Estudio de la evolución de contaminantes atmosféricos basado en variables meteorológicas

Alejandro López Gómez

Tesis presentada en conformidad con los requisitos del Máster de Economía, Finanzas y Computación

Universidad de Huelva & Universidad Internacional de Andalucía

Septiembre 2023

Estudio de la evolución de contaminantes atmosféricos basado en variables meteorológicas

Alejandro López Gómez

Máster en Economía, Finanzas y Computación

Supervisado por

Antonio Peregrín Rubio

Francisco Alfredo Márquez Hernández

Universidad de Huelva y Universidad Internacional de Andalucía

2023

Resumen

La contaminación atmosférica supone un problema para el medio ambiente y la salud de las personas cuando superan determinados umbrales de concentración. Estos niveles de concentración dependen de las distintas fuentes de emisión, como los vehículos o la actividad industrial, pero también de las condiciones meteorológicas. Con la aplicación de algoritmos de aprendizaje automático y las técnicas adecuadas se puede desarrollar modelos predictivos que, además de anticipar los niveles de contaminación con horas de antelación, aporten información sobre la influencia de estas variables y de qué manera afecta al modelo.

Palabras clave: Contaminación, aprendizaje automático, serie temporal, modelo predictivo.

Abstract

Air pollution supposes a problem for the environment and human health when it exceeds certain concentration thresholds. These concentration levels depend on the different emission sources, such as vehicles or industrial activity, but also on the weather conditions. By applying machine learning algorithms and the appropriate techniques, predictive models can be developed that, in addition to anticipating the pollution levels with hours in advance, provide information about the influence of these variables and how they affect the model.

Keywords: Pollution, machine learning, time series, predictive model.

Índice

Índice de Tablas 6

Índice de Figuras 8

1 **INTRODUCCIÓN** 11

1.1 Contaminantes presentes en el aire 11

1.1.1 Materia Particulada 12

1.1.2 Ozono 12

1.1.3 Dióxido de nitrógeno 12

1.1.4 Dióxido de azufre 13

1.2 Principales focos de emisión en Huelva 13

1.3 Índice de calidad del aire 14

1.4 Objetivos 16

2 **APLICACIÓN: DESCRIPCION DE LA BASE DE DATOS** 17

2.1 Datos de contaminación ambiental 17

2.2 Datos meteorológicos 20

2.3 Combinación de las bases de datos 21

3 **PROPUESTA TÉCNICA** 23

3.1 Modelos de aprendizaje automático 23

3.1.1 Modelos lineales 23

3.1.2 kNN 24

3.1.3 Árboles de decisión 24

3.1.4 Random Forest 24

3.1.5 XGBoost 25

3.2 Modelo autorregresivo 25

4 **ESTUDIO EXPERIMENTAL** 29

4.1 Preprocesamiento 29

4.2 Comparativa de modelos 30

4.3 Optimización de hiperparámetros 32

4.3.1 NO_2 34

4.3.2 SO_2 36

4.3.3 PM2,5 38

4.3.4 PM10 41
4.3.5 O_3 43
4.4 Creación de la nueva variable ‘Festivo’ 45
4.4.1 NO_2 45
4.4.2 SO_2 46
4.4.3 PM2,5 47
4.4.4 PM10 48
4.4.5 O_3 49
4.5 Importancia de los atributos meteorológicos 50
4.6 Modelo predictivo sobre la estación Campus el Carmen 54
4.6.1 NO_2 54
4.6.2 SO_2 55
4.6.3 PM2,5 56
4.6.4 PM10 57
4.6.5 O_3 59

5 **RESULTADOS Y DISCUSIÓN** 61
5.1 Búsqueda de hiperparámetros 62
5.2 Nueva variable ‘Festivo’ 64
5.3 Importancia de atributos 66
5.4 Estudio de la estación Campus el Carmen 68

6 **CONCLUSIONES Y TRABAJOS FUTUROS** 72
6.1 Trabajos futuros 72

7 **BIBLIOGRAFÍA** 74

APÉNDICE A 77

APÉNDICE B 83

Índice de Tablas

Tabla 1. Cantidad de contaminantes emitidos en la localidad de Huelva en el año 2020 agrupados por foco emisor. 14
Tabla 2. Valores límites del Índice de calidad del aire para cada contaminante. 14
Tabla 3. Mensajes para la salud según el índice de calidad del aire. 15
Tabla 4. Contaminantes medidos por cada estación. 17
Tabla 5. Descriptivos de la base de datos. 18
Tabla 6. Valores nulos y descripción de cada atributo. 20
Tabla 7. Descriptivos de la base de datos. 21
Tabla 8. Evaluación de los diferentes algoritmos empleados para cada contaminante. 30
Tabla 9. Algoritmos seleccionados para cada contaminante. 32
Tabla 10. Búsqueda de hiperparámetros de XGBoost para NO_2. 34
Tabla 11. Búsqueda de hiperparámetros de kNN para NO_2. 35
Tabla 12. Búsqueda de hiperparámetros de XGBoost para SO_2. 36
Tabla 13. Búsqueda de hiperparámetros de kNN para SO_2. 37
Tabla 14. Búsqueda de hiperparámetros de XGBoost para PM2,5. 38
Tabla 15. Búsqueda de hiperparámetros de Random Forest para PM2,5. 39
Tabla 16. Búsqueda de hiperparámetros de Ridge para PM2,5. 40
Tabla 17. Búsqueda de hiperparámetros de XGBoost para PM10. 41
Tabla 18. Búsqueda de hiperparámetros de Ridge para PM10. 42
Tabla 19. Búsqueda de hiperparámetros de XGBoost para O3. 43
Tabla 20. Búsqueda de hiperparámetros de Ridge O_3. 44
Tabla 21. Búsqueda de hiperparámetros de XGBoost para NO_2 incluyendo 'Festivo'. 45
Tabla 22. Búsqueda de hiperparámetros de XGBoost para SO_2 incluyendo 'Festivo'. 46
Tabla 23. Búsqueda de hiperparámetros de RandomForest para PM2,5 incluyendo 'Festivo'. 47
Tabla 24. Búsqueda de hiperparámetros de XGBoost para PM10 incluyendo 'Festivo'. 48
Tabla 25. Búsqueda de hiperparámetros de Ridge para O3 incluyendo 'Festivo'. 49
Tabla 26. Búsqueda de hiperparámetros de XGBoost para NO_2 en la estación Campus el Carmen. 54

Tabla 27. Búsqueda de hiperparámetros de XGBoost para SO_2 en la estación Campus el Carmen. 55
Tabla 28. Búsqueda de hiperparámetros de XGBoost para PM2,5 en la estación Campus el Carmen..... 56
Tabla 29. Búsqueda de hiperparámetros de XGBoost para PM10 en la estación Campus el Carmen..... 57
Tabla 30. Búsqueda de hiperparámetros de XGBoost para O_3 en la estación Campus el Carmen. 59
Tabla 31. Resultados de los mejores algoritmos para cada contaminante..... 61
Tabla 32. Resultado de los mejores modelos..... 63
Tabla 33. Comparación de los modelos cuando se introduce la variable 'Festivo'. 66
Tabla 34. Errores de los mejores modelos de cada contaminante desglosados en las distintas estaciones (error en MAPE). 69
Tabla 35. Descriptivos de los contaminantes desglosados por estaciones. 69
Tabla 36. Comparación de errores de Campus del Carmen con los otros modelos. 71

Índice de Figuras

Figura 1. Gráfico de cajas por meses del ozono. 18
Figura 2. Gráfico de cajas por horas del día del ozono. 19
Figura 3. Gráfico de cajas por horas del día del dióxido de nitrógeno. 19
Figura 4. Gráfico de cajas por horas del día del dióxido de azufre. 20
Figura 5. Correlación de Pearson de las diferentes variables. 22
Figura 6. Representación gráfica de Random Forest. 25
Figura 7. Funcionamiento del modelo autorregresivo. 26
Figura 8. Transformación de una serie de tiempo con variables exógenas en una matriz. 26
Figura 9. Representación de la transformación de diferentes series temporales en una sola matriz. 27
Figura 10. Ilustración gráfica del backtesting empleado en este estudio. 28
Figura 11. Diagrama de Paretto del proceso de optimización de los modelos **1** y **2**..... 35
Figura 12. Diagrama de Paretto del proceso de optimización de los modelos **3** y **4**..... 36
Figura 13. Diagrama de Paretto del proceso de optimización de los modelos **5** y **6**..... 37
Figura 14. Diagrama de Paretto del proceso de optimización de los modelos **7** y **8**..... 38
Figura 15. Diagrama de Paretto del proceso de optimización de los modelos **9** y **10**... 39
Figura 16. Diagrama de Paretto del proceso de optimización de los modelos **11** y **12**. 40
Figura 17. Diagrama de Paretto del proceso de optimización de los modelos **13** y **14**. 41
Figura 18. Diagrama de Paretto del proceso de optimización de los modelos **15** y **16**. 42
Figura 19. Diagrama de Paretto del proceso de optimización de los modelos **17** y **18**. 43
Figura 20. Diagrama de Paretto del proceso de optimización de los modelos **19**..... 44
Figura 21. Diagrama de Paretto del proceso de optimización de los modelos **20** y **21**. 44
Figura 22. Diagrama de Paretto del proceso de optimización de los modelos **22** y **23**. 46
Figura 23. Diagrama de Paretto del proceso de optimización de los modelos **24** y **25**. 47
Figura 24. Diagrama de Paretto del proceso de optimización de los modelos **26** y **27**. 48
Figura 25. Diagrama de Paretto del proceso de optimización de los modelos **28** y **29**. 49
Figura 26. Diagrama de Paretto del proceso de optimización del modelo **30**..... 50
Figura 27. Gráfica del efecto de las variables exógenas del modelo **23** para el contaminante NO_2..... 51

Figura 28. Gráfica del efecto de las variables exógenas del modelo **25** para el contaminante SO_2. 51
Figura 29. Gráfica del efecto de las variables exógenas del modelo **26** para el contaminante PM2,5. 52
Figura 30. Gráfica del efecto de las variables exógenas del modelo **28** para el contaminante PM10. 52
Figura 31. Gráfica del efecto de las variables exógenas del modelo **30** para el contaminante O_3. 53
Figura 32. Diagrama de Paretto del proceso de optimización de los modelos **31** y **32**. 55
Figura 33. Diagrama de Paretto del proceso de optimización de los modelos **33** y **34**. 56
Figura 34. Diagrama de Paretto del proceso de optimización de los modelos **35** y **36**. 57
Figura 35. Diagrama de Paretto del proceso de optimización de los modelos **37** y **38**. 58
Figura 36. Diagrama de Paretto del proceso de optimización de los modelos **39** y **40**. 59
Figura 37. Predicción realizada en la estación de La Orden para el contaminante SO_2 entre 17/03/2023 y 21/03/2023 usando los modelos **5** y **6**. 62
Figura 38. Predicción realizada en la estación de Campus el Carmen para el contaminante NO_2 entre 15/04/2023 y 30/04/2023 usando el modelo **2**. 64
Figura 39. Predicción realizada en la estación de La Orden para el contaminante NO_2 entre 17/03/2023 y 21/03/2023 con la variable 'Festivo', usando el modelo **23**..... 65
Figura 40. Porcentaje de las emisiones totales de cada contaminante debidas al tráfico rodado. 66
Figura 41. Porcentaje de las emisiones totales de cada contaminante debidas a la industria y producción de energía. 67
Figura 42. Situación geográfica de la actividad industrial en Huelva y Palos de la Fronteara (marcado rojo) y de las diferentes estaciones de monitorización de contaminantes distribuidas por la capital de Huelva (puntos azules). 68

Figura A1. Predicción NO_2 en Campus el Carmen con el modelo **23**. 77
Figura A2. Predicción NO_2 en Marismas del Titán con el modelo **23**. 77
Figura A3. Predicción SO_2 en Los Rosales con el modelo **25**. 78
Figura A4. Predicción PM2,5 en Pozo Dulce con el modelo **26**. 78
Figura A5. Predicción PM10 en Pozo Dulce con el modelo **28**. 79
Figura A6. Predicción O_3 en Campus el Carmen con el modelo **30**. 79
Figura A7. Predicción NO_2 en Campus el Carmen con el modelo **32**. 80

Figura A8. Predicción SO_2 en Campus el Carmen con el modelo **33**. 80
Figura A9. Predicción PM2,5 en Campus el Carmen con el modelo **35**. 81
Figura A10. Predicción PM10 en Campus el Carmen con el modelo **38**. 81
Figura A11. Predicción O_3 en Campus el Carmen con el modelo **39**. 82

Figura B1. Importancias de variables en modelo **23** de NO_2. 83
Figura B2. Importancias de variables en modelo **25** de SO_2. 84
Figura B3. Importancias de variables en modelo **26** de PM2,5. 85
Figura B4. Importancias de variables en modelo **28** de PM10. 86
Figura B5. Importancias de variables en modelo **30** de O_3. 87

1 INTRODUCCIÓN

La contaminación ambiental se define como la presencia de sustancias en la atmósfera que pueden causar efectos adversos en las personas o en el medio ambiente. Los vehículos de combustión, la actividad industrial o los incendios forestales son algunas de las fuentes más comunes de los contaminantes ambientales. Estos causan numerosos efectos en la salud, sobre todo en aquellos individuos más susceptibles. Varios estudios han relacionado determinados problemas de corazón y cerebrovasculares con algunos contaminantes atmosféricos, de la misma forma que se relaciona con otras enfermedades respiratorias como la EPOC, el agravamiento de los síntomas alérgicos o un efecto nocivo para el funcionamiento correcto de los pulmones en los niños[1].

Huelva es una ciudad con una industria básica y química muy activa donde se elaboran productos muy diversos como combustibles, cobre, ácido sulfúrico, amoníaco, etc.[2] Esta actividad industrial, junto a las emisiones asociadas a los vehículos de combustión, pueden aumentar la cantidad de contaminantes atmosféricos. Para el control de estas sustancias la Junta de Andalucía dispone de distintas estaciones de medición de los principales contaminantes atmosféricos distribuidas por toda Andalucía, incluyendo Huelva. Además, permite el acceso a esta información a cualquier ciudadano a través de su página web[3].

Por otro lado, es bien conocido como determinados factores ambientales pueden variar los niveles de contaminación, como es el caso del viento y su dirección. Para su medición, la AEMET también dispone de distintas estaciones distribuidas por todo el país[4]. De esta forma, se puede crear un base de datos combinando los datos de contaminación y las variables meteorológicas para poder crear un modelo predictivo que permitiese conocer horas antes el nivel de contaminación ambiental de la ciudad de Huelva.

1.1 Contaminantes presentes en el aire

Entre los diferentes contaminantes presentes en la atmósfera, la Organización Mundial de la Salud (OMS) considera como principales: la materia particulada, el ozono troposférico y los óxidos de azufre y de nitrógeno[5].

1.1.1 Materia Particulada

La materia particulada se forma habitualmente en la atmósfera como resultado de reacciones químicas entre otros contaminantes y se define como cualquier sustancia presente en ella, a excepción del agua pura, cuyo tamaño está comprendido entre los 0,002-100 micrómetros de diámetro. La materia particulada incluye las partículas con un diámetro de 10 micrómetros o menos (PM10) y aquellas con un diámetro de 2,5 o menos (PM2,5)[6]. Estas partículas contienen líquidos o sólidos que pueden ser inhalados, causando diversos efectos negativos en la salud de las personas. Los niveles de estas partículas presentan una relación directa con el número de fallecimientos y hospitalización a causa de enfermedades cardiovasculares y pulmonares, siendo las PM2,5 las que presentan mayor riesgo para la salud debido a su menor tamaño[7].

1.1.2 Ozono

El ozono es un gas que se encuentra en la estratosfera (a una altura de 12 a 40 km), pero que también puede surgir en niveles más bajo de la atmósfera como consecuencia de reacciones fotoquímicas entre los óxidos de nitrógeno y los compuestos orgánicos volátiles (COVs, tales como el tolueno, el benceno o el xileno) en unas condiciones meteorológicas concretas de altas temperaturas e intensa radiación solar, lo que lo convierte en un contaminante secundario. Por lo tanto, el foco de emisión son aquellas actividades que emiten contaminantes relacionadas con su formación, como es el tráfico y las industrias emisoras de óxidos de nitrógenos y COVs[8]. Debido a su baja solubilidad en el agua, el ozono tiene la capacidad de penetrar profundamente en los pulmones, causando diversos problemas en el organismo, en especial a aquellas personas con enfermedades respiratorias crónicas[6].

1.1.3 Dióxido de nitrógeno

El dióxido de nitrógeno (NO_2) está muy relacionado con el tráfico, ya que es emitido principalmente por los sistemas de combustión, como consecuencia de la reacción del nitrógeno y el oxígeno en condiciones de alta presión y temperatura. La fabricación de ácido nítrico y otros procesos de nitración industrial también son otras fuentes importantes de este contaminante[9]. Es un gas muy irritante para el sistema

respiratorio con una alta capacidad para penetrar en el pulmón, lo que conduce a diversos problemas respiratorios como asma o EPOC[6].

1.1.4 Dióxido de azufre

El dióxido de azufre (SO_2) está fuertemente relacionado con el consumo de combustibles fósiles debido a la presencia de compuestos de azufre en el petróleo, siendo los principales focos emisores: las centrales térmicas, refinerías de petróleo, la industria del cobre y los vehículos de combustión, entre otros. Los niveles elevados de este gas están asociados con irritación respiratoria o bronquitis, sobre todo en las personas más susceptibles como las personas mayores de edad y los niños. También tiene efectos negativos sobre el medio ambiente, como es la lluvia ácida[6].

1.2 Principales focos de emisión en Huelva

El origen de la contaminación en Huelva, como en toda Andalucía, viene determinada por la actividad industrial, que se corresponde principalmente a la industria petroquímica y química, metalúrgica y alimentaria, entre otras (**Tabla 1**). Aunque es importante resaltar que el transporte sigue siendo una de las principales causas del deterioro de la calidad del aire, siendo responsable de un alto porcentaje de la emisión de muchos contaminantes, como el NO_2. A esto hay que sumarle algunas condiciones naturales que favorecen la concentración de contaminantes, como el alcance de altas concentraciones de partículas procedentes de los desiertos del norte de África, y la alta intensidad solar y temperatura en algunas épocas del año que incrementan los niveles de ozono[10].

Tabla 1. Cantidad de contaminantes emitidos en la localidad de Huelva en el año 2020 agrupados por foco emisor. [11]

Actividad	COVs	NO_x	PM10	PM2,5	SO_2	Total
Agricultura, ganadería y alimentación	106,9	73,4	8,3	1,7	0,6	190,9
Energía	432,0	928,7	29,1	25,7	101,0	1516,6
Industria	3002,1	1499,6	59,9	7,8	5641,2	10210,7
Otros	1622,7	47,9	12,5	10,5	38,2	1731,8
Transporte	72,8	851,9	63,7	50,0	93,5	1131,9
Total	5236,5	3401,5	173,6	95,7	5874,5	14781,8

* Las cantidades se muestran toneladas.

1.3 Índice de calidad del aire

El índice de calidad del aire (ICA) utiliza los datos en tiempo real de las estaciones de medición y son complementados por datos modelizados si fuera necesario. El valor del índice lo determina las concentraciones de hasta cinco contaminantes: PM10, PM2,5, el ozono troposférico, el dióxido de nitrógeno y el dióxido de azufre.

Este índice refleja el potencial impacto que puede tener la calidad del aire sobre la salud, por lo que se le asigna la peor categoría en términos de calidad del aire de cualquiera de los contaminantes. Las bandas del ICA se han establecido en función de los riesgos relativos asociados a la exposición a corto plazo a PM2,5, O_3 y NO_2 según lo establecido por la OMS, y en el caso del SO_2 los valores límites están establecidos por la UE (**Tabla 2**)[12].

Tabla 2. Valores límites del Índice de calidad del aire para cada contaminante.

SO_2		PM2,5		PM10		O_3		NO_2		CATEGORÍA ÍNDICE
0	100	0	10	0	20	0	50	0	40	**BUENA**
101	200	11	20	21	40	51	100	41	90	**RAZONABLEMENTE BUENA**

SO_2		PM2,5		PM10		O_3		NO_2		CATEGORÍA ÍNDICE
201	350	21	25	41	50	101	130	91	120	**REGULAR**
351	500	26	50	51	100	131	240	121	230	**DESFAVORABLE**
501	750	51	75	101	150	241	380	231	340	**MUY DESFAVORABLE**
751-1250		76-800		151-1200		381-800		341-1000		**EXTREMADAMENTE DESFAVORABLE**

* Los límites se expresan en μ/m^3.

El ICA incorpora, además, recomendaciones sanitarias para la población en general y para aquellas personas más sensibles a la contaminación, que incluye tanto a personas con problemas respiratorios, sean adultos o niños, y a adultos con afecciones cardíacas (**Tabla 3**)[12].

Tabla 3. Mensajes para la salud según el índice de calidad del aire.

CALIDAD AIRE	MENSAJES PARA LA SALUD
Buena	Calidad satisfactoria
Razonablemente buena	Calidad aceptable, la contaminación no supone un riesgo para la salud
Regular	La calidad del aire puede presentar un riesgo moderado para los grupos de riesgo.
Desfavorable	Toda la población puede experimentar efectos negativos sobre su salud.
Muy desfavorable	Condiciones de emergencia para la salud pública, la población entera puede verse seriamente afectada.
Extremadamente desfavorable	Condiciones de emergencia para la salud pública, la población entera puede verse gravemente afectada.

1.4 Objetivos

El objetivo de este Trabajo de Fin de Máster es la creación de un modelo predictivo de la contaminación ambiental en Huelva para las próximas 24h, que presente el mínimo error posible haciendo uso de los datos de la medición de los diferentes contaminantes presentes en el aire. Como variables exógenas se usará las condiciones meteorológicas y así, poder conocer también el grado de implicación de cada una de ellas.

2 APLICACIÓN: DESCRIPCION DE LA BASE DE DATOS

Para la realización de este estudio ha sido necesario el uso de dos bases de datos: los datos de contaminación ambiental proporcionados por la Junta de Andalucía y los datos meteorológicos proporcionados por la AEMET[3,4].

2.1 Datos de contaminación ambiental

Esta base de datos se compone de datos horarios desde el 1 de enero de 2022 hasta el 30 de abril de 2023, con un total de 69828 instancias. Formada por 7 atributos que incluye: la estación de medición, la fecha y hora y los niveles de concentración de PM10, PM2,5, NO_2, O_3 y SO_2, que son medidos en μ/m^3. Los niveles de concentración son de tipo numérico, las estaciones de medición son atributos nomínales y la Fecha tipo fecha, y no todas las estaciones registran todos los contaminantes (**Tabla 4**). Para cada valor fecha existe un registro para cada estación. Hay un total de 131596 valores nulos.

Tabla 4. Contaminantes medidos por cada estación.

Estación	Contaminantes medidos
Campus el Carmen	PM10, PM2,5, NO_2, O_3 y SO_2
Los Rosales	PM10, NO_2 y SO_2
La Orden	PM10, NO_2, O_3 y SO_2
Pozo Dulce	PM10, PM2,5, NO_2 y SO_2
Romeralejo	PM10 y SO_2
Marismas del Titán	PM10, NO_2 y SO_2

Tabla 5. Descriptivos de la base de datos.

	PM10	PM2,5	NO_2	O_3	SO_2
Cuenta	62566	22123	47279	21412	64164
Media	25,64	8,10	10,58	59,83	5,49
Desv.	15,46	5,04	12,39	21,40	7,64
Mín.	-29,76	-4,49	-13,14	4,62	0,00
Q1	15,93	4,86	2,00	44,88	2,00
Mediana	22,31	7,07	7,00	59,46	3,00
Q3	31,37	10,15	14,00	73,88	7,00
Máx.	219,41	61,76	173,00	140,38	373,00

Analizando los datos, en el ozono se puede encontrar una clara estacionalidad (**Figura 1**), encontrándose mayores concentraciones en los meses centrales del año. Este comportamiento se debe a que es un contaminante secundario formado a partir de otros compuestos cuando interactúan con la radiación solar, siendo los meses de primavera y verano los que presentan mayor intensidad y, por lo tanto, mayor contaminación.

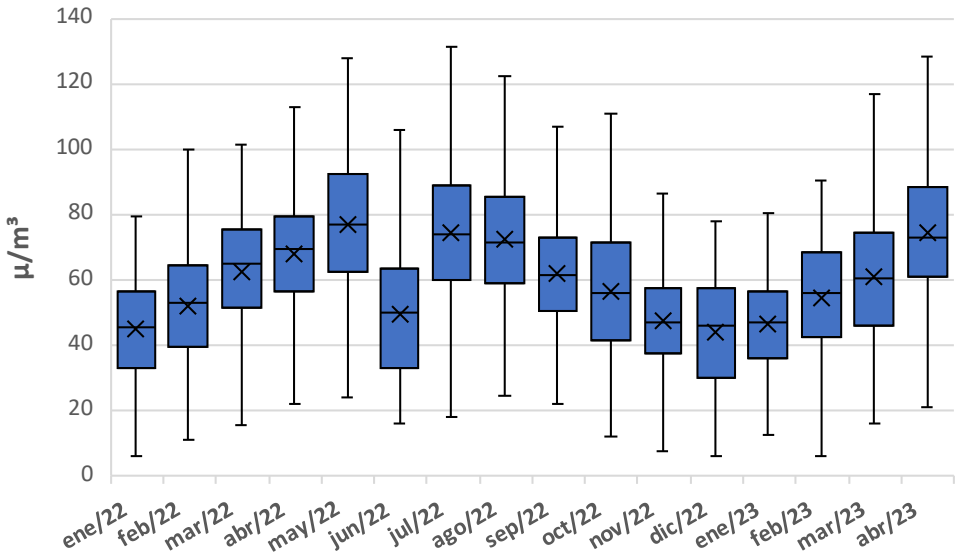

Figura 1. Gráfico de cajas por meses del ozono.

Por horas se observa un comportamiento similar en el caso del ozono, con un aumento de la concentración en las horas con mayor actividad solar (**Figura 2**). Por otro lado, el NO_2 presenta los mayores picos de concentraciones entre las 7:00-8:00 y las 19:00-20:00 (**Figura 3**). El SO_2 presenta las mayores concentraciones entre las 12:00-17:00 (**Figura 4**).

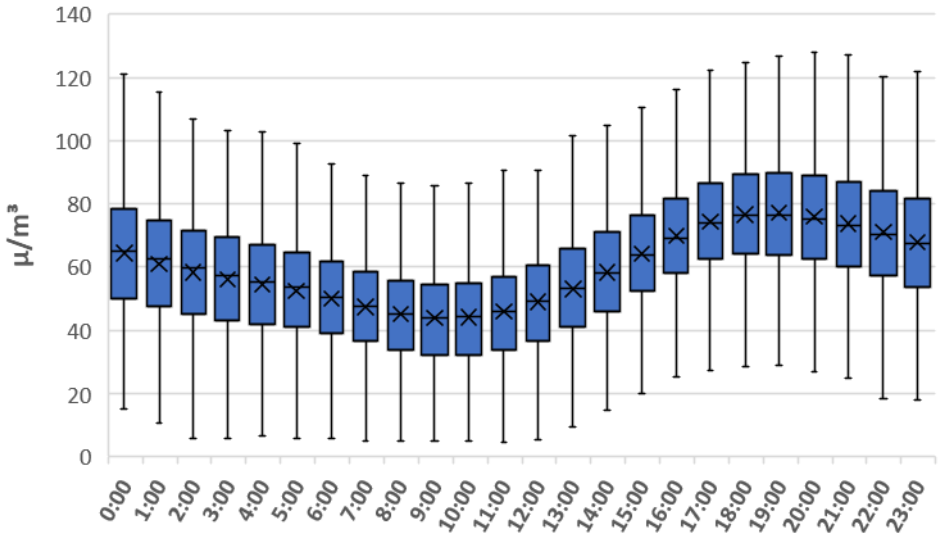

Figura 2. Gráfico de cajas por horas del día del ozono.

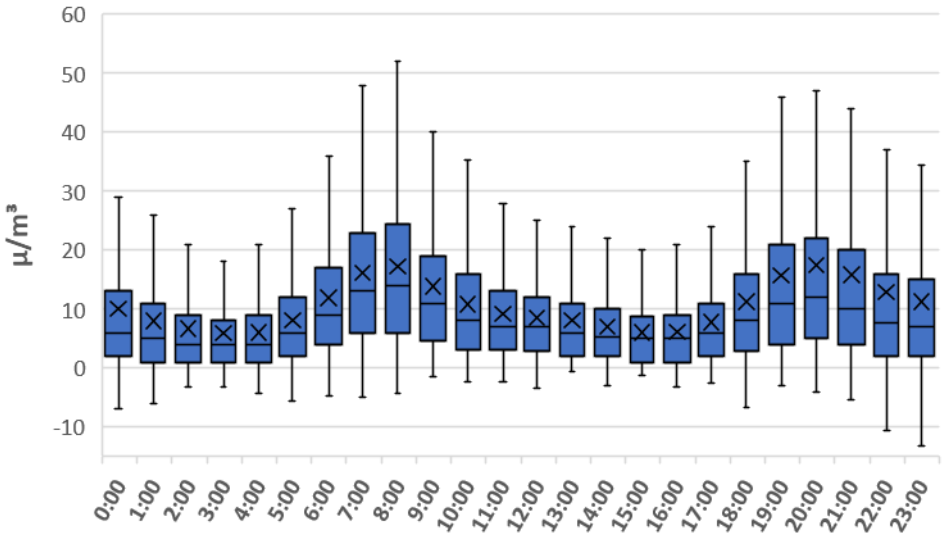

Figura 3. Gráfico de cajas por horas del día del dióxido de nitrógeno.

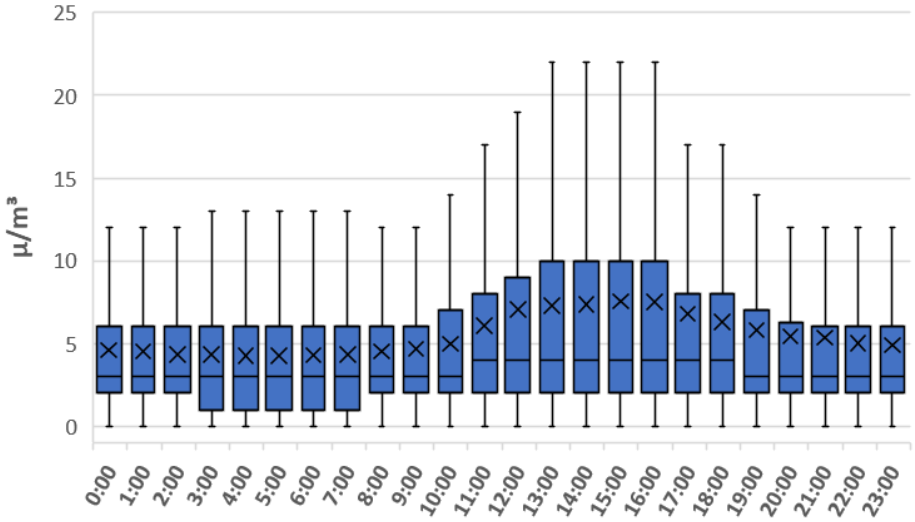

Figura 4. Gráfico de cajas por horas del día del dióxido de azufre.

2.2 Datos meteorológicos

Esta base de dato recoge datos horarios desde el 1 de enero de 2022 al 30 de abril de 2023, con un total de 11640 instancias. Se compone de 7 atributos: fecha y hora, temperatura, precipitación, humedad, velocidad del viento, dirección del viento y la presión atmosférica. Todos son atributos numéricos a excepción de la fecha, que es tipo fecha. Hay un total de 981 valores nulos, que incluye la falta de mediciones de algunos días completos.

Tabla 6. Valores nulos y descripción de cada atributo.

Atributo	Valores nulos	Descripción
Fecha	0	Fecha y hora
Temperatura	137	Temperatura en °C
Precipitación	255	Precipitación en mm
Humedad	137	Humedad relativa %
Viento	146	Velocidad del viento en m/s
Dirección	146	Dirección en grados
pres	160	Presión en hPa

Tabla 7. Descriptivos de la base de datos.

	Temperatura	Precipitación	Humedad	Viento	pres
Cuenta	11503	11385	11503	11494	11480
Media	17,95	0,04	67,33	2,81	1016,13
Desv.	6,93	0,40	20,97	1,57	5,64
Mín.	0,90	0,00	7,00	0,00	997,80
Q1	12,90	0,00	52,00	1,60	1012,20
Mediana	17,50	0,00	70,00	2,40	1015,60
Q3	22,40	0,00	85,00	3,80	1020,20
Máx.	43,60	13,20	100,00	12,40	1030,70

2.3 Combinación de las bases de datos

Al analizar la correlación entre las distintas variables, se observa que las precipitaciones no muestran correlación para ninguno de los contaminantes (**Figura 5**). También confirma la relación positiva del ozono con las épocas del año y las horas más calurosas, debido a la mayor intensidad de la radiación solar. Por otro lado, se observa una correlación negativa por parte de la concentración de NO_2 con la temperatura. Este comportamiento puede ser explicado también por la intensidad de la radiación solar que, al actuar como catalizador en reacciones fotoquímicas entre el NO_2 y otros compuestos, provoca una disminución de sus niveles. Además, se puede apreciar la fuerte correlación entre PM10 y PM2,5, siendo ambos compuesto el mismo tipo de contaminantes, pero diferenciados por su tamaño.

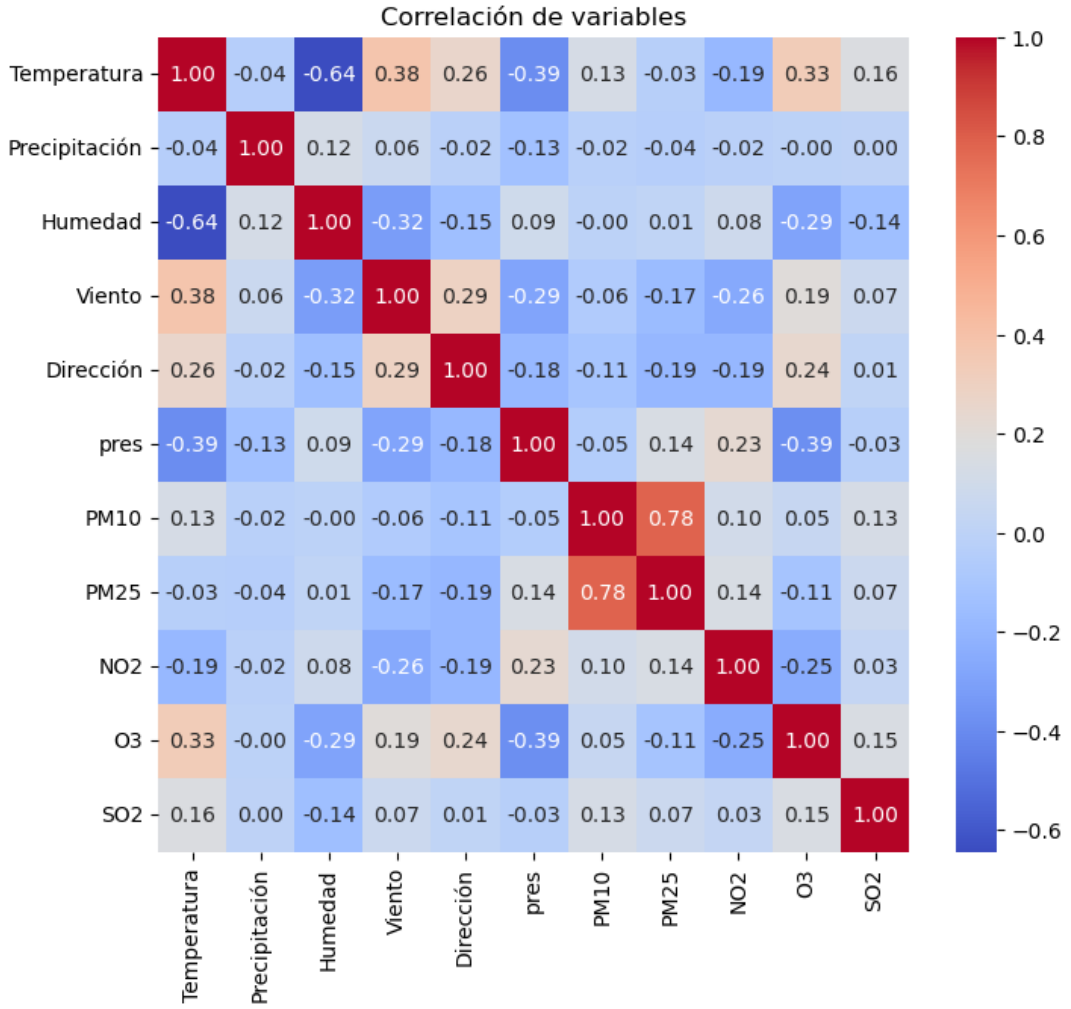

Figura 5. Correlación de Pearson de las diferentes variables.

3 PROPUESTA TÉCNICA

Todas las estaciones de medición distribuidas por todo el territorio monitorizan de forma constante los niveles de contaminación, por lo que es interesante la propuesta de un modelo predictivo, empleando algoritmos de aprendizaje automático, para la determinación de la contaminación en las siguientes horas y reconocer que variables aumentan o disminuyen la concentración de estos.

3.1 Modelos de aprendizaje automático

Entre los métodos de aprendizaje automático para regresión más populares y que podrían aportar buenos resultados se encuentran: las regresiones lineales, kNN, árboles de decisión, Random Forest y Gradient Boosting[13,14].

3.1.1 Modelos lineales

Los modelos lineales son los métodos paramétricos más simples y son muy usados en muchos problemas ya que pueden ser resueltos fácilmente por este modelo, incluso aquellos no lineales. El más sencillo de todos es el método de los mínimos cuadrados y es posible aplicarlo en problemas con múltiples predictores.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$L = \frac{1}{2}\sum_{i=1}^{n} \left|\left| y_i - \hat{y}_i \right|\right|^2$$

Otros modelos lineales son la regresión Ridge y Lasso. La regresión Ridge impone una penalización en el tamaño de los coeficientes. Estos coeficientes minimizan el residuo de mínimos cuadrados. Por otro lado, la regresión Lasso tiende a preferir soluciones con menos coeficientes distintos de cero, reduciendo así el número de características de las que depende la solución.

3.1.2 kNN

Entre los métodos no paramétricos, el algoritmo más sencillo es el k-vecinos más cercano (kNN de sus siglas en inglés) que puede usarse para clasificación o regresión. Dado un valor k y un punto predictivo x_0, este algoritmo identifica las k observaciones entrenadas más cercanas a x_0, representadas como N_0. La métrica más común para medir la separación entre puntos es la distancia Euclídea, que es la distancia en línea recta entre dos puntos. Luego, estima el valor de salida como una media de todos los valores de N_0. Esta media puede ser aritmética o pondera en función de su distancia. En los problemas de clasificación la salida es la clase mayoritaria de N_0.

3.1.3 Árboles de decisión

Los árboles de decisión son otro método no paramétrico que tiene como objetivo crear un modelo para predecir el valor de una variable objetivo mediante el aprendizaje de una serie de reglas de decisión simples. Son muy útiles cuando es necesario mostrar cómo funciona un proceso de decisión ya que su estructura puede ser mostrada fácilmente, la cual se basa en una secuencia de decisiones. Comienza en la raíz, donde una variable es evaluada, y una de las dos ramas que surgen es seleccionada. Este proceso se repite hasta la rama final. Puede ser usado tanto en clasificación como en regresión.

3.1.4 Random Forest

El algoritmo de Random Forest es un método de ensamblado bastante empleado en los problemas de clasificación. Este modelo consiste en un conjunto de árboles de decisión creados a partir de muestras aleatorias con diferentes formas para dividir el nodo. En lugar de buscar la mejor opción, se usa un subconjunto aleatorio de características, tratando de encontrar el umbral que mejor separare los datos, obteniendo como resultado muchos árboles entrenados de forma más débil y cada uno de ellos proporcionara una predicción diferente. El valor final, en el caso de la regresión es el promedio de todos ellos, y en clasificación es la clase más votada.

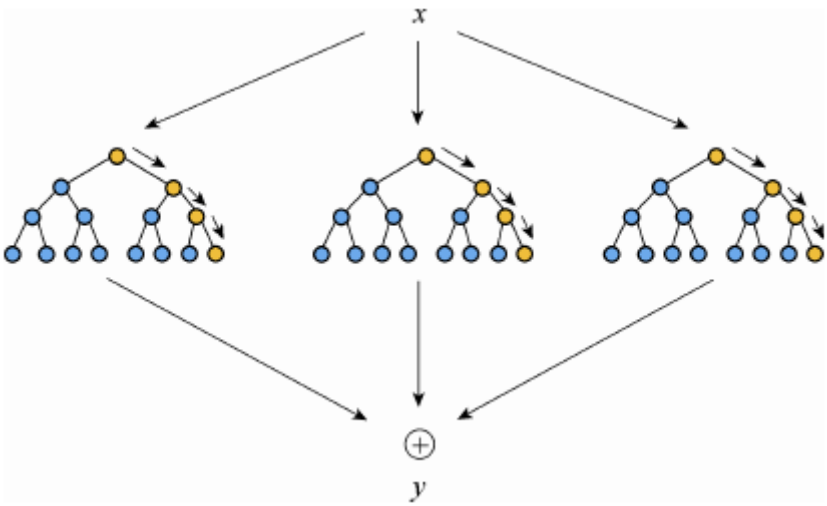

Figura 6. Representación gráfica de Random Forest.

3.1.5 XGBoost

Actualmente, uno de los algoritmos de ensamblado más usado es la potenciación del gradiente (XGBoost). Este algoritmo está diseñado para ser más rápido, creando secuencialmente un conjunto de árboles. Cada uno de ellos es creado con la información previa de los árboles construidos. Los pesos de cada nuevo árbol pueden reducirse para reducir el impacto de un solo árbol en la puntuación final, dejando espacio para que los siguientes arboles mejoren el modelo. Además, contiene un término de penalización que evita el sobreajuste y simplifica los modelos producidos[15]. Este método se puede aplicar a números tipos de problemas y suele ofrecer buenos resultados, incluyendo modelos predictivos de sustancias contaminantes[16].

3.2 Modelo autorregresivo

Para la predicción de los valores futuros, se puede modelar la serie en función de su comportamiento pasado (autorregresivo, **Figura 7**) y empleado variables externas que puedan afectar a la concentración de los contaminantes, como puede ser el viento y su dirección. Como existen diferentes contaminantes, se construirá un modelo predictivo para cada uno de ellos. Para obtener la mejor predicción posible, se comparará los diferentes modelos para seleccionar el más eficiente, usando una ventana temporal de 24h para todos ellos. Para poder aplicar estos algoritmos de aprendizaje automático, es

necesario transformar la serie temporal en una matriz, en la cual cada valor estará relacionado con la ventana temporal (o *lags*) que le precede (**Figura 8**). De esta forma, también se podrá incluir las variables exógenas a la serie. Las predicciones se realizarán para las próximas 24h, ya que se usa como variable exógena las condiciones meteorológicas y las predicciones realizadas por la AEMET tienen un alcance menor a las 48 horas.

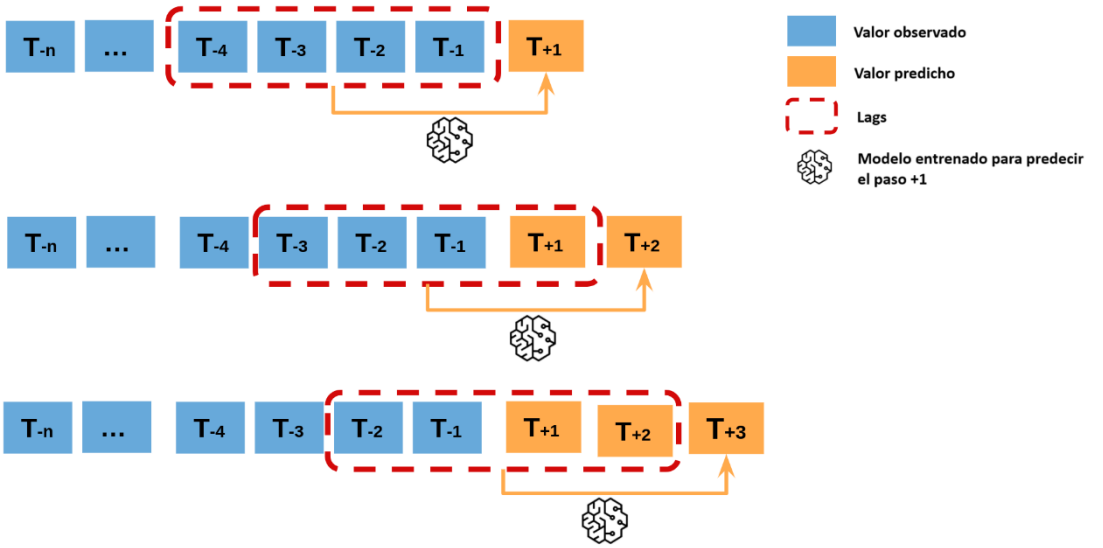

Figura 7. Funcionamiento del modelo autorregresivo.

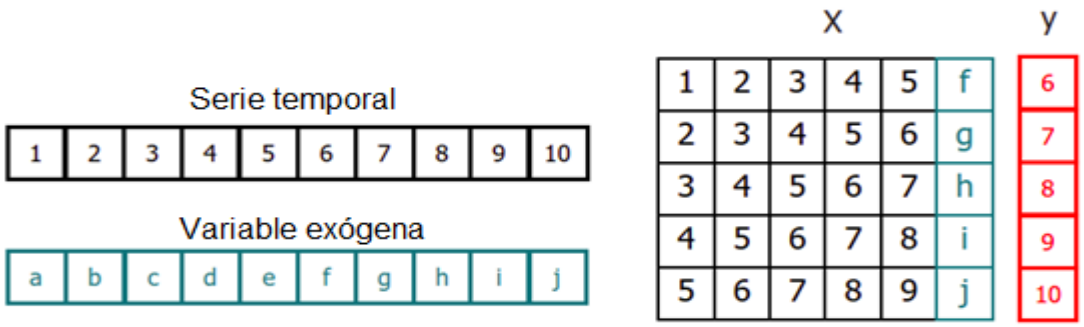

Figura 8. Transformación de una serie de tiempo con variables exógenas en una matriz.

Como para cada contaminante existen diferentes estaciones de medición, pero que están relativamente cerca, ya que todas las estaciones seleccionadas para este estudio se encuentran en la capital, se entrenará los modelos con los datos de todas las estaciones

juntas. Para mantener la independencia de las distintas estaciones y no usar como predictores los valores pasados de otra serie temporal, cada estación será una nueva columna que tendrá valor de 1 cuando los datos se correspondan a esa estación y 0 en caso contrario (**Figura 9**). Este tipo de entrenamiento es útil cuando las distintas series temporales siguen la misma dinámica[17], ya que pueden seguir el mismo patrón con respecto a sus valores pasados y futuros al tratarse todas ellas del mismo contaminante.

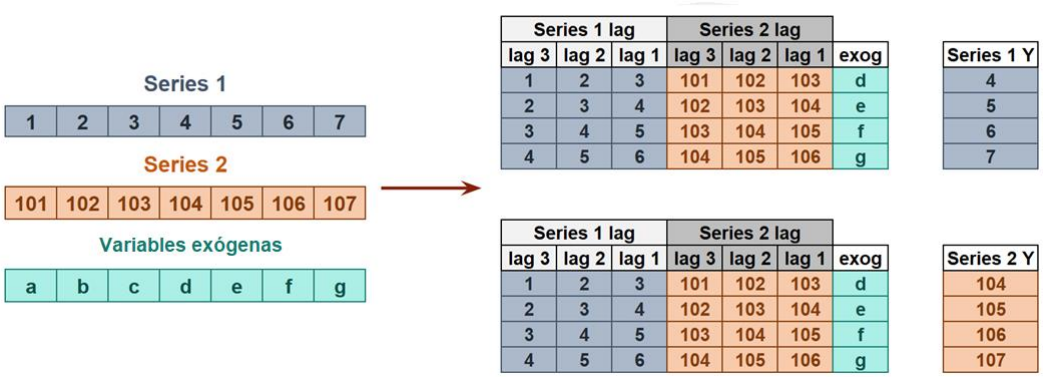

Figura 9. Representación de la transformación de diferentes series temporales en una sola matriz.

La forma correcta de evaluar los distintos modelos es mediante la partición de los datos en un conjunto de entrenamiento y test. Como conjunto de entrenamiento se usará los datos del año 2022 y como test los datos de enero a abril de 2023. De esta forma se usa el 75% de los datos para entrenamiento y el 25% para test. Como método de evaluación se realizará un *backtesting*, que consiste en la evaluación del modelo aplicándolo a datos históricos (**Figura 10**). Estas predicciones realizadas se evalúan usando el error porcentual absoluto medio (MAPE) y r^2 (para medir el ajuste del modelo a los datos).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

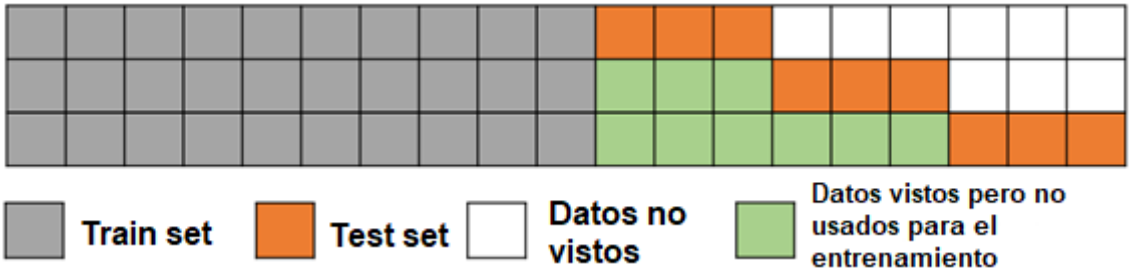

Figura 10. Ilustración gráfica del backtesting empleado en este estudio.

4 ESTUDIO EXPERIMENTAL

El estudio experimental consta de diferentes partes, comenzando por adaptar la base de datos a las necesidades del estudio, como la partición de la base de datos en los diferentes contaminantes, la imputación de sus valores nulos y la búsqueda de los mejores algoritmos para cada uno de ellos. Seguido de un proceso de optimización de hiperparámetros, la creación de nuevas variables y la importancia de las distintas variables exógenas, así como un estudio sobre una estación concreta.

4.1 Preprocesamiento

Para la imputación de nulos, lo primero que se realizó es una interpolación lineal, estableciendo como límite 5 valores nulos consecutivos. De esta forma se evita la imputación de valores completamente lineales en la temperatura cuando la ausencia de valores es elevada. Para solucionar el problema de ausencias de mediciones mayores en determinados días, se imputan los datos de la estación meteorológica más cercana, que se encuentra en Gibraleón. Tras estas medidas, sigue existiendo valores nulos de precipitación y de presión. La precipitación se imputa como 0, ya que es el valor más habitual, y la presión se vuelve a imputar mediante interpolación lineal, en este caso sin límite. La dirección del viento, que viene marcado en grados, es transformada en 8 puntos cardinales: norte (N), sur (S), este (E), oeste (O), noreste (NE), noroeste (NO), sureste y suroeste (SO). Tras esto, esta columna es transformada en columnas *dummies*, obteniendo el valor 1 cuando tiene esa dirección y 0 en los demás casos.

Los datos de contaminación son separados en 5 conjunto de datos, uno para cada contaminante. Para la imputación de los nulos, primero fue necesario la unión de cada base de datos de cada contaminante con los datos climatológicos mediante la fecha. Tras la observación de la estacionalidad de algunos contaminantes atmosféricos, se construye dos nuevas características temporales para poder imputar estos datos nulos: el mes y la hora. De esta manera, se evita que, por ejemplo, los días con temperaturas bajas durante la noche en las épocas más calurosas sean imputados de igual manera que los días de invierno durante el día. Los conjuntos de datos fueron separados en conjunto de entrenamiento (01/01/2022 hasta 31/12/2022) y test (01/01/2023 hasta 31/04/2023). Se

usó el algoritmo de kNN con k=5, entrenado con el conjunto de entramiento, y aplicándolo a ambos conjuntos para imputar todos los valores nulos presentes.

4.2 Comparativa de modelos

Una vez realizado el procesamiento, cada uno de los algoritmos propuestos es entrenado con los diferentes conjuntos de datos haciendo uso de la biblioteca Sklearn de Python, con el objetivo de encontrar los modelos con mayor precisión en las predicciones. Para encontrar que modelo puede ser el más adecuado, primero se entrenan los modelos con los hiperparámetros establecidos por defecto en cada uno de ellos y se usa 24 periodos como ventana temporal para obtener un modelo base de todos ellos (**Tabla 8**).

Tabla 8. Evaluación de los diferentes algoritmos empleados para cada contaminante.

Contaminante	**Algoritmo**	**Error**
NO_2	AdaBoostRegressor	29,5265
	kNN	4,6638
	Lasso	5,0309
	Linear Regression	5,0233
	Random Forest	5,2659
	Ridge	5,0233
	XGBoost	4,6644
SO_2	AdaBoostRegressor	17,7093
	kNN	2,4850
	Lasso	2,9902
	Linear Regression	2,6124
	Random Forest	4,4770
	Ridge	2,6124
	XGBoost	2,5909

Contaminante	Algoritmo	Error
PM2,5	AdaBoostRegressor	0,9936
	kNN	0,5477
	Lasso	0,5334
	Linear Regression	0,3322
	Random Forest	0,3274
	Ridge	0,3322
	XGBoost	0,3556
PM10	AdaBoostRegressor	0,2628
	kNN	0,2105
	Lasso	0,1710
	Linear Regression	0,1622
	Random Forest	0,1773
	Ridge	0,1622
	XGBoost	0,1627
O_3	AdaBoostRegressor	0,2316
	kNN	0,1887
	Lasso	0,1469
	Linear Regression	0,1250
	Random Forest	0,1446
	Ridge	0,1250
	XGBoost	0,1360

* El error es el MAPE expresado en tanto por uno.

4.3 Optimización de hiperparámetros

Una vez entrenados los diferentes algoritmos se seleccionan los mejores para la realización de una optimización de sus hiperparámetros y de la ventana temporal. Debido a la posibilidad de entrenar el algoritmo de XGBoost empleando aceleradores de GPU, que reduce el tiempo necesario del entrenamiento, se usa para todos los contaminantes[18,19]. El resto de los algoritmos seleccionados se realiza en base a sus resultados (**Tabla 9**).

Para la búsqueda de hiperparámetros se hará uso de Optuna[20], un software de optimización de hiperparámetros de código abierto. Este realiza la optimización de hiperparámetros como si se tratase de un proceso de minimización o maximización de una función objetivo que toma como valores de entradas los hiperparámetros y devuelve la puntuación del modelo. A cada proceso de optimización se le conoce como *study* y cada evaluación de la función objetivo como *trial*, y mediante la interacción de los distintos *trials*, Optuna va construyendo gradualmente la función objetivo[21]. Las métricas usadas para la optimización son el MAPE (minimización) y el r^2 (maximización).

Tabla 9. Algoritmos seleccionados para cada contaminante.

Contaminante	Algoritmo
NO_2	XGBoost, kNN
SO_2	XGBoost, kNN
PM2,5	XGBoost, Random Forest, Ridge
PM10	XGBoost, Ridge
O_3	XGBoost, Ridge

Para el algoritmo XGBoost, fueron optimizados los hiperparámetros más típicos[19,22-24], empleando 200 *trials* con el objetivo de minimizar el MAPE y de maximizar el r^2:

- learning_rate: la reducción del tamaño de paso se usa para evitar el sobreajuste, con un rango [0,1], los valores más típicos son 0,01-0,1.
- n_estimators: número de árboles empleados.
- max_depth: la máxima profundidad del árbol. El incremento de este parámetro hace el modelo más complejo y es más probable el sobreentrenamiento. Con un rango de [0, +inf], los valores típicos son 3-10.
- min_child_weight: la suma mínima de pesos de todas las instancias requeridas para mantener un nodo secundario. En regresión, esto solo es la suma requerida de instancias mínimas para realizar la partición.
- subsample: la fracción de observaciones necesarias para ser aleatorizadas en cada árbol. Valores más bajos hacen el algoritmo más conservador frente al sobre entrenamiento. Los valores típicos se encuentran entre 0,5-1.
- colsample_bytree: representa la fracción de columnas que son aleatorizadas en cada árbol. Valores típicos entre 0,5-1.
- gamma: especifica la reducción mínima requerida para realizar una partición. Hace más conservador el modelo.

Para el algoritmo Random Forest se eligieron los hiperparámetros que más comúnmente son optimizados según la bibliografía[23-26], empleando 75 *trials*:

- n_estimators: el número de árboles usados por el algoritmo.
- max_depth: profundidad máxima que presenta los árboles.
- min_samples_split: el número mínimo de muestras requeridas para poder dividir un nodo.
- min_samples_leaf: el número mínimo de muestras para estar en un nodo de la hoja.

Para el modelo kNN los hiperparámetros entrenados fueron los siguientes[27], empleando 75 *trials*:

- n_neighbors: números de vecinos empleados.
- weights: peso usado en cada predicción.

- metric: métrica usada para la distancia.

Finalmente, para la regresión Ridge, el único valor optimizado es *alpha*, una constante que multiplica el término de regularización[24], empleando también 75 *trials*.

4.3.1 NO_2

Para este contaminante se obtiene dos modelos óptimos, uno con el menor MAPE y otro con el mayor r^2 (**Tabla 10** y **Tabla 11**).

Tabla 10. Búsqueda de hiperparámetros de XGBoost para NO_2.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **1**	Modelo **2**
learning_rate	0,3	0-0,5	0,0012	0,0204
n_estimators	100	50-700	490	380
max_depth	6	2-11	3	3
min_child_weight	1	1-7	2	3
subsample	1	0,3-1	0,3647	0,3648
colsample_bytree	1	0,3-1	0,9789	0,5776
gamma	0	0-1,5	0,0842	0,6863
lag	24	12-60	27	48
MAPE			1,6996	4,1010
r^2			-0,6686	0,4238

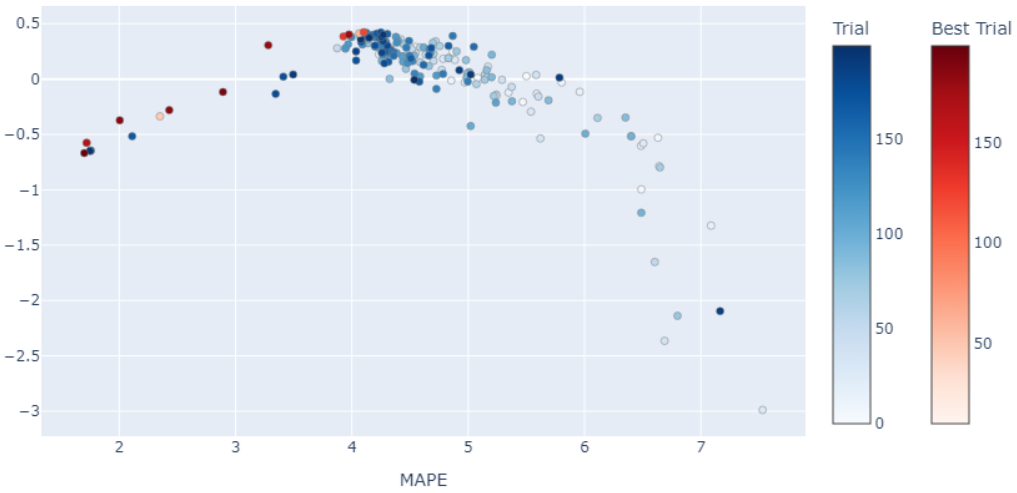

Figura 11. Diagrama de Paretto del proceso de optimización de los modelos **1** y **2**.

Tabla 11. Búsqueda de hiperparámetros de kNN para NO_2.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **3**	Modelo **4**
n_neighbors	5	1-18	5	18
weights	uniform	uniform', 'distance'	‘uniform’	‘distance’
metric	minkowski	‘minkowski', 'euclidean', 'manhattan'	‘manhattan’,	‘manhattan’,
lag	24	12-60	30	51
MAPE			3,7228	3,8374
r^2			0,1883	0,3576

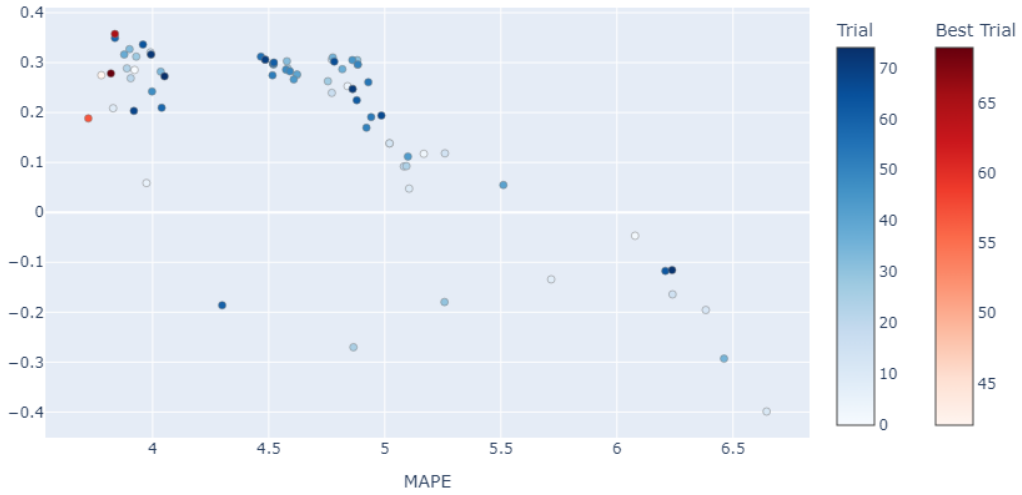

Figura 12. Diagrama de Paretto del proceso de optimización de los modelos **3** y **4**.

A pesar de que los modelos **1** y **3** son los que presentan mayor precisión, el escaso ajuste del modelo a los datos no permite una buena interpretación de estos. Por lo tanto, el mejor modelo para este contaminante es el **2**, ya que posee mayor r^2 y permite una mejor interpretación de los datos, aunque implique tener mayor error porcentual.

4.3.2 SO_2

Para este contaminante se obtiene también dos modelos óptimos, uno con el menor MAPE y otro con el mayor ajuste (**Tabla 12** y **Tabla 13**).

Tabla 12. Búsqueda de hiperparámetros de XGBoost para SO_2.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **5**	Modelo **6**
learning_rate	0,3	0-0,5	0,0028	0,0172
n_estimators	100	50-700	630	500
max_depth	6	2-11	5	7
min_child_weight	1	1-7	1	7
subsample	1	0,3-1	0,6936	0,7218

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **5**	Modelo **6**
colsample_bytree	1	0,3-1	0,8011	0,4621
gamma	0	0-1,5	0,0685	0,7953
lag	24	12-60	18	30
MAPE			1,2406	2,5926
r^2			-0,2017	0,3010

Figura 13. Diagrama de Paretto del proceso de optimización de los modelos **5** y **6**.

Tabla 13. Búsqueda de hiperparámetros de kNN para SO_2.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **7**	Modelo **8**
n_neighbors	5	1-18	11	16
weights	uniform	uniform', 'distance'	‘distance’	‘distance’
metric	minkowski	‘minkowski', 'euclidean', 'manhattan'	‘manhattan’	‘euclidean’
lag	24	12-60	45	30

MAPE	1,7887	2,4757
r^2	0,1790	0,2452

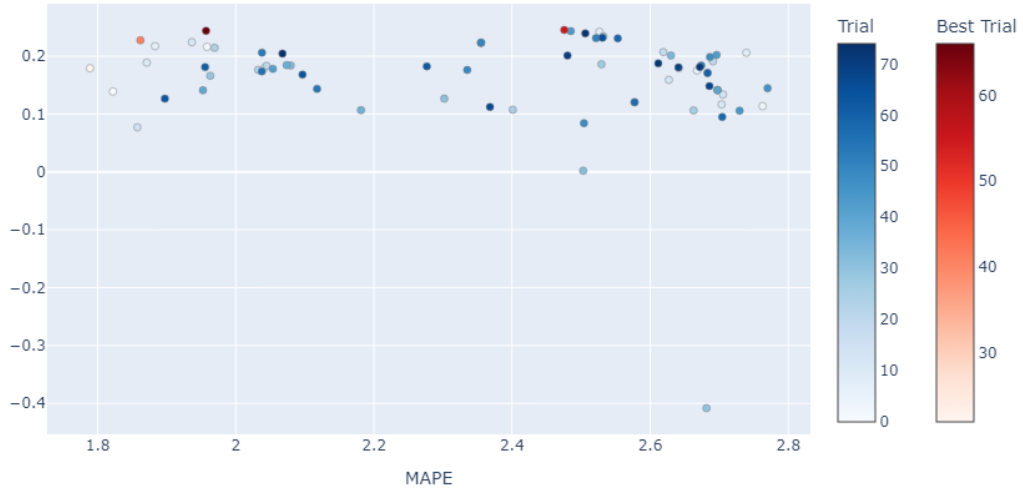

Figura 14. Diagrama de Paretto del proceso de optimización de los modelos **7** y **8**.

Como en el caso anterior, los modelos con menor MAPE para el SO_2 presentan un bajo ajuste que no permitirían una buena interpretación de los modelos. Por lo tanto, el mejor algoritmo y sus mejores hiperparámetros lo presenta el modelo **6**.

4.3.3 PM2,5

Tabla 14. Búsqueda de hiperparámetros de XGBoost para PM2,5.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **9**	Modelo **10**
learning_rate	0,3	0-0,5	0,1595	0,2177
n_estimators	100	50-700	510	620
max_depth	6	2-11	7	7
min_child_weight	1	1-7	5	1

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **9**	Modelo **10**
subsample	1	0,3-1	0,6353	0,7378
colsample_bytree	1	0,3-1	0,6223	0,6816
gamma	0	0-1,5	0,5256	1,4756
lag	24	12-60	57	60
MAPE			0,3029	0,3140
r^2			0,5667	0,5794

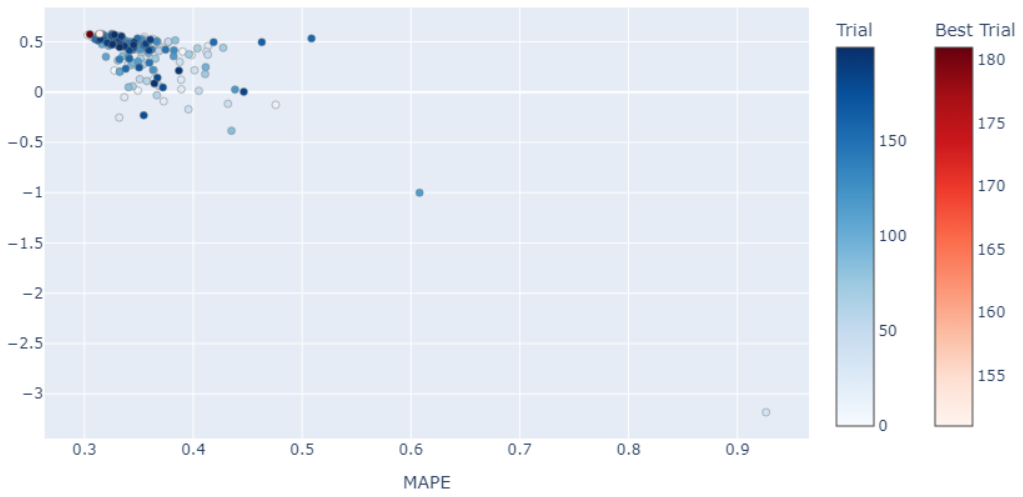

Figura 15. Diagrama de Paretto del proceso de optimización de los modelos **9** y **10**.

Tabla 15. Búsqueda de hiperparámetros de Random Forest para PM2,5.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **11**	Modelo **12**
n_estimators	100	50-700	130	320
max_depth	None	2-11	10	10
min_sample_split	2	2-12	4	2
min_sample_leaf	1	1-7	5	7
lag	24	12-60	30	27

MAPE	0,2955	0,2981
r^2	0,5215	0,5845

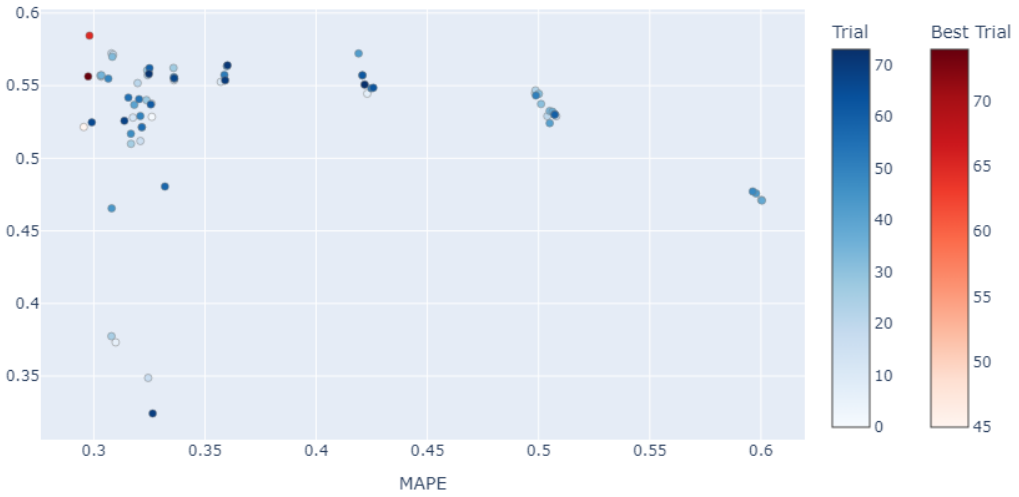

Figura 16. Diagrama de Paretto del proceso de optimización de los modelos **11** y **12**.

Tabla 16. Búsqueda de hiperparámetros de Ridge para PM2,5.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **13**	Modelo **14**
alpha	1	0-2,5	1,5667	2,2802
lag	24	12-60	18	15
MAPE			0,3227	0,3256
r^2			0,4873	0,4950

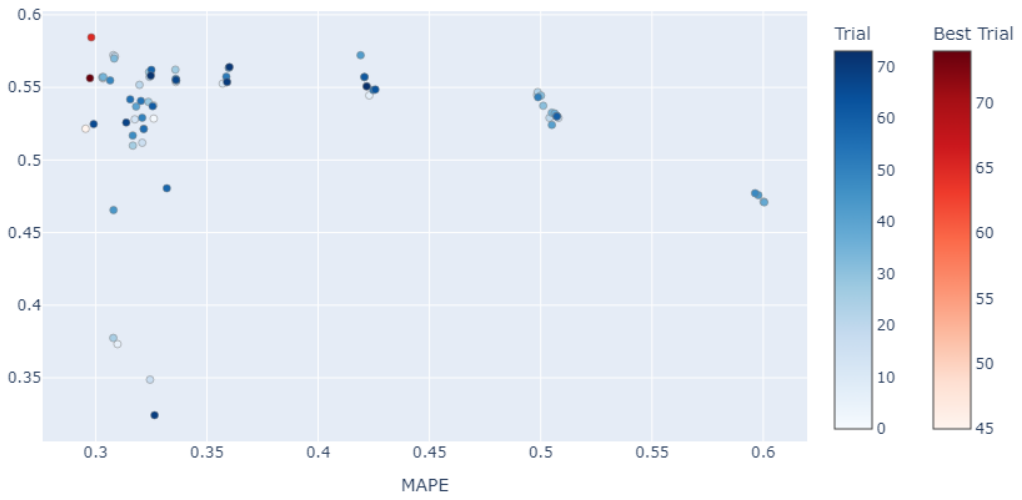

Figura 17. Diagrama de Paretto del proceso de optimización de los modelos **13** y **14**.

En general, este contaminante presenta un buen ajuste en todos los algoritmos empleados, mejorando el ajuste a medida que el error disminuye, siendo el r^2 mayor a 0,40 en todos los casos. Además, la diferencia entre los modelos con mayor ajuste y los modelos con mayor precisión son muy pequeñas. Por lo tanto, el modelo elegido para este contaminante es que el presenta menor el error, el modelo **11**.

4.3.4 PM10

Tabla 17. Búsqueda de hiperparámetros de XGBoost para PM10.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **15**	Modelo **16**
learning_rate	0,3	0-0,5	0,2044	0,0886
n_estimators	100	50-700	210	400
max_depth	6	2-11	6	3
min_child_weight	1	1-7	7	7
subsample	1	0,3-1	0,8733	0,8733
colsample_bytree	1	0,3-1	0,6925	0,3509
gamma	0	0-1,5	0,9507	0,9188
lag	24	12-60	18	15

MAPE	0,1539	0,1547
r^2	0,7257	0,7401

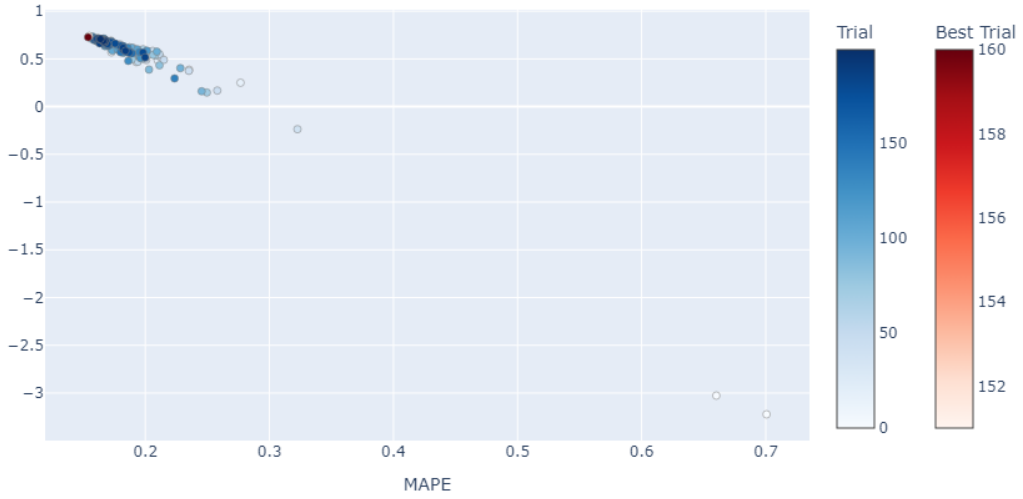

Figura 18. Diagrama de Paretto del proceso de optimización de los modelos **15** y **16**.

Tabla 18. Búsqueda de hiperparámetros de Ridge para PM10.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **17**	Modelo **18**
alpha	1	0-2,5	0,086	0,1222
lag	24	12-60	33	51
MAPE			0,1650	0,1658
r^2			0,6555	0,6610

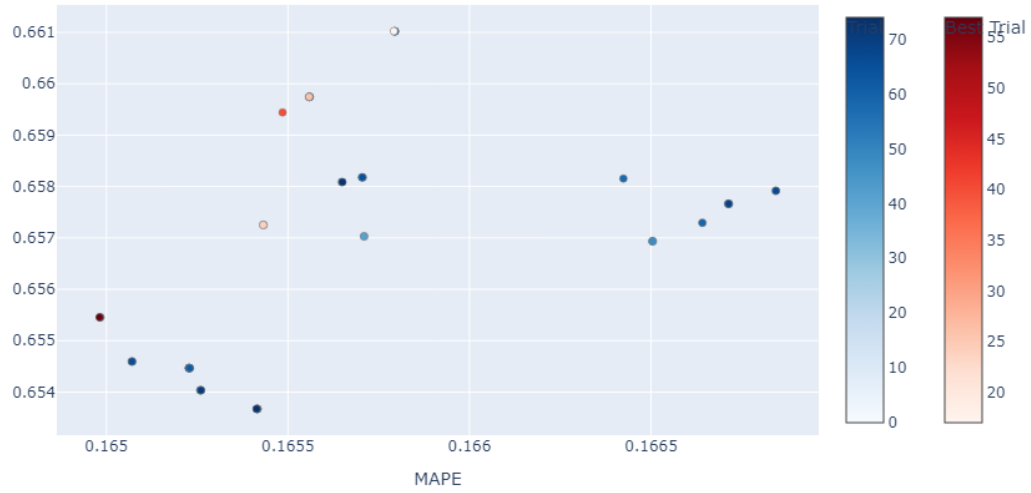

Figura 19. Diagrama de Paretto del proceso de optimización de los modelos **17** y **18**.

Como el caso anterior, el modelo **15** presenta el MAPE más bajo y además un buen r^2. Por lo tanto, el modelo presenta una buena interpretabilidad y precisión.

4.3.5 O_3

Tabla 19. Búsqueda de hiperparámetros de XGBoost para O3.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **19**
learning_rate	0,3	0-0,5	0,052
n_estimators	100	50-700	470
max_depth	6	2-11	5
min_child_weight	1	1-7	7
subsample	1	0,3-1	0,5760
colsample_bytree	1	0,3-1	0,9643
gamma	0	0-1,5	0,7040
lag	24	12-60	48
MAPE			0,1236
r^2			0,8174

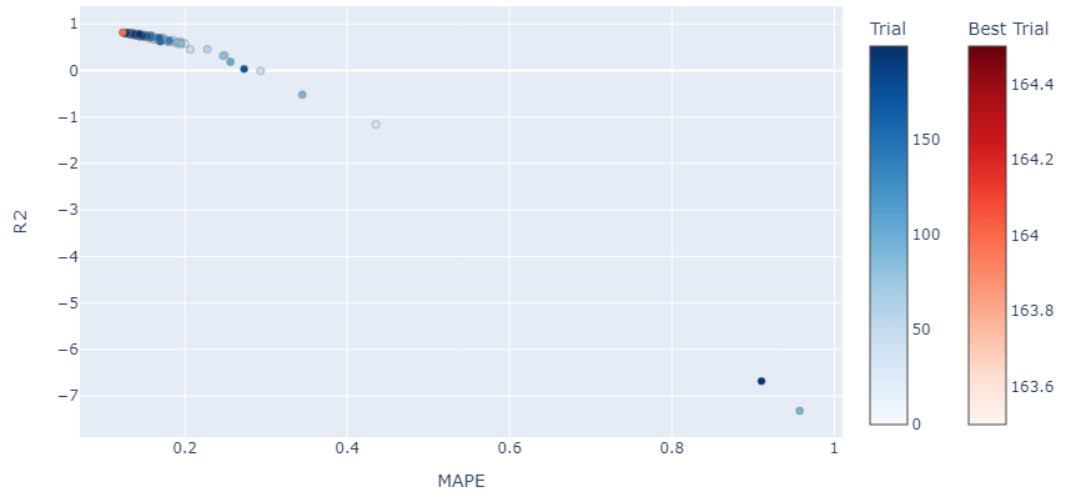

Figura 20. Diagrama de Paretto del proceso de optimización de los modelos **19**.

Tabla 20. Búsqueda de hiperparámetros de Ridge O_3.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **20**	Modelo **21**
alpha	1	0-2,5	1,7751	0,3903
lag	24	12-60	45	54
MAPE			0,1230	0,1233
r^2			0,8203	0,8214

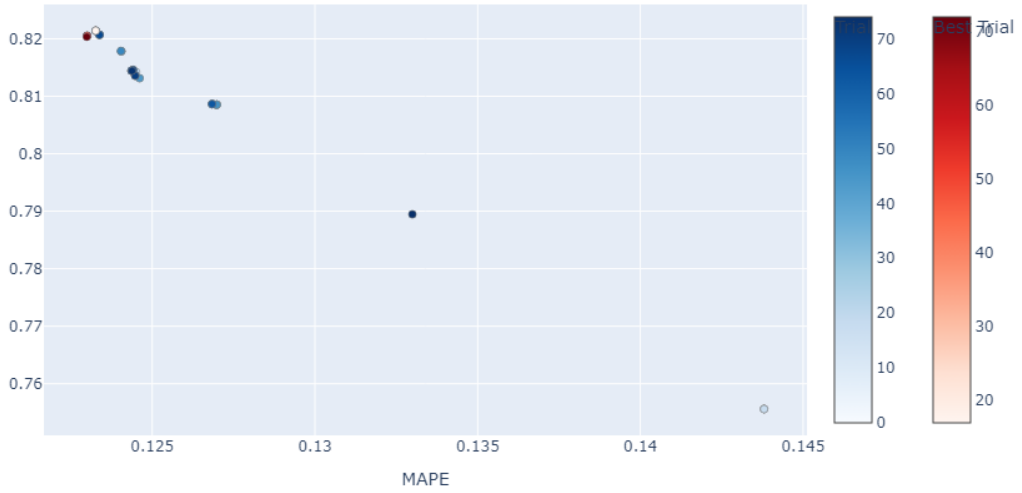

Figura 21. Diagrama de Paretto del proceso de optimización de los modelos **20** y **21**.

Para el ozono se localiza un modelo de XGBoost que presenta tanto el menor error como el mayor ajuste. En cambio, con Ridge se obtiene dos modelos, pero debido a la escasa diferencia entre los resultados de ambos y al gran valor del r^2, se selecciona el modelo **20** que presenta una mayor precisión.

4.4 Creación de la nueva variable 'Festivo'

Con el objetivo de obtener una mejor precisión y un mayor ajuste de los modelos, se introduce además una nueva variable llamada 'Festivo' que obtiene el valor de 1 en caso de ser fin de semana o festivo, basándose en el hecho de que algunos de estos contaminantes están fuertemente relacionados con el uso de vehículos a motor y estas fechas pueden estar relacionadas con el aumento o disminución del tráfico. Empleando los algoritmos con los mejores resultados, se realiza de nuevo una búsqueda de hiperparámetros usando esta nueva variable exógena llamada 'Festivo'.

4.4.1 NO_2

Tabla 21. Búsqueda de hiperparámetros de XGBoost para NO_2 incluyendo 'Festivo'.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **22**	Modelo **23**
learning_rate	0,3	0-0,5	0,0012	0,0572
n_estimators	100	50-700	60	50
max_depth	6	2-11	3	8
min_child_weight	1	1-7	2	4
subsample	1	0,3-1	0,8691	0,7116
colsample_bytree	1	0,3-1	0,9763	0,7786
gamma	0	0-1,5	0,9035	0,0909
lag	24	12-60	54	60
MAPE			1,4883	3,8413
r^2			-0,7483	0,4417

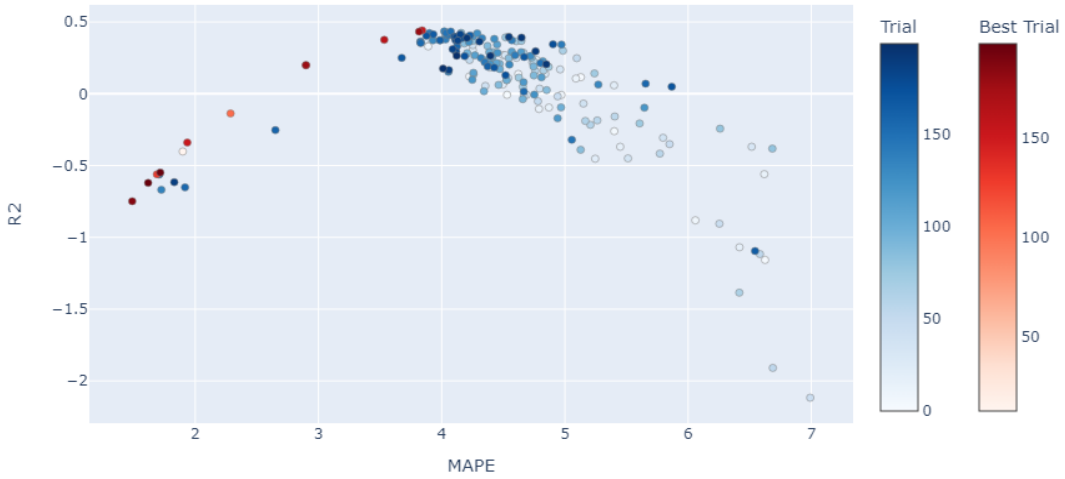

Figura 22. Diagrama de Paretto del proceso de optimización de los modelos **22** y **23**.

Continuando con los criterios anteriores, para el NO_2 el mejor modelo es el que presenta mejor ajuste (modelo **23**). Este modelo mejora los resultados del modelo **2** tras la inserción de esta nueva variable.

4.4.2 SO_2

Tabla 22. Búsqueda de hiperparámetros de XGBoost para SO_2 incluyendo 'Festivo'.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **24**	Modelo **25**
learning_rate	0,3	0-0,5	0,0090	0,009
n_estimators	100	50-700	50	640
max_depth	6	2-11	7	7
min_child_weight	1	1-7	4	7
subsample	1	0,3-1	0,5970	0,7755
colsample_bytree	1	0,3-1	0,7748	0,3082
gamma	0	0-1,5	1,1181	1,0262
lag	24	12-60	33	30
MAPE			1,1047	2,5716
r^2			-0,6112	0,2955

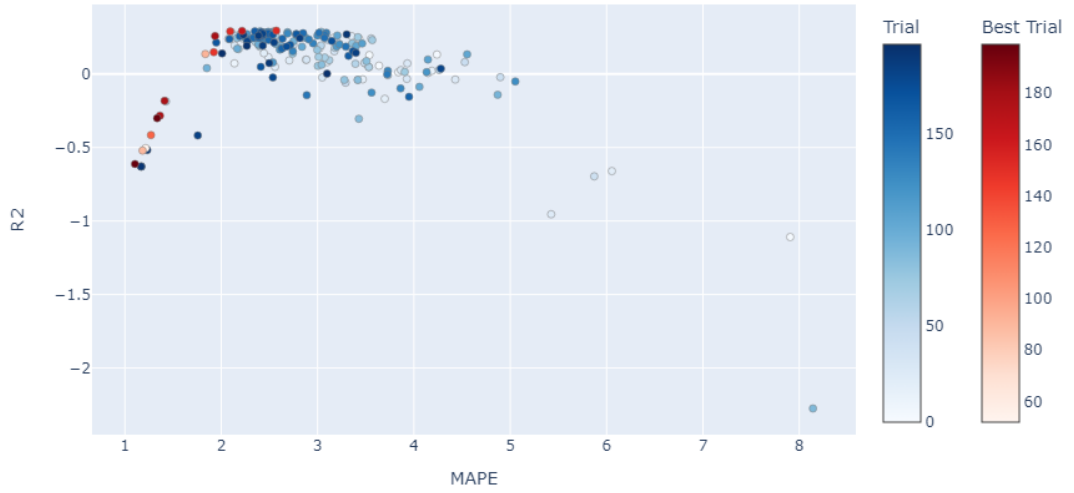

Figura 23. Diagrama de Paretto del proceso de optimización de los modelos **24** y **25**.

Empleando la nueva variable se sigue observando la pérdida de interpretabilidad del modelo a partir de cierto límite en el error, por lo tanto, el mejor modelo es el que presenta mejor r^2 (modelo **25**) y que, en este caso, la nueva variable no presenta una mejora notable con respecto al modelo **6**.

4.4.3 PM2,5

Tabla 23. Búsqueda de hiperparámetros de RandomForest para PM2,5 incluyendo 'Festivo'.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **26**	Modelo **27**
n_estimators	100	50-700	110	630
max_depth	None	2-11	11	10
min_sample_split	2	2-12	10	10
min_sample_leaf	1	1-7	5	7
lag	24	12-60	15	18
MAPE			0,2802	0,2940
r^2			0,5932	0,6015

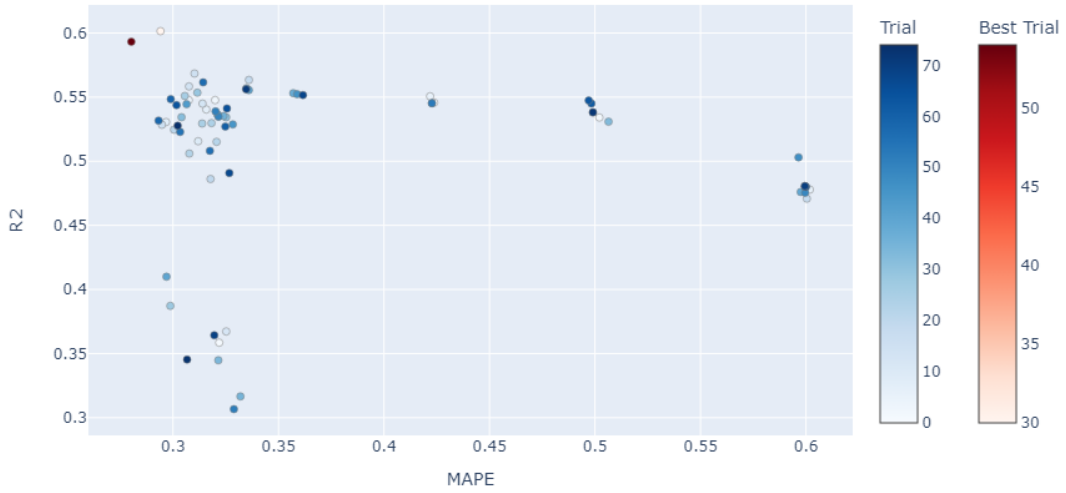

Figura 24. Diagrama de Paretto del proceso de optimización de los modelos **26** y **27**.

Los resultados de ambos modelos son similares, por lo que se selecciona el modelo con menor error (modelo **26**). Para PM2,5, esta nueva variable mejora el ajuste y disminuye ligeramente el error en comparación con el modelo **11**.

4.4.4 PM10

Tabla 24. Búsqueda de hiperparámetros de XGBoost para PM10 incluyendo 'Festivo'.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **28**	Modelo **29**
learning_rate	0,3	0-0,5	0,0636	0,0205
n_estimators	100	50-700	380	680
max_depth	6	2-11	8	9
min_child_weight	1	1-7	1	6
subsample	1	0,3-1	0,7094	0,7815
colsample_bytree	1	0,3-1	0,8614	0,8614
gamma	0	0-1,5	1,0230	0,3130
lag	24	12-60	15	15
MAPE			0,1512	0,1529
r^2			0,7409	0,7468

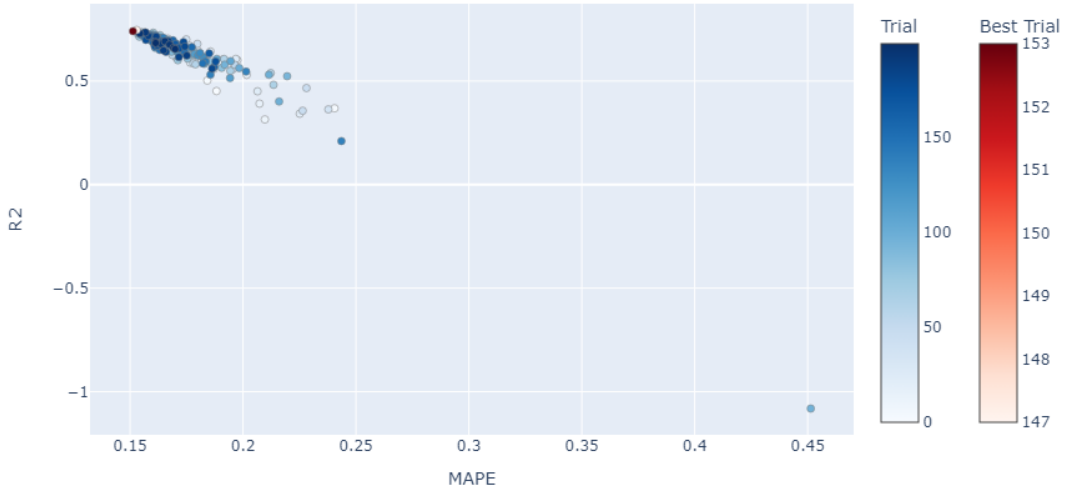

Figura 25. Diagrama de Paretto del proceso de optimización de los modelos **28** y **29**.

Seleccionando el modelo con menor error (modelo **28**), se aprecia un error ligeramente más bajo y un mejor ajuste en comparación con el modelo **15**.

4.4.5 O_3

Tabla 25. Búsqueda de hiperparámetros de Ridge para O3 incluyendo 'Festivo'.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **30**
alpha	1	0-2,5	0,0349
lag	24	12-60	54
MAPE			0,1235
r^2			0,8176

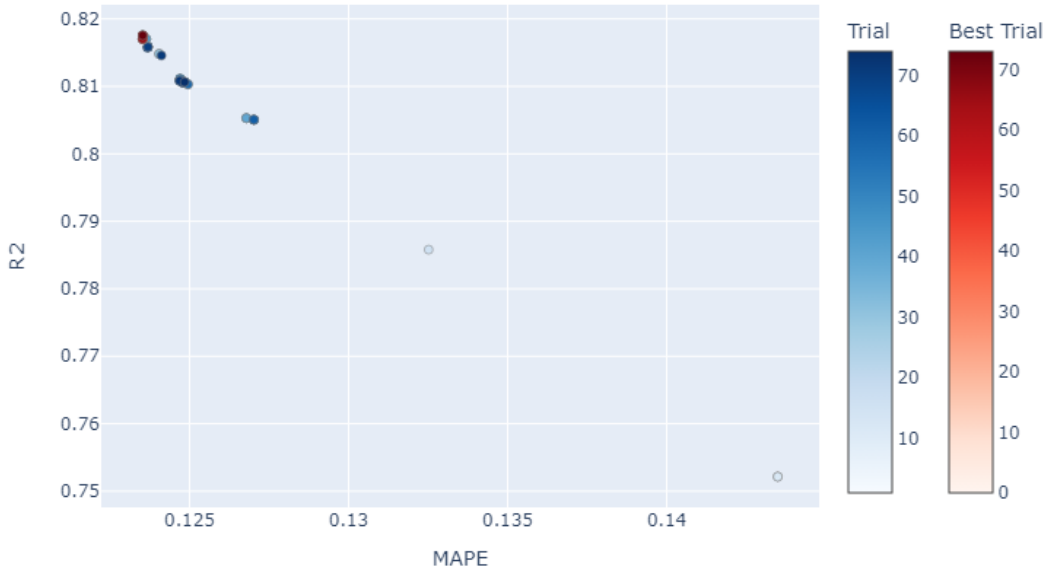

Figura 26. Diagrama de Paretto del proceso de optimización del modelo **30**.

En el modelo **30** se observa un pequeño aumento del error al incorporar la nueva variable con respecto al modelo **20**, aunque sin ningún cambio importante. El comportamiento durante el proceso de optimización es similar, aumentando el ajuste a medida que disminuye el error del modelo.

4.5 Importancia de los atributos meteorológicos

Con el fin de comprender como afectan las variables meteorológicas a los niveles de contaminación atmosférica en Huelva, se realiza un estudio de la importancia que tiene estas variables exógenas en el modelo. Para ello, se hace uso de SHAP[28], una biblioteca de Python que permite la visualización de los efectos de las variables en el resultado final en función del valor que tomen estas. Este estudio se realiza en los mejores modelos encontrados que emplean la variable 'Festivo' (modelos **23**, **25**, **26**, **28** y **30**), ya que su incorporación permite la creación de modelos más precisos en algunos contaminantes o, en otros, no produce cambios negativos apreciables.

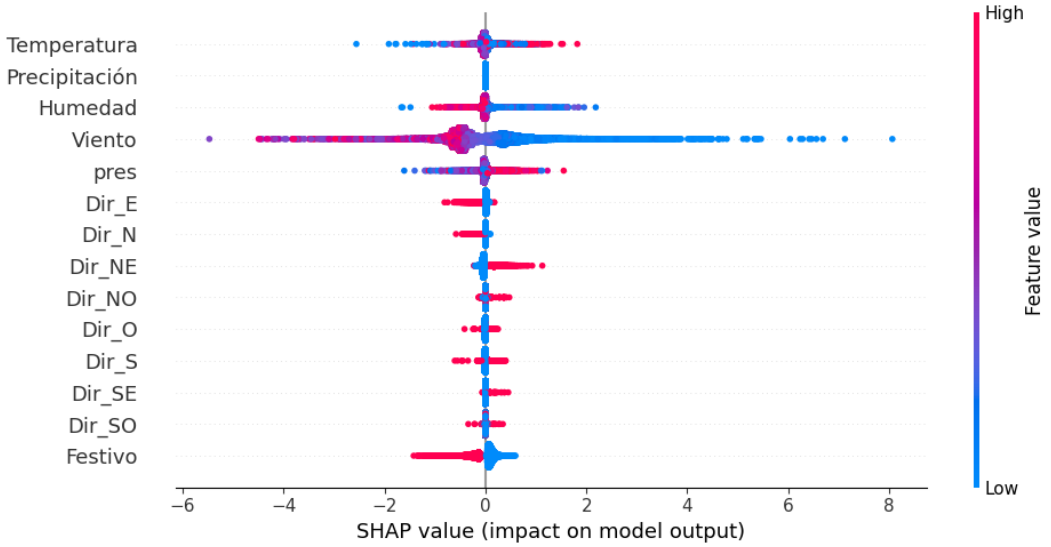

Figura 27. Gráfica del efecto de las variables exógenas del modelo **23** para el contaminante NO_2.

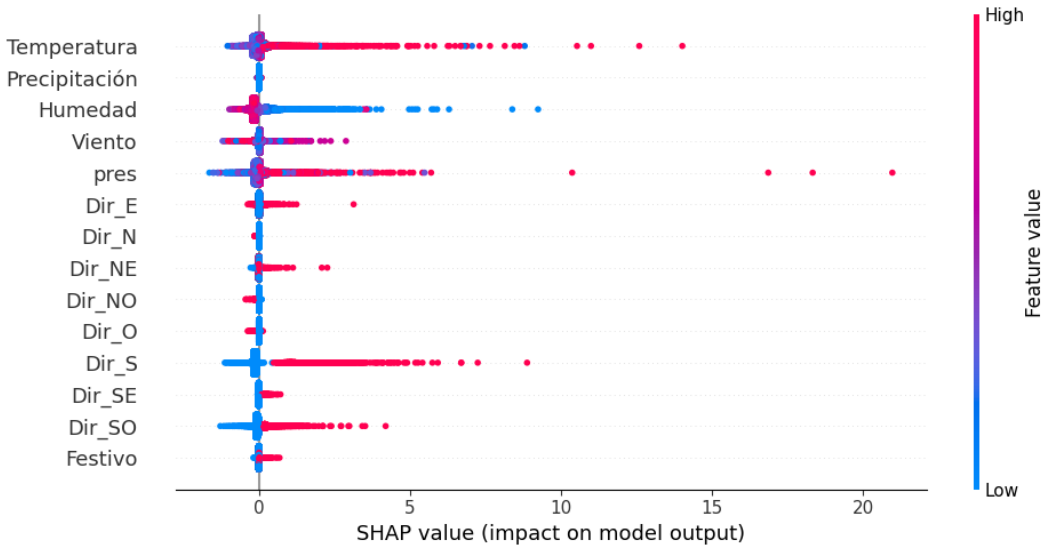

Figura 28. Gráfica del efecto de las variables exógenas del modelo **25** para el contaminante SO_2.

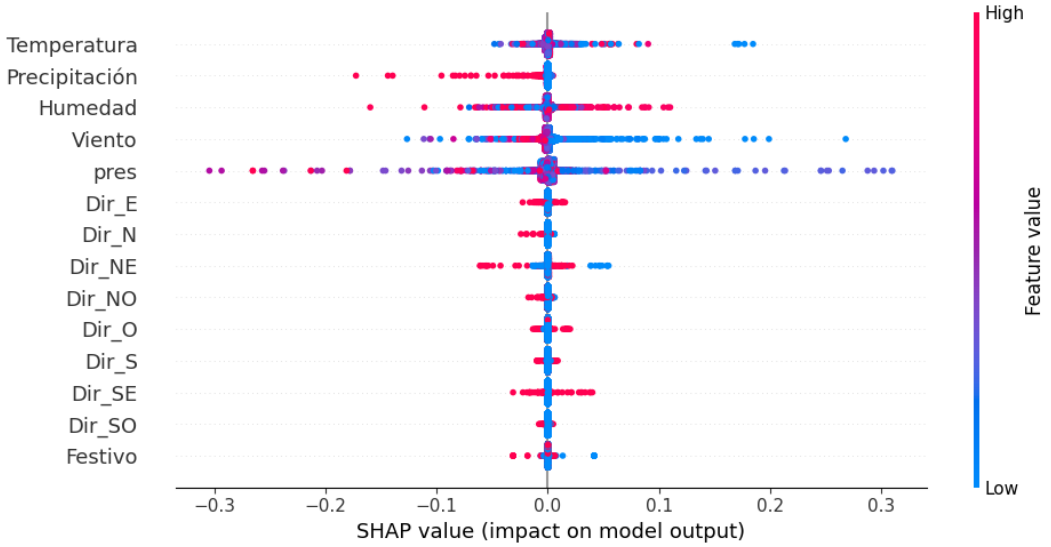

Figura 29. Gráfica del efecto de las variables exógenas del modelo **26** para el contaminante PM2,5.

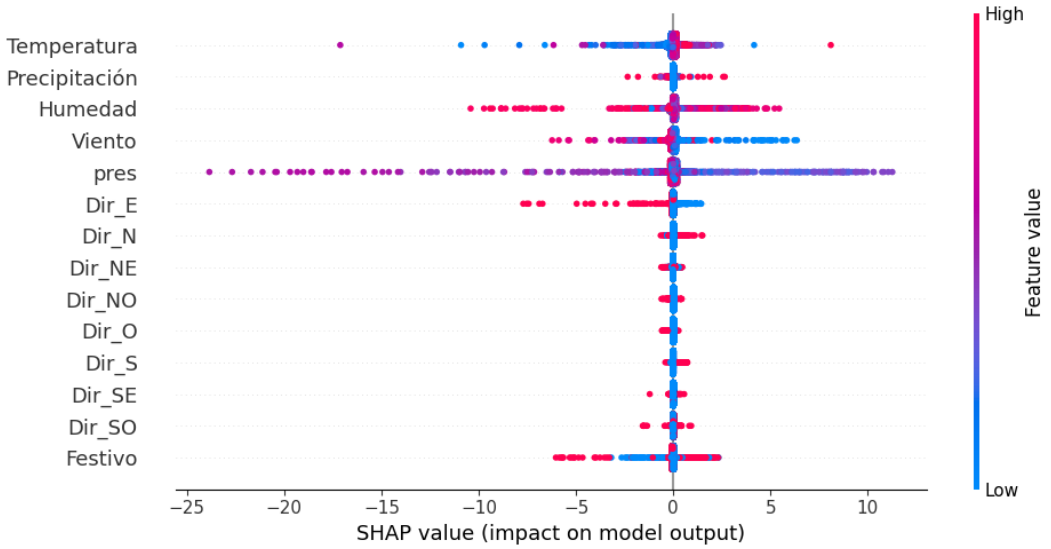

Figura 30. Gráfica del efecto de las variables exógenas del modelo **28** para el contaminante PM10.

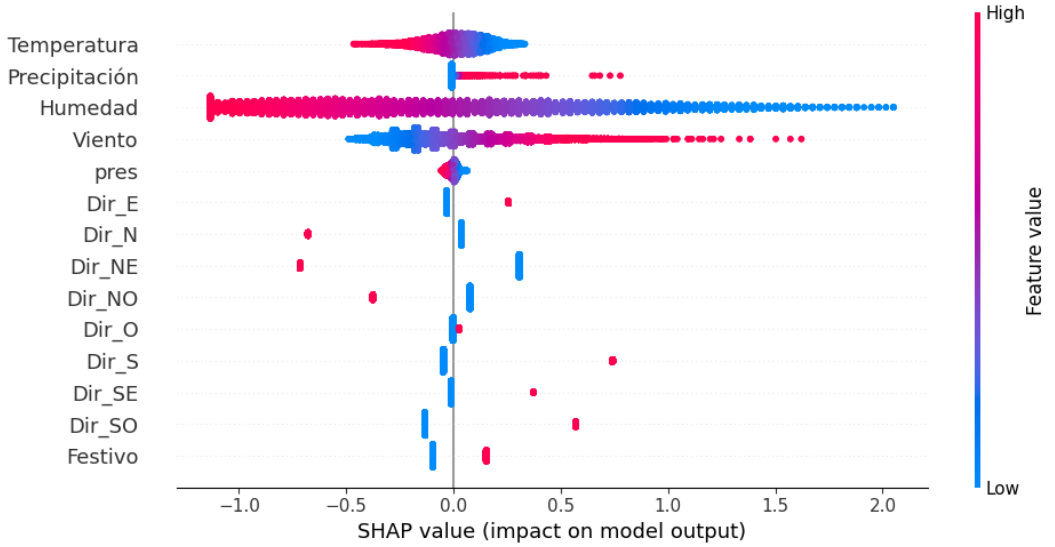

Figura 31. Gráfica del efecto de las variables exógenas del modelo **30** para el contaminante O_3.

En la variable temperatura no hay una única tendencia: en el caso del SO_2 a medida que aumenta la temperatura aumentan los niveles de concentración de este; en cambio, en el ozono se observa que el aumento de la temperatura provoca una disminución de los niveles. La precipitación solo tiene importancia con los PM2,5 y el ozono, disminuyendo cuando llueve los primeros y aumentando el segundo. Niveles bajos de humedad también están relacionados con valores más bajos de NO_2, SO_2 y O_3. La variable viento tiene un gran impacto en todos los modelos, aumentando los niveles de concentración cuando la velocidad de este es bajo y disminuyéndolo a medida que aumenta la velocidad. La dirección del viento también tiene impacto en los contaminantes, las direcciones N, NE y E disminuyen las concentraciones de NO_2, PM10 y O_3, pero aumenta las de SO_2. Las direcciones S, SO y SE producen aumentos de la concentración de SO_2 y O_3. Por último, la variable ‘Festivo’ tiene un claro impacto negativo en el NO_2.

4.6 Modelo predictivo sobre la estación Campus el Carmen

Finalmente, se elabora un modelo predictivo para cada contaminante para la estación del Campus del Carmen, ya que monitoriza todos los contaminantes atmosféricos, y así comprender mejor los resultados obtenidos en este estudio y determinar si es mejor el entramiento para una estación concreta o usar todas las disponibles. En este caso, se realiza una búsqueda de hiperparámetros empleando únicamente el algoritmo de XGBoost para establecer una comparativa con los modelos anteriores que empleaban este algoritmo.

4.6.1 NO_2

Tabla 26. Búsqueda de hiperparámetros de XGBoost para NO_2 en la estación Campus el Carmen.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **31**	Modelo **32**
learning_rate	0,3	0-0,5	0,3106	0,0694
n_estimators	100	50-700	170	390
max_depth	6	2-11	4	8
min_child_weight	1	1-7	5	4
subsample	1	0,3-1	0,3200	0,8247
colsample_bytree	1	0,3-1	0,8159	0,8698
gamma	0	0-1,5	1,1847	1,1917
lag	24	12-60	36	60
MAPE			4,8704	4,9161
r^2			0,4535	0,6054

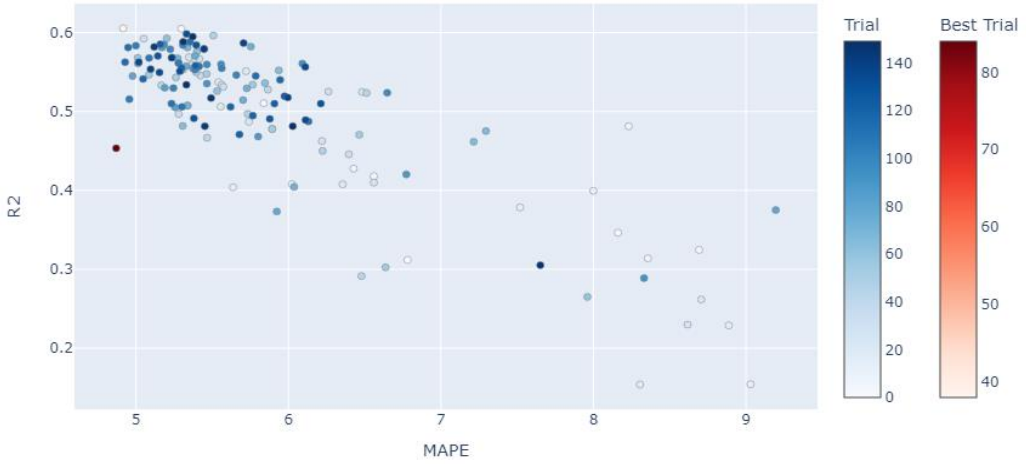

Figura 32. Diagrama de Paretto del proceso de optimización de los modelos **31** y **32**.

El MAPE en esta estación es muy elevado, pero en general, al aumentar el r^2 no hay un gran aumento del error. El modelo **32** presenta mejor ajuste respecto al **31** con una variación mínima en el error, por lo que es el modelo más adecuado.

4.6.2 SO_2

Tabla 27. Búsqueda de hiperparámetros de XGBoost para SO_2 en la estación Campus el Carmen.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **33**	Modelo **34**
learning_rate	0,3	0-0,5	0,0066	0,1621
n_estimators	100	50-700	330	690
max_depth	6	2-11	10	7
min_child_weight	1	1-7	3	7
subsample	1	0,3-1	0,7515	0,8247
colsample_bytree	1	0,3-1	0,3989	0,9415
gamma	0	0-1,5	0,0059	0,1602
lag	24	12-60	42	54
MAPE			0,2855	0,4325
r^2			0,2199	0,2432

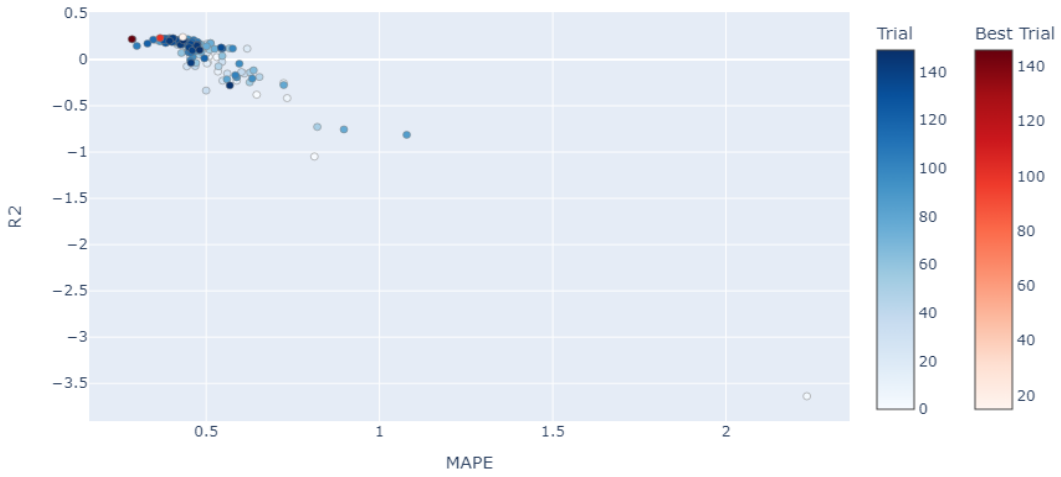

Figura 33. Diagrama de Paretto del proceso de optimización de los modelos **33** y **34**.

El error obtenido para el SO_2 es relativamente bajo, pero se observa un gran incremento en término porcentuales cuando aumenta ligeramente el ajuste. En cambio, no se observa ninguna tendencia a empeorar el ajuste a medida que disminuye el error. En este caso, el modelo más adecuado sería el modelo **33**, ya que presenta menor error y la pérdida de la interpretabilidad es muy baja.

4.6.3 PM2,5

Tabla 28. Búsqueda de hiperparámetros de XGBoost para PM2,5 en la estación Campus el Carmen.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **35**	Modelo **36**
learning_rate	0,3	0-0,5	0,2842	0,4250
n_estimators	100	50-700	140	610
max_depth	6	2-11	7	8
min_child_weight	1	1-7	4	4
subsample	1	0,3-1	0,9731	0,7584
colsample_bytree	1	0,3-1	0,3386	0,9462
gamma	0	0-1,5	0,6970	0,9880
lag	24	12-60	39	45

MAPE	0,2913	0,3270
r^2	0,5738	0,6001

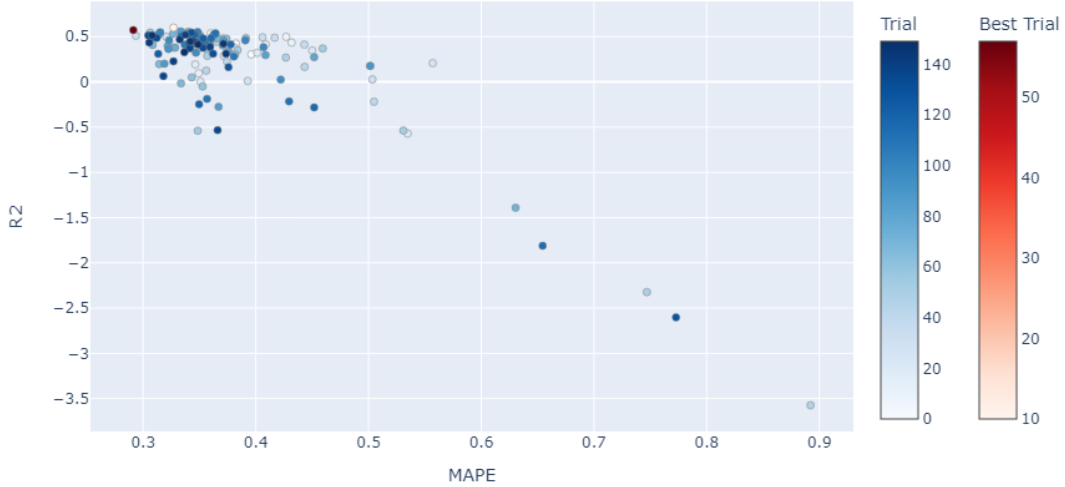

Figura 34. Diagrama de Paretto del proceso de optimización de los modelos **35** y **36**.

La diferencia de resultados entre los dos modelos es mínima, por lo que el más adecuado es el modelo **35** ya que presenta el menor error.

4.6.4 PM10

Tabla 29. Búsqueda de hiperparámetros de XGBoost para PM10 en la estación Campus el Carmen.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **37**	Modelo **38**
learning_rate	0,3	0-0,5	0,2607	0,2607
n_estimators	100	50-700	390	460
max_depth	6	2-11	6	9
min_child_weight	1	1-7	2	1
subsample	1	0,3-1	0,7907	0,6920

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **37**	Modelo **38**
colsample_bytree	1	0,3-1	0,7591	0,5278
gamma	0	0-1,5	1,1731	1,1731
Lag	24	12-60	15	21
MAPE			0,1925	0,1948
r^2			0,6264	0,7210

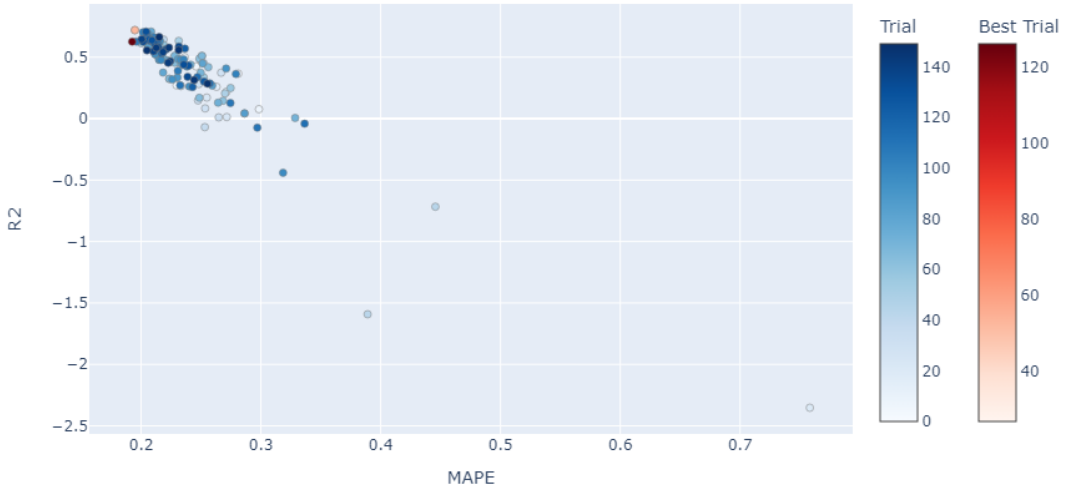

Figura 35. Diagrama de Paretto del proceso de optimización de los modelos **37** y **38**.

El ajuste y la precisión de ambos modelos es muy alta, aunque el modelo **38** presenta mayor r^2, pero el aumento del error es despreciable en comparación con el aumento del ajuste.

4.6.5 O_3

Tabla 30. Búsqueda de hiperparámetros de XGBoost para O_3 en la estación Campus el Carmen.

Hiperparámetros	Valor por defecto	Rango búsqueda	Modelo **39**	Modelo **40**
learning_rate	0,3	0-0,5	0,0640	0,0400
n_estimators	100	50-700	410	240
max_depth	6	2-11	6	6
min_child_weight	1	1-7	1	5
subsample	1	0,3-1	0,7153	0,4002
colsample_bytree	1	0,3-1	0,7311	0,6961
gamma	0	0-1,5	1,3278	0,6708
lag	24	12-60	36	39
MAPE			0,1340	0,1344
r^2			0,7987	0,8046

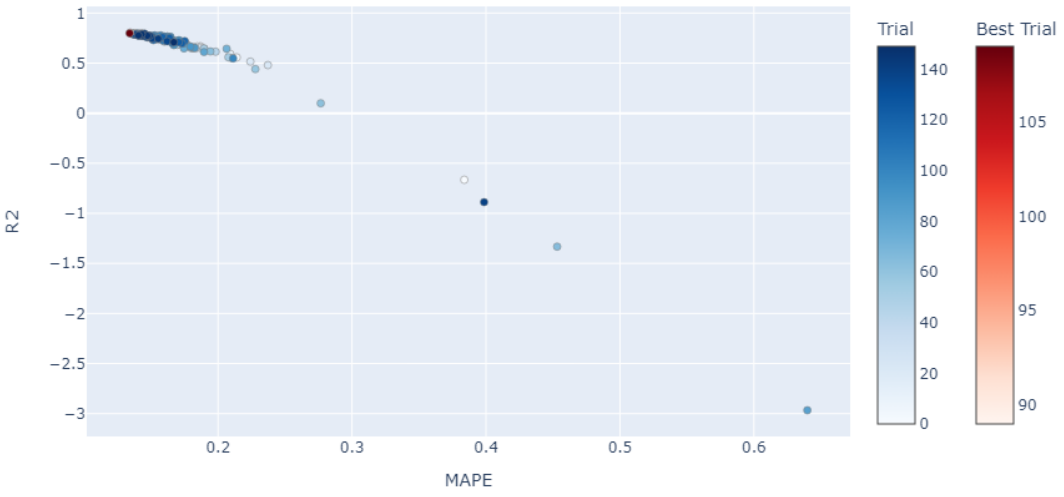

Figura 36. Diagrama de Paretto del proceso de optimización de los modelos **39** y **40**.

Como en caso anteriores, el ozono no presenta grandes diferencias entre modelos, encontrando una gran relación entre la precisión y el ajuste del modelo en el proceso de optimización y siendo indiferente la elección de uno de los dos modelos, aunque se selecciona el modelo **39** al presentar menor MAPE.

5 RESULTADOS Y DISCUSIÓN

Para la construcción del modelo predictivo para cada uno de los contaminantes, se entrenaron todos los algoritmos usando los parámetros por defecto con el objetivo de construir un modelo base y encontrar aquellos con menor error. En cuanto a la selección de los mejores algoritmos para los estudios posteriores, XGBoost se aplica a todos los contaminantes debido a que muestra buenos resultados y a la posibilidad de entrenar el modelo empleando GPU, reduciendo el tiempo necesario, lo que permite una búsqueda mayor de hiperparámetros y, por lo tanto, una mejor optimización de los modelos. Con respecto al resto de algoritmos, se seleccionan aquellos con menor MAPE (**Tabla 31**). Los mejores modelos base para los contaminantes ozono, PM10 y PM2,5 ya presentan un error relativamente bajo antes de optimizar sus hiperparámetros.

Tabla 31. Resultados de los mejores algoritmos para cada contaminante.

Contaminante	**Algoritmo**	**Error***
NO_2	kNN	4,6638
	XGBoost	4,6644
SO_2	kNN	2,4850
	XGBoost	2,5909
	Ridge	2,6124
PM2,5	Random Forest	0,3274
	Ridge	0,3322
	XGBoost	0,3556
PM10	Ridge	0,1622
	XGBoost	0,1627
O_3	Ridge	0,1250
	XGBoost	0,1360

* El error se muestra en MAPE.

5.1 Búsqueda de hiperparámetros

En el proceso de optimización de parámetros se observó que cuando se emplea XGBoost en los contaminantes NO_2 y SO_2 se puede lograr un error porcentual relativamente bajo, pero con una gran pérdida del ajuste del modelo. En la **Figura 37** se muestra la predicción de los niveles de SO_2 en La Orden con los modelos **5** y **6**, apreciándose como el modelo **5**, que presenta el MAPE más bajo, tiene un comportamiento casi lineal, minimizando de esta manera el error de predicción, pero mostrando una escasa interpretabilidad, sin capacidad de anticipación ante las variaciones más agresivas de este contaminante. En cambio, el modelo **6**, que obtiene el mayor r^2 a pesar de tener un error porcentual relativamente alto, es capaz de predecir las fluctuaciones que presenta.

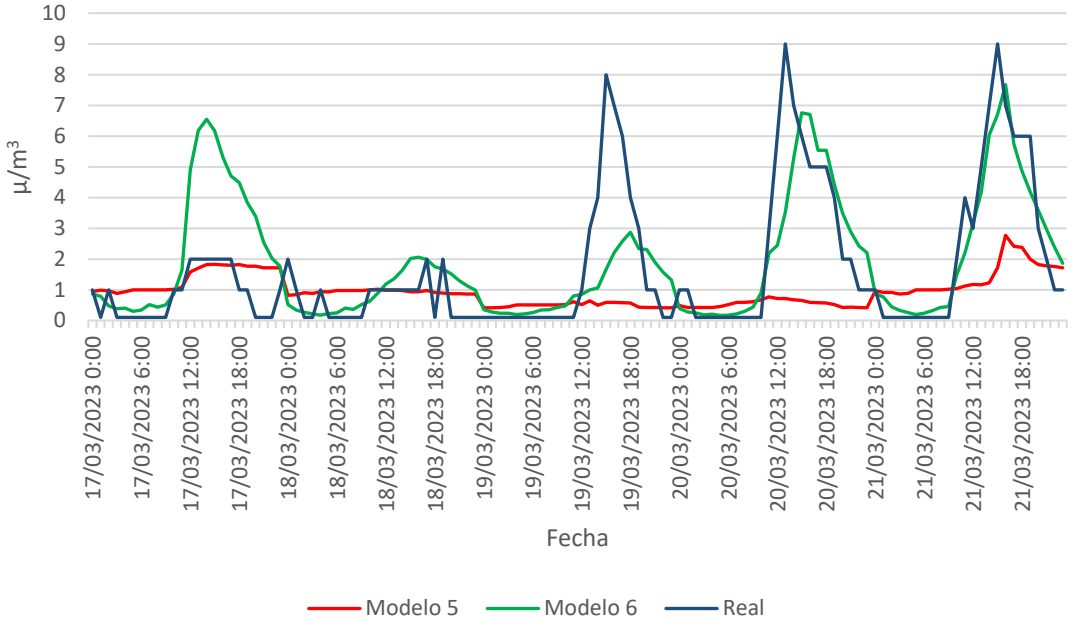

Figura 37. Predicción realizada en la estación de La Orden para el contaminante SO_2 entre 17/03/2023 y 21/03/2023 usando los modelos **5** y **6**.

Esta situación que experimenta los contaminantes NO_2 y SO_2 con el algoritmo de XGBoost, en la cual el modelo con menor ajuste no es aquel que presenta el menor error de predicción, se puede apreciar mejor en las gráficas de Paretto donde se muestra el proceso de optimización (**Figura 11** y **Figura 13**). A medida que el MAPE disminuye, el ajuste para estos modelos también lo hace, por lo que se pierde precisión de los resultados

cuando se intenta conseguir una mayor interpretabilidad del modelo. En cambio, para los contaminantes PM10 y PM2,5 (**Figura 15** y **Figura 18**), cuanto menor es el error, mayor es el ajuste de los modelos, siendo aún más evidente con el contaminante O_3, en el cual se obtiene un modelo con el mínimo error y el mejor el ajuste (**Figura 20**).

En los otros modelos optimizados, no se observa esta tendencia a la pérdida de interpretabilidad de los modelos a medida que se aumenta la precisión de los resultados. Los modelos que emplean kNN para NO_2 y SO_2 no muestran una tendencia clara, pero no presenta una gran pérdida del ajuste a medida que disminuye el error de predicción con la modificación de los hiperparámetros (**Figura 12** y **Figura 14**). Con el algoritmo de Random Forest para PM2,5 se observa que a medida que disminuye el error aumenta el ajuste (**Figura 16**), de la misma forma que ocurre con el modelo **9** y **10** (XGBoost). Finalmente, para los modelos de Ridge, la variación del error y de la interpretabilidad es muy pequeña en cada uno de los *trials*, pero el error y el ajuste mejoran conjuntamente (**Figura 17**, **Figura 19** y **Figura 21**).

Como la interpretación del modelo es una característica importante a tener en cuenta para un buen modelo predictivo de la contaminación atmosférica, ya que la principal finalidad es predecir las variaciones de concentración del contaminante para poder evitar determinados riesgos a la población, para los contaminante NO_2 y SO_2 el criterio más adecuado es el r^2, ya que permite interpretar la evolución del contaminante a lo largo de las horas aunque eso implique aumentar el error de predicción para estas sustancias. En cambio, para PM10, PM2,5 y O_3 se selecciona el que tenga menor error ya que la pérdida de ajuste del modelo es despreciable, sin que pueda mostrar una gran variación en la interpretación de este, siendo más interesante un modelo con mayor precisión (**Tabla 32**).

Tabla 32. Resultado de los mejores modelos.

Contaminante	**Modelo**	**Algoritmo**	**MAPE**	**r^2**
NO_2	Modelo **2**	XGboost	4,1010	0,4238
SO_2	Modelo **6**	XGBoost	2,5926	0,3010
PM2,5	Modelo **11**	Random Forest	0,2955	0.5215

Contaminante	Modelo	Algoritmo	MAPE	r^2
PM10	Modelo **15**	XGBoost	0,1539	0,7257
O_3	Modelo **20**	Ridge	0,1230	0,8203

5.2 Nueva variable 'Festivo'

Analizando detenidamente las predicciones de estos modelos junto a sus valores reales, se puede apreciar como en algunos contaminantes hay una disminución de su concentración durante los fines de semana (**Figura 38**). Debido a que una de las fuentes de emisión de los contaminantes son los vehículos a motor y bajo la hipótesis de que el tráfico en la ciudad es menor durante los días festivos y los fines de semana, se propone introducir una nueva variable llamada 'Festivo' a cada uno de los modelos seleccionados, que toma el valor de 1 los fines de semana y los días festivos, y 0 en caso contrario. Como se puede observar en la **Figura 39**, tras insertar esta nueva variable, la predicción de los picos de mayor concentración mejora, siendo más intensos los días laborables y menos intensos los fines de semana (día 29 de abril en **Figura 39**)[i].

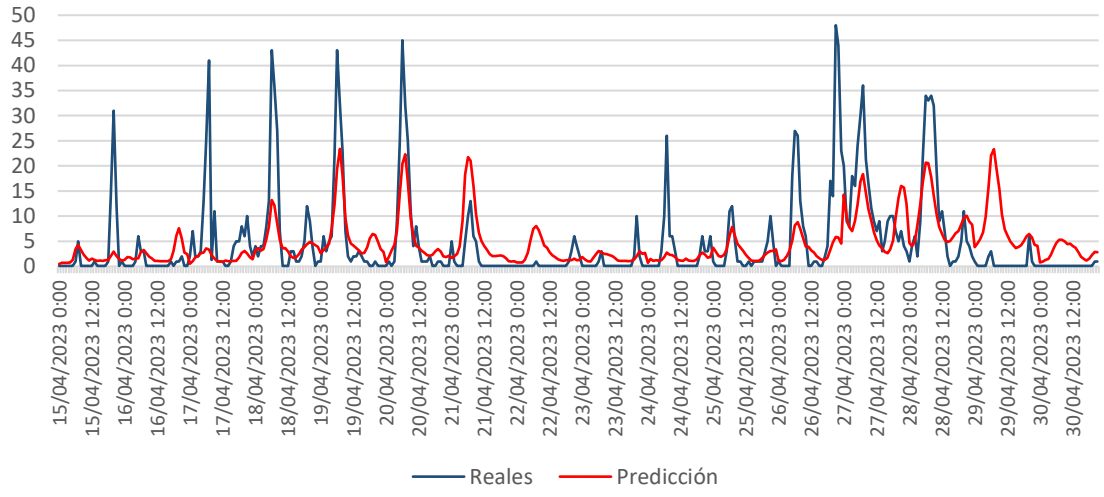

Figura 38. Predicción realizada en la estación de Campus el Carmen para el contaminante NO_2 entre 15/04/2023 y 30/04/2023 usando el modelo **2**.

[i] Ver *Apéndice A* para consultar otras gráficas.

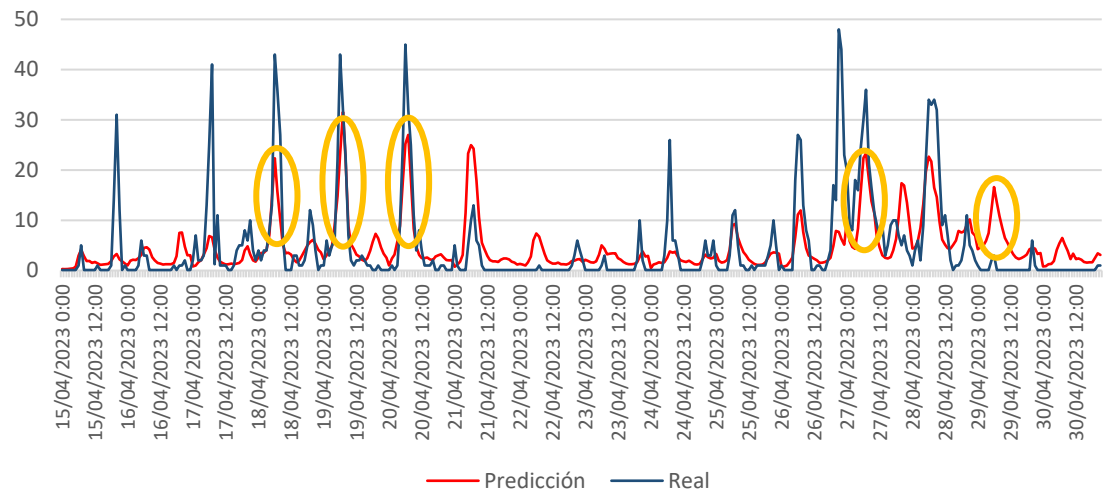

Figura 39. Predicción realizada en la estación de La Orden para el contaminante NO_2 entre 17/03/2023 y 21/03/2023 con la variable 'Festivo', usando el modelo **23**.

En el proceso de optimización de los modelos usando la nueva variable, se observó en las gráficas de Paretto el mismo comportamiento que cuando no se empleaba, por lo que el criterio de selección de los mejores modelos es el mismo. Al comparar los resultados de los modelos antes y después de introducir la variable 'Festivo' (**Tabla 33**), se aprecia una mayor mejora en términos de precisión y de ajuste en el contaminante NO_2, cuya fuente de emisión mayoritaria son los vehículos a motor. En cambio, para el resto de los contaminantes, las emisiones debidas al tráfico representan un porcentaje menor y tienen menos influencia en sus niveles de concentración (**Figura 40**)[11]. Esta relación demuestra que el tráfico es una variable exógena importante a tener en cuenta para los modelos de contaminación ambiental, sobre todo para determinados contaminantes, como el NO_2 o el SO_2, y en zonas donde el tráfico sea una de las principales fuentes de emisión, pudiendo mejorar la precisión del modelo. Esta variable se ha tratado de introducir con la variable 'Festivo', pero no mediante una variable más precisa como es el volumen del tráfico en determinados días y fechas ya que se carece de esta información en Huelva.

Tabla 33. Comparación de los modelos cuando se introduce la variable 'Festivo'.

		Sin 'Festivo'			'Con Festivo'		
Contaminante	**Algoritmo**	**Modelo**	**MAPE**	**r^2**	**Modelo**	**MAPE**	**r^2**
NO_2	XGboost	**2**	4,1010	0,4238	**23**	3,8413	0,4417
SO_2	XGBoost	**6**	2,5926	0,3010	**25**	2,5716	0,2955
PM2,5	RandomForest	**11**	0,2955	0.5215	**26**	0,2802	0,5932
PM10	XGBoost	**15**	0,1539	0,7257	**28**	0,1512	0,7409
O_3	Ridge	**20**	0,1230	0,8203	**30**	0,1235	0,8176

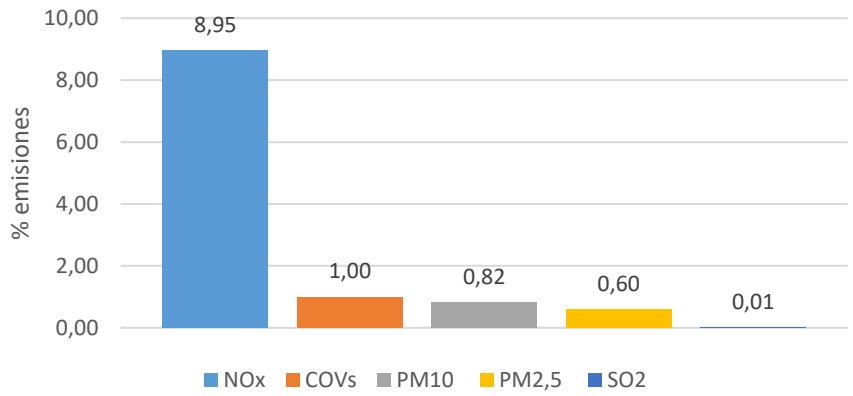

Figura 40. Porcentaje de las emisiones totales de cada contaminante debidas al tráfico rodado.

5.3 Importancia de atributos

Para comprender mejor la influencia de los factores meteorológicos en la contaminación atmosférica de Huelva, se realiza un estudio de la importancia de las diferentes variables meteorológicas[ii]. Respecto a la variable 'Temperatura', solo muestra una tendencia en el caso del SO_2, que aumenta los niveles con los momentos de mayor calor; y con el O_3, que disminuye los valores de concentración cuando esta aumenta. Las precipitaciones solo tienen importancia en PM2,5 y el O_3, disminuyendo sus

[ii] Para consultar la importancia del resto de variables, ver *Apéndice B.*

concentraciones cuando llueve. Por otro lado, niveles más elevados de humedad disminuye las concentraciones de NO_2, SO_2 y O_3.

La variable 'Festivo' que fue introducida con el objetivo de mejorar la precisión de los modelos, solo tiene efecto en NO_2, el único modelo que presentó mayor mejora en el ajuste al introducir la variable y el contaminante más relacionado con el tráfico. De esta forma se confirma la importancia de la introducción de una variable que controle el tráfico rodado, siendo importante únicamente cuando la fuente de emisión principal de la zona sean los vehículos, en caso contrario parece no tener ningún tipo de influencia, como es el caso del SO_2.

El viento es la variable con mayor importancia en la mayoría de ellos, disminuyendo la concentración atmosférica cuando aumenta la velocidad de este debido a que favorece la dispersión de los contaminantes. Aunque no solo la velocidad tiene impacto sobre el modelo, sino también su dirección. Los contaminantes SO_2, PM10 y O_3 se ven muy afectados por la dirección S, aumentando los niveles de estos. Estos contaminantes tienen como principal fuente de emisión la industria y la producción de energía (**Figura 41**)[11], cuyas actividades tienen una gran presencia en el sur de Huelva, principalmente en la Avda. Francisco Montenegro (Huelva) y en Palos de la Frontera (**Figura 42**). Estos gases y partículas pueden ser dirigidos desde esta zona cuando el viento sopla desde el sur, y de la misma forma también se ven reducidos cuando el viento proviene del norte, siendo más notable en el caso del SO_2 (**Figura 28**).

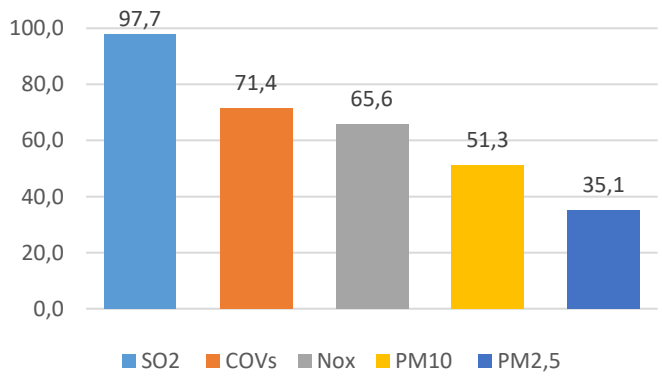

Figura 41. Porcentaje de las emisiones totales de cada contaminante debidas a la industria y producción de energía.

Figura 42. Situación geográfica de la actividad industrial en Huelva y Palos de la Fronteara (marcado rojo) y de las diferentes estaciones de monitorización de contaminantes distribuidas por la capital de Huelva (puntos azules).

5.4 Estudio de la estación Campus el Carmen

Por último, la **Tabla 34** muestra el error de los mejores modelos de cada contaminante desglosado en las diferentes estaciones de medición de cada uno de ellos. Algunos modelos presentan una gran variación del error dependiendo de la estación donde se aplica el modelo. Estos resultados se deben a la gran diferencia de niveles de concentración de los contaminantes en cada zona (**Tabla 35**), como es el caso del NO_2 que tiene una concentración media de 6,17 μ/m^3 en la estación Campus el Carmen, mientras que en la estación la Marismas del Titán la concentración supera los 14 μ/m^3. Este hecho se aprecia también en el contaminante SO_2. En cambio, la materia particulada

y el ozono presentan valores similares en todas las estaciones. Esta situación podría ser la causa del gran error porcentual que presenta algunos de los modelos y de su gran falta de ajuste. Por ello, se desarrolla un modelo predictivo para una única estación, en este caso la estación Campus del Carmen, ya que registra los niveles de concentración de todos los contaminantes. El algoritmo elegido es XGBoost, ya que es el más rápido de entrenar y el que ha mostrado buenos resultados en todos los contaminantes.

Tabla 34. Errores de los mejores modelos de cada contaminante desglosados en las distintas estaciones (error en MAPE).

	NO_2	**SO_2**	**PM2,5**	**PM10**	**O_3**
Modelo	**23**	**25**	**26**	**28**	**30**
Campus el Carmen	5,2632	0,3779	0,2993	0,1896	0,1333
La Orden	0,8796	7,8123	-	0,1045	0,1138
Los Rosales	4,4253	5,9650	-	0,1247	-
Marimas del Titán	0,3743	0,1692	-	0,2077	-
Pozo Dulce	8,2640	0,6460	0,2612	0,1571	-
Romelejo	-	0,4591	-	0,1240	-
Media	3,8413	2,5716	0,2802	0,1513	0,1235

Tabla 35. Descriptivos de los contaminantes desglosados por estaciones.

Contaminante	**Estaciones**	**Media**	**Mínimo**	**Máximo**
NO_2	CAMPUS EL CARMEN	6,17	0,10	103,00
	LA ORDEN	7,77	0,10	173,00
	LOS ROSALES	9,50	0,10	89,00
	MARISMAS DEL TITÁN	14,87	1,00	101,00
	POZO DULCE	10,60	0,10	170,00

Contaminante	Estaciones	Media	Mínimo	Máximo
SO_2	CAMPUS EL CARMEN	3,47	0,10	84,00
	LA ORDEN	5,89	0,10	373,00
	LOS ROSALES	5,35	0,10	172,00
	MARISMAS DEL TITÁN	4,29	0,10	229,00
	POZO DULCE	3,54	0,10	216,00
	ROMERALEJO	9,91	1,00	122,00
PM2,5	CAMPUS EL CARMEN	7,89	0,10	61,76
	POZO DULCE	8,26	0,10	53,91
PM10	CAMPUS EL CARMEN	21,10	3,44	219,41
	LA ORDEN	24,96	8,66	180,39
	LOS ROSALES	26,07	8,78	205,84
	MARISMAS DEL TITÁN	23,56	3,71	175,99
	POZO DULCE	24,61	0,10	198,98
	ROMERALEJO	33,83	5,27	93,75
O_3	CAMPUS EL CARMEN	57,40	4,63	121,63
	LA ORDEN	61,44	6,25	140,38

* Datos expresados en μ/m^3.

Con la búsqueda de hiperparámetros enfocada a disminuir el MAPE y maximizar el r^2 de esta única estación, se obtiene resultados similares, aunque en la mayoría de ellos presentan un mejor de ajuste del modelo a los datos reales (**Tabla 36**). El NO_2 es el contaminante que más se beneficia de esta estrategia, mejorando tanto el error como el ajuste. Para el SO_2 se consigue una mejor precisión de la predicción, pero se ve afectado ligeramente la interpretabilidad del modelo. La predicción de la materia particulada presenta el mismo error, aunque la interpretabilidad del modelo, y en el caso del ozono la diferencia es inapreciable. Por otro lado, las gráficas de Paretto de todos los contaminantes muestran una mejora de la precisión a medida que aumenta el r^2 del

modelo, incluidos NO_2 y SO_2, que anteriormente presentaban una gran pérdida de la interpretabilidad cuando el error era mínimo[iii].

Tabla 36. Comparación de errores de Campus del Carmen con los otros modelos.

	Todas las estaciones[1]			**Campus el Carmen[2]**		
	Modelo	**MAPE**	**r^2**	**Modelo**	**MAPE**	**R^2**
NO_2	**23**	5,2632	0,5589	**32**	4,9161	0,6054
SO_2	**25**	0,3779	0,2460	**33**	0,2855	0,2199
PM2,5	**26**	0,2993	0,5141	**35**	0,2913	0,5738
PM10	**28**	0,1896	0,6755	**38**	0,1948	0,7210
O_3	**30**	0,1333	0,8095	**39**	0,1340	0,7987

[1] MAPE que presenta la estación Campus del Carmen con los modelos creados enfocados a disminuir el error y maximizar el ajuste en conjunto.
[2] MAPE que presenta la estación Campus el Carmen con los modelos creados enfocados a disminuir el error y maximizar el ajuste de esta estación.

Estos resultados demuestran que una diferencia elevada en los niveles de contaminación entre dos puntos de la ciudad puede perjudicar la eficacia de los modelos, ya que los modelos más afectados en este experimento son el NO_2 y el SO_2, los cuales presentan una mayor diferencia de concentración entre diferentes estaciones. Por otro lado, en el resto de los modelos no se aprecia grandes diferencias en los resultados, por lo que indica que la creación de distintos modelos para las diferentes zonas de la ciudad es una buena estrategia que permite obtener predicciones más precisas y que se ajusten mejor a las futuras variaciones de concentración.

[iii] Ver *Apéndice A* para consultar gráficas de predicción.

6 CONCLUSIONES Y TRABAJOS FUTUROS

- Se logra encontrar los mejores algoritmos para cada uno los contaminantes atmosféricos. Los mejores algoritmos ya muestran buena precisión sin ser optimizados para la materia particulada (PM2,5 y PM10) y el ozono.
- La optimización de hiperparámetros proporciona mejoras en la precisión y ajuste de todos los modelos, aunque en el proceso de optimización para el dióxido de nitrógeno y dióxido de azufre se observan problemas de interpretabilidad a medida que disminuye el error, que son solucionados con la elección de los modelos con mayor r^2.
- La creación de la nueva variable 'Festivo' permite mejorar la precisión y el ajuste del modelo diseñado para el dióxido de nitrógeno, un gas muy relacionado con el tráfico rodado.
- Los mejores modelos encontrados son: **23**, **25**, **26**, **28** y **30**.
- El efecto de las variables exógenas varía en función del contaminante, pero la velocidad del viento es común en todos, reduciendo los niveles cuando la velocidad es mayor. La dirección de este también toma protagonismo en los contaminantes SO_2, PM10 y O_3, siendo las direcciones sur las que más influyen, aumentando los niveles de estos contaminantes, encontrándose así relación con la ubicación de la zona industrial en la provincia de Huelva.
- El desarrollo de modelos predictivos para una única estación mejora la precisión y el ajuste de los modelos en la mayoría de los contaminantes, ya que se eliminan las diferencias existentes en la concentración media de cada estación a causa de su localización.

6.1 Trabajos futuros

La adición de la nueva variable 'Festivo', que guarda relación con el tráfico rodado, supone un aumento de la precisión del modelo aplicado a los óxidos de nitrógeno, por lo que contar con una variable exógena que proporcione información del estado del

tráfico y, mediante las estimaciones oportunas, del nivel de desplazamientos en coche a determinadas horas y días del año, podría permitir el desarrollo de mejores modelos predictivos de contaminación ambiental, sobre todo en aquellas ciudades donde los principales focos de emisión sean la quema de combustibles. Aun así, determinadas variables exógenas pueden ser complicadas de controlar, como es el caso de las emisiones provenientes de la industria, ya que no son necesariamente constantes.

Por otro lado, las diferencias de precisión del mismo modelo para determinadas zonas, ya sea por la cercanía a zonas de actividad industrial o por un mayor volumen de tráfico rodado, hace que el desarrollo especifico de diferentes modelos predictivos para regiones más localizadas de la ciudad mejore la precisión y ajuste de los modelos. Además, se podría obtener información sobre que variables influyen en mayor medida a los niveles de concentración, permitiendo el desarrollo de un plan de acción para la reducción de la contaminación en la zona y la localización de puntos de alta calidad del aire en la ciudad que pueden ser de interés para las personas más sensibles, como los niños y personas con determinadas patologías.

Finalmente, hay diferentes estudios publicados en los que se emplean redes neuronales, las cuales no han sido estudiadas en este Trabajo de Fin de Máster, pero están dando buenos resultados en diferentes proyectos, incluido la elaboración de modelos predictivos de la calidad del aire[29], por lo que la aplicación de estos algoritmos podría aportar mayor precisión a las predicciones.

7 BIBLIOGRAFÍA

[1] Mannucci, P. M., Harari, S., Martinelli, I., & Franchini, M. (2015). Effects on health of air pollution: a narrative review. *Internal and emergency medicine*, *10*, 657-662.

[2] https://aiqbe.es/produccion-aiqbe

[3] https://www.juntadeandalucia.es/medioambiente/portal/areas-tematicas/atmosfera/la-calidad-del-aire

[4] https://www.aemet.es/es/eltiempo/observacion/ultimosdatos

[5] https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[6] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. *Frontiers in public health*, *8*, 14.

[7] Cheung, K., Daher, N., Kam, W., Shafer, M. M., Ning, Z., Schauer, J. J., & Sioutas, C. (2011). Spatial and temporal variation of chemical composition and mass closure of ambient coarse particulate matter (PM10–2.5) in the Los Angeles area. *Atmospheric environment*, *45*(16), 2651-2662.

[8] Villányi, V., Turk, B., Batic, F., & Csintalan, Z. (2010). *Ozone pollution and its bioindication* (p. 4622). London: Intech Open.

[9] Richmond-Bryant, J., Chris Owen, R., Graham, S., Snyder, M., McDow, S., Oakes, M., & Kimbrough, S. (2017). Estimation of on-road NO2 concentrations, NO2/NOX ratios, and related roadway gradients from near-road monitoring data. *Air Quality, Atmosphere & Health*, *10*, 611-625.

[10] https://www.juntadeandalucia.es/medioambiente/portal/fuentes-emisoras-antropogenicas

[11] https://www.juntadeandalucia.es/medioambiente/portal/acceso-rediam

[12] BOE-A-2020-10426

[13] Bonaccorso, G. (2017). Machine learning algorithms.

[14] Packt Publishing Ltd.; James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

[15] Luckner, M., Topolski, B., & Mazurek, M. (2017, May). Application of XGBoost algorithm in fingerprinting localisation task. In *IFIP International Conference on Computer Information Systems and Industrial Management* (pp. 661-671). Cham: Springer International Publishing.

[16] Pan, B. (2018, February). Application of XGBoost algorithm in hourly PM2. 5 concentration prediction. In *IOP conference series: earth and environmental science* (Vol. 113, p. 012127). IOP publishing.

[17] https://skforecast.org/0.10.0/user_guides/independent-multi-time-series-forecasting

[18] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*, 1937-1967.

[19] https://xgboost.readthedocs.io/

[20] https://optuna.org/

[21] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

[22] https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[23] Callens, A., Morichon, D., Abadie, S., Delpey, M., & Liquet, B. (2020). Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. *Applied Ocean Research, 104*, 102339.

[24] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

[25] Yu, Z., Zhang, C., Xiong, N., & Chen, F. (2022). A New Random Forest Applied to Heavy Metal Risk Assessment. *Comput. Syst. Sci. Eng.*, *40*(1), 207-221.

[26] Charoen-Ung, P., & Mittrapiyanuruk, P. (2019). Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning. In *Recent Advances in Information and Communication Technology 2018: Proceedings of the 14th International Conference on Computing and Information Technology (IC2IT 2018)* (pp. 33-42). Springer International Publishing.

[27] Marco, R., Ahmad, S. S. S., & Ahmad, S. (2022). Bayesian hyperparameter optimization and Ensemble Learning for Machine Learning Models on software effort estimation. *International Journal of Advanced Computer Science and Applications*, *13*(3).

[28] https://shap.readthedocs.io/en/latest/

[29] a) Kök, İ., Şimşek, M. U., & Özdemir, S. (2017, December). A deep learning model for air quality prediction in smart cities. In *2017 IEEE international conference on big data (big data)* (pp. 1983-1990). IEEE. b) Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, *23*, 22408-22417. c) Navares, R., & Aznarte, J. L. (2020). Predicting air quality with deep learning LSTM: Towards comprehensive models. *Ecological Informatics*, *55*, 101019.

APÉNDICE A

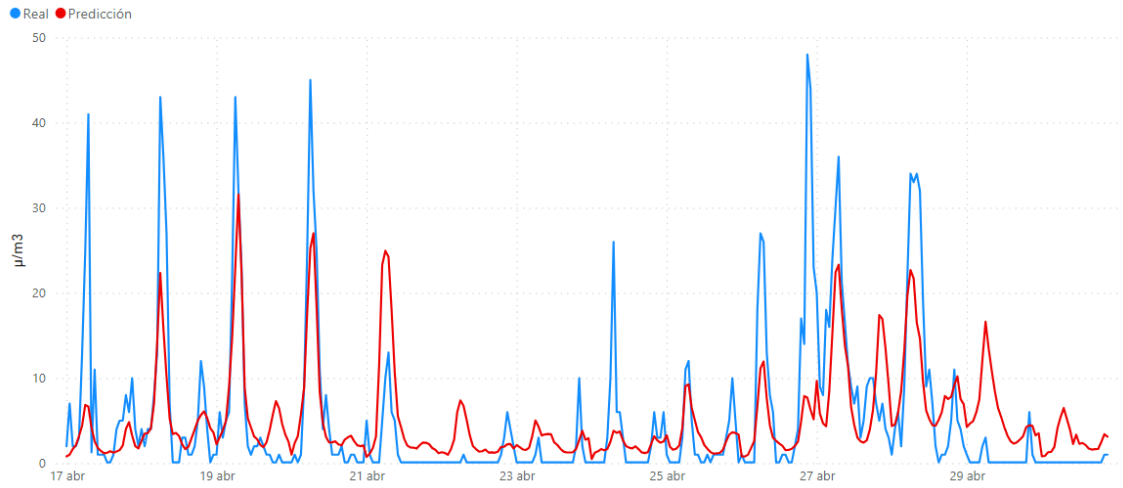

Figura A1. Predicción NO_2 en Campus el Carmen con el modelo **23**.

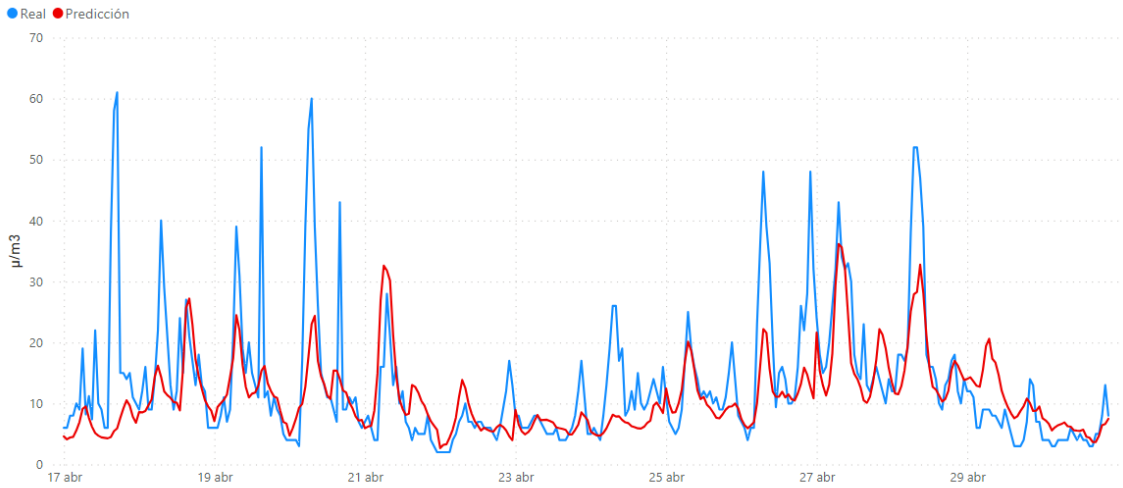

Figura A2. Predicción NO_2 en Marismas del Titán con el modelo **23**.

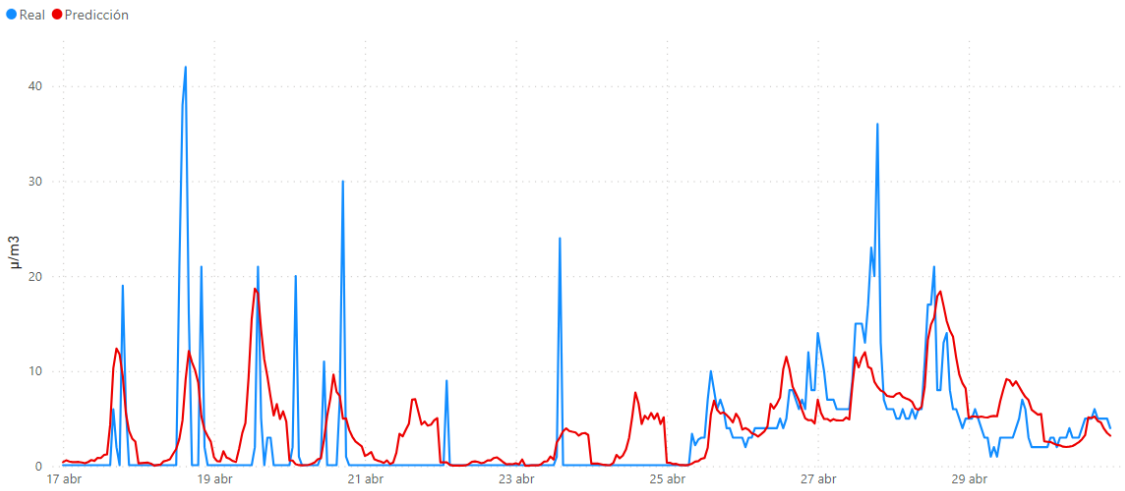

Figura A3. Predicción SO_2 en Los Rosales con el modelo **25**.

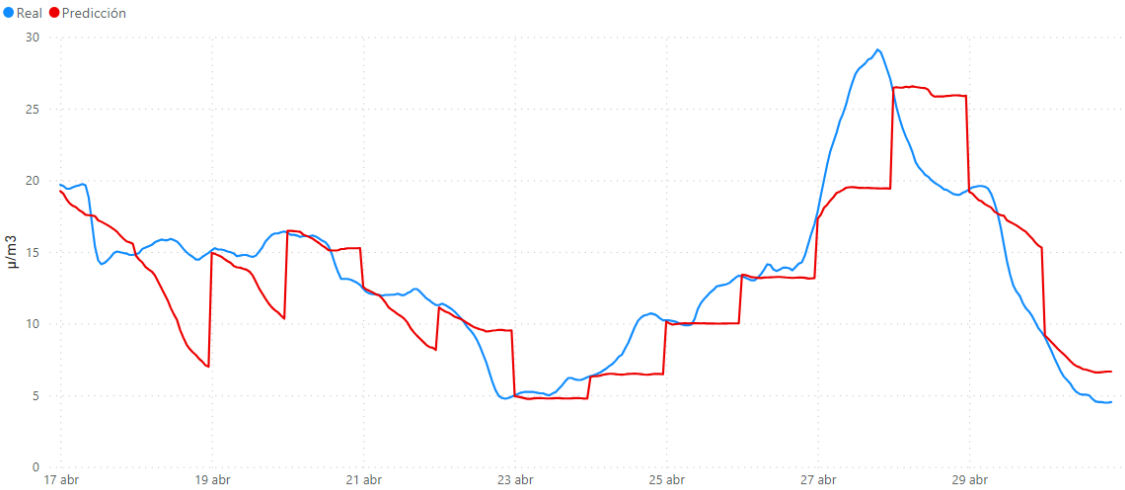

Figura A4. Predicción PM2,5 en Pozo Dulce con el modelo **26**.

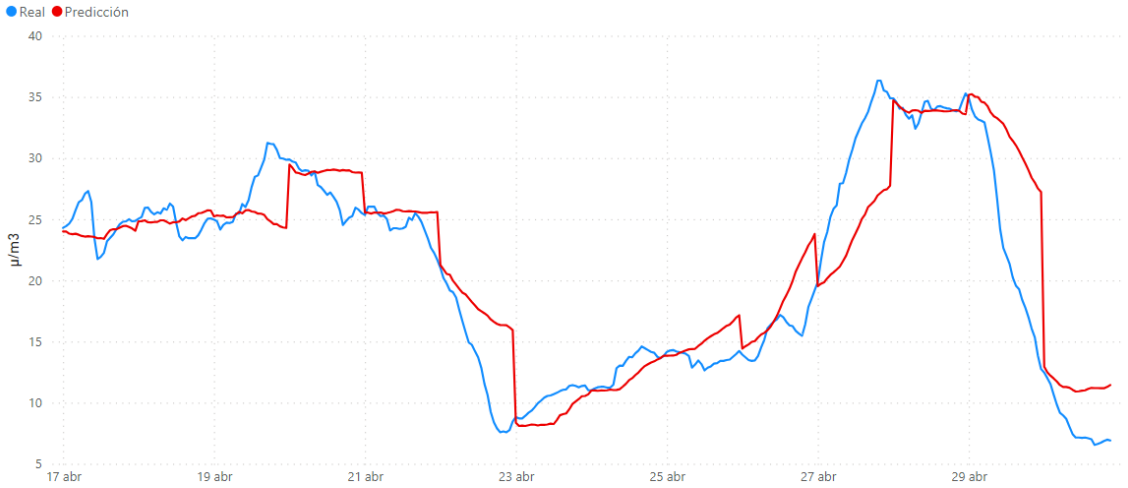

Figura A5. Predicción PM10 en Pozo Dulce con el modelo **28**.

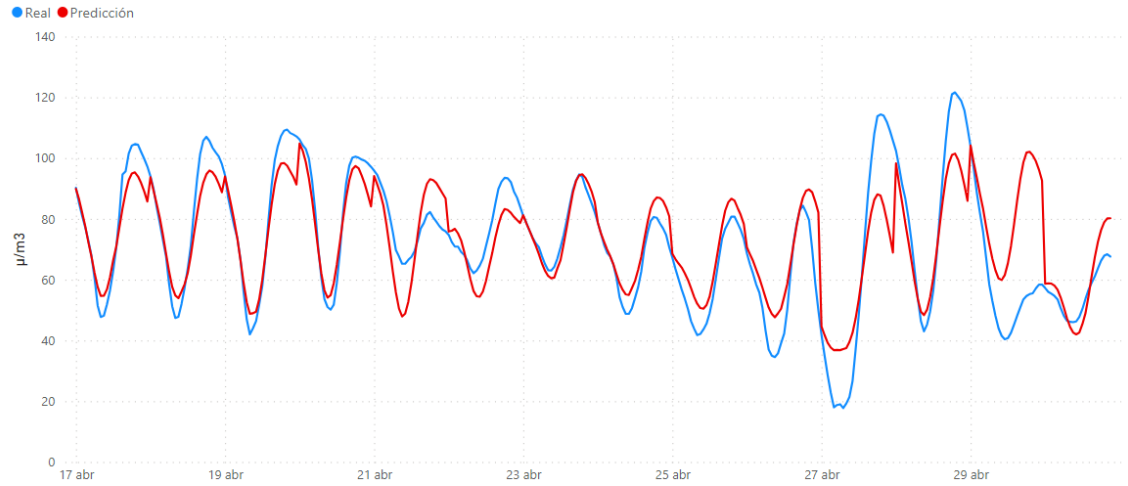

Figura A6. Predicción O_3 en Campus el Carmen con el modelo **30**.

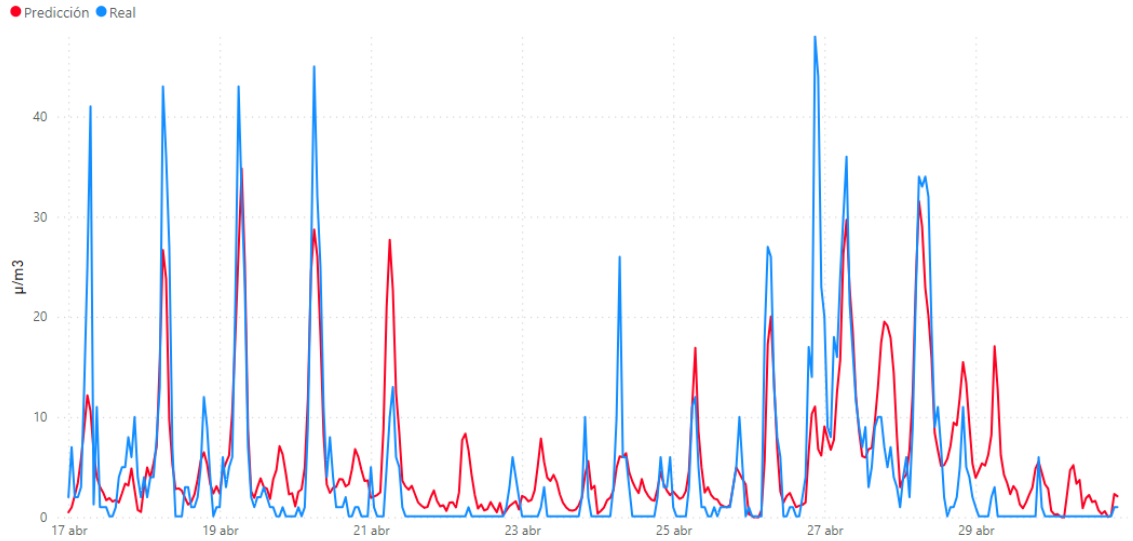

Figura A7. Predicción NO_2 en Campus el Carmen con el modelo **32**.

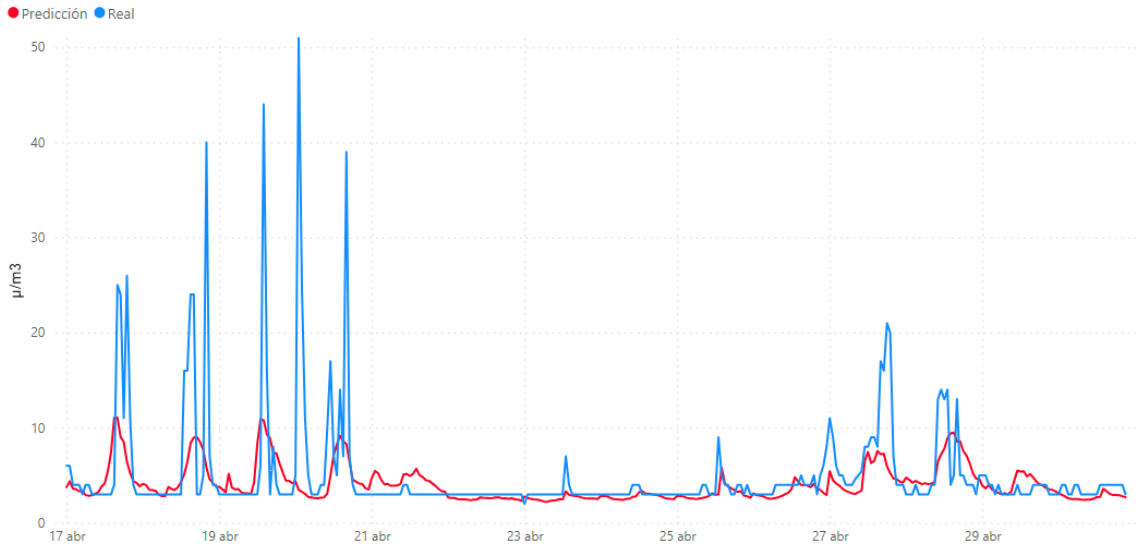

Figura A8. Predicción SO_2 en Campus el Carmen con el modelo **33**.

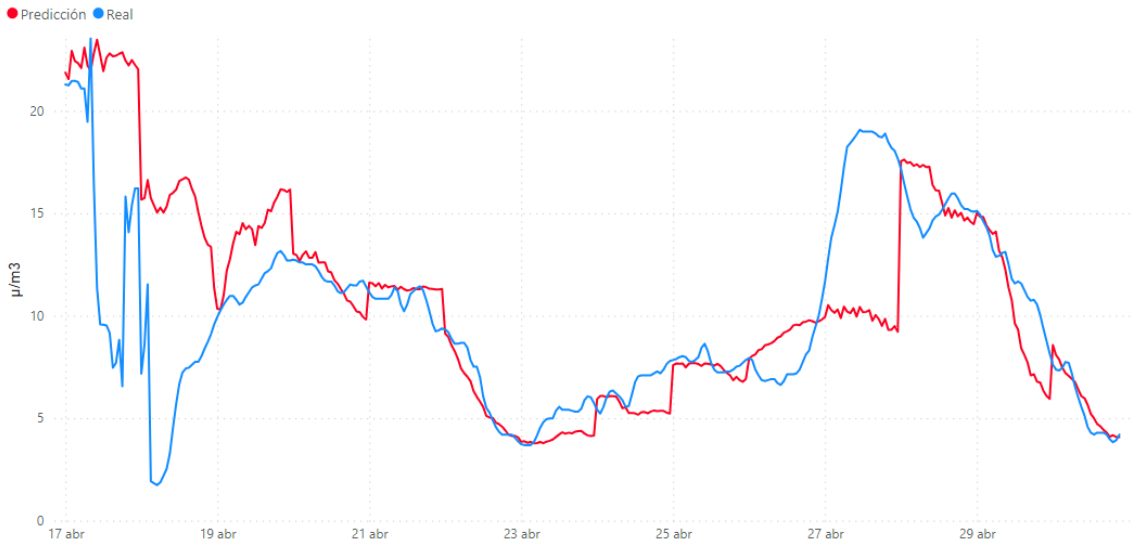

Figura A9. Predicción PM2,5 en Campus el Carmen con el modelo **35**.

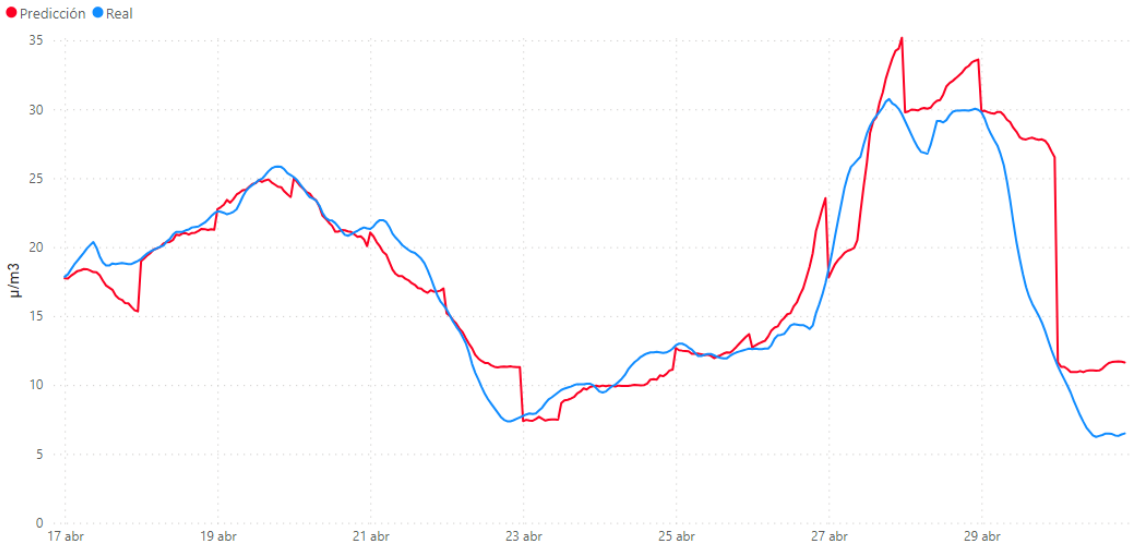

Figura A10. Predicción PM10 en Campus el Carmen con el modelo **38**.

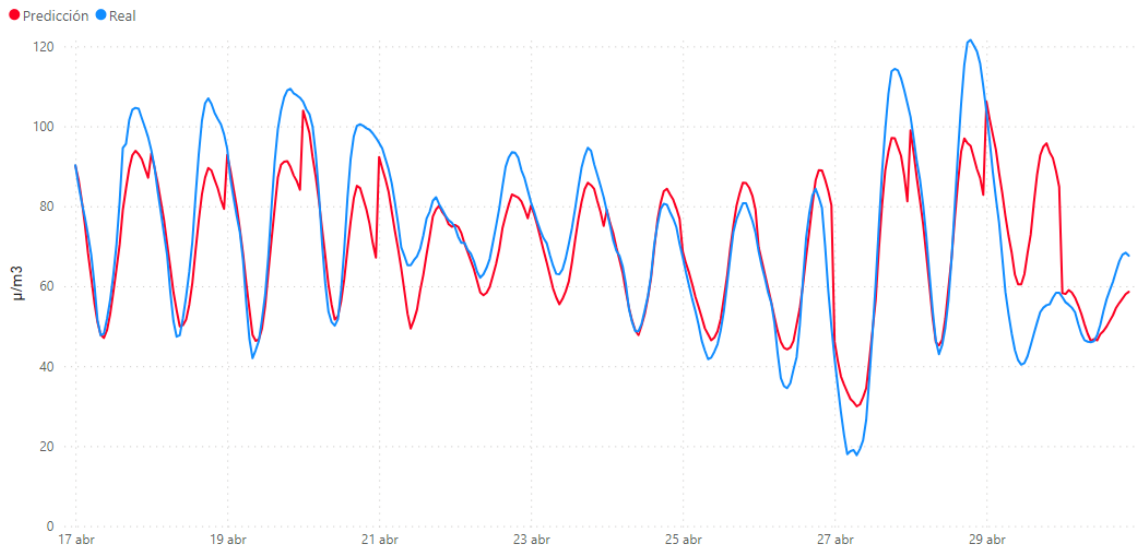

Figura A11. Predicción O_3 en Campus el Carmen con el modelo **39**.

APÉNDICE B

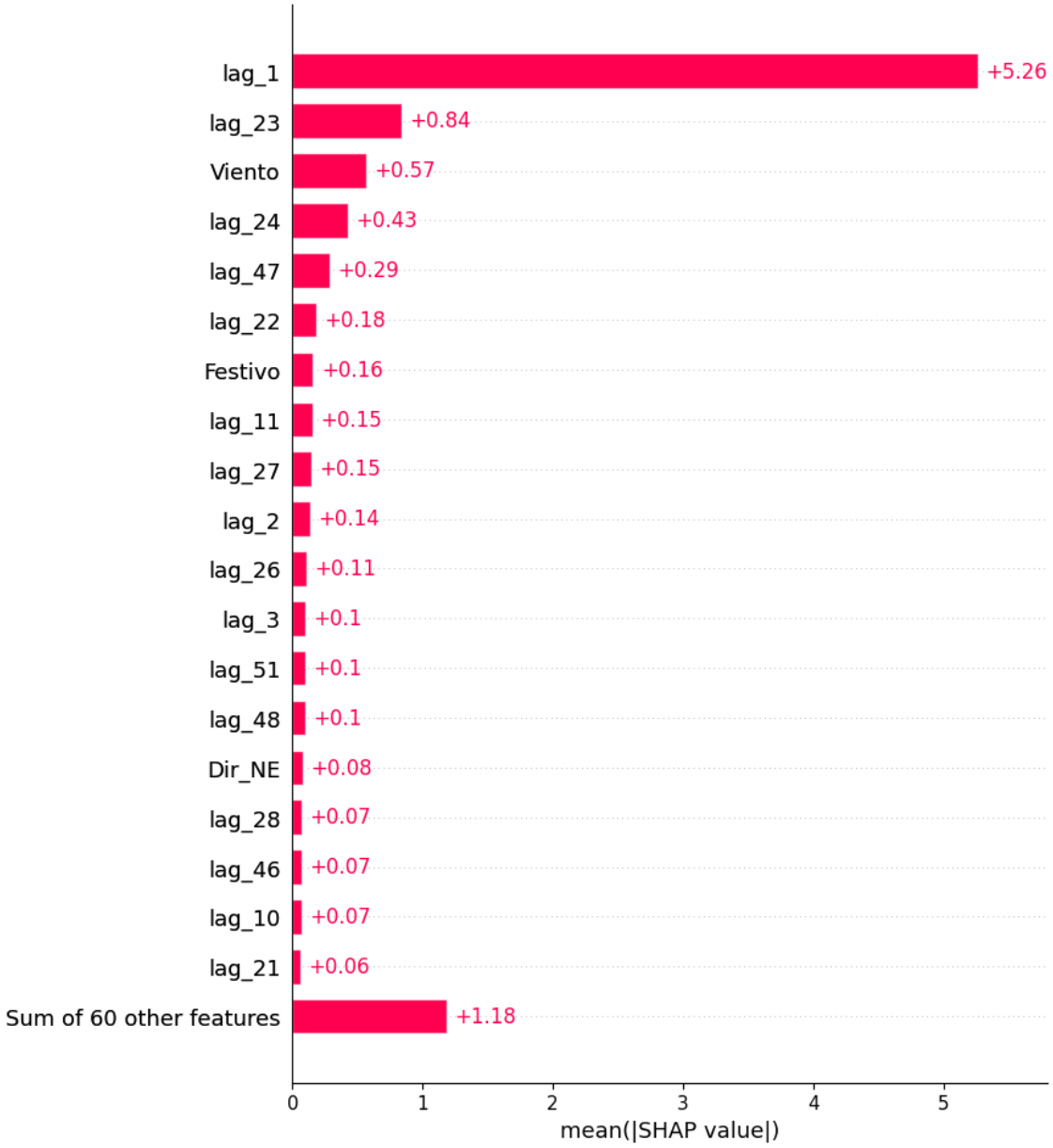

Figura B1. Importancias de variables en modelo **23** de NO_2.

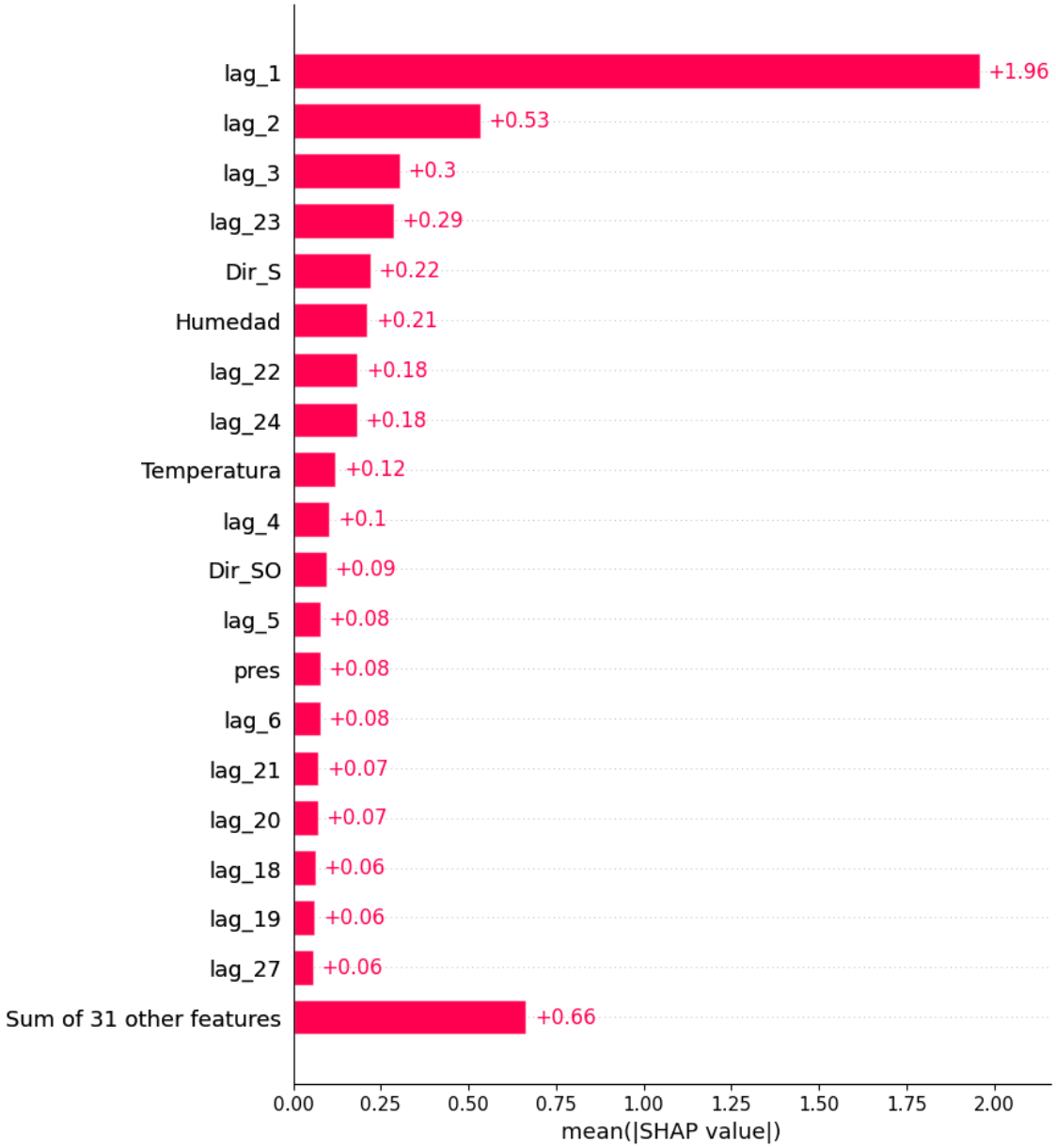

Figura B2. Importancias de variables en modelo **25** de SO_2.

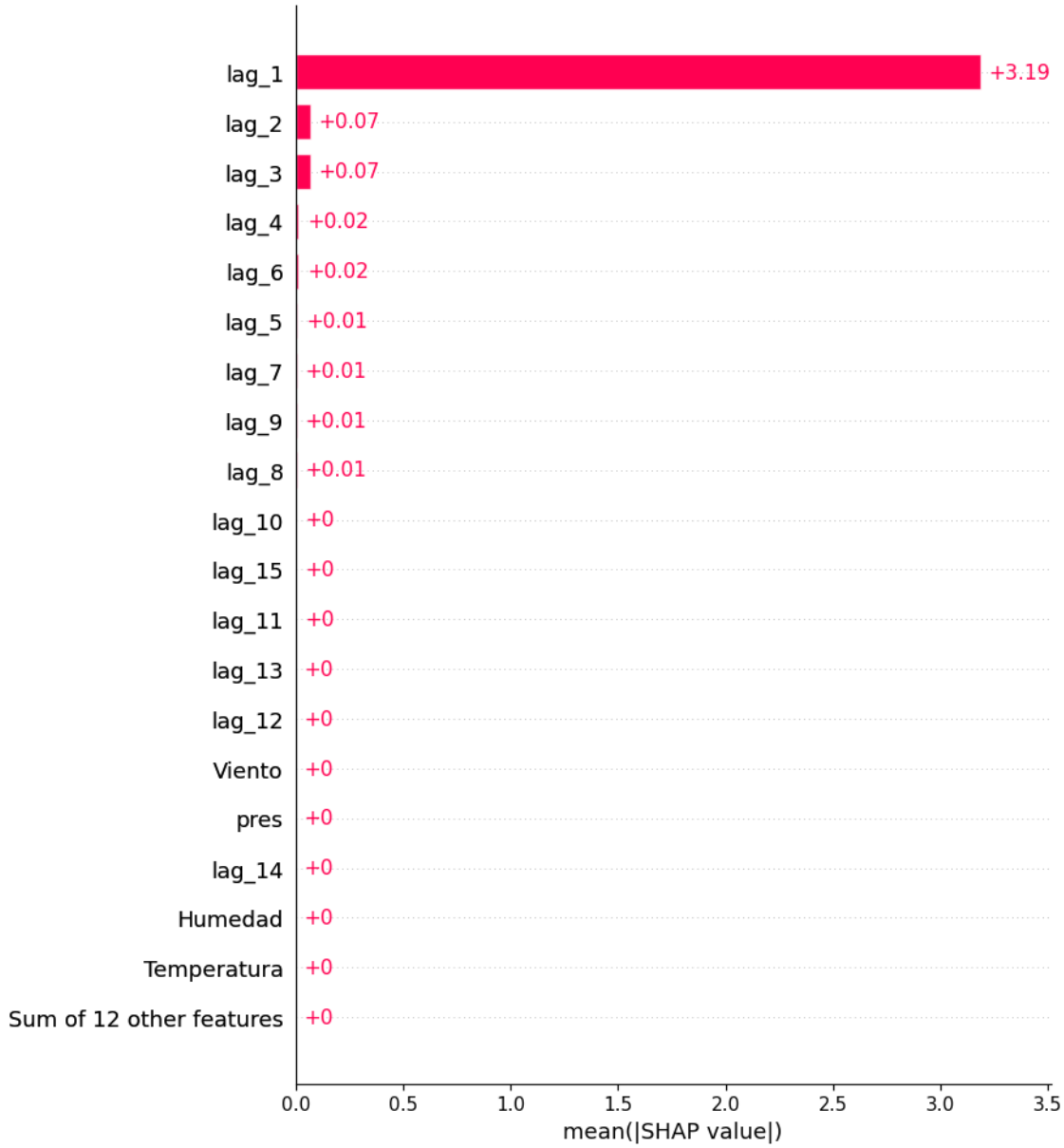

Figura B3. Importancias de variables en modelo **26** de PM2,5.

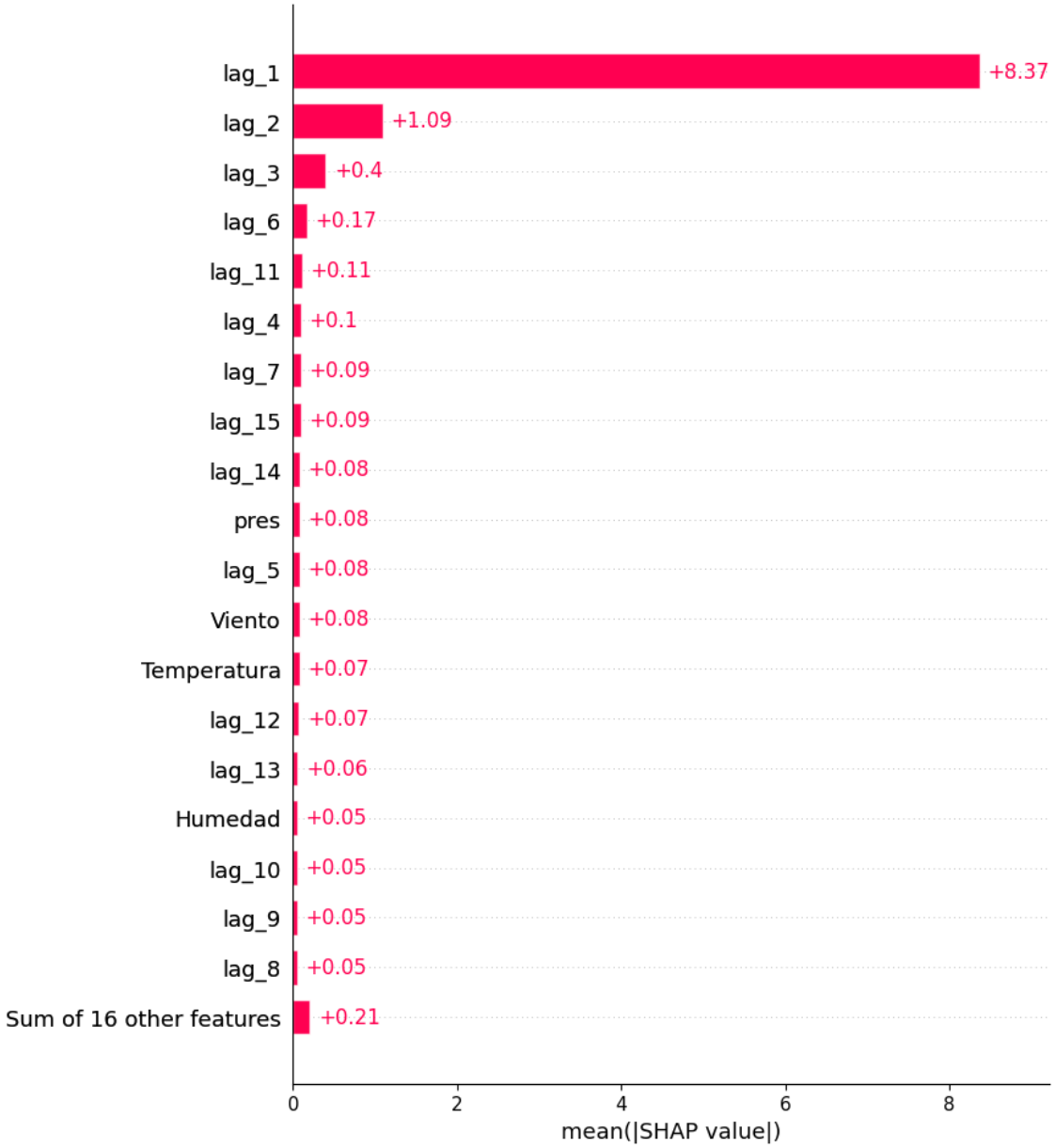

Figura B4. Importancias de variables en modelo **28** de PM10.

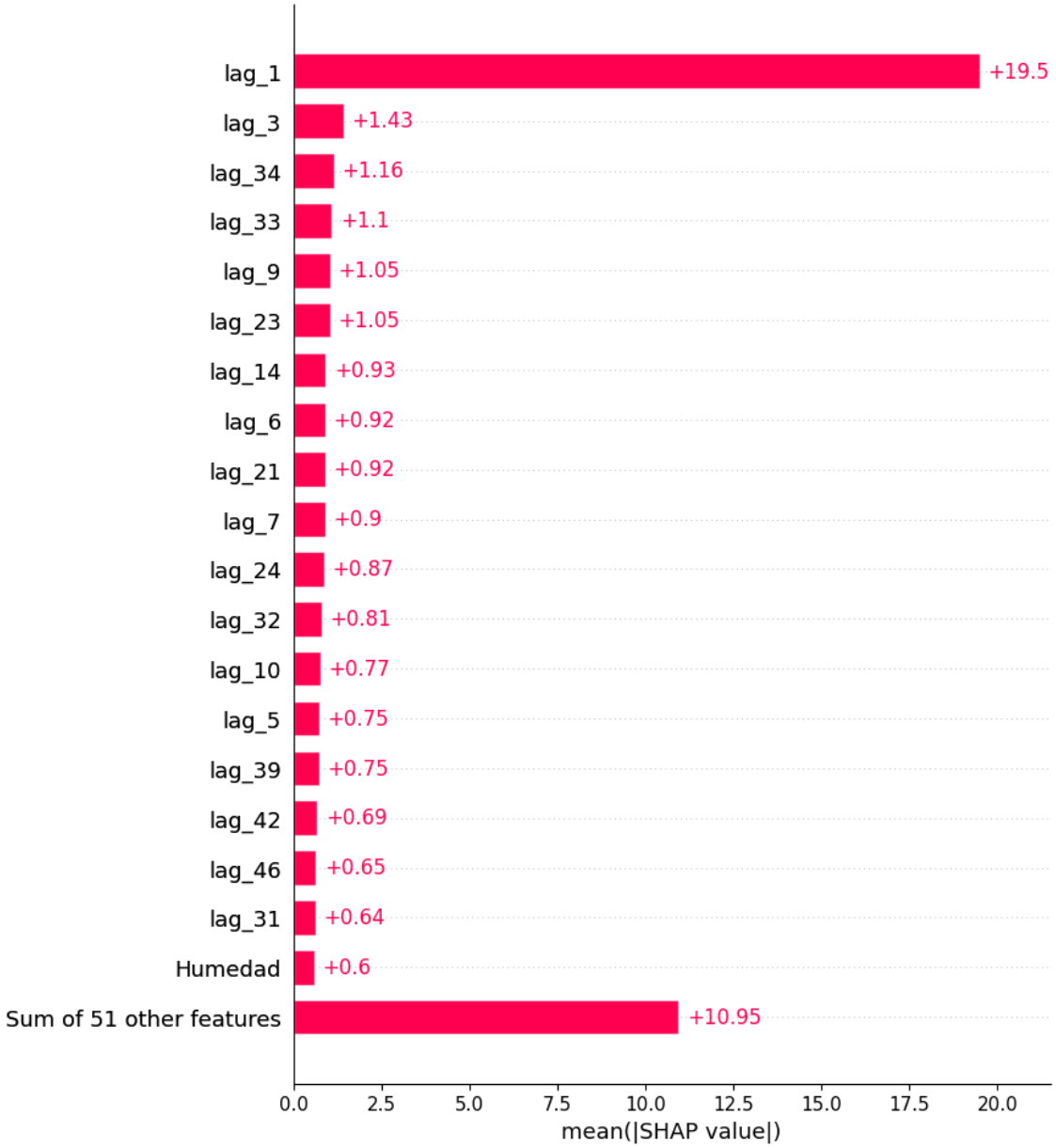

Figura B5. Importancias de variables en modelo **30** de O_3.