SISTEMA PREDICTIVO DE POLUCIÓN

Alfredo
Daniela Echeverría García
Antonio García Toro
Miguel Ruiz Rivilla
Alejandro Álvarez Tenedor

Madrid, 11 de junio de 2024

Contenido

1. Introducción

1.1 Contexto y Motivación

La contaminación del aire es uno de los principales desafíos ambientales y de salud pública que enfrentan las ciudades modernas. Las áreas urbanas, con su alta densidad de población, tráfico vehicular intenso, y actividades industriales, son particularmente vulnerables a los niveles elevados de contaminantes atmosféricos. El estudio de la contaminación en estas áreas es crucial por varias razones:

- Salud pública: La exposición a altos niveles de contaminantes del aire, como el dióxido de nitrógeno (NO2), partículas en suspensión (PM10 y PM2.5), monóxido de carbono (CO), y ozono (O3), está asociada con una amplia gama de problemas de salud. Estos incluyen enfermedades respiratorias y cardiovasculares, cáncer de pulmón, y mortalidad prematura. La Organización Mundial de la Salud (OMS) estima que la contaminación del aire es responsable de millones de muertes prematuras anualmente.

- Impacto ambiental: La contaminación del aire no solo afecta la salud humana, sino que también tiene efectos adversos sobre el medio ambiente. Puede dañar la vegetación, contaminar cuerpos de agua, y contribuir al cambio climático. Por ejemplo, el dióxido de carbono (CO2) y el metano (CH4) son gases de efecto invernadero que exacerban el calentamiento global.

- Calidad de vida: La calidad del aire es un factor determinante de la calidad de vida en las ciudades. Niveles elevados de contaminación pueden reducir la visibilidad, generar olores desagradables y afectar negativamente la estética urbana. Además, la contaminación puede limitar las actividades al aire libre y reducir el atractivo turístico de las ciudades.

Un estudio europeo concluye que Madrid es la urbe de Europa con niveles más altos de dióxido de nitrógeno en el aire y la mayor proporción de muertes por número de habitantes relacionada con sus efectos en la salud. [1]

Figura 1. Contaminación del cielo de Madrid

1.2 Objetivos del Estudio

El objetivo principal de este proyecto es desarrollar un modelo predictivo que pueda prever los niveles de contaminación en Madrid con precisión. Este modelo servirá como una herramienta invaluable para los habitantes de la ciudad, permitiéndoles tomar decisiones informadas sobre sus actividades al aire libre, especialmente aquellas relacionadas con la salud, como el ejercicio físico.

A través del desarrollo de dicho proyecto investigaremos y cuantificaremos la influencia de diversas variables sobre los niveles de contaminación en Madrid. Este análisis es esencial para comprender las dinámicas subyacentes que afectan la calidad del aire y, por ende, la salud pública y el medio ambiente.

2. Modelo predictivo de concentración de NO_2

En este apartado se va a explicar todo el proceso referente al desarrollo del modelo predictivo de concentración de NO2 en la Comunidad de Madrid. Pese a que la polución está relacionada con diferentes tipos de partículas y gases, en este caso vamos a centrar el estudio en la partícula de NO2 debido al gran impacto negativo que tiene en la salud humana una alta concentración de dicho contaminante.

Dicho proceso involucra todo lo relacionado a un modelo CRISP-DM el cual involucra la extracción de los datos y su posterior comprensión. Tras esto se aplica una limpieza y transformación de los datos, así como la generación de nuevas variables. Finalmente, se entrenarán los diferentes modelos de Machine Learning, se evaluarán los resultados y finalmente se pondrá en producción el modelo más robusto y preciso.

2.1 Extracción de los datos

Para el desarrollo de dicho proyecto se han utilizado 4 fuentes de datos:

Datos meteorológicos: se han utilizado datos horarios recogidos desde el año 2019 hasta 2024 procedente del portal de datos abiertos de Madrid [2].

- Velocidad del viento
- Dirección del viento
- Temperatura
- Humedad relativa
- Presión Barométrica
- Precipitación

Polución del aire: los datos de contaminación atmosférica han sido medidos automáticamente en las estaciones de medición de diferentes puntos de la Comunidad de Madrid [3]. La calidad del aire se encuentre influenciada por la presencia en el aire de los siguientes materiales contaminantes:

- CO: El monóxido de carbono es un gas inodoro, incoloro y tóxico que se genera como producto de los dispositivos de combustión. En consecuencia, se encuentra en niveles elevados en lugares y horas del día con alto tráfico.

  Respirar niveles altos de CO puede causar daños permanentes del corazón y el cerebro.
- NO: el monóxido de carbono es un compuesto químico originado por la combustión incompleta de carburantes fósiles y de biocombustibles.
- NO2: El dióxido de nitrógeno es una sustancia química emitida en los motores de combustión, especialmente de diésel. Su exposición se encuentra asociada con enfermedades respiratorias crónicas y al envejecimiento prematuro de los pulmones.

- NOX: Los óxidos de nitrógeno (NOx) son una mezcla de gases compuestos por monóxido de nitrógeno (NO) y dióxido de nitrógeno (NO2). Estos gases son producidos principalmente por la combustión de combustibles fósiles en vehículos y plantas industriales.

- Partículas por millón 2.5 (PM2,5): las partículas en suspensión son una mezcla de sustancias químicas orgánicas, polvo, hollín y metales. Estas partículas, como su nombre lo indica, tienen un diámetro de 2,5 micras, por lo que resultan más finas que un cabello humano, lo que implica que pueden pasar al torrente sanguíneo a través de la respiración pudiendo provocar enfermedades respiratorias y cardiovasculares.

- Partículas por millón 10: son partículas en suspensión con un diámetro menor de 10 micras. Al igual que las PM2.5, las PM10 son una mezcla de polvo, hollín, metales y sustancias químicas orgánicas. Estas partículas pueden ser inhaladas profundamente en los pulmones y están asociadas con enfermedades respiratorias y cardiovasculares, aunque no penetran tan profundamente en el sistema respiratorio como las PM2.5.

Pese a que se muestran diferentes tipos de partículas relacionadas con la polución, como hemos comentado previamente, se va a analizar únicamente la partícula NO2 resultante de la combustión de motores diésel o gasolina, así como del funcionamiento de sistemas de calefacción.

Datos de intensidad de tráfico: se han tomado datos referentes a la intensidad de tráfico medida en vehículos/hora en diferentes puntos de Madrid [4]. Estos datos también han sido extraídos sobre el repositorio abierto de información que pone a disposición el Ayuntamiento de Madrid.

Festividades de la Comunidad de Madrid: se ha extraído un calendario de festividades de la Comunidad de Madrid desde el año 2019 a la actualidad [5].

La extracción de los datos se ha realizado de forma manual directamente desde la página del Ayuntamiento de Madrid para el caso de los datos de polución y festividades. Para el caso de los datos de polución e intensidad de tráfico fue necesario aplicar técnicas de Web Scrapping dado que los datos venían de forma mensual y por lo tanto la extracción manual se volvía costosa.

2.2 Preprocesamiento de Datos

Una vez extraídos todos los datos en crudo, el primer paso consiste en procesarlos, mergearlos y finalmente limpiarlos.

Dado que disponemos de diferentes puntos de medida, a lo largo de los diferentes distritos de la Comunidad de Madrid, se decidió centrar el análisis en uno de ellos situado en la Calle de Alcalá.

Los datasets relacionados con la polución y la información meteorológica presentaban un formato tabular con una estructura similar. Se indicaba la provincia, el municipio, la estación desde la que se realiza la medida y finalmente la magnitud medida. Además, se indica el año, mes y día y a continuación 24 columnas referentes a las 24 horas del día con cada una de las medidas de la magnitud analizada.

	PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	H02	V02
0	28	79	102	81	28079102_81_98	2019	1	1	0.65	V	0.95	V
1	28	79	102	81	28079102_81_98	2019	1	2	0.50	V	0.95	V
2	28	79	102	81	28079102_81_98	2019	1	3	2.22	V	2.53	V
3	28	79	102	81	28079102_81_98	2019	1	4	0.87	V	0.77	V
4	28	79	102	81	28079102_81_98	2019	1	5	0.57	V	1.82	V
5	28	79	102	81	28079102_81_98	2019	1	6	1.08	V	1.33	V
6	28	79	102	81	28079102_81_98	2019	1	7	1.88	V	1.55	V
7	28	79	102	81	28079102_81_98	2019	1	8	1.28	V	1.22	V
8	28	79	102	81	28079102_81_98	2019	1	9	1.17	V	0.67	V
9	28	79	102	81	28079102_81_98	2019	1	10	1.67	V	2.28	V

Figura 2. Datos atmosféricos en crudo.

Dado que dichos datasets contienen datos diarios, fue necesario aplicar dos transformaciones a los datos, la primera de ellas para trasponer las columnas de las horas en filas de forma que tuviésemos una hora del día por fila. La segunda transformación consistió en trasponer el dato de la magnitud de forma que tuviésemos en cada fila las diferentes magnitudes meteorológicas y de polución medidas. A continuación, se muestra cómo quedó el dataset de variables meteorológicas tras aplicar las diversas transformaciones:

datetime	vel_viento	dir_viento	temperatura	humedad_relativa	presion_barometrica	precipitacion
2019-01-01 00:00:00	0.68	48.0	2.6	71.0	959.0	0.0
2019-01-01 01:00:00	0.69	32.0	2.5	71.0	959.0	0.0
2019-01-01 02:00:00	0.70	43.0	1.7	74.0	959.0	0.0
2019-01-01 03:00:00	0.70	74.0	1.1	75.0	959.0	0.0
2019-01-01 04:00:00	0.66	67.0	1.2	74.0	959.0	0.0
2019-01-01 05:00:00	0.67	31.0	0.2	78.0	959.0	0.0
2019-01-01 06:00:00	0.79	67.0	-0.3	80.0	959.0	0.0
2019-01-01 07:00:00	0.77	62.0	-0.3	79.0	959.0	0.0
2019-01-01 08:00:00	0.85	78.0	-1.0	81.0	960.0	0.0
2019-01-01 09:00:00	0.71	82.0	-0.3	77.0	960.0	0.0

Figura 3. Datos atmosféricos procesados.

En lo que respecta a los datos de intensidad de tráfico, los datos se encuentran recogidos con una frecuencia de 15 minutos donde se extraen las variables de intensidad de tráfico, ocupación, carga y velocidad media. Además, se indican variables de calidad del dato como error o periodo de integración. Es importante destacar que las variables ocupación y carga son complementarias a la intensidad, dando indicativos de como de ocupada se encentra la vía durante ese periodo de 15 minutos.

En el caso del dataset de intensidad de tráfico, se han seleccionado únicamente las horas puntas para poder así mergearlo con los datasets anteriormente comentados. En lo que respecta al dataset de festivos, dado que el dato es diario, se ha expandido a 24 horas, para facilitar el mergeo con el resto de datasets.

Finalmente, una vez que disponíamos de todos los datasets correctamente procesados y transformados, el paso final fue juntarlos en un único dataset y a partir del cual trabajaríamos. El aspecto final de dicho dataset es el siguiente:

	datetime	vel_viento	dir_viento	temperatura	humedad_relativa	presion_barometrica	precipitacion	NO2	intensidad	ocupacion	carga
0	2019-01-01 00:00:00	0.68	48.0	2.6	71.0	959.0	0.0	73.0	135	0.0	7
1	2019-01-01 01:00:00	0.69	32.0	2.5	71.0	959.0	0.0	82.0	690	4.0	27
2	2019-01-01 02:00:00	0.70	43.0	1.7	74.0	959.0	0.0	72.0	691	35.0	51
3	2019-01-01 03:00:00	0.70	74.0	1.1	75.0	959.0	0.0	66.0	507	3.0	20
4	2019-01-01 04:00:00	0.66	67.0	1.2	74.0	959.0	0.0	64.0	371	2.0	14
5	2019-01-01 05:00:00	0.67	31.0	0.2	78.0	959.0	0.0	56.0	296	2.0	11
6	2019-01-01 06:00:00	0.79	67.0	-0.3	80.0	959.0	0.0	56.0	318	0.0	12
7	2019-01-01 07:00:00	0.77	62.0	-0.3	79.0	959.0	0.0	52.0	309	2.0	13
8	2019-01-01 08:00:00	0.85	78.0	-1.0	81.0	960.0	0.0	49.0	209	1.0	7
9	2019-01-01 09:00:00	0.71	82.0	-0.3	77.0	960.0	0.0	46.0	147	1.0	6

Figura 4. Dataset con todas las variables predictoras y la variable objetivo NO2.

Dicho dataset contiene un total de 43.847 registros. A continuación, se van a exponer las unidades de medida de cada una de las variables predictoras:

Variable	*Unidad de medida*
vel_viento	m/s
dir_viento	-
temperatura	$C°$
humedad_relativa	%
presion_barometrica	mb
precipitacion	l/m^2
intensidad	$vehículos/hora$
ocupacion	%
carga	%
NO2	$\mu g/m^3$

Tabla 1. Unidades de medida de las variables predictoras y objetivo.

2.2.1 Tratamiento de Valores Faltantes

Una vez formado el dataset sobre el que trabajaremos, el siguiente paso será analizar y tratar aquellos valores faltantes. En el gráfico 5 se observa el porcentaje de valores faltantes en cada una de las variables:

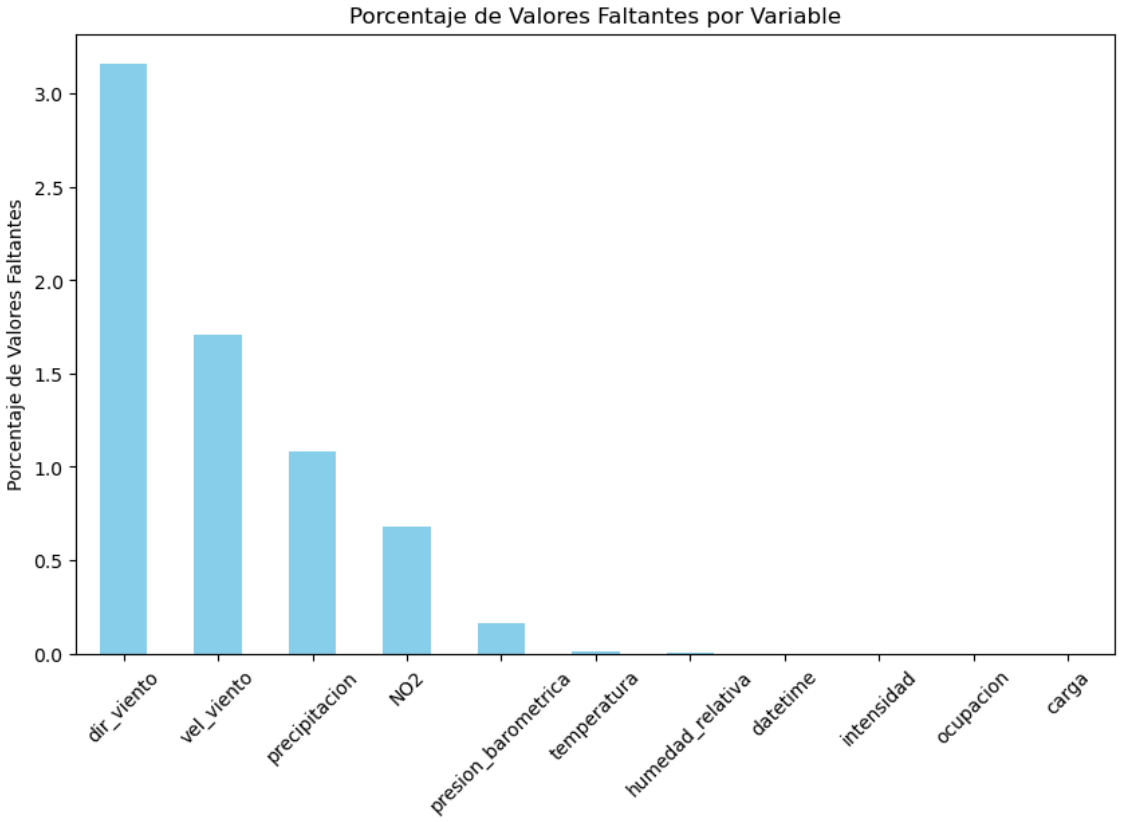


Figura 5. Porcentaje de valores faltantes por variable.

Como se puede observar, el dataset tiene un porcentaje bajo de valores faltantes no superando el 4% en el peor de los casos. En este caso podríamos optar por eliminarlos o por imputarlos. Dado que no se desea perder datos, se ha decidido imputar realizando una interpolación lineal (esto es debido a la estacionalidad que presentan las variables).

2.2.2 Tratamiento de Outliers

Otro factor a tener en cuenta sobre los datos de entrada son los outliers los cuales hacen referencia a valores atípicos debido a situaciones especiales o errores en la toma de datos. Como se puede observar en la figura 6, podemos decir que los datos presentan gran cantidad de outliers. Sin embargo, en nuestro caso únicamente eliminaremos aquellos referentes a errores de medida como pueden ser temperaturas de -55° o presiones extremadamente bajas.

El resto de 'outliers' son consideraros interesantes para el modelo, como por ejemplo días con altas precipitaciones o vientos considerablemente fuertes.

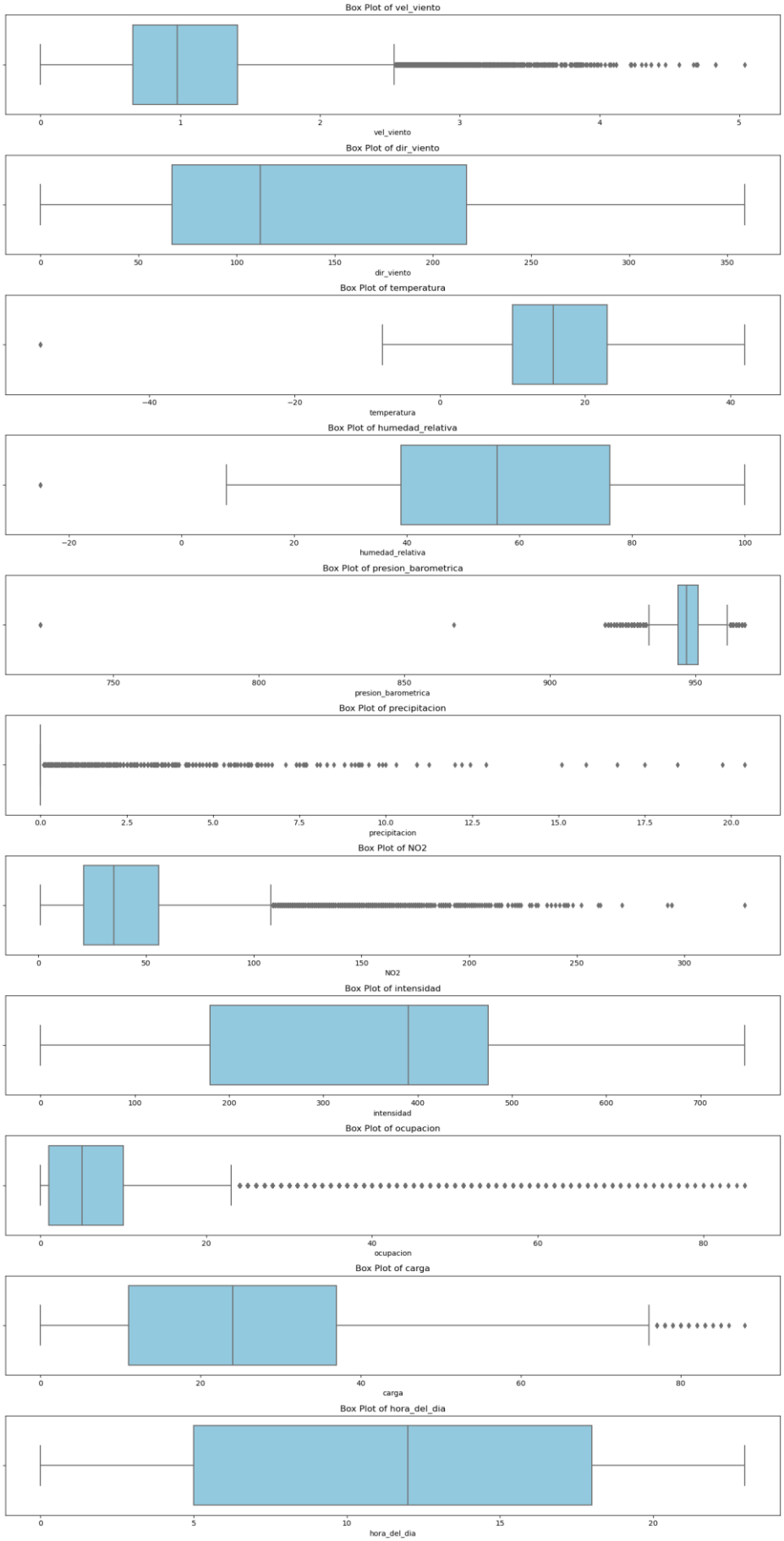


Figura 6. Outliers.

2.3 Feature Engineering

2.3.1 Generación de Nuevas Variables

Es de vital importancia generar nuevas variables a partir de las ya disponibles que puedan capturar relaciones no lineales con la variable objetivo, así como una posible mejora en la interpretación del problema. Las variables generadas son las siguientes:

- Hora del día
- Franja horaria
- Fin de semana
- Estación
- Festivo
- Confinamiento

Como podemos observar la mayoría de las variables hacen referencia a datos temporales. La variable confinamiento hace referencia al periodo de confinamiento sufrido entre marzo y junio de 2020.

2.3.2 Estandarización de Variables

Estandarizar las variables numéricas es un paso crucial en el preprocesamiento de datos para los modelos de aprendizaje automático ya que hará que todos nuestros datos tengan una escala uniforme teniendo media 1 y desviación estándar 0. Esto facilitará mejor convergencia en los modelos de Machine Learning así como una mejor interpretabilidad de los coeficientes de cada característica.

A continuación, se pueden observar las diferentes escalas de los datos numéricos de nuestro dataset:

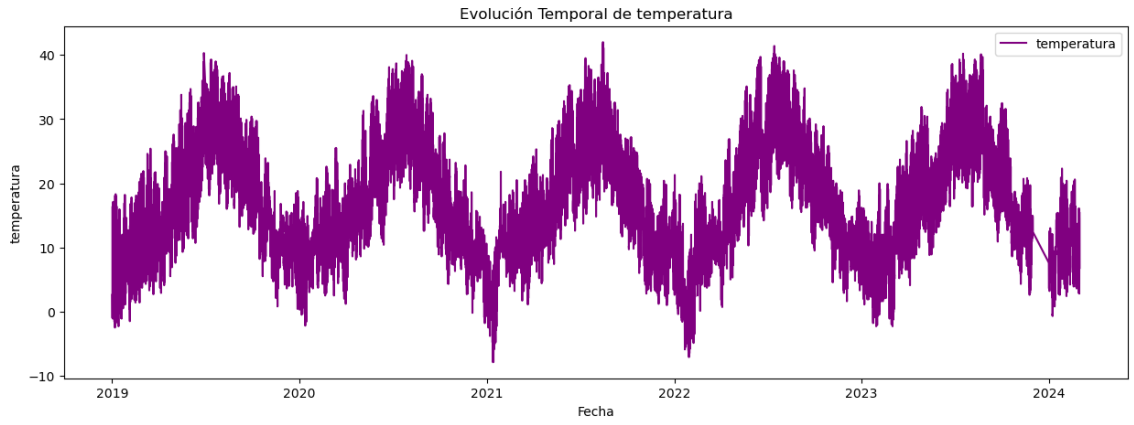
	vel_viento	dir_viento	temperatura	humedad_relativa	presion_barometrica	precipitacion	NO2	intensidad	ocupacion	carga
count	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000	43747.0000000000
mean	1.1117060598	138.4040734222	16.6323130729	57.3709511509	947.2925457746	0.0467663855	41.8488924955	341.4414016961	8.4781356436	25.2737330560
std	0.5512173895	82.8821758058	8.7556247209	21.9701839309	5.8735630348	0.4482376620	28.8464713857	172.6393491258	11.2994366948	15.9304216857
min	0.0000000000	0.0000000000	-7.9000000000	8.0000000000	921.0000000000	0.0000000000	1.0000000000	0.0000000000	0.0000000000	0.0000000000
25%	0.6700000000	67.0000000000	10.0000000000	39.0000000000	944.0000000000	0.0000000000	21.0000000000	180.0000000000	1.0000000000	11.0000000000
50%	0.9900000000	113.9058295964	15.6000000000	56.0000000000	947.0000000000	0.0000000000	35.0000000000	390.0000000000	5.0000000000	24.0000000000
75%	1.4100000000	215.0000000000	23.1000000000	76.0000000000	951.0000000000	0.0000000000	56.0000000000	475.0000000000	10.0000000000	37.0000000000
max	5.0400000000	359.0000000000	42.0000000000	100.0000000000	967.0000000000	20.4000000000	328.0000000000	747.0000000000	85.0000000000	88.0000000000

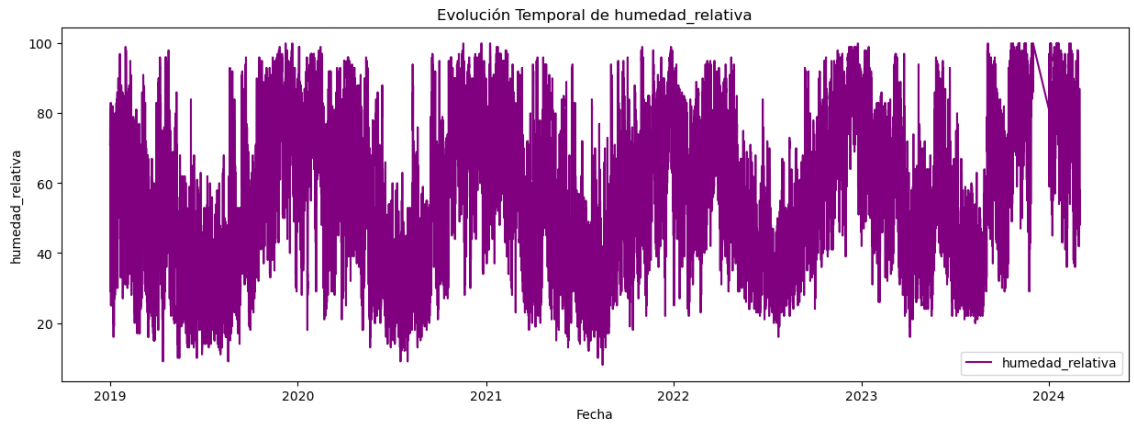
Figura 7. Datos descriptivos de las variables numéricas.

Se ha aplicado un *StandardScaler* a todas las variables numéricas que conforman el dataset, incluida la variable objetivo.

2.3.3 Codificación de Variables Categóricas

Los modelos de Machine Learning requieren de datos de entrada numéricos y por lo tanto es necesario aplicarle una transformación One Hot Encoding a todas las variables de tipo categóricas. De esta forma, se generarán tantas variables dummies como n-1 categorías tenga la variable:





franja_horaria_mañana	franja_horaria_medio_dia	franja_horaria_noche	franja_horaria_tarde	estacion_otoño	estacion_primavera	estacion_verano	confinamiento_si	festivo_si	fin_de_semana_si
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Figura 8. One Hot Enconding sobre las variables categóricas.

2.4 Exploración y Visualización de Datos

En este apartado se van a mostrar las diferentes visualizaciones sobre los datos, así como diagramas de correlaciones con la variable objetivo.

Comenzamos visualizando las variables meteorológicas desde 2019 a 2024:

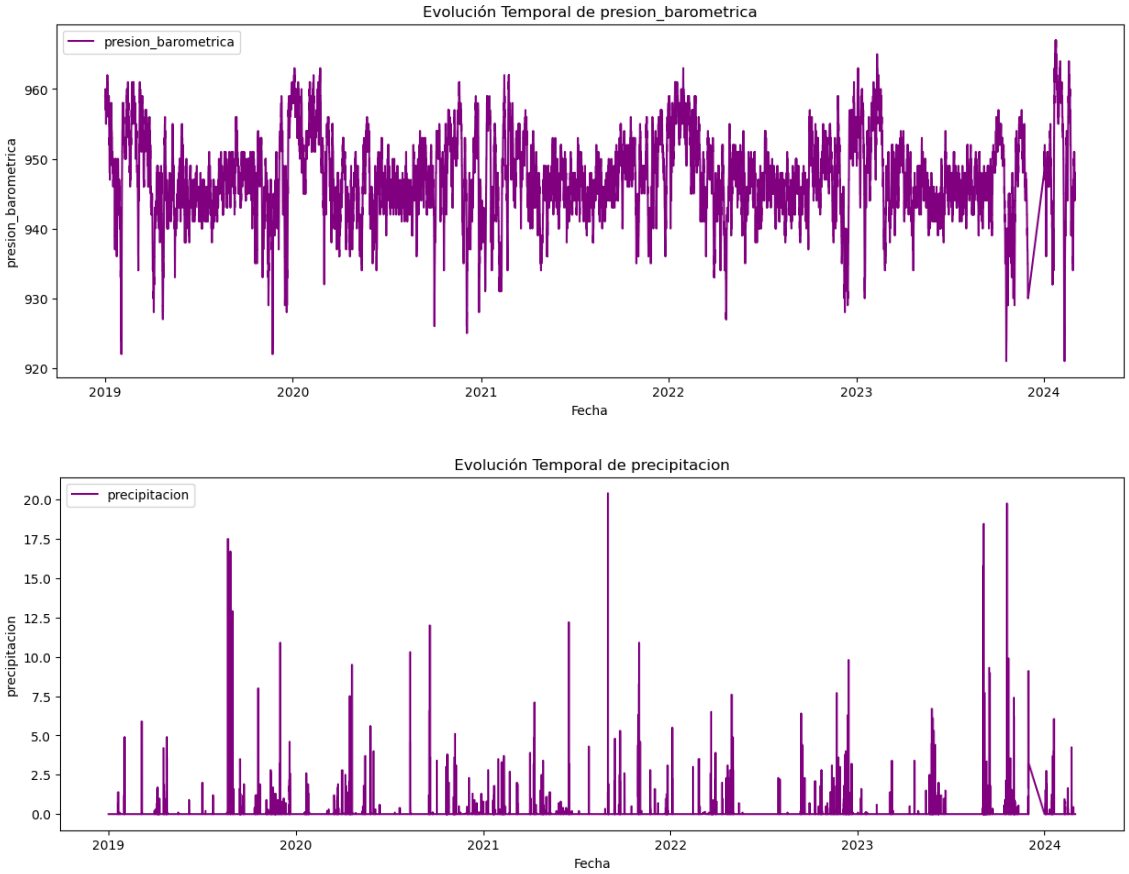


Figura 9. Evolución de las variables meteorológicas.

Como podemos observar, la gran mayoría de ellas presentan cierta estacionalidad (con forma senoidal) exceptuando la variable referente a las precipitaciones la cual presenta picos elevados en ciertos días puntuales.

Si pasamos a analizar la evolución de la variable referente a la intensidad de tráfico obtenemos lo siguiente:

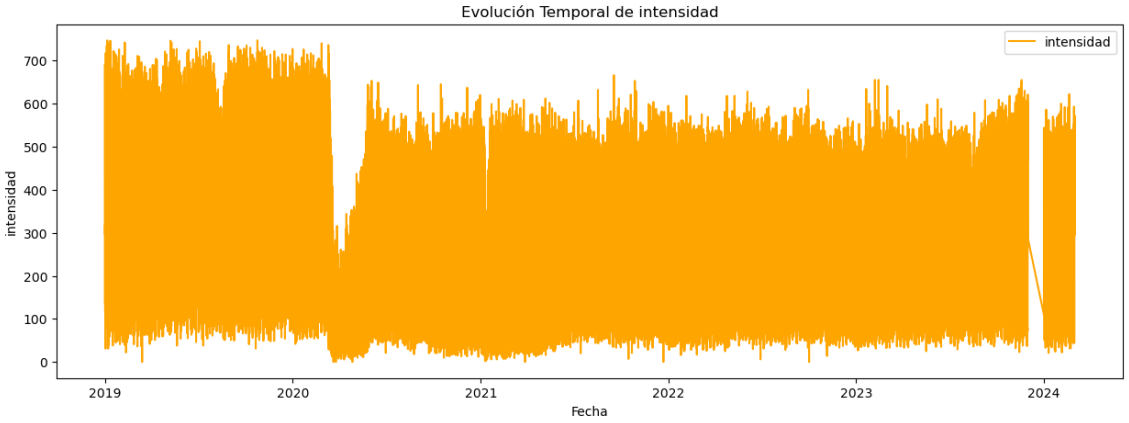


Figura 10. Evolución variable intensidad de tráfico.

Se observa la disminución en el tráfico en el periodo referente a la cuarentena. En lo que a finales de 2023 se refiere, el dataset no disponía de dicha información.

Finalmente se analiza la evolución de la variable objetivo NO2:

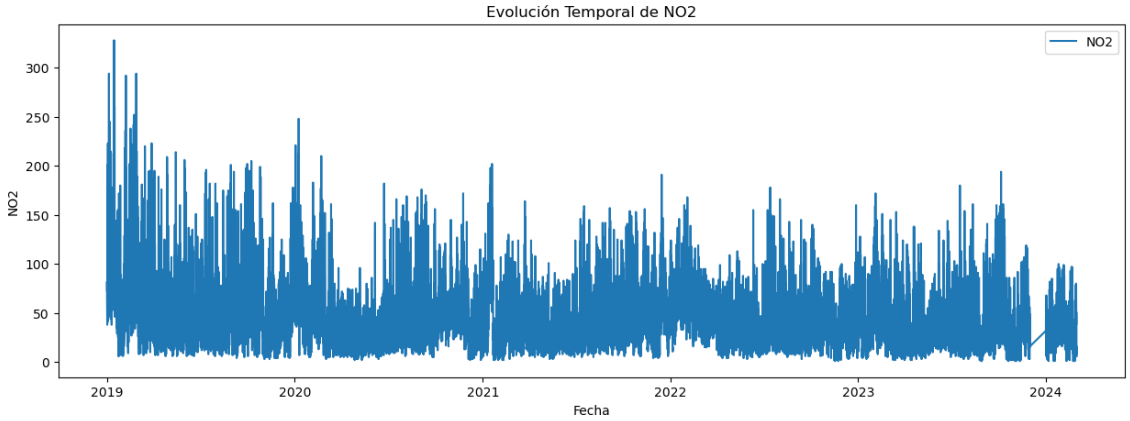


Figura 11. Evolución del NO2.

En este caso observamos que presenta cierta estacionalidad o tendencia. El objetivo será poder predecir con exactitud dicha evolución a lo largo del tiempo.

Tras analizar la evolución de las variables temporales, podemos analizar cómo varía la polución respecto a nuestras variables categóricas como puede ser la estación del año o la hora del día:

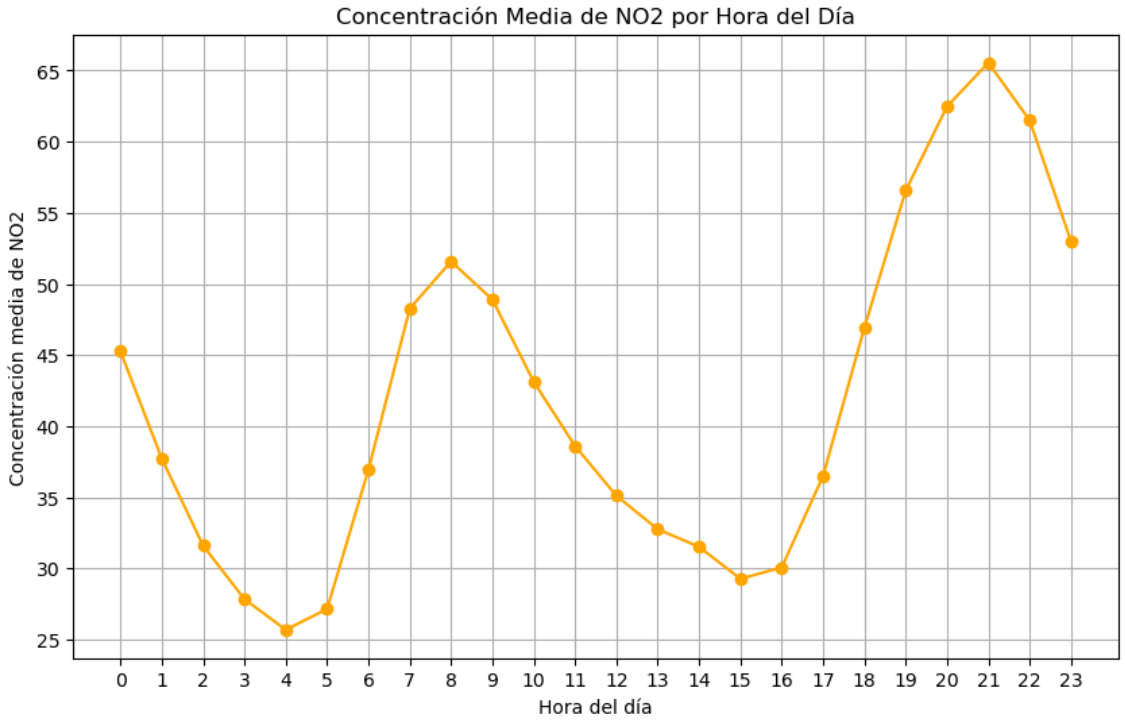


Figura 11. Evolución de la concentración de NO_2 a lo largo del día.

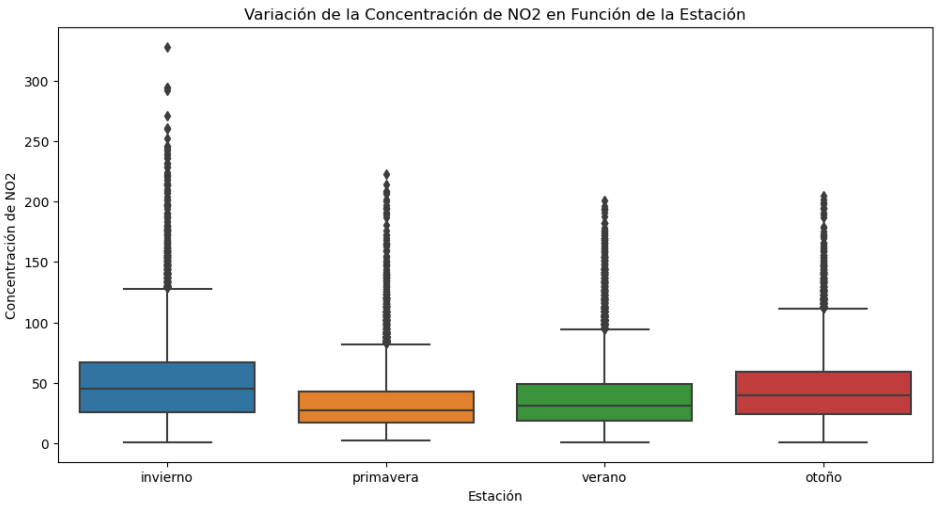


Figura 12. Distribución de la concentración de NO2 en las estaciones del año.

Analizando los gráficos anteriores, observamos la evolución del NO_2 a lo largo del día, observando mínimos durante la madrugada o a las 3 de la tarde debido seguramente, a una disminución en el tráfico. En lo que a la distribución del NO_2 por estaciones, no se observan grandes diferencias.

Tras analizar la evolución de las variables, es interesante analizar las posibles correlaciones entre las variables numéricas y la variable objetivo:

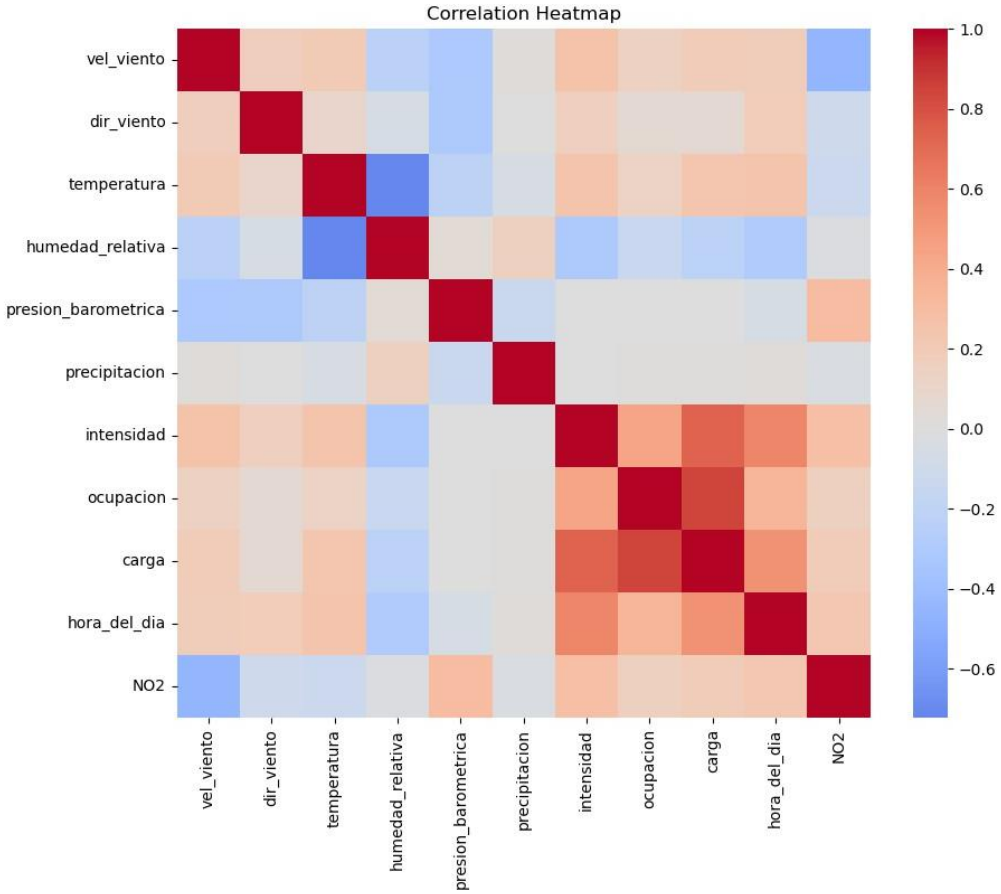


Figura 13. Mapa de correlación.

En el mapa de correlación de la figura 13 se puede observar que variables como la presión barométrica o la intensidad de vehículos están positivamente correladas con la variable de concentración de NO_2. En el caso de variables como la velocidad del viento, la correlación es inversa lo cual tiene sentido ya que un viento con elevada potencia puede dispersar la polución y con ello reducir la concentración de NO_2.

Podemos observar también un cuadrado de correlaciones entre las variables intensidad, ocupación, carga y hora del día. Una buena práctica es aplicar PCA a este conjunto de variables.

Si analizamos directamente la correlación de la concentración de NO_2 y el resto de variables obtenemos lo siguiente:

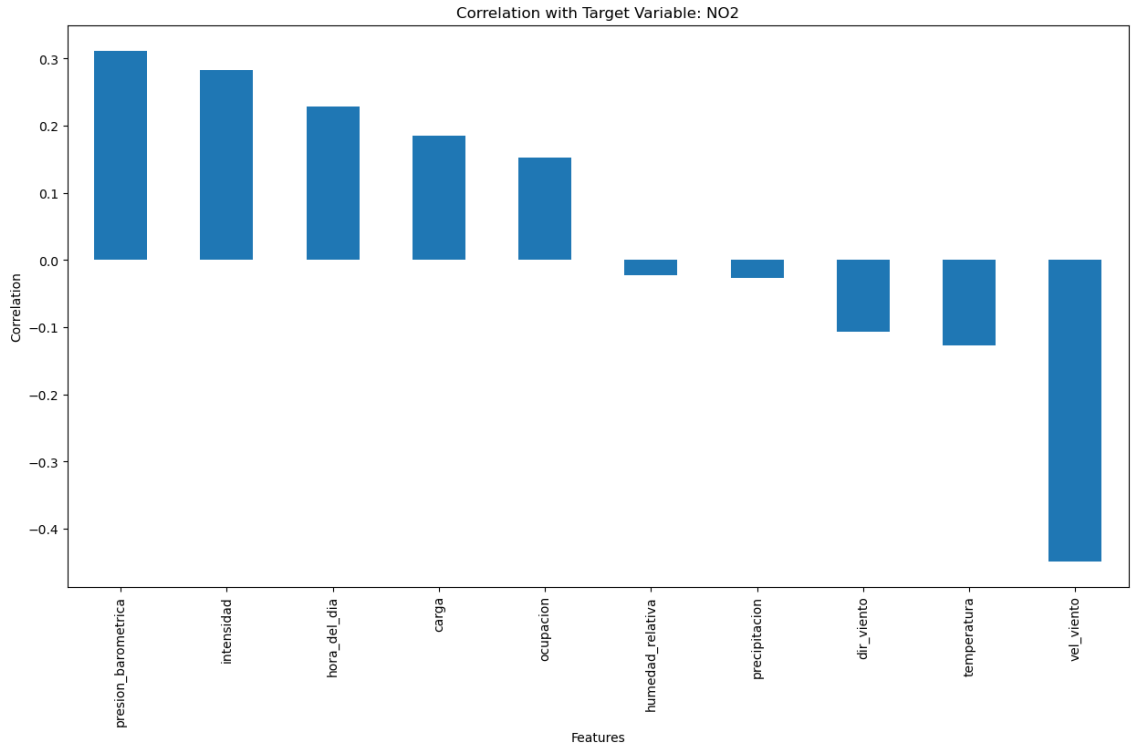


Figura 14. Gráfico de correlaciones con la concentración de NO_2.

En este gráfico se corroboran las correlaciones explicadas anteriormente. Variables como la humedad o la precipitación influyen en una mínima medida sobre la variable objetivo.

2.5Modelos de Predicción

2.5.1 Selección de Modelos

Una vez recogidos, limpiado y transformado los datos, ya se puede pasar a la etapa de entrenamiento. Para ello vamos a entrenar un total de 6 modelos de regresión:

- Linear Regression
- Random Forest
- Gradient Boosting
- SVM
- Decission Tree
- XGBoost

Para el entrenamiento y validación de dichos modelos se ha utilizado la técnica de validación cruzada dividiendo en un 80% - 20% los datos de entrenamiento y validación. En dicha técnica se realizarán 5 entrenamientos por modelo modificando el set de train y validación. A continuación, se muestran los resultados de los entrenamientos de cada uno de los modelos:

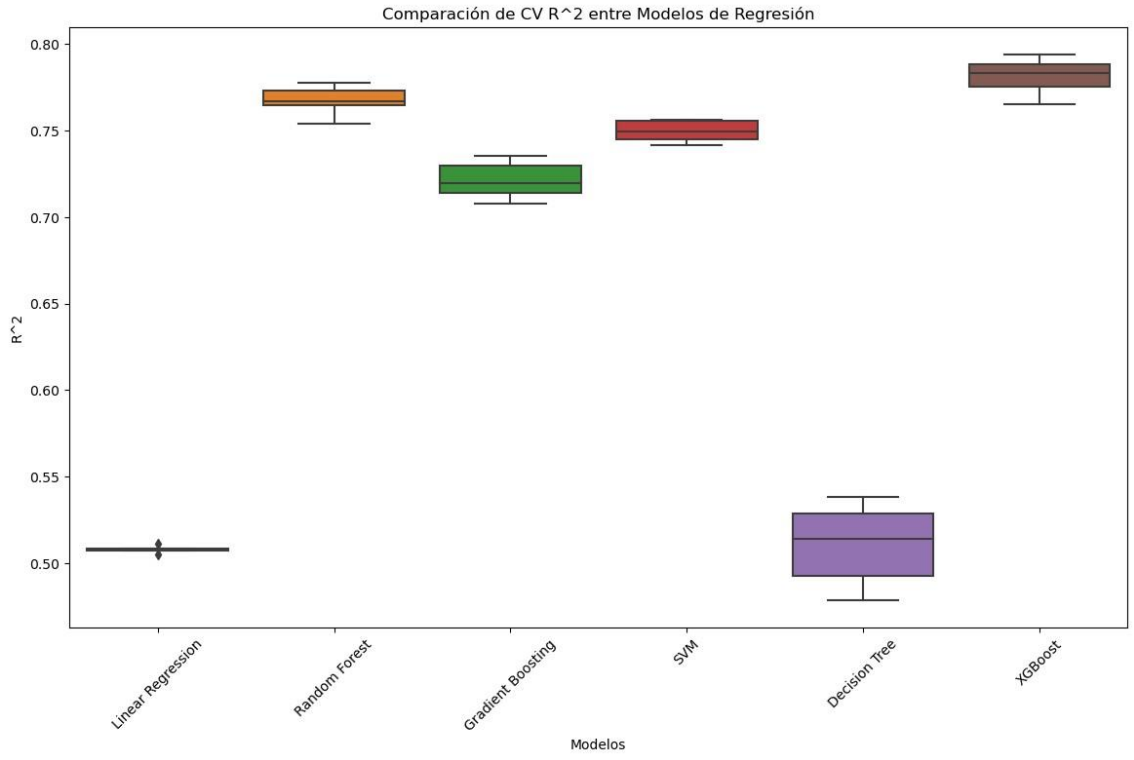


Figura 15. R^2 resultante de la validación cruzada de cada uno de los modelos.

El R^2 indica la proporción de varianza de la variable dependiente que es explicada por cada uno de los modelos. Cuanto más próxima a 1, el modelo explicará una mayor varianza sobre la concentración de NO_2. En este caso se puede observar que el XGBoost es el que mejor responde a la hora de predecir la concentración de NO_2 obteniendo un $\boldsymbol{R^2 = 0,78}$.

Si analizamos el RMSE (error cuadrático medio):

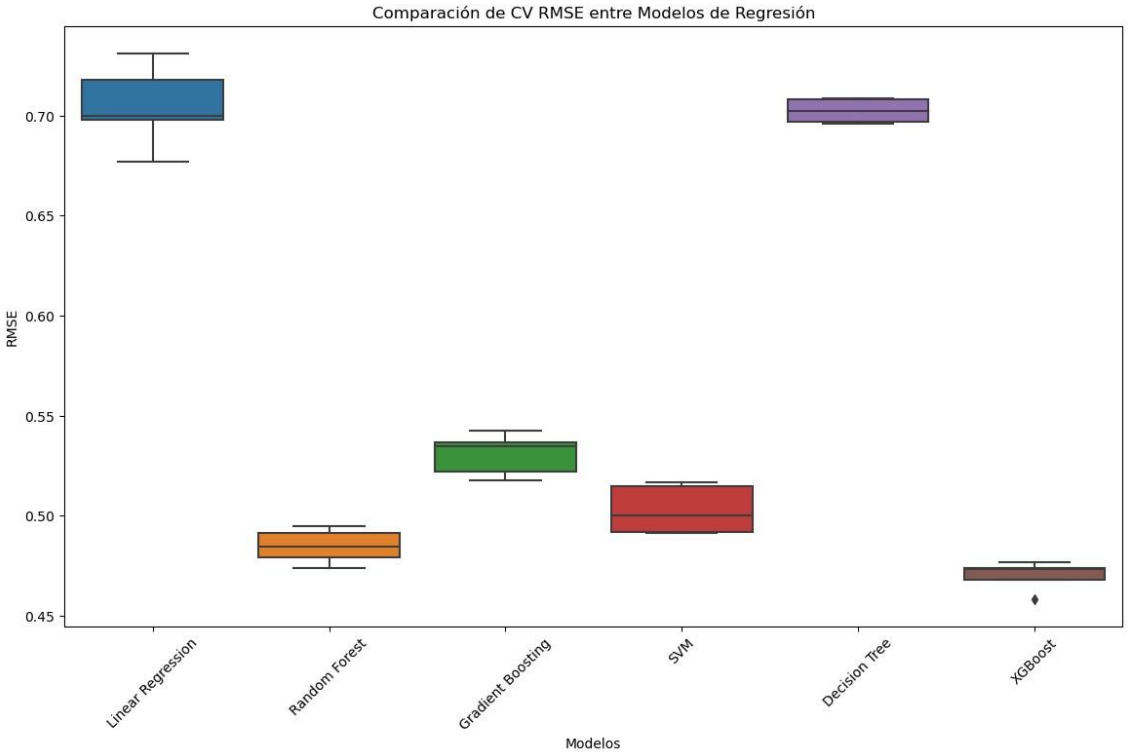


Figura 16. $RMSE$ resultante de la validación cruzada de cada uno de los modelos.

El *RMSE* es una métrica que nos indicará cuán dispersos están los puntos de los datos observados alrededor de la línea de regresión. Se puede observar que el XGBoost es el modelo que menor *RMSE* presenta y por lo tanto el que mejores predicciones arrojará.

Con esto se concluye el XGBoost como modelo con mejores predicciones y por lo tanto el elegido para predecir la concentración de NO_2.

2.5.2 Optimización de Hiperparámetros

Una vez entrenado y seleccionado el mejor modelo, el siguiente paso consiste en hiperparametrizar dicho modelo para maximizar la precisión en las predicciones. Para ello se ha aplicado un *RandomizedGridSearch.*

El *RandomizedSearchCV* es una técnica de optimización de hiperparámetros utilizada para encontrar la mejor combinación de hiperparámetros para un modelo de aprendizaje automático. A diferencia del *GridSearchCV*, que explora todas las combinaciones posibles en un espacio de búsqueda predefinido, el *RandomizedSearchCV* selecciona aleatoriamente una muestra de combinaciones de hiperparámetros para evaluar. Esto hace que *RandomizedSearchCV* sea más eficiente en términos de tiempo y recursos computacionales, especialmente cuando el espacio de búsqueda es grande.

Tras aplicar dicha técnica se obtuvo un R^2 del 0,8. Dicho modelo será el que se despliegue en producción.

2.5.3 Validación del modelo

Finalmente, con la finalidad de validar la precisión y robusted del modelo a la hora de realizar predicciones sobre la concentración de NO_2, se van a realizar predicciones sobre el conjunto de datos disponible donde, además de disponer de datos de entrenamiento, disponemos de un conjunto de datos completamente ajeno al de entrenamiento:

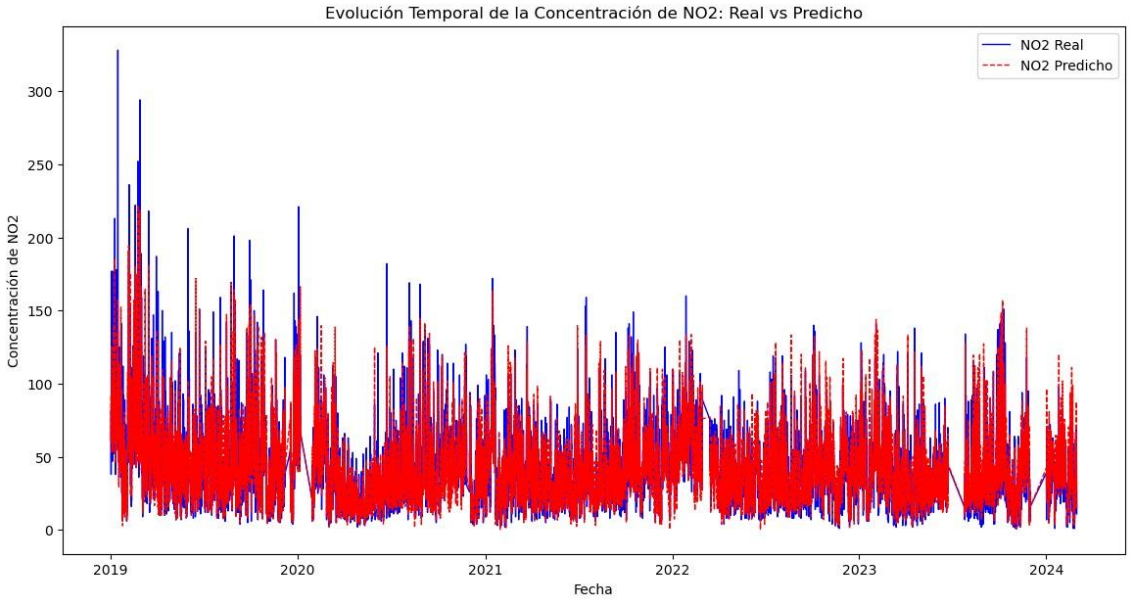


Figura 17. Predicción de la concentración de NO_2.

Como se puede observar, la predicción de los niveles de NO_2 se ajusta de forma adecuada a los niveles de NO_2 reales, mostrando así la precisión del modelo entrenado.

3. Aplicación ADAMA Project

Una vez desarrollado el modelo predictivo, el último punto consiste en ponerlo en producción. Para ello, hemos desarrollado una aplicación en Streamlit con la finalidad de ayudar a los deportistas a determinar si es seguro o no hacer deporte al aire libre en función de la concentración de NO_2.

Pueden acceder a dicha aplicación mediante el siguiente enlace:

ADAMA · Streamlit (adamasaturdaysai.streamlit.app)

La aplicación dispone de 4 pestañas principales:

- Introducción: donde se presenta el proyecto, así como las diferentes fuentes de datos utilizadas.
- Visualizaciones: en esta pestaña seremos capaces de visualizar las diferentes variables atmosféricas y de polución desde el año 2019 a la actualidad.
- Modelo predictivo de NO_2: donde el usuario será capaz de realizar predicciones y recibir una recomendación sobre la seguridad de hacer deporte al aire libre.

- Detección automática de vehículos: sistema para detectar y determinar la intensidad de tráfico mediante YOLO. De esta forma, se podría a su vez predecir la concentración de NO_2 utilizando dicha intensidad de tráfico como variable de entrada del modelo junto con el resto de variables de entrada siendo capaces de obtener una predicción de polución utilizando únicamente una cámara.

A continuación, se muestran de forma gráfica las diferentes ventanas de la aplicación desarrollada.

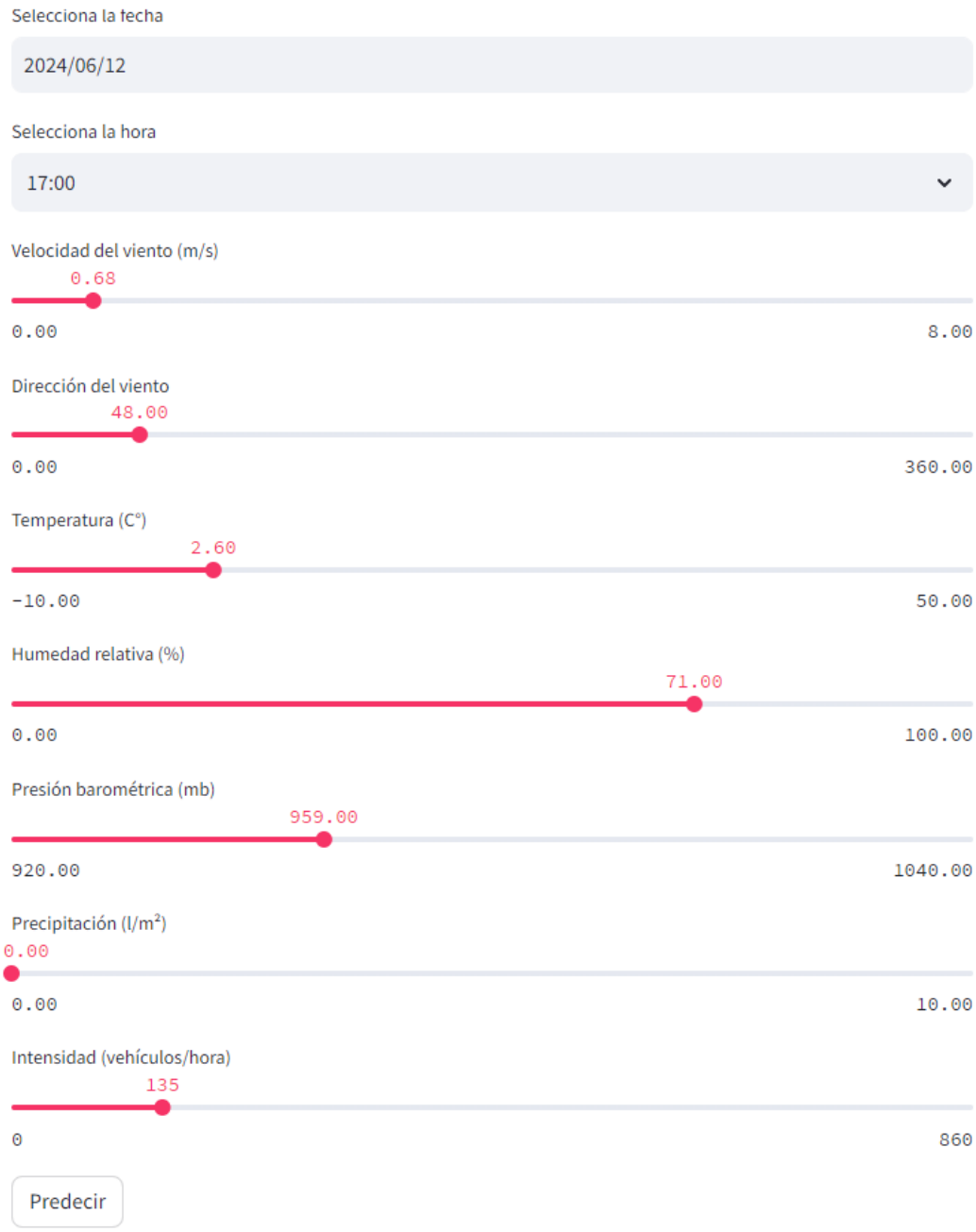


Figura 18. Predicción de la concentración de NO_2.

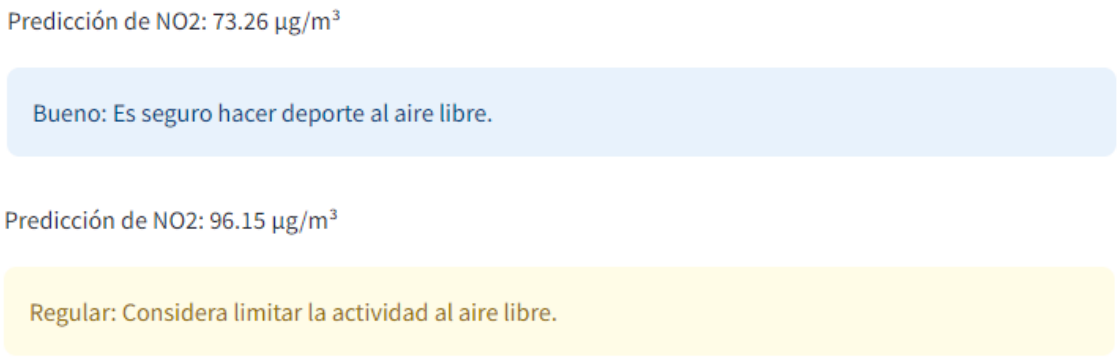


Figura 19. Predicción de la concentración de NO_2 y recomendación para el deportista.

Figura 20. Detección automática de vehículos mediante YOLO.

Finalmente disponemos de otro apartado relacionado con las visualizaciones de las variables meteorológicas y de polución:

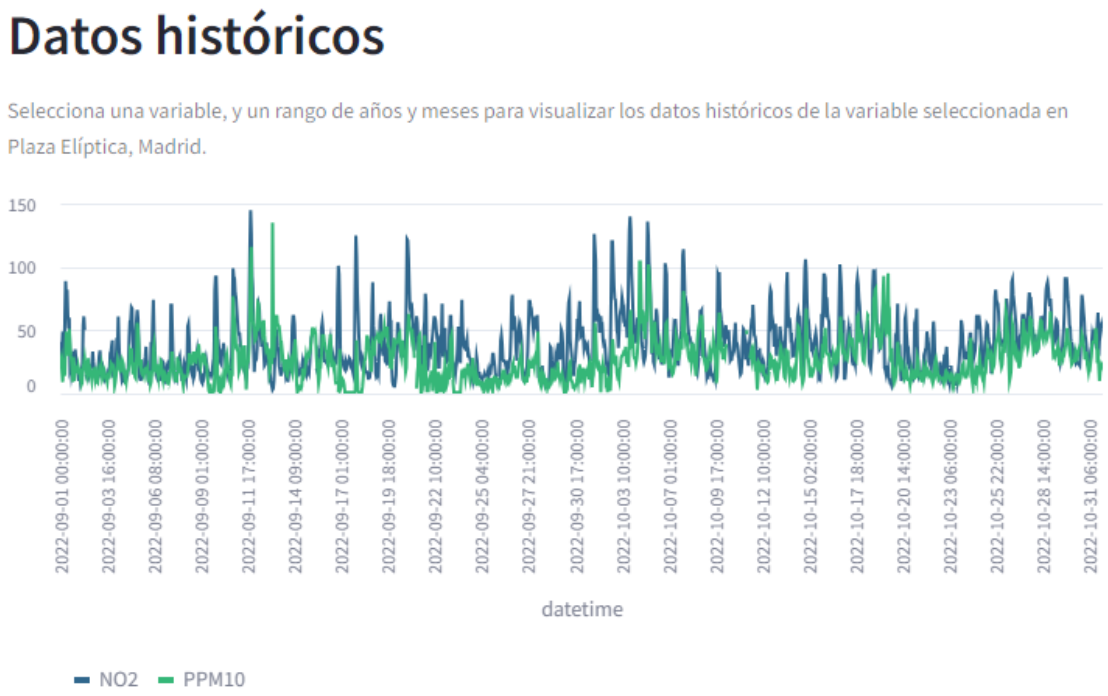


Figura 21. Visualización de variables atmosféricas y de polución.

4. Conclusiones

En este proyecto se ha demostrado con éxito el desarrollo de un proyecto End-to-End de Machine Learning, desarrollando una aplicación que ayude a los deportistas a determinar si es saludable hacer deporte al aire libre o no en función de la concentración de NO_2.

Se puede concluir que los resultados del modelo predictivo son precisos, con un *accuracy* del 80%. Dicha precisión podría mejorarse incrementando la cantidad de datos ya que, como se ha comentado previamente, los datos fueron recogidos de una única estación de control.

Otra posible mejora consistiría en alimentar al modelo de predictivo de NO_2 con la intensidad de tráfico detectada por las cámaras, siendo así posible determinar la polución en determinados puntos únicamente con el uso de una cámara. Esto supondría un ahorro económico para el Ayuntamiento de Madrid, integrando nuestro sistema en las cámaras de tráfico ya implantadas y evitando así el uso de sensores de polución y su costo asociado.

Para concluir, agradecer al grupo de *Saturdays AI* por el desarrollo del curso y los conceptos de ML y DL aprendidos en él.

REFERENCIAS

[1]
Madrid cumple por segundo año con los límites europeos de dióxido de nitrógeno en el aire | Noticias de Madrid | EL PAÍS (elpais.com)

[2]
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=2ac5be53b4d2b610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

[3]
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=41e01e007c9db410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

[4]
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=02f2c23866b93410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

[5]
https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=9f710c96da3f9510VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default