación. De esta forma, el coeficiente de determinación r2 aumenta en un 0.01 respecto del modelo con el dataset completo, logrando un valor de 0.7547. A diferencia del óxido nítrico, la predominancia del modelo *Ensamble* respecto del resto de modelos es clara. Así, este es seleccionado con el más adecuado.

En el caso de la predicción de PM10 el mejor valor de predicción se obtuvo con el modelo *Random forest* sin eliminación de variables, con un coeficiente r2 de 0.8936. Es importante mencionar que los niveles de precisión logrados para este tipo de partículas son mayores. Así, múltiples modelos como *Ensamble*, *Xgboost* y *Random Forest* poseen un coeficiente de determinación superior a 0.85. Sin embargo, como se ha mencionado previamente, la interpretabilidad del modelo es de gran importancia a la hora de su selección, así como su parsimoniosidad, siempre y cuando no suponga una penalización importante en la precisión. Es por ello que nuevamente seleccionamos el modelo Random forestçomo el mas adecuado en este caso.

A la vista de estos resultados, podemos concluir que estos modelos pueden resultar de mucha utilidad para las autoridades locales tanto a la hora de establecer políticas de restricciones de tráfico, como a la hora de tomar otro tipo de decisiones de planificación urbana. Esto se desarrolla más en el siguiente capítulo.

Capítulo 7

Líneas Futuras

En este apartado vamos a identificar una serie de puntos de mejora del modelo propuesto en este trabajo. Los resultados han sido muy positivos, se han logrado obtener medidas de precisión por encima del 75 % para los tres contaminantes considerados. Para mejorar los resultados obtenidos en este documento, se proponen las siguientes líneas de trabajo.

La primera medida sería el evaluar modelos más avanzados para el análisis de series temporales como las redes neuronales recurrentes, lo cual puede suponer una mejora significativa en la precisión de las predicciones. Estos modelos son especialmente adecuados para capturar patrones y tendencias en datos que proceden de secuencias temporales, lo que los hace ideales para el análisis de la acumulación de partículas contaminantes a lo largo del tiempo.

También puede suponer una mejora importante ampliar el conjunto de variables explicativas consideradas en el estudio. La inclusión de un mayor número de variables puede proporcionar una visión más completa y precisa de los factores influyentes en la contaminación. Así, es recomendable investigar e identificar variables adicionales con una correlación significativa con la acumulación de partículas contaminantes. Entre estas variables podrían incluirse, por ejemplo, los límites de velocidad en las vías, las restricciones de tráfico y las actividades industriales cercanas.

Este estudio se puede aplicar a otro tipo de contaminantes como por ejemplo a la predicción de CO2, dada su relevancia en temas de contaminación urbana. Para la recopilación de datos en este caso se podría establecer una colaboración con empresas como OPUS RSE. Empresas capaces de desplegar medidores de CO2 por las ciudades para un seguimiento exhaustivo de sus niveles, que permitirían implementar un esquema de entrenamiento de modelos de predicción similar al descrito en este documento.

La disposición de medidas reales junto con la posibilidad de generar nuevos valores estimados con estos modelos en cualquier punto de la ciudad, permitiría

construir un mapa de calor interactivo y a tiempo real que estime la contaminación en todas aquellas vías de Madrid para las que hay medidas de tráfico. Este tipo de mapas no solamente resultaría de utilidad para las autoridades locales sino que se podría compartir con la población, simplificando notablemente el acceso a estos datos que serían de mucha utilidad para enfermos con patologías respiratorias. Adicionalmente, esta medida puede ayudar a concienciar a la población sobre la concentración de contaminantes en las calles de su ciudad y animarles a participar activamente en su reducción, como por ejemplo a nivel de tráfico vehicular.

Capítulo 8

Discusión

Después de profundizar en las variables predictoras y en su eficacia para la predicción de métricas de contaminación, creemos que queda lugar para la investigación y mejora de los resultados obtenidos en este trabajo.

Los resultados para los modelos seleccionados han superado el 75 % de precisión en todos ellos, y en el caso de la predicción de NO y PM2.5, la precisión ha sido superior al 80 %. Todo esto teniendo en cuenta lo limitado del acceso a variables predictoras o a más estaciones de medida. Es cierto que la Comunidad de Madrid tiene muchas más estaciones de medida de calidad del aire de las empleadas en este estudio, pero sólo las utilizadas aquí tenían una estación de medida de tráfico en la misma vía.

Los resultados tan optimistas nos animan a pensar a que este tipo de modelos pueden llegar a sustituir a los equipos de medidas, permitiendo un ahorro importante en infraestructura y su mantenimiento. Esta sustitución no puede ser definitiva porque los patrones de comportamiento de los madrileños pueden variar, así como la sustitución de la tecnología de combustión por vehículos eléctricos. Por ello, seguiría siendo necesario re-entrenar estos modelos periódicamente. Sin embargo, si que pueden permitir obtener una mayor granularidad de medidas de emisiones sin que esto suponga un aumento en la inversión.

El tema del aumento de la granularidad busca mejorar la resolución de las medidas para lograr un mapa más detallado. La mejora de la granularidad tiene mucho que ver con los resultados de este estudio, en el que hemos comprobado que variables como el ancho de la calle, la altura de los edificios, o el volumen de tráfico tienen una relación directa con los niveles de emisiones. Esto puede suponer que en una determinada calle haya unos niveles de contaminación tolerables, mientras que en una paralela, los niveles de contaminación sean peligrosos para la salud. Actualmente hay 38 estaciones de medida, claramente insuficientes para poder detectar calles con niveles de contaminación peligrosos.

Por otro lado, la capacidad de estimar las emisiones a partir de variables ya

disponibles, todas ellas obtenidas en tiempo real, permite a las autoridades locales no sólo tener un mapa de contaminación actualizado, sino también ser capaces de reaccionar a picos de contaminación en zonas de especial sensibilidad, como pueden ser zonas en las que haya población sensible (niños o ancianos) o zonas de alta densidad.

Los resultados de este trabajo muestran que la posibilidad de predecir la contaminación urbana es factible y se puede utilizar para intervenir en la planificación de zonas urbanas. Este tipo de modelos pueden utilizarse para simular nuevos escenarios y optar por aquellos que permiten minimizar la concentración de partículas contaminantes. Es decir, mejorar la planificación urbana de zonas mediante la simulación de diferentes condiciones de contaminación en función de variables sobre las que las autoridades tengan capacidad de actuación. Un ejemplo de esto pueden ser la construcción de zonas verdes, el número de carriles de una carretera o la limitación de las alturas de los edificios.

Por último, cabe destacar que la mejora de la concentración de emisiones contaminantes tiene una repercusión directa en el gasto en salud de una comunidad. Estudios indican que en países desarrollados, un 20 % de la incidencia total de enfermedades puede atribuirse a factores medioambientales [1]. La posibilidad de minimizar estas emisiones implicaría una disminución en el número de afectados, resultando en un ahorro muy importante en el sistema sanitario actual.

Bibliografía

[1] Francisco Vargas Marcos. *La contaminación ambiental como factor determinante de la salud*. 2005.

[2] Yasuo Yoshikawa, Hitoshi Kunimi y Shizuo Ishizawa. "Numerical simulation model for predicting air quality along urban main roads: first report, development of atmospheric diffusion model". En: *Heat Transfer-Japanese Research: Co-sponsored by the Society of Chemical Engineers of Japan and the Heat Transfer Division of ASME* 27.7 (1998), págs. 483-496.

[3] Parlamento Europe. *Reducir las emisiones de carbono: objetivos y políticas de la UE*. 2024. URL: `https://www.europarl.europa.eu/topics/es/article/20180305STO99003/reducir-las-emisiones-de-carbono-objetivos-y-politicas-de-la-ue`.

[4] Ch Vlachokostas et al. "Combining regression analysis and air quality modelling to predict benzene concentration levels". En: *Atmospheric environment* 45.15 (2011), págs. 2585-2592.

[5] Paul E Bieringer et al. "A method for targeting air samplers for facility monitoring in an urban environment". En: *Atmospheric Environment* 80 (2013), págs. 1-12.

[6] Jasleen Kaur Sethi y Mamta Mittal. "Ambient air quality estimation using supervised learning techniques". En: *EAI Endorsed Transactions on Scalable Information Systems* 6.22 (2019), e8-e8.

[7] Yuelai Su. "Prediction of air quality based on Gradient Boosting Machine Method". En: *2020 International Conference on Big Data and Informatization Education (ICBDIE)*. IEEE. 2020, págs. 395-397.

[8] Mauro Castelli et al. "A machine learning approach to predict air quality in California". En: *Complexity* 2020 (2020).

[9] Meng Gao, Liting Yin y Jicai Ning. "Artificial neural network model for ozone concentration estimation and Monte Carlo analysis". En: *Atmospheric Environment* 184 (2018), págs. 129-139.

[10] Yuan Huang et al. "Air quality prediction using improved PSO-BP neural network". En: *Ieee Access* 8 (2020), págs. 99346-99353.

[11] Xavier Querol. *La calidad del aire en las ciudades*. 2018.

[12] Xavier Querol et al. "Calidad del aire urbano, salud y tráfico rodado". En: *Fundación Gas Natural* (2006).

[13] J Baldasano et al. *16 MEJORA DE LA CALIDAD DEL AIRE POR CAMBIO DE COMBUSTIBLE A GAS NATURAL EN AUTOMOCIÓN. APLICACIÓN A MADRID Y BARCELONA*. 2009.

[14] Emiliano Aránguez et al. "Contaminantes atmosféricos y su vigilancia". En: *Revista española de salud pública* 73 (1999), págs. 123-132.

[15] Ferran Ballester. "Contaminación atmosférica, cambio climático y salud". En: *Revista Española de salud pública* 79 (2005), págs. 159-175.

[16] Laura Fernández Rivas. "TRABAJO FIN DE GRADO El papel de los óxidos de nitrógeno en el Cambio Climático. Efectos sobre la salud." Tesis doct. UNIVERSIDAD COMPLUTENSE, 2015.

[17] scikit-learn. *Machine Learning in python*. URL: `https://scikit-learn.org/stable/`.

[18] Will Koehrsen. *Random Forest in Python*. 2017. URL: `https://towardsdatascience.com/random-forest-in-python-24d0893d51c0`.

[19] Xiaotong Yao, Xiaoli Fu y Chaofei Zong. "Short-Term Load Forecasting Method Based on Feature Preference Strategy and LightGBM-XGboost". En: *IEEE Access* 10 (ene. de 2022), págs. 1-1. DOI: `10.1109/ACCESS.2022.3192011`.

[20] Joaquín Amat Rodrigo. *Redes neuronales con Python*. 2021. URL: `https://cienciadedatos.net/documentos/py35-redes-neuronales-python`.

[21] Dennis Bakhuis. *k-nearest-neighbor (KNN) algorithm*. 2020. URL: `https://towardsdatascience.com/k-nearest-neighbor-knn-algorithm-3d16dc5c45ef`.

[22] Gang Luo. "A review of automatic selection methods for machine learning algorithms and hyper-parameter values". En: *Network Modeling Analysis in Health Informatics and Bioinformatics* 5 (2016), págs. 1-16.

[23] Gautam Kumar. *How to identify the outliers in your Data??* URL: `https://kmr-gautam2893.medium.com/how-to-identify-the-outliers-in-your-data-ee9c28b42fc3`.

[24] Ajuntamiento de Madrid. *Portal de datos abiertos del Ayuntamiento de Madrid*. URL: `https://datos.madrid.es/portal/site/egob/`.

[25] Naciones Unidas. *Paz, dignidad e igualdad en un planeta sano*. URL: `https://www.un.org/es/`.

Anexo I

ALINEACIÓN DEL PROYECTO CON LOS ODS

Como se ha mencionado anteriormente, uno de los objetivos principales de este proyecto es contribuir a la reducción, en un 55 %, de las emisiones netas de gases de efecto invernadero para el año 2030, contribuyendo así, con la Ley Europea del Clima aprobada por el Parlamento Europeo. Este objetivo sin duda se alinea con dos de los Objetivos de Desarrollo Sostenible establecidos por las Naciones Unidas para abordar desafíos globales. El primero de ellos es "Ciudades y comunidades sostenibles", correspondiente al Objetivo de Desarrollo Sostenible 11. El segundo es "Acción por el clima", Objetivo de Desarrollo Sostenible 13.

EL ODS número 11 pretende lograr que las ciudades sean inclusivas, seguras, resilientes y sostenibles. En un contexto en el que el crecimiento urbano es exponencial, la contaminación atmosférica y la escasez de espacios abiertos suponen un desafío [25], dificultando alcanzar el desarrollo sostenible. Los modelos de predicción de calidad del aire que se desarrollan en este proyecto buscan identificar zonas de alta contaminación y la magnitud de las variables influyentes en ellas. Con esta información se podrán tomar medidas de prevención y mitigación tales como la reducción del tráfico vehicular en las zonas urbanas más problemáticas, o la expansión de zonas verdes y áreas peatonales. Medidas que contribuyen con el ODS 11.

El ODS número 13 se centra en tomar medidas urgentes para combatir el cambio climático y sus impactos, consecuencia directa de los contaminantes presentes en nuestro planeta. "Para limitar el aumento global de la temperatura muy por debajo de los 2 °C, o incluso de 1,5 °C, el mundo debe transformar sus sistemas energéticos, industriales, de transporte, alimentarios, agrícolas y forestales"[25]. Este proyecto contribuye a dicha transformación. El uso de los modelos de predicción desarrollados permite un monitoreo efectivo de las emisiones contaminantes que facilita su reducción y consecuentemente los efectos del cambio climático.

A modo de conclusión, el desarrollo de un modelo de predicción de calidad del aire tiene un impacto significativo en la creación de ciudades más sostenibles y en la lucha contra el cambio climático. Las medidas basadas en este modelo no solo mejoran la calidad del aire, sino que también promueven un desarrollo urbano más saludable, alineándose plenamente con los Objetivos de Desarrollo Sostenible 11 y 13.