Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN

ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

PREDICCIÓN DE LA CONTAMINACIÓN DEL AIRE EN MADRID

NOMBRE:

ANDREA PAOLA IANNELLI

CURSO 2021-2022

TÍTULO: PREDICCIÓN DE LA CONTAMINACIÓN DEL AIRE EN MADRID

AUTOR: ANDREA PAOLA IANNELLI

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

DIRECTOR DEL PROYECTO: DANIEL DOMÍNGUEZ

FECHA: OCTUBRE DE 2022

RESUMEN

La contaminación del aire resulta cada vez un tema más preocupante para la sociedad debido a los efectos negativos que es capaz de producir en la salud de los seres humanos y los animales a corto y largo plazo. La predicción de la polución en el aire permite anticipar la calidad que tendrá el aire en nuestro día a día para tomar las previsiones necesarias requeridas.

El portal de datos abiertos del ayuntamiento de Madrid nos proporciona los datos históricos referentes a distintos contaminantes del aire como monóxido de carbono, dióxido de azufre, monóxido de nitrógeno y partículas en el aire, de igual forma nos proporciona datos históricos meteorológicos, y datos históricos demográficos, como población, número de locales y número de turismos registrados en los distintos barrios de Madrid.

A través del análisis de los datos y la implementación de diversos modelos de regresión se fue capaz de determinar que definitivamente existe una relación entre los datos meteorológicos (temperatura, precipitación, presión) y los datos demográficos con los valores de las distintas moléculas consideradas como contaminantes del aire.

Palabras clave: Predicción, regresión, contaminación del aire.

ABSTRACT

Air pollution is becoming an increasingly worrying issue for society due to the negative effects it can produce on the health of humans and animals in the short and long term. The prediction of air pollution allows us to anticipate the quality that the air will have in our day to day to make the necessary forecasts required.

The open data portal of the Madrid City Council provides us with historical data regarding different air pollutants such as carbon monoxide, sulfur dioxide, nitrogen monoxide and particles in the air, in the same way it provides us with

historical meteorological data, and historical demographics data, such as population, number of premises and number of passenger cars registered in the different neighborhoods of Madrid.

Through the analysis of the data and the implementation of various regression models, it was possible to determine that there is definitely a relationship between the meteorological data (temperature, precipitation, pressure) and the demographic data with the values of the different molecules considered as air pollutants.

Keywords: Prediction, regression, air pollution.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1. INTRODUCCIÓN

1.1 Planteamiento del problema

La calidad y la contaminación en el aire en cualquier ciudad del mundo ha sido un problema cada vez más preocupante para los habitantes de esta, y ser capaces de registrarla, estudiarla y calcularla nos permite obtener registros a través del tiempo y ser capaces de tomar acciones para nuestro día a día y el futuro. La contaminación de aire o atmosférica se define por la proporción de ciertas moléculas o partículas en el aire de un sitio determinado, en donde dichas partículas tienen un efecto negativo en la salud de los seres humanos y animales, siendo unos de los gases contaminantes más conocidos el dióxido de carbono, el monóxido de carbono y el ozono.

Los efectos de una alta proporción de contaminantes en el aire de una zona pueden variar desde enfermedades respiratorias hasta problemas cardiacos y derrames cerebrales, en donde la población vulnerable, como niños y personas de tercera edad, son los más propensos a sufrir las consecuencias de una mala calidad el aire en la zona donde viven. Según la OMS (Organización Mundial de la Salud) cerca de 7 millones de personas mueren cada año a causa de las partículas contenidas en el aire. A causa de esto, la calidad del aire se ha convertido en una de las mayores preocupaciones de la salud pública a nivel mundial como se observa en múltiples ciudades de China, India y Pakistán, en donde la contaminación de la atmosfera se ha convertido en un tema preocupante para distintas organizaciones como la OMS y la ONU (Organización de las Naciones Unidas).

Existen múltiples factores que perjudican la calidad del aire, en donde los más evidentes son aquellos producidos por acciones humanas como la agricultura y ganadería masiva, actividades industriales de cualquier tipo (especialmente aquellas relacionadas con la quema de combustibles fósiles), incorrecta gestión de los residuos, entre otras. Sin embargo, los factores climatológicos como la temperatura, la precipitación o la humedad, tiene, un gran efecto, el cual puede ser negativo o positivo, en las proporciones de los

contaminantes presentes en el aire. Este se debe a que dichos factores son capaces de generar reacciones químicas en la atmosfera capaces de influenciar en la calidad del aire.

1.2 Objetivos del proyecto

- Obtención de los datos históricos de la ciudad de Madrid relacionados con la calidad del aire, datos climatológicos y datos demográficos.
- Procesamiento y visualización de los datos para una compresión profunda de las correlaciones existentes entre las variables meteorológicas y demográficas, y los distintos contaminantes en la ciudad de Madrid.
- Implementación y mejora de diversos algoritmos de Machine Learning para la creación de un modelo de predicción de las diversas moléculas contaminantes a partir de los datos anteriormente mencionados.
- Análisis e interpretación de los resultados obtenidos a través del modelo de predicción.

Capítulo 2. OBTENCIÓN Y PROCESAMIENTO DE LOS DATOS

En primer lugar, los datos utilizados para la obtención del modelo de predicción fueron obtenidos en su mayoría del portal de datos abiertos del ayuntamiento de Madrid. Dicho portal lo describen en su página oficial (Ayuntamiento de Madrid, 2022) como “una iniciativa global, ligada a las políticas de Gobierno Abierto, que persigue que los datos y la información, especialmente las que poseen las administraciones públicas, se publiquen de forma abierta, regular y reutilizable para todo el mundo, sin restricciones de acceso, copyright, patentes u otros mecanismos de control. La filosofía origen de estas iniciativas es fomentar la transparencia, la eficiencia, la participación ciudadana y el desarrollo económico”.

2.1 Datos meteorológicos de Madrid

Los datos meteorológicos disponibles en el portal anteriormente mencionados corresponden a los datos de los años 2019, 2020, 2021 y 2022 (hasta el mes de Julio), dichos datos están disponibles en dos formatos, registros por hora y por día. En el caso de este trabajo, se tomó la decisión de trabajar con los registros diarios para disminuir el coste computacional que implicaba trabajar con los datos horarios.

A continuación, se muestran la estructura de los datos meteorológicos, la cual se mantienen para todos los años antes mencionados.

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	D01	V01	D02	V02
28	79	4	1	28079004_82_98	2019	1	18	V	20	V

Tabla 2.1 Estructura de los datos meteorológicos de Madrid

Los datos correspondientes a la provincia y municipio serán siempre 28 y 79 respectivamente, debido a que los datos trabajados corresponden a medidas tomadas solamente en la comunidad autónoma de Madrid y el municipio Madrid. “ESTACION” y “PUNTO_MUESTREO” corresponden al centro en donde se

midió el dato climatológico identificado en la columna “MAGNITUD”. Por último, tendremos tantos pares de columnas como días en el mes plasmado en la columna “MES”, en donde la columna “Dn” corresponde al valor del parámetro medido (temperatura, presión barométrica, radiación solar, etc.) en su dimensión correspondiente, por otro lado, la columna “Vn”, la cual es una variable categórica que representa si el valor obtenido ha sido validado o no.

2.1.1 Datos meteorológicos medidos

El portal de datos en donde se obtuvieron los datos facilita una tabla en donde se puede conocer el dato climatológico medido señalado en la variable categórica nominal “MAGNITUD” a través de valores numéricos, de la misma manera, señalan la dimensión de medida correspondiente y el método de medición de cada parámetro.

CÓDIGO	PARÁMETRO	UNIDAD DE MEDIDA	TÉCNICA DE MEDIDA
80	RADIACIÓN ULTRAVIOLETA	Mw/m2	98
81	VELOCIDAD VIENTO	m/s	98
82	DIR. DE VIENTO	-	98
83	TEMPERATURA	ºC	98
86	HUMEDAD RELATIVA	%	98
87	PRESION BARIOMETRICA	mb	98
88	RADIACION SOLAR	W/m2	98
89	PRECIPITACIÓN	l/m2	98

Tabla 2.2 Parámetros meteorológicos medidos

2.1.2 Estaciones meteorológicas

De igual forma, el portal de datos abiertos del ayuntamiento de Madrid proporciona todos los datos necesarios con respecto a las estaciones meteorológicas encargadas de medir los datos climatológicos. A continuación, se muestra la estructura de datos de la tabla con dichos datos.

CÓDIGO
CÓDIGO_CORTO
ESTACIÓN
DIRECCION
LONGITUD_ETRS89
LATITUD_ETRS89
ALTITUD
VV (81)
DV (82)
T (83)
HR (86)
PB (87)
RS (88)
P (89)
COD_VIA
VIA_CLASE
VIA_PAR
VIA_NOMBRE
NUM_VIA
COORDENADA_X_ETRS89
COORDENADA_Y_ETRS89
LONGITUD
LATITUD

Tabla 2.3 Estructura de datos de las estaciones meteorológicas

La columna “CODIGO_CORTO” corresponde a la columna “ESTACION” en la tabla 2.1, de la misma forma, los números en las columnas “VV”, “DV”, “T”, “HR”, “PB”, “RS”, “P” corresponden con los números en la tabla 2.2, que muestran cuales parámetros se miden en cada una de las estaciones.

2.2 Datos de la calidad del aire en Madrid

Los datos que hacen referencia a la calidad del aire siguen la misma estructura que los datos meteorológicos, en este caso, hay datos disponibles a partir del año 2001, sin embargo, debido a que el ayuntamiento de Madrid no tiene registro de los datos meteorológicos antes del 2019 se trabajara solamente con los datos correspondientes a la calidad del aire a partir del 2019 como se señaló anteriormente.

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANO	MES	D01	V01	D02	V02
28	79	4	1	28079004_1_38	2019	1	18	V	20	V

Tabla 2.4 Estructura de los datos de la calidad del aire en Madrid

2.2.1 Magnitudes medidas

De la misma manera, se nos facilita una tabla con todos aquellos "contaminantes" mesurados que corresponden a la variable categórica "MAGNITUD".

Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO_2	µg/m ³	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m³	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m³	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO_2	µg/m ³	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m³	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m³	47	Id.
12	Óxidos de Nitrógeno	NOx	µg/m³	08	Quimioluminiscencia
14	Ozono	O_3	µg/m ³	06	Absorción ultravioleta
20	Tolueno	TOL	µg/m³	59	Cromatografía de gases
30	Benceno	BEN	µg/m³	59	Id.
35	Etilbenceno	EBE	µg/m³	59	Id.
37	Metaxileno	MXY	µg/m³	59	Id.
38	Paraxileno	PXY	µg/m³	59	Id.
39	Ortoxileno	OXY	µg/m³	59	Id.
42	Hidrocarburos totales (hexano)	TCH	mg/m³	02	Ionización de llama
43	Metano	CH4	mg/m³	02	Id.
44	Hidrocarburos no metánicos (hexano)	NMHC	mg/m³	02	Id.

Tabla 2.5 Magnitudes de la calidad del aire medidos

2.2.2 Estaciones encargadas de medir la calidad del aire

Es importante mencionar que la estaciones en donde se toman los datos meteorológicos y las estaciones donde se miden los "contaminantes" no son las mismas en todos los casos.

CODIGO
CODIGO_CORTO
ESTACION
DIRECCION
LONGITUD_ETRS89
LATITUD_ETRS89
ALTITUD
COD_TIPO
NOM_TIPO
NO2
SO2
CO
PM10
PM2_5
O3
BTX
COD_VIA
VIA_CLASE
VIA_PAR
VIA_NOMBRE
Fecha alta
COORDENADA_X_ETRS89
COORDENADA_Y_ETRS89
LONGITUD
LATITUD

Tabla 2.6 Estructura de datos de las estaciones para la medición de la calidad de aire en Madrid

2.3 Datos sobre los distintos barrios de Madrid

Al igual que los datos ya mencionados, los datos acerca de los distintos barrios de también fueron obtenidos a través del portal de datos abiertos del ayuntamiento de Madrid, sin embargo, los datos requeridos se encontraban en distintos ficheros y al cruzar las distintas tablas se obtuvo la siguiente estructura:

id
Año
Distrito
Barrio
Poblacion
Area HA
Densidad Poblacional
Turismos
Turismos por persona
Locales
Densidad locales

Tabla 2.7 Estructura de datos acerca de los barrios de Madrid

Con respecto a la columna “Turismos” hace referencia al número de turismos registrados en el barrio y en cuanto a la columna “Locales” se refiere al total de locales en el mismo (incluye locales a pie de calle y locales agrupados)

El dato de la columna “Densidad Poblacional” se calculó a partir de la población y el área en hectáreas, dicho dato fue calculado ya que es un dato más certero en cuanto a la manera en cómo se distribuye la población en la ciudad de Madrid y resulta más fácil comparar barrios de distintos tamaños entre sí, de igual forma se obtuvieron las columnas “Densidad locales”, la cual indica la densidad de locales en un barrio determinado y “Turismos por persona”, la cual fue obtenida mediante la relación entre número de turismos registrados en un barrio y la población del mismo.

2.4 Procesamiento de los datos

Para el correcto análisis y entrenamiento del modelo fue necesario el procesamiento de los datos para cambiar la estructura original de los mismos, el cruce de estos, y el tratamiento de los valores nulos. El procesamiento de los datos fue realizado en su mayoría a través de un jupyter notebook con Python 3 y la librería Pandas.

2.4.1 Reestructuración de los datos

La estructura original de los datos señalada en los puntos anteriores, en donde cada instancia representa todos los valores obtenidos para cada día de un mes para una misma molécula o dato climatológico en una de las estaciones, recordando que cada día viene acompañado con una columna que muestra si el dato fue validado o no. Para el correcto entrenamiento del modelo se requiere una estructura en donde cada instancia muestra todos los datos disponibles (contaminantes, climatológicos y demográficos) para una misma estación y un mismo día, para esto se definió la siguiente función que permite obtener la estructura deseada, filtrar los valores no validados y obtener una columna con la fecha de la toma de los datos.

```
def Reshape(df):
    for i,row in df.iterrows():
        df.loc[i,'01'] = str(row['D01']) + row['V01']
        df.loc[i,'02'] = str(row['D02']) + row['V02']
        df.loc[i,'03'] = str(row['D03']) + row['V03']
        df.loc[i,'04'] = str(row['D04']) + row['V04']
        df.loc[i,'05'] = str(row['D05']) + row['V05']
        df.loc[i,'06'] = str(row['D06']) + row['V06']
        df.loc[i,'07'] = str(row['D07']) + row['V07']
        df.loc[i,'08'] = str(row['D08']) + row['V08']
        df.loc[i,'09'] = str(row['D09']) + row['V09']
        df.loc[i,'10'] = str(row['D10']) + row['V10']
        df.loc[i,'11'] = str(row['D11']) + row['V11']
        df.loc[i,'12'] = str(row['D12']) + row['V12']
        df.loc[i,'13'] = str(row['D13']) + row['V13']
        df.loc[i,'14'] = str(row['D14']) + row['V14']
        df.loc[i,'15'] = str(row['D15']) + row['V15']
        df.loc[i,'16'] = str(row['D16']) + row['V16']
        df.loc[i,'17'] = str(row['D17']) + row['V17']
        df.loc[i,'18'] = str(row['D18']) + row['V18']
        df.loc[i,'19'] = str(row['D19']) + row['V19']
        df.loc[i,'20'] = str(row['D20']) + row['V20']
        df.loc[i,'21'] = str(row['D21']) + row['V21']
        df.loc[i,'22'] = str(row['D22']) + row['V22']
        df.loc[i,'23'] = str(row['D23']) + row['V23']
        df.loc[i,'24'] = str(row['D24']) + row['V24']
        df.loc[i,'25'] = str(row['D25']) + row['V25']
        df.loc[i,'26'] = str(row['D26']) + row['V26']
        df.loc[i,'27'] = str(row['D27']) + row['V27']
        df.loc[i,'28'] = str(row['D28']) + row['V28']
        df.loc[i,'29'] = str(row['D29']) + row['V29']
        df.loc[i,'30'] = str(row['D30']) + row['V30']
        df.loc[i,'31'] = str(row['D31']) + row['V31']

    df = df.drop(['D01', 'V01', 'D02', 'V02', 'D03', 'V03', 'D04',
       'V04', 'D05', 'V05', 'D06', 'V06', 'D07', 'V07', 'D08', 'V08', 'D09',
       'V09', 'D10', 'V10', 'D11', 'V11', 'D12', 'V12', 'D13', 'V13', 'D14',
       'V14', 'D15', 'V15', 'D16', 'V16', 'D17', 'V17', 'D18', 'V18', 'D19',
       'V19', 'D20', 'V20', 'D21', 'V21', 'D22', 'V22', 'D23', 'V23', 'D24',
       'V24', 'D25', 'V25', 'D26', 'V26', 'D27', 'V27', 'D28', 'V28', 'D29', 'V29'
        , 'D30', 'V30', 'D31', 'V31'],axis = 1)

    df = pd.melt(df, id_vars=['PROVINCIA', 'MUNICIPIO', 'ESTACION', 'MAGNITUD', 'PUNTO_MUESTREO',
       'ANO', 'MES'],
        var_name='DIA',
        value_name='valor')

    df = df[~((df['MES'].isin([4,6,9,11])) & (df['DIA']==str(31)))]
    df = df[~((df['MES'] == 2)& (df['DIA'].isin([str(29),str(30),str(31)])) & (df['ANO']==2019))]
    df = df[~((df['MES'] == 2) & (df['DIA'].isin([str(30),str(31)])) & (df['ANO']==2020))]
    df = df[~((df['MES'] == 2) & (df['DIA'].isin([str(29),str(30),str(31)])) & (df['ANO']==2021))]
    df = df[~((df['MES'] == 2) & (df['DIA'].isin([str(29),str(30),str(31)])) & (df['ANO']==2022))]

    df.rename(columns = {'ANO':'year', 'MES':'month', 'DIA':'day'}, inplace = True)
    df['year'] = df['year'].astype(int)
    df['month'] = df['month'].astype(int)
    df['day'] = df['day'].astype(int)

    df['Datetime'] = ''
    for i,row in df.iterrows():
        df.at[i,'Datetime'] = datetime.date(year = row['year'], day = row['day'], month = row['month'])

    for i,row in df.iterrows():
        df.loc[i,'validacion'] = row['valor'][len(row['valor'])-1]
        df.loc[i,'valor'] = row['valor'][0:len(row['valor'])-1]

    df = df[df['validacion'] == 'V']
    df['valor'] = df['valor'].astype('float')
    for i,row in df.iterrows():
        df.loc[i,'id'] = str(row['Datetime']) + '-' + str(row['ESTACION'])

    return df
```

Ilustración 1. Función encargada de reestructurar los datos

Esta función fua aplicada para todos los ficheros de cada año, tanto de los datos meteorológicos como los datos de la calidad del aire. A través de esta función obtuvimos una estructura en donde cada instancia representa un valor (molécula contaminante o data climatológico) medido en una estación en una fecha dada, para luego aplicar la función *'pivot_table'* que nos permite agrupar los datos con la estructura deseada, obteniendo así una tabla en donde cada

instancia constituye los datos meteorológicos o los datos de la calidad del aire para una misma estación y un mismo día.

```
df_pivot = df.pivot_table('valor', ['ESTACION','Datetime'], 'MAGNITUD')
```

Ilustración 2. Funcion 'pivot_table' de la librería Pandas

Finalmente, se procedió a unir las distintas tablas de los distintos años obteniendo así una sola tabla con los datos meteorológicos y otra con los datos de la calidad del aire para facilitar el manejo de estas.

2.4.2 Cruce de datos meteorológicos y datos de la calidad del aire

Las estaciones encargadas de tomar los datos meteorológicos y las encargadas de medir los distintos contaminantes que determinan la calidad del aire no son siempre las mismas ni están ubicadas en los mismos lugares en todos los casos. Es por esto, por lo que fue necesario determinar cuál estación (para la medición de la calidad del aire) se encuentra más cercana de cada una de las estaciones meteorológicas. Dicho proceso se realizó de manera manual,

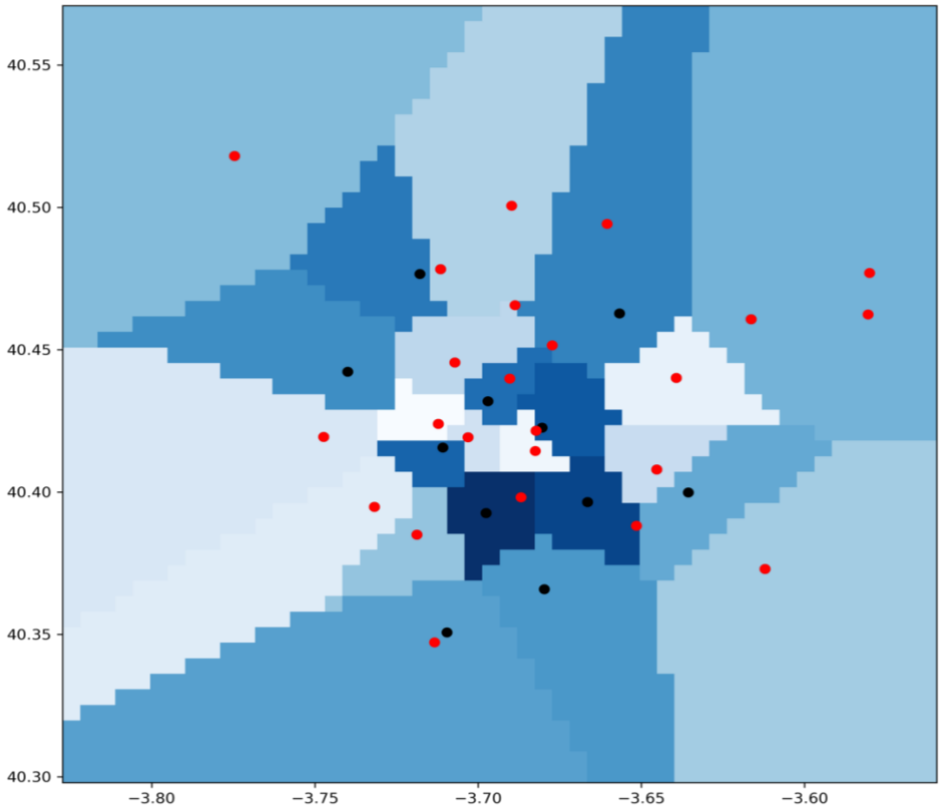

Ilustración 3. Mapa de la estaciones metereologicas de Madrid.

determinando en primer lugar en que barrio de Madrid se ubican las estaciones meteorológicas y posteriormente determinar cuáles estaciones encargadas de medir los contaminantes se encontraba más cerca de esta a través de la siguiente ilustración, en donde los puntos negros representan estaciones meteorológicas y los puntos rojos representan estaciones encargadas de medir la calidad del aire.

Una vez que se han determinado cuales fueron las estaciones cercanas entre sí, se pudo realizar el cruce de los datos obteniendo una tabla única en donde cada instancia representa todos los datos meteorológicos y moléculas contaminantes disponibles para un mismo día y una misma estación con ubicación especifica.

2.4.3 Cruce de datos (Barrios de Madrid)

Los datos demográficos disponibles de los barrios de Madrid señalados en los puntos anteriores se muestran para cada uno de los barrios en distintos años (en este caso del 2019 al 2022), dichos datos se encuentran presentes en la tabla obtenida anteriormente y fueron los utilizados para realizar el cruce de datos.

2.4.4 Tratamiento de valores nulos

Dado que las estaciones no miden todos los mismos datos nos encontramos con un dataset con un gran número de valores nulos. En este caso el eliminar todas aquellas columnas con una gran parte de valores faltantes o eliminar filas con valores faltantes conllevaría a la eliminación de una gran parte de los datos.

Por consiguiente, se decidió imputar los datos faltantes con la media del valor faltante medido en las otras estaciones en el mismo día. Dicha imputación se realizó a través de la siguiente función a las columnas correspondientes a las moléculas contaminantes y datos climatológicos.

```
def fill_na(row,col):
    if math.isnan(row[col]):
        value = df_pivot[df_pivot['Datetime'] == row['Datetime']][col].mean()
    else:
        value = row[col]
    return value
```

Ilustración 4. Función para la imputación de los datos

Una vez realizada la imputación, las columnas correspondientes a la radiación ultravioleta, hidrocarburos totales, hidrocarburos no metánicos y metano continuaban presentando valores nulos, esto es debido a que para ciertos días no fueron tomados en ninguna de las estaciones datos correspondientes a estas variables.

2.5 Tratamiento de Outliers

Analizando las distribuciones de las variables se encontraron valores e instancias que se podrían considerar como *outliers*, sin embargo, se tomó la decisión de no eliminar dichas instancias, ya que las mismas pueden representar valores extremos registrados, recordando que todos los valores no validados fueron eliminados.

Aunque en la documentación correspondiente a lo datos utilizados no se menciona las razones por la cual un dato se considera 'No validado', podríamos suponer que se deben a errores en las en los instrumentos o métodos de medición o condiciones climáticas extremas que pudieran haber afectado a los equipos.

2.6 Almacenamiento de los datos

Debido al pequeño tamaño de los ficheros con los que se trabajan (9.4 MB una vez procesados) y el uso de estos, no resulta pertinente el empleo de una base de datos para la lectura de los datos. Es por esto por lo que el almacenamiento de los ficheros se realizó en local.

Capítulo 3. ANALISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos representa un paso fundamental antes de crear cualquier modelo predictivo ya que nos permite conocer con profundidad los datos disponibles con los que estamos trabajando. Este proceso ayuda a identificar patrones y tendencias, anomalías y verificar suposiciones previamente planteadas a través de visualizaciones y estadísticos que facilitan el análisis.

3.1 Las variables

Primeramente, resulta pertinente analizar, estudiar y entender cada una de las columnas con las que se está trabajando. A continuación, se muestra una tabla en donde se puede visualizar una breve descripción, la unidad de medida (si aplica) y el tipo de variable de cada de las columnas. De igual forma, resulta pertinente destacar que se trabajara con un data set de 3587 instancias.

Nombre de la columna	Descripción	Unidad	Tipo de variable
CODIGO_CORTO	Código que identifica a la estación	N/A	Variable categórica nominal
Datetime	Fecha de la toma de los datos	N/A	Variable númerica discreta
Año	Año de donde se tomo el dato	N/A	Variable categórica ordinal
Barrio	Barrio en donde se ubica la estación	N/A	Variable categórica nominal
Poblacion	Poblacion del barrio	N/A	Variable númerica discreta
Area HA	Area del barrio	Ha (Hectareas)	Variable númerica continua
Densidad Poblacional	Densidad poblacional del barrio	población/ha	
Turismos	Número de turismos registrados en el barrio	N/A	Variable númerica discreta
Turismos por persona	Número de turismos por persona registrados en el barrio	Turismo/persona	Variable númerica continua
Locales	Número de locales en el barrio	N/A	Variable númerica discreta
Densidad locales	Densidad de los locales en el barrio	Locales/ha	Variable númerica continua
RU	Radiación Ultravioleta	Mw/m2	
VV	Velocidad del viento	m/s	
DV	Dirección del viento	N/A	
TEMP	Temperatura	ºC	
HR	Humedad Relativa	%	
PB	Presión bariometrica	mb	
RS	Radiación solar	W/m2	
PREC	Precipitación	l/m2	
NO2	Dioxido de nitrogeno	µg/m 3	
SO2	Dioxido de azufre	µg/m 3	
CO	Monoxido de carbono	mg/m3	
PM_10	Particulas menores a 10 mm	µg/m 3	
PM_25	Particulas menores a 2,5 mm	µg/m 3	
O3	Ozono	µg/m 3	
NO	Monóxido de Nitrógeno	µg/m3	
NOx	Óxidos de Nitrógeno	µg/m 3	
TOL	Tolueno	µg/m 3	
BEN	Benceno	µg/m 3	
EBE	Etilbenceno	µg/m 3	
TCH	Hidrocarburos totales (hexano)	mg/m3	
CH4	Metano	mg/m3	
NMHC	Hidrocarburos no metánicos (hexano)	mg/m3	

Ilustración 5. Descipción de las variables de entrada del modelo

3.2 Valores nulos

Luego del procesamiento de los datos relatado en el capítulo anterior, se fue capaz de permutar aquellos valores nulos, sin embargo, luego de dicha permutación se siguen manteniendo cuatro filas con valores nulos, estas columnas son: RU, TCH, CH4, NMHC. Como se había mencionado, estos valores nulos se deben a que existían múltiples días en donde no se midieron valores para estas variables. Como consecuencia de esto, no se considerarán estas cuatro variables para el entrenamiento del modelo.

3.3 Correlación entre las variables

Los coeficientes de correlación existentes entre las variables pueden ser fácilmente analizados a través de una matriz de correlación, para este análisis fue utilizado el coeficiente de correlación de Pearson, el cual se encarga de medir la dependencia lineal existente entre dos variables numéricas. En este caso, los recuadros en azul oscuro representan una correlación directamente proporcional entre ambas variables, y en el caso de los recuadros azul claro, representan una relación inversamente proporcional entre dichas variables.

A través de la matriz se puede observar cómo existen altas relaciones positivas (directamente proporcionales) entre la mayoría de los contaminantes, por otro parte, se puede visualizar como el ozono tiene relaciones negativas (inversamente proporcionales) con las distintas moléculas, esto se podría deber a que el ozono, cuando se encuentra en la estratosfera, protege de la radicación solar que evita que se generen reacciones químicas encargadas de producir moléculas contaminantes.

Un hecho evidente a observar en la matriz es las altas relaciones positivas que existen entre las variables demográficas, por ejemplo, entre las variables población y número de turismos. De la misma manera, se observa como existen más relaciones, tanto positivas o negativas, entre las variables meteorologías y nuestras variables a predecir (moléculas contaminantes) que entre las variables demográficas y las variables que se buscan pronosticar.

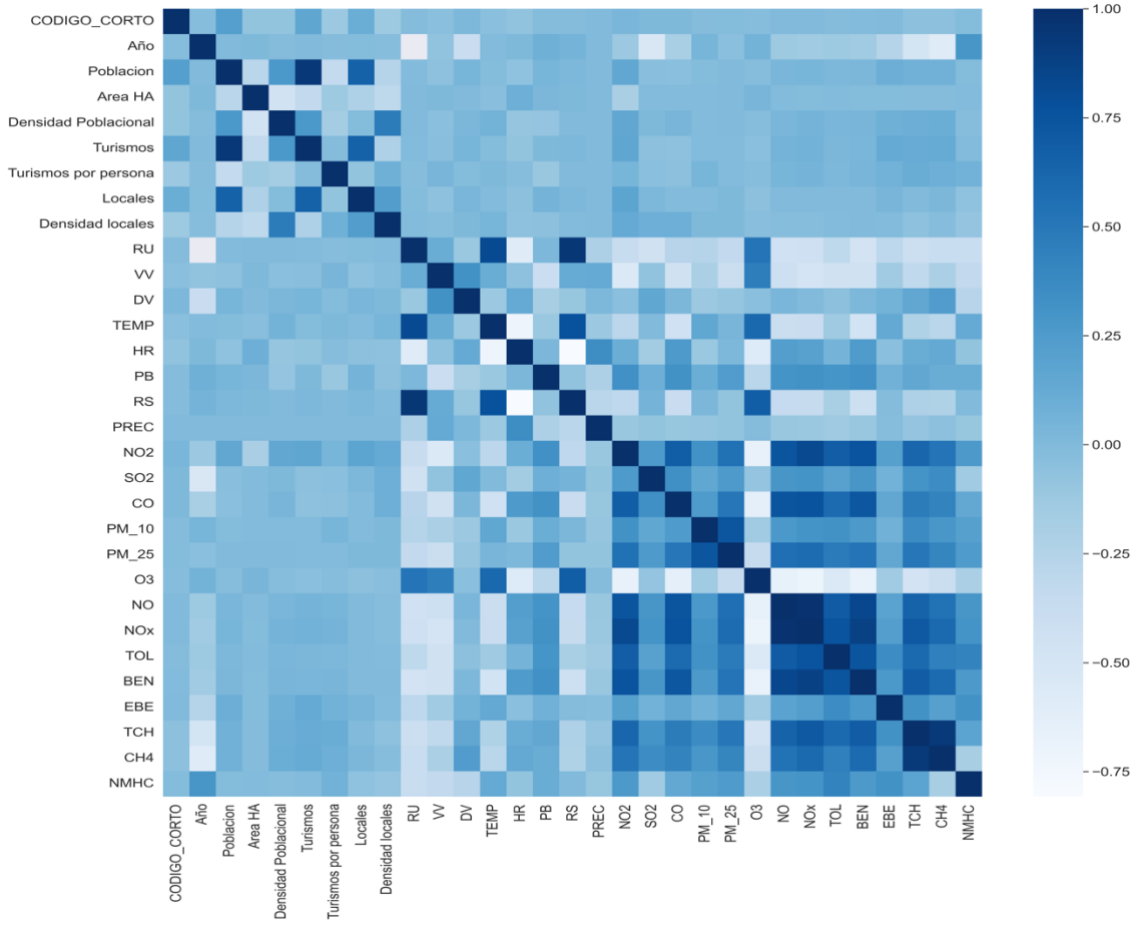

Ilustración 6. Matriz de correlación de las variables

3.4 Ubicación geográfica de las estaciones

Las diversas estaciones tanto meteorológicas o encargas de medir la calidad del aire se encuentran distribuidas en Madrid, ubicándose la mayoría en la zona central de la ciudad. En la siguiente imagen se puede visualizar su ubicación, en la cual los puntos morados representan estaciones meteorológicas (26 estaciones) y los puntos verdes las estaciones que miden las moléculas contaminantes (24 estaciones). Resulta importante recordar que en muchos casos ambos tipos de estaciones se encuentra en el mismo lugar.

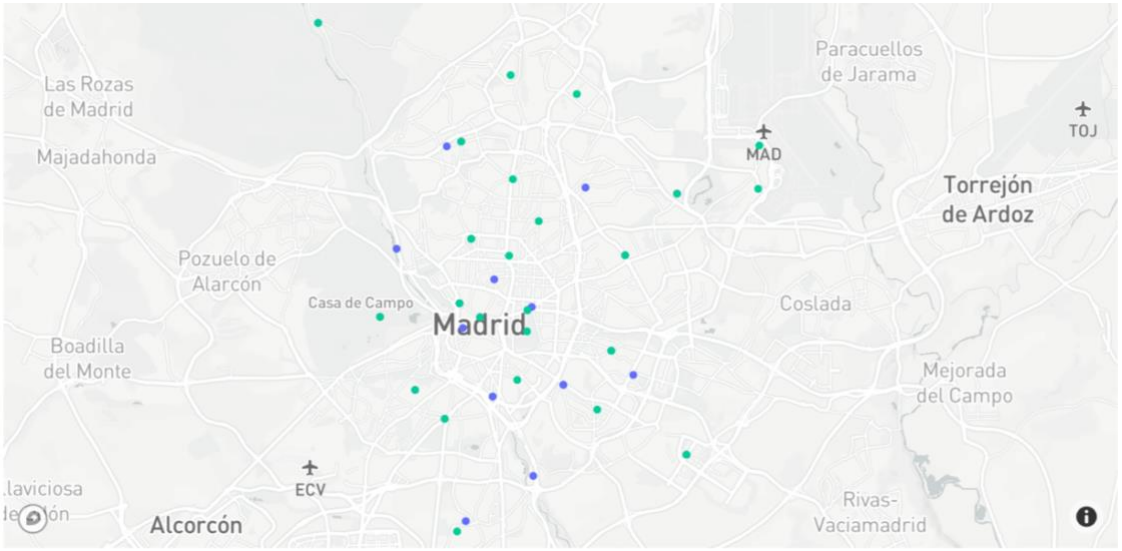

Ilustración 7. Mapa de las estaciones meteorológicas de Madrid

3.5 Análisis de los barrios de Madrid

En este trabajo, solo se analizarán los barrios los cuales tienen algún tipo de estación. A través de este análisis se podrá comprender cuales son los barrios de Madrid con mayor afluencia de personas, locales y vehículos (Turismos). Para este análisis se hace uso de distintas graficas de barras, en donde se muestran los valores correspondientes a las columnas: densidad poblacional, locales, densidad de locales, turismos y turismos por persona.

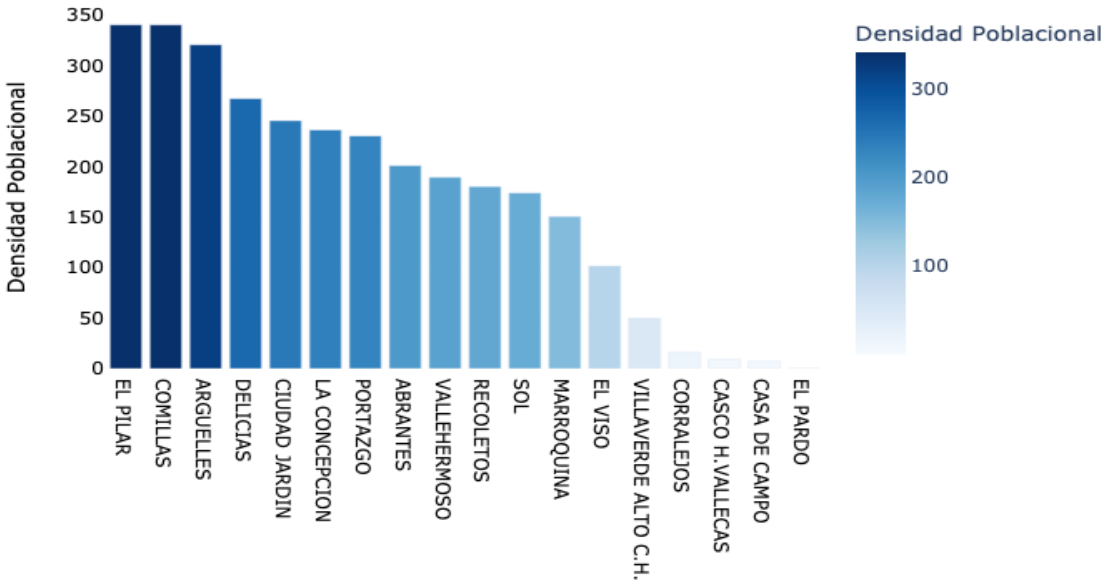

Ilustración 8. Densidad Poblacional de los barrios de Madrid

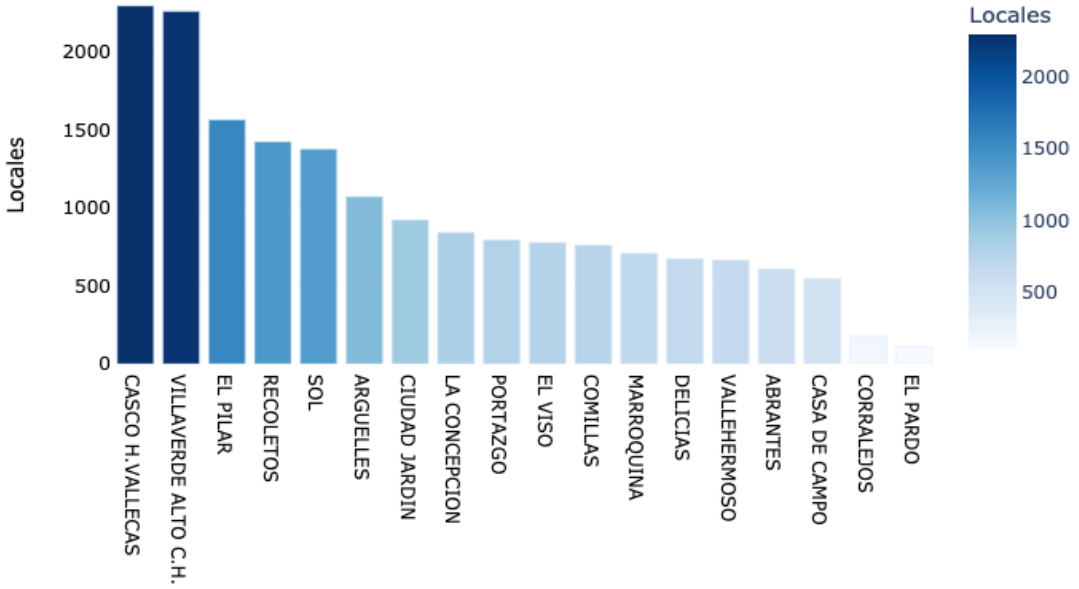

Ilustración 9. Número de locales de los barrios de Madrid

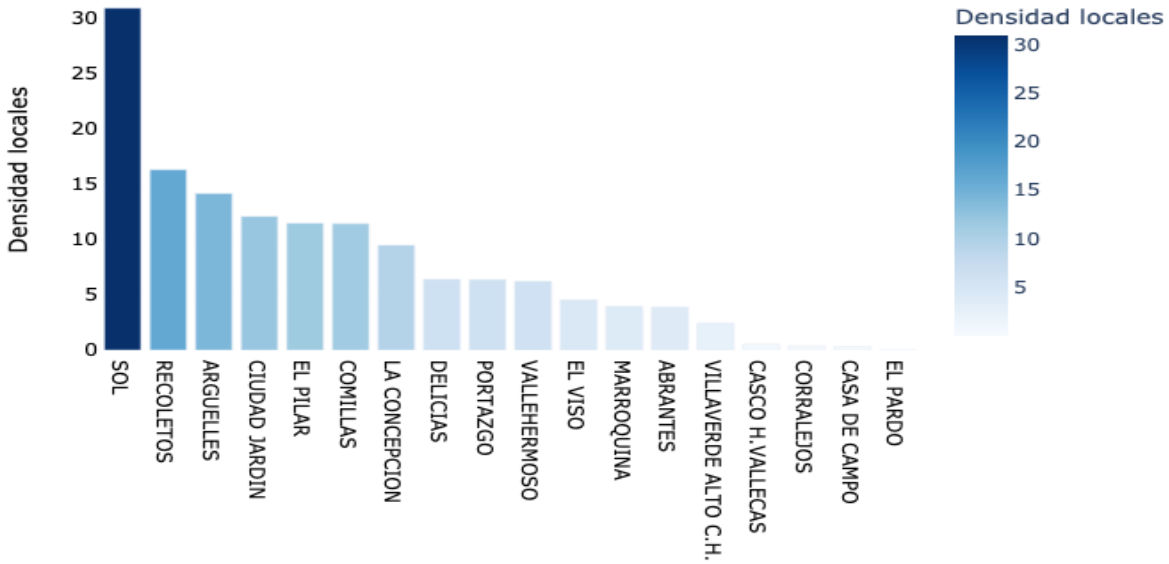

Ilustración 10. Densidad de locales en los barrios de Madrid

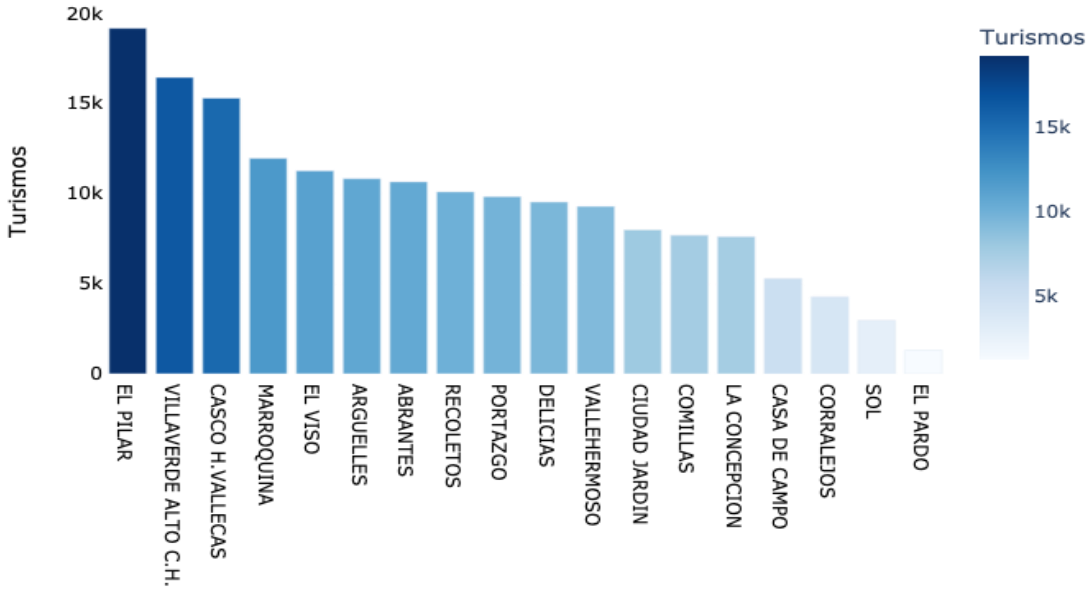

Ilustración 11. Número de turismos en los barrios de Madrid

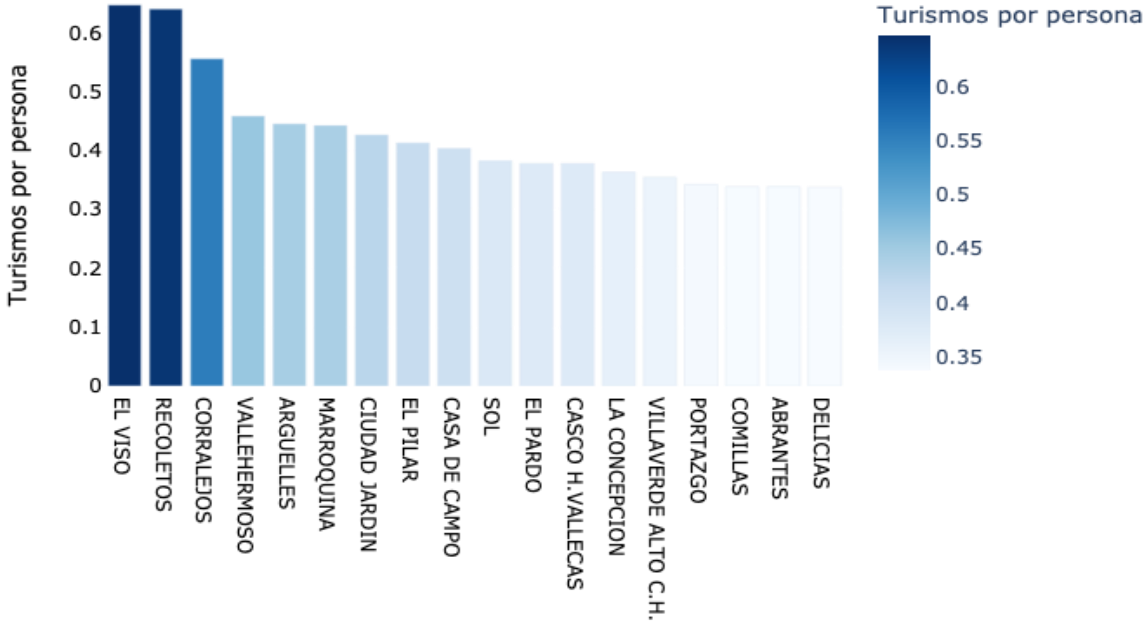

Ilustración 12. Turismos por persona en los barrios de Madrid

Entre los barrios que poseen algún tipo de estación (meteorológica o de calidad del aire), los tres barrios con mayor densidad poblacional son: barrio el pilar, comillas y arguelles, tres barrios residenciales de la ciudad. De igual forma, los barrios con mayor número de locales son casco histórico de Vallecas, y casco histórico de Villaverde, esto no se debe a que estos barrios se consideren como importantes zonas comerciales de Madrid, al contrario, dichos valores se obtuvieron debido al extenso tamaño de estos barrios y es por este hecho que la variable 'Densidad locales' da una visión más certera de cuáles son los puntos más comerciales de la ciudad. Observando la ilustración 11, se determina que el barrio con la mayor densidad de locales es Sol, barrio con alta densidad de comercios y sobre todo puntos turísticos.

Un aspecto importante observado en las gráficas es la ubicación de los barrios El pardo y casa de campo, los cuales se posicionan en los últimos lugares, ese posicionamiento de debe a que ambos barrios representan en su mayoría zonas verdes.

En cuanto al número de turismos, los barrios con mayor número de turismos son el barrio el pilar, casco histórico de Villaverde y casco histórico de Vallecas, siendo todos estos barrios mayormente residenciales, no obstante, al momento de conocer cuáles son los barrios con mayor número de turismos por persona se puede observar cómo los barrios de el viso y recoletos toman los primeros lugares, hecho que se debe al alto poder adquisitivo de sus habitantes, siendo estas dos zonas de los barrios con mayores renta per cápita de toda España.

3.6 Datos meteorológicos de Madrid

En este apartado se observan distintas series temporales de diversos datos climatológicos en la ciudad de Madrid. A través de estas graficas se pueden observar la estacionalidad de los datos como también las tendencias de aumento o disminución.

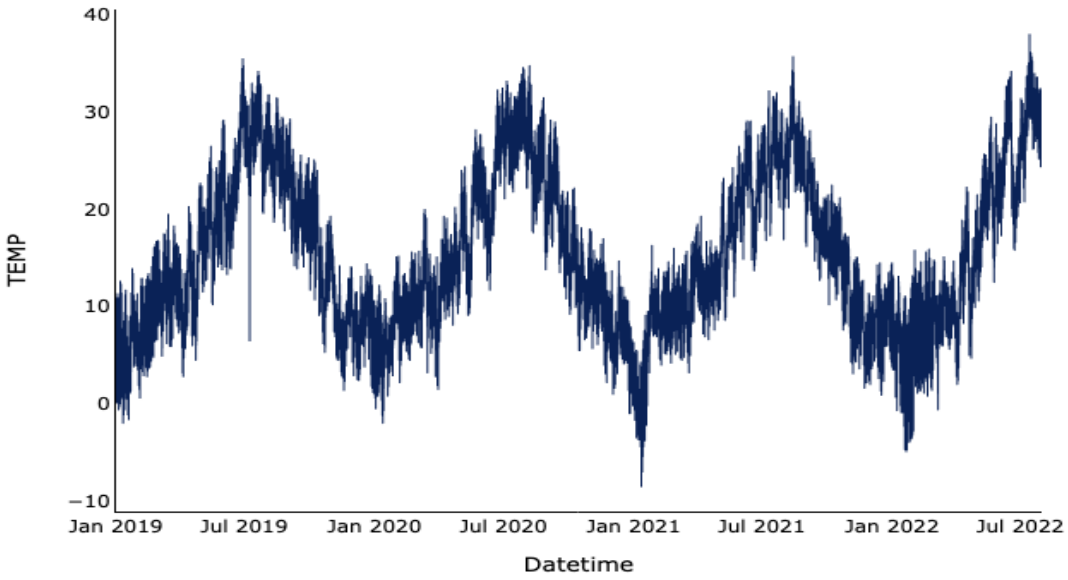

Ilustración 13. Serie temporal para la temperatura (°C)

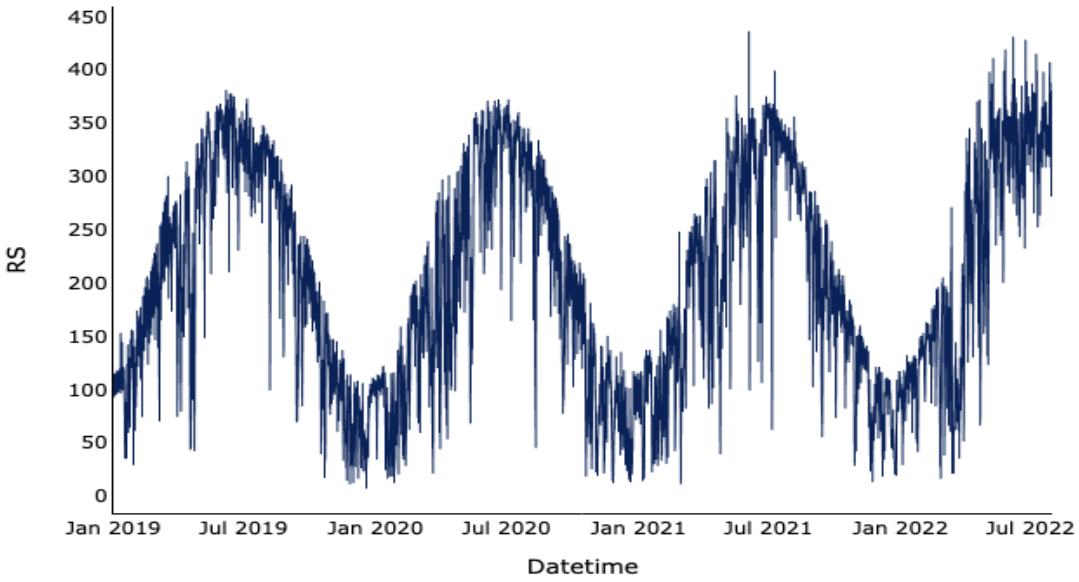

Ilustración 14. Serie temporal para radiación solar (W/m^2)

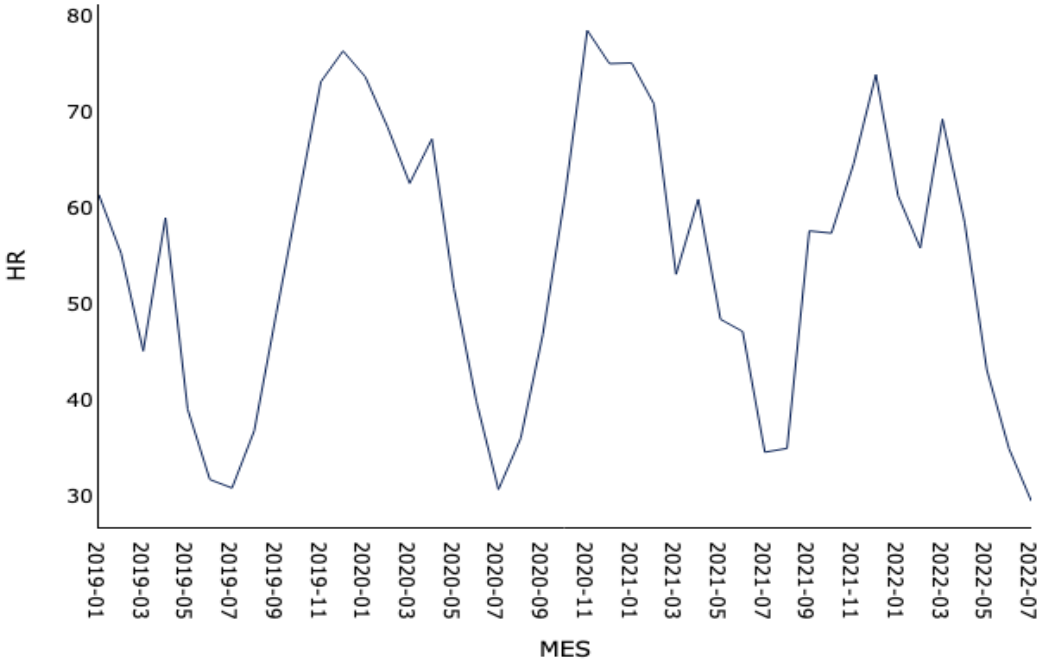

Ilustración 15. Serie temporal para la humedad relativa (%)

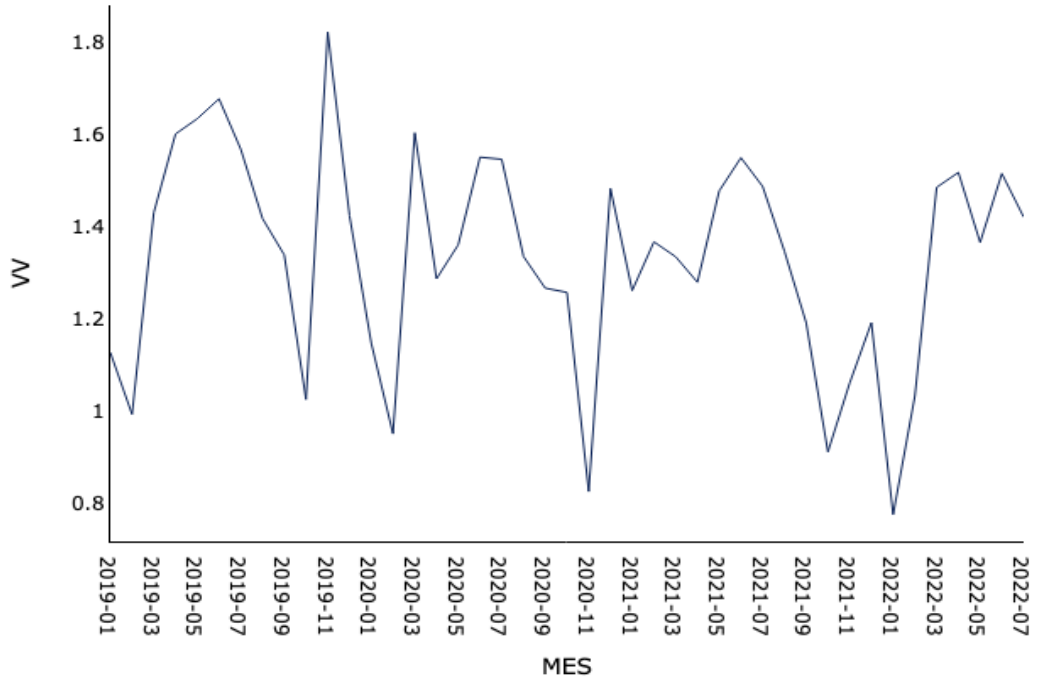

Ilustración 16. Serie temporal para la velocidad del viento (m/s)

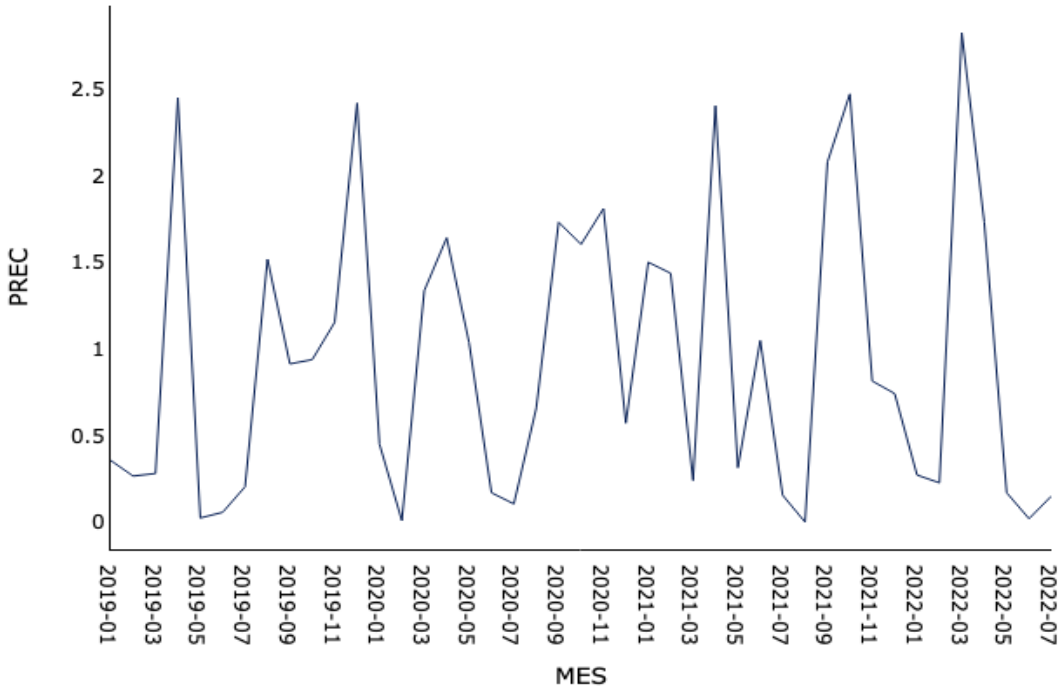

Ilustración 17. Serie temporal para la precipitacion (l/m2)

A través de estas graficas se puede visualizar, las relaciones existentes entre la temperatura, la radiación solar y la humedad relativa. Por un lado, la temperatura y la radiación solar tienen una relación directa y se puede observar fácilmente la estacionalidad en ambos conjuntos de datos, con máximos en los meses de verano y mínimos en los meses de invierno. Por el contrario, la humedad es inversa a la temperatura y la radiación solar, con valores mal altos en los meses de invierno y valores más bajos en los meses de verano.

En cuanto a la velocidad del viento, notamos como se registran valores contenidos dentro de un rango en la mayoría de los días, con ciertos picos en los meses de enero. Por otro lado, referente a la precipitación, se registran múltiples días con cero precipitación y días con máxima precipitación, especialmente en los meses de otoño y primavera.

3.7 Distribuciones de las variables de entrada

En el caso de este trabajo, conocer las distribuciones de las variables utilizadas en el modelo de predicción permite analizar si luego de la permutación

realizada en el capítulo anterior, en donde se completaron los valores nulos con la media del día, obtuvimos variables en donde sus valores se acercan siempre a la media. Este caso tendría como consecuencia un modelo de predicción con muy alta precisión pero que no sería capaz de predecir con nuevos valores que se alejen de esta media.

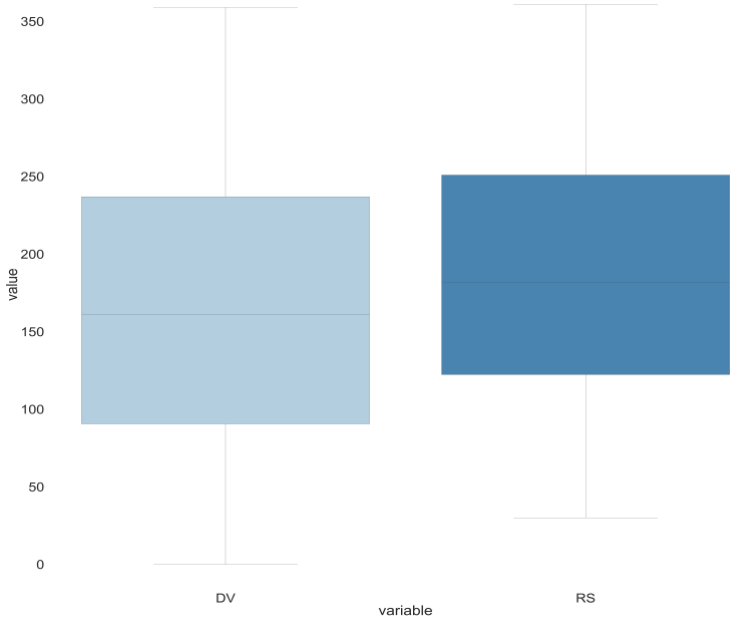

Ilustración 18. Distribución de las variables dirección del viento y radiación solar

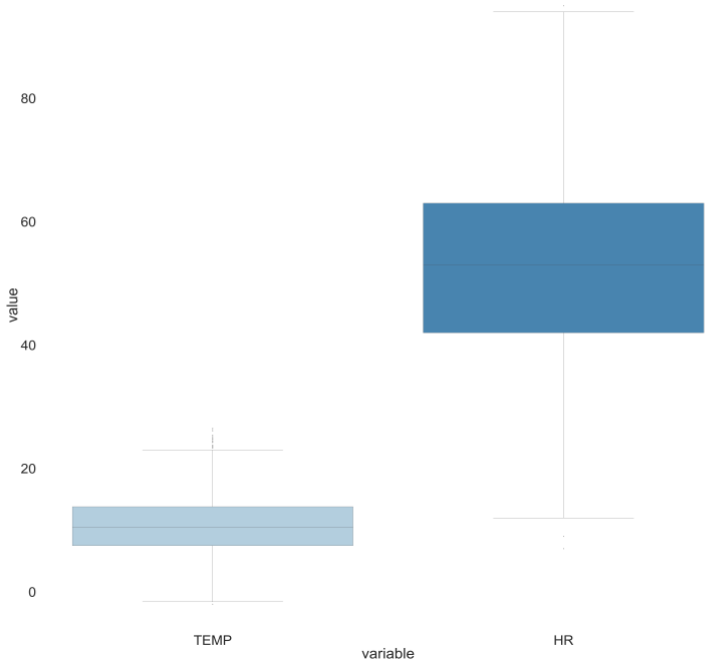

Ilustración 19. Distribución de las variables temperatura y humedad relativa

En las gráficas anteriores se muestran las distribuciones de algunas de las variables utilizadas en el modelo para las cuales fue necesario realizar permutación de los datos, aun habiendo realizado este procesamiento vemos como estas variables mantienen una distribución en donde no se encuentran todos los valores relativamente cercanos a la media.

3.8 Distribuciones de las variables de salida

Conocer las distribuciones de nuestras variables de salida nos permite conocer si existe una varianza en los valores de cada variable. En el caso contrario, obtendríamos modelos de predicción con muy buenos resultados. en las métricas de evaluación, sin embargo, el mismo podría estar simplemente prediciendo un valor cercano a la media. A continuación, se muestran graficas con la distribución de varias variables de salida.

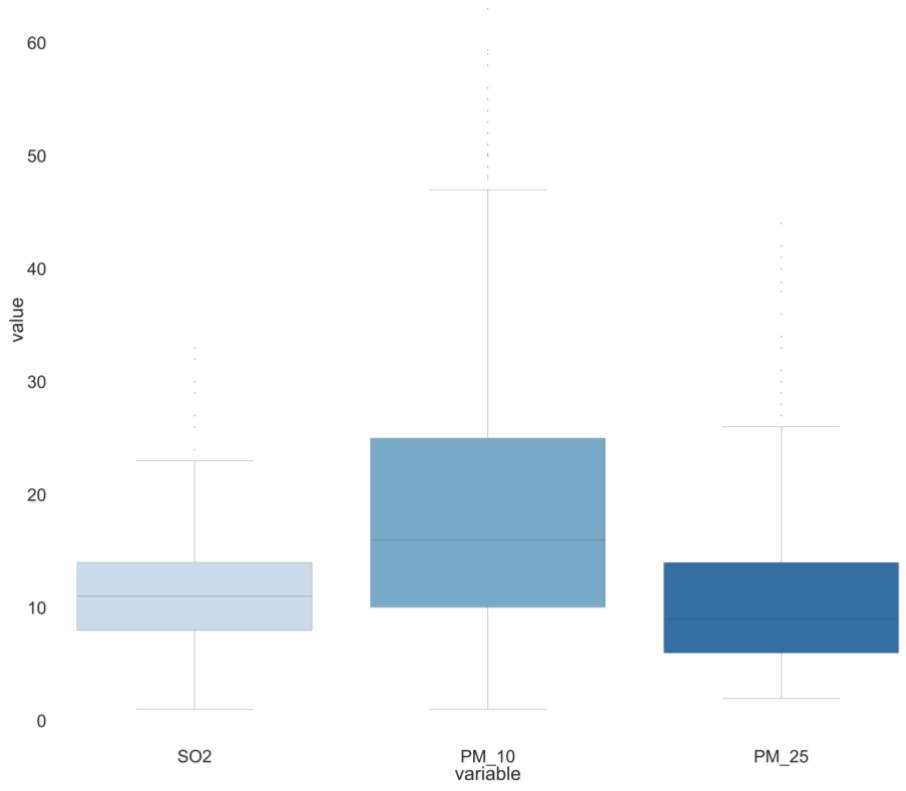

Ilustración 20. Distribución de las variables de salida

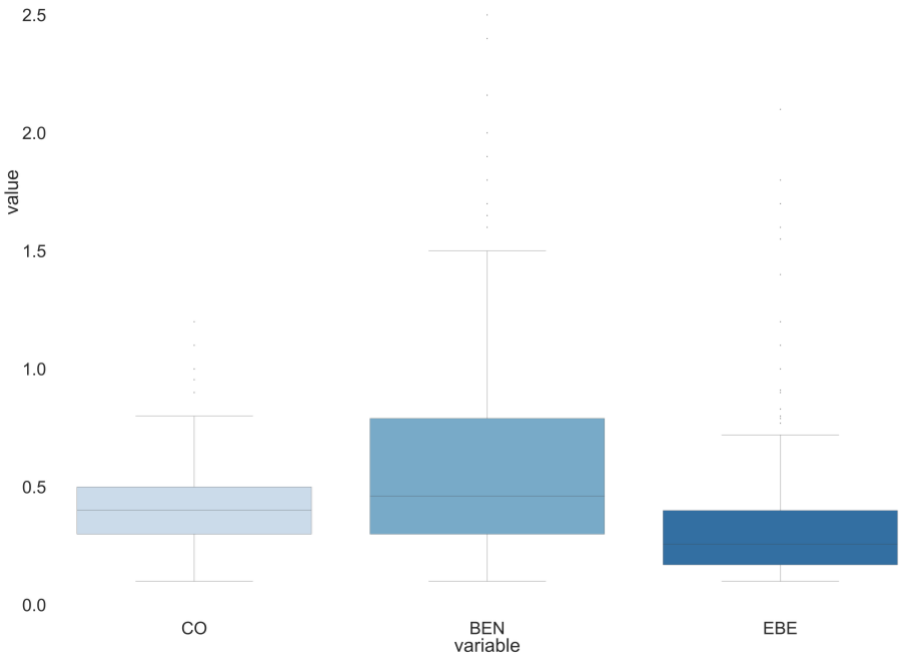

Ilustración 21. Distribución de las variables de salida

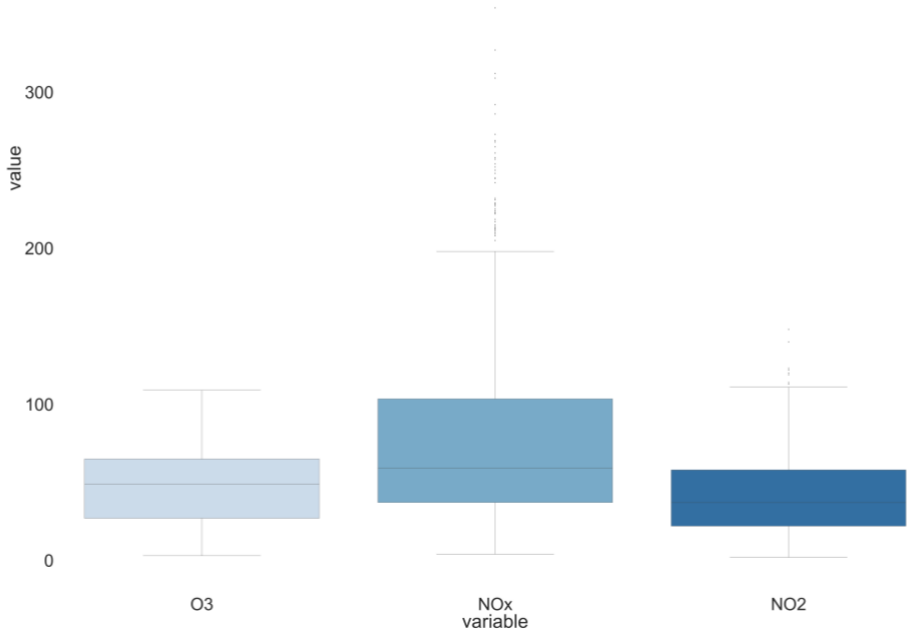

Ilustración 22. Distribución de las variables de salida

Capítulo 4. MODELOS DE PREDICCIÓN

En este capítulo se explicará todo lo relacionado al modelo de predicción, como serán evaluados y la interpretación de los resultados obtenidos con el modelo. El entrenamiento de los modelos fue realizado a través de la librería de Python, Scikit-Learn (Pedregosa, 2011).

4.1 Aprendizaje supervisado regresión

Primeramente, resulta primordial definir el tipo de modelo que se busca implementar. En el caso de este trabajo se implementa un modelo de aprendizaje supervisado, en donde los datos con los que estamos trabajando están etiquetados, es decir, poseen un "target" que sería la variable que se busca predecir a partir del resto de las variables. Existen dos tipos de problemas que resuelven los modelos de aprendizaje supervisado, de regresión y de clasificación, en esta investigación se entrena un modelo de regresión ya que buscamos predecir un valor numérico continuo.

4.2 Métricas de evaluación

Para la evaluación y comparación de los distintos modelos implementados, se utilizó el coeficiente 'R2', el cual es comúnmente utilizado para la evaluación de los modelos predictivos de regresión. Dicho valor es capaz de proporcionar una medida que representan que tan ajustados se encuentran los valores predichos de los valores reales, este coeficiente conocido como coeficiente de determinación puede tomar un valor de 0 a 1, siendo 1 una predicción perfecta. Resulta importante aclarar que valor de 'R2' obtenido a través de Scikit-Learn es capaz de regresar valores negativos. En resumen, 'R2' es el porcentaje de la variación en la variable dependiente que es explicado por el modelo utilizado. A continuación, se muestra la ecuación de dicho coeficiente.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$$

Ecuación 1. R2

En donde $\sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2$ representa la suma residual de cuadrados y, $\sum_{i=0}^{n}(y_i - \bar{y})^2$, representa la varianza de los valores reales.

Existen distintas métricas, aparate de R2, que pueden ser utilizadas para evaluar modelos de regresión como MAE, RSE o MRSE, no obstante, el resultado de esto últimos dependerá del rango de valores que en el que se encuentre nuestra variable de salida, en el caso de este trabajo, al estar prediciendo distintas variables de salida (todas con distintos rangos de valores distintos), resulta más conveniente la utilización de una métrica, como es el caso de R2, que pueda ser comparable entre distintos modelos encargados de evaluar distintas variables.

4.3 Algoritmos para implementar

En el siguiente apartado se explica una breve descripción de los algoritmos que fueron utilizados para resolver nuestro el problema de regresión de este trabajo.

4.3.1 Regresión lineal

La regresión o ajuste lineal es uno de los algoritmos más conocidos y utilizados para resolver problemas de regresión. El análisis de regresión es un método que permite analizar la variabilidad de una determinada variable en función de la información que le proporcionan una o más variables (Pedhazur, 1982).

Para evitar un problema conocido como '*overfitting'*, en donde el modelo es capaz de predecir muy bien los valores con los que fue entrenado, pero no es capaz de predecir nuevos valores, se utiliza regularización Ridge o Lasso. En el caso de la regularización Ridge, la función de coste utilizada en el modelo (función encargada de medir el error del modelo) es penalizada cuando el modelo asigna pesos muy grandes a algunas de las variables. La regularización Lasso funciona de manera similar, solo que en este ultimo los pesos de las variables pueden ser igual a cero.

4.3.2 Support Vector Machine

Las máquinas de vectores de soporte (para problemas de regresión) son un conjunto de algoritmos que buscan conseguir un hiperplano que junto con las líneas limites (dos líneas definidas a una distancia definida alrededor del hiperplano) que sea capaz de encerrar el mayor número de instancias.

4.3.3 Arboles de decisión

Los árboles de decisión son algoritmos que buscan generar reglas o decisiones hasta ser capaz de predecir un valor. Existen múltiples modelos de predicción basadas en los árboles de decisión, en el caso de este trabajo se aplicaron los siguientes: *random forest regressor, adaboost, gradient boosting regressor y bagging regressor.*

4.3.4 Nearest-Neighbors

El algoritmo *k-nearest-neighbor* o k-vecinos es un modelo no paramétrico utilizado tanto para clasificación como regresión en donde el modelo busca conseguir cuales son las instancias que se encuentran más cercanas para poder realizar la predicción o clasificación.

4.4 Datos de entrenamiento y testeo

Para el entrenamiento de los modelos y el muestreo de los datos para dicho entrenamiento se utilizó el método de *cross valiadation*, el cual consiste en dividir los datos en un número de grupos predeterminado, llamado comúnmente *'folds'*, y reentrenar el modelo tantas veces como *folds* se hayan determinado, utilizando en cada entrenamiento un grupo de testeo distinto, por lo tanto, la métrica obtenida al final consiste en la media de todas las métricas obtenidas en cada uno de los entrenamientos. En el caso de este trabajo se utilizaron cinco *'folds'*, es decir, en cada 'entrenamiento de los datos' se utilizó el 80% de los datos para el entrenamiento y el otro 20% para el testeo.

4.5 Preprocesamiento

Una de las técnicas de procesamiento empleadas previamente al entrenamiento fue la normalización. Esta consiste en ajustar las distintas escalas de todas las variables a una misma escala común. No todos los modelos se ven beneficiados por dicho preprocesamiento (como los árboles de decisión) sin embargo, en una primera instancia se evalúan los distintos modelos bajo las mismas condiciones.

4.6 Resultados

En primer lugar, los hiperparametros utilizados para la implementación de los modelos corresponden a los valores asignados por defecto por la librería Scikit-Learn. Seguidamente, se muestra una tabla en donde se observan los parámetros más importantes utilizados en cada uno de los modelos.

Modelo	Parámetro	Descripción	Valor
Regresión Lineal	n/a	n/a	n/a
Regresión Lineal (Ridge)	Alpha	Valor encargado de penalizar la función de coste	1
Regresión Lineal (Lasso)	Alpha	Valor encargado de penalizar la función de coste	1
Adaboost	base_estimator	Modelo de predicción utilizado	Árbol de decisión
	n_estimators	Número de estimadores creados	50
Gradient Boosting	max_depth	Máxima profundidad de los arboles	3
	n_estimators	Número de estimadores creados	100
Random forest	max_depth	Máxima profundidad de los arboles	3
	n_estimators	Número de estimadores creados	100
Bagging Regressor	base_estimator	Modelo de predicción utilizado	Árbol de decisión
	bootstrap	Técnica de muestreo sin reemplazamiento	True
	n_estimators	Número de estimadores creados	10
K-nearest-neighbors	n_neighbors	Número de 'vecinos' a considerar en la predicción	5
Support Vector Machine con kernerl lineal	epsilon	Determina el margen de tolerancia donde no se otorga penalidad a los errores	0.1
	kernel	El tipo de kernel utilizado en el algoritmo	linear
Support Vector Machine con kernerl RBF	epsilon	Determina el margen de tolerancia donde no se otorga penalidad a los errores	0.1
	kernel	El tipo de kernel utilizado en el algoritmo	rbf

Tabla 4.1 Parámetros de los modelos

En la siguientes graficas se muestran los resultados de la métrica de evaluación ‘R2’ para los distintos contaminantes con cada uno de los modelos utilizados.

En donde:

- RL: regresión lineal.
- RIDGE: regresión lineal con regularización ridge.
- LASSO: regresión lineal con regularización lasso.
- ADABOOST: árbol de decisión adaboost.
- GB: Gradient boosting.
- BR: Bagging regressor.
- RF: random forest.
- KNR: K-nearest-neighbor regressor
- SVR LINEAR: Support vector machine con kernel lineal.
- SVR RBF: Support vector machine con kernel rbf.

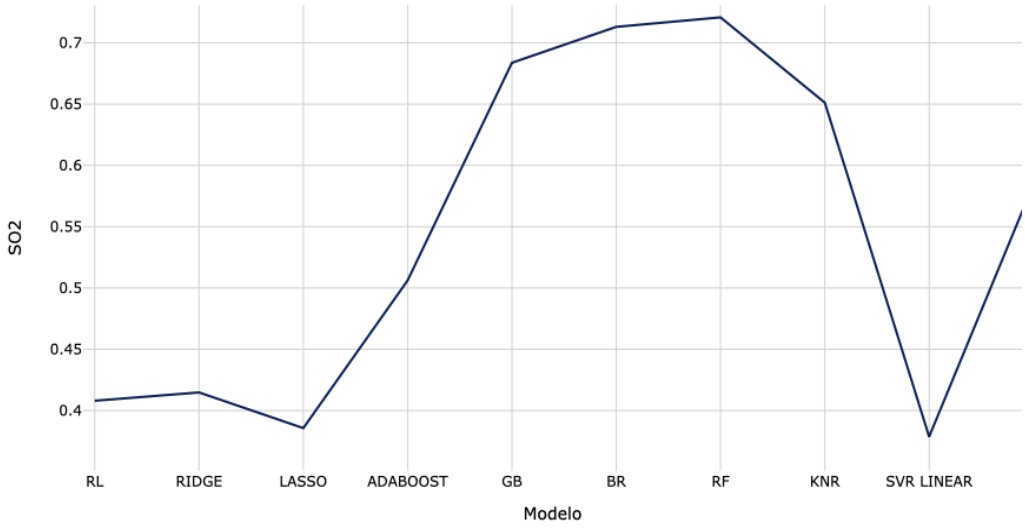

Ilustración 23. R2 obtenido de los modelos al predecir SO2

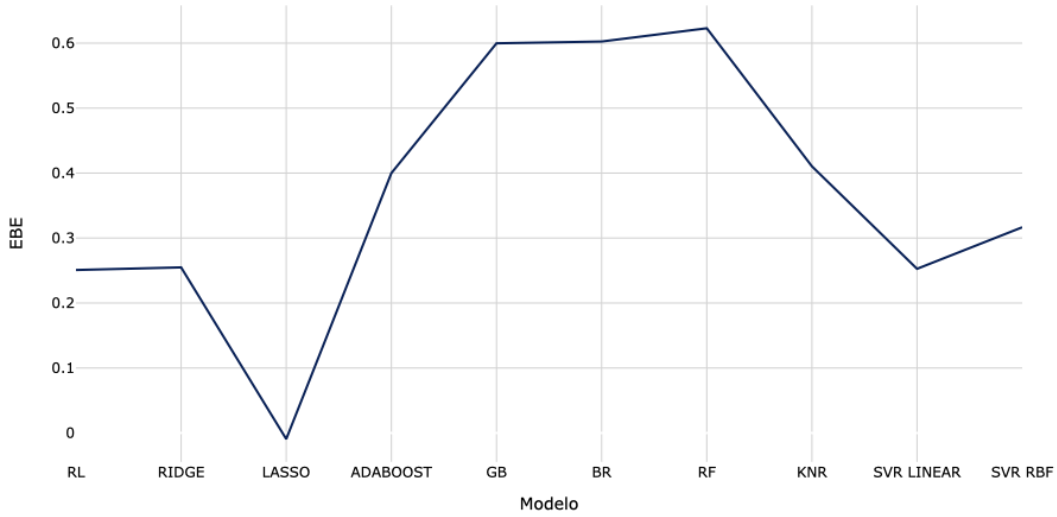

Ilustración 24. R2 obtenido de los modelos al predecir EBE

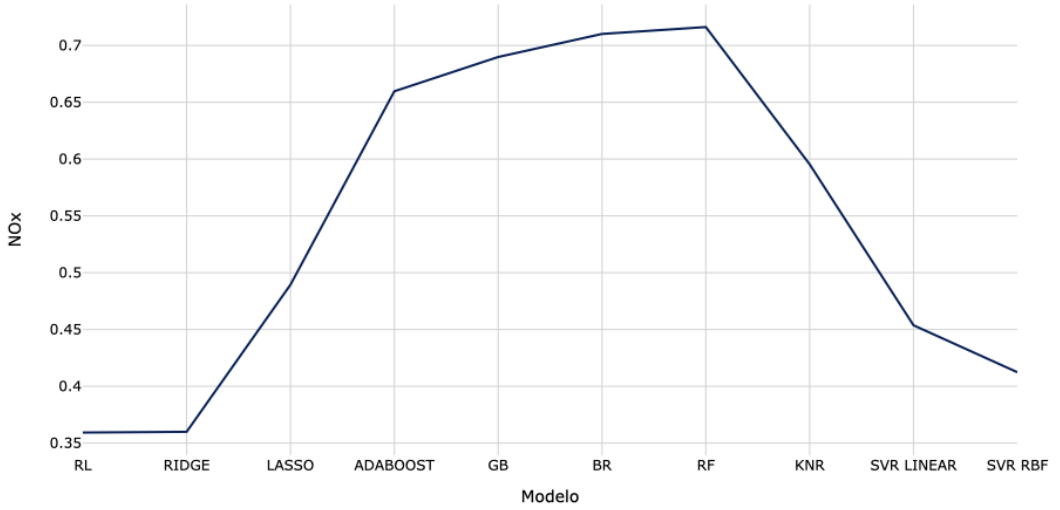

Ilustración 26. R2 obtenido de los modelos al predecir NOx

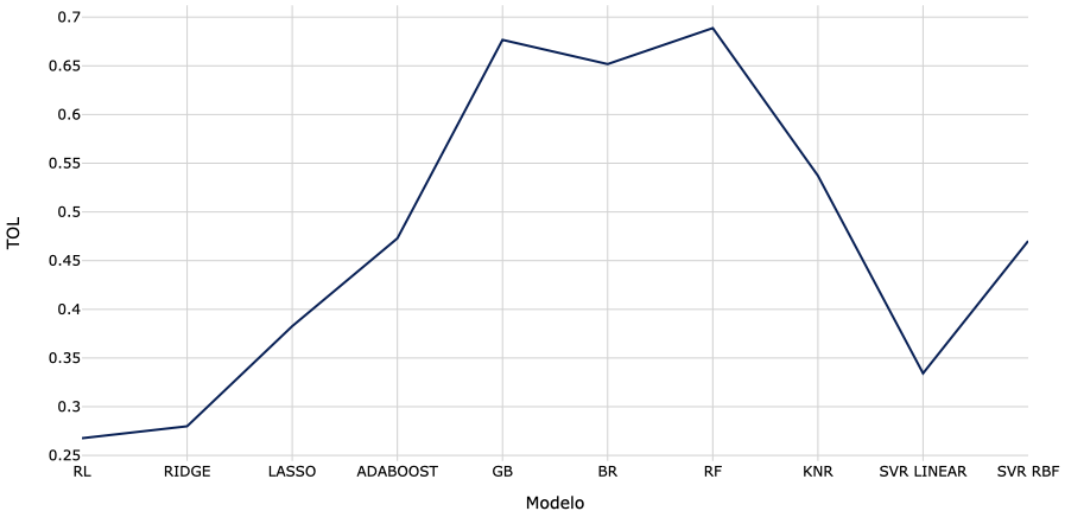

Ilustración 25. R2 obtenido de los modelos al predecir TOL

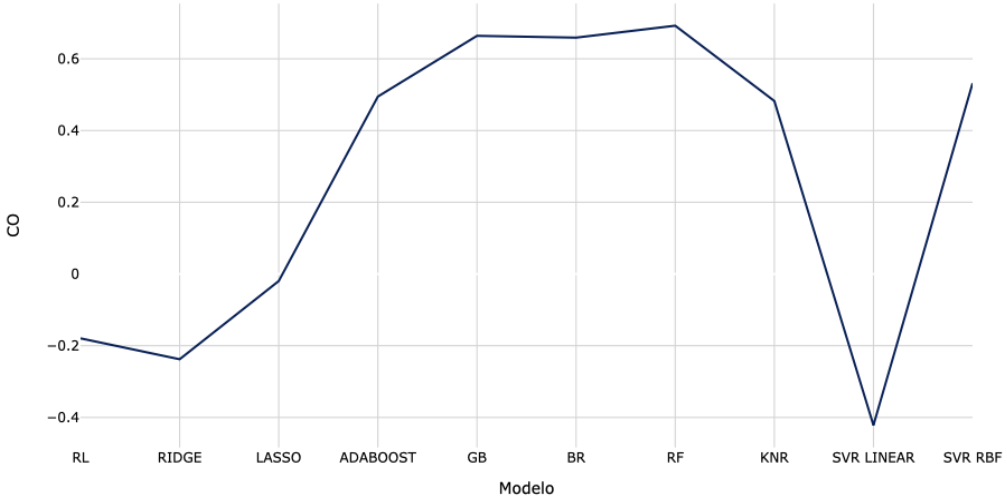

Ilustración 29. R2 obtenido de los modelos al predecir CO

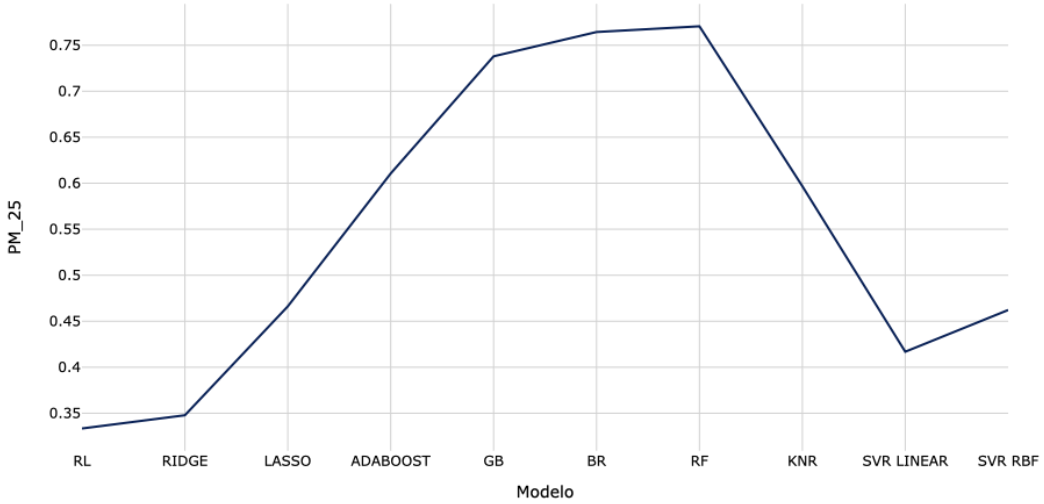

Ilustración 28. R2 obtenido de los modelos al predecir PM_25

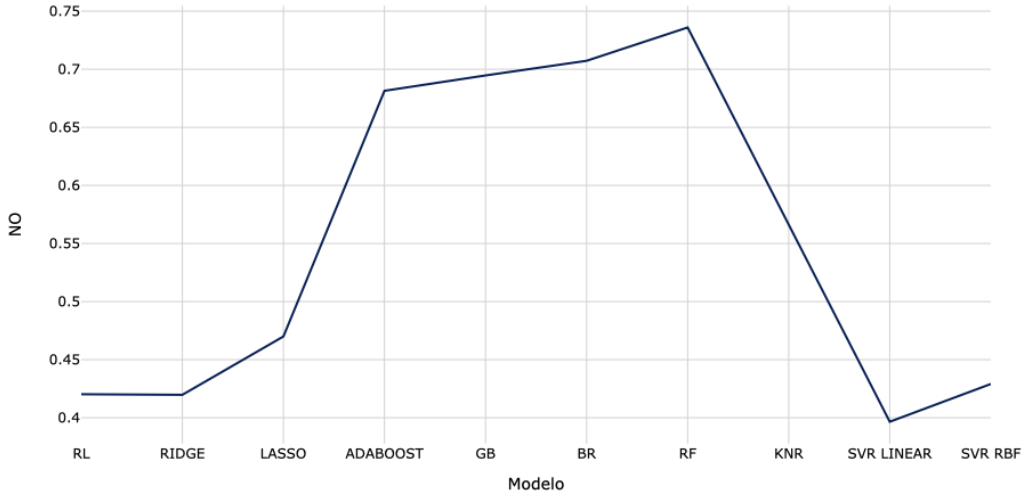

Ilustración 27. R2 obtenido de los modelos al predecir NO

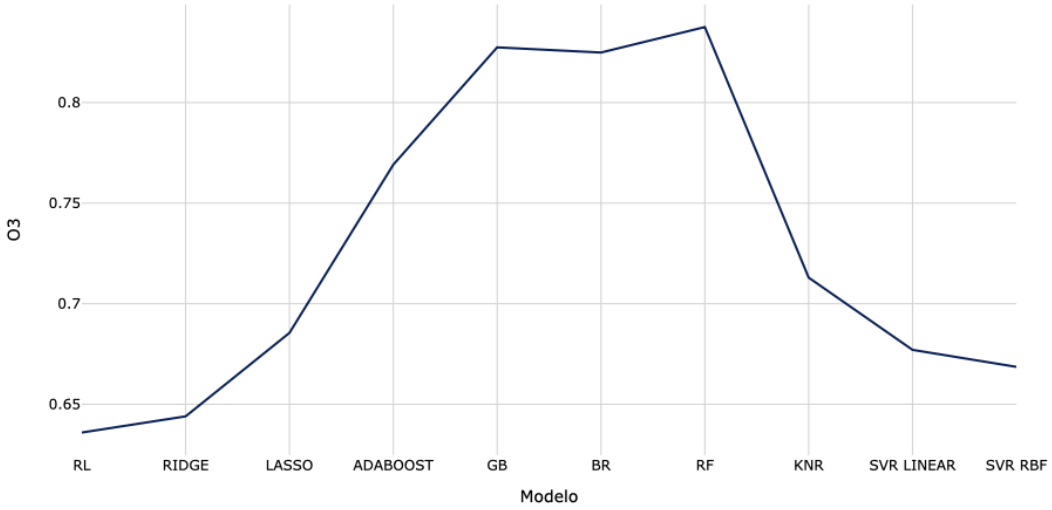

Ilustración 32. R2 obtenido de los modelos al predecir O3

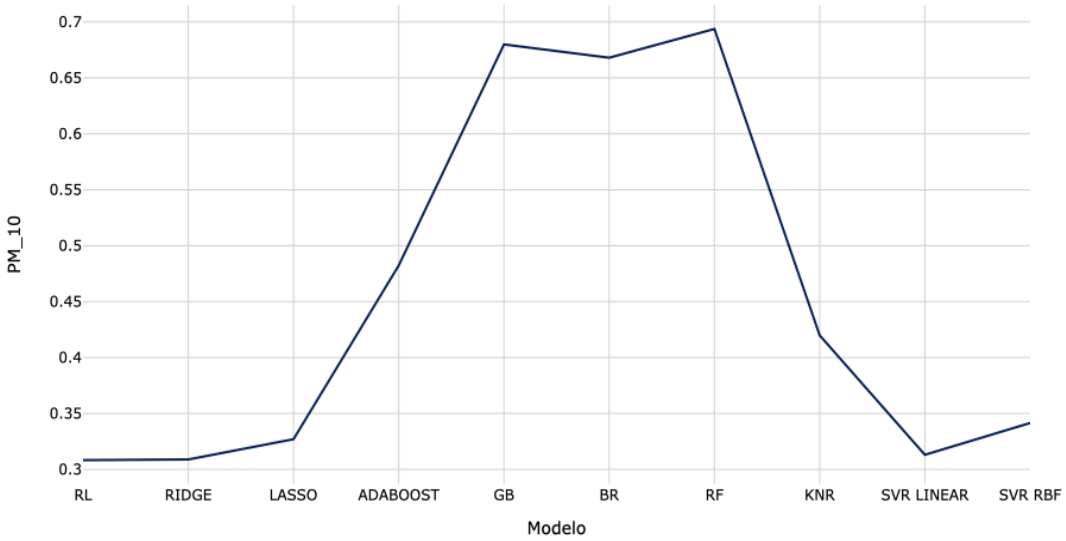

Ilustración 31. R2 obtenido de los modelos al predecir PM_10

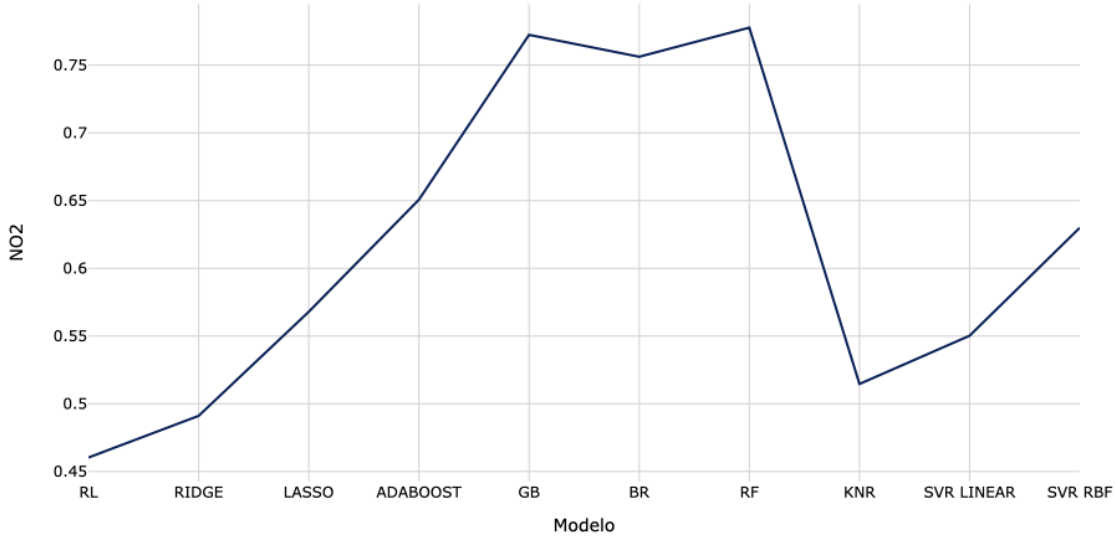

Ilustración 30. R2 obtenido de los modelos al predecir NO2

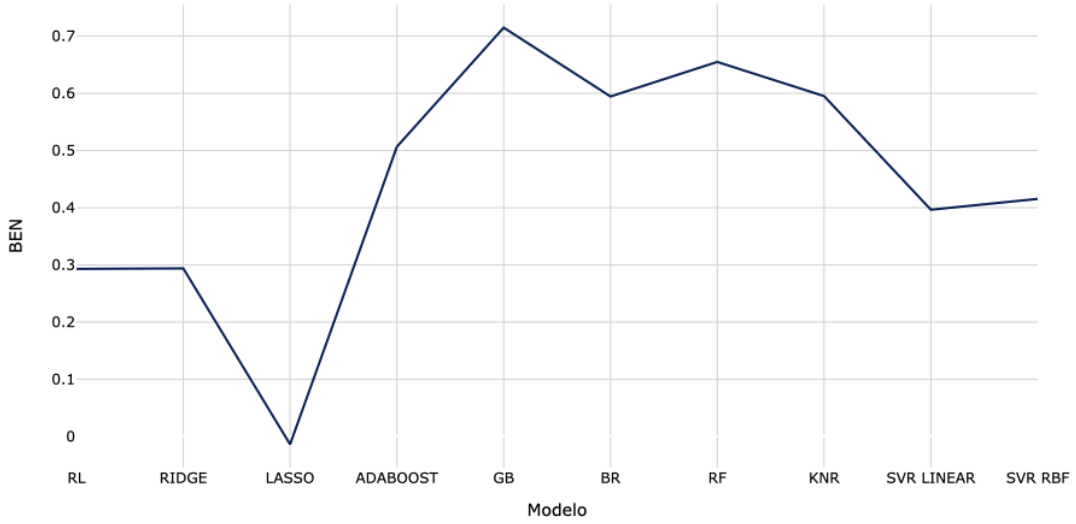

Ilustración 33. R2 obtenido de los modelos al predecir BEN

A través de estas graficas podemos observar cómo los mejores resultados para todos los contaminantes fueron obtenidos con los modelos basados en árboles de decisiones, específicamente *Gradient Boosting, Bagging Regressor y Random Forest.*

De igual manera, la siguiente tabla nos permite determinar cuál fue el contaminante predicho con el cual se obtuvieron mejores métricas, el cual fue el ozono. Este resultado nos puede llevar a la conclusión de que dicho contaminante se ve altamente influenciado por las variables de entrada utilizadas en este trabajo, es decir, una gran proporción de los valores del ozono en la ciudad de Madrid pueden ser predichos a través de datos meteorológicos y demográficos de la zona.

Modelo	NO2	SO2	CO	PM_10	PM_25	O3	NO	NOx	TOL	BEN	EBE
RL	0.460	0.408	-0.180	0.308	0.333	0.636	0.420	0.359	0.268	0.293	0.251
RIDGE	0.491	0.415	-0.238	0.309	0.348	0.644	0.420	0.360	0.280	0.294	0.255
LASSO	0.568	0.386	-0.020	0.327	0.466	0.686	0.470	0.489	0.383	-0.014	-0.009
ADABOOST	0.651	0.506	0.495	0.482	0.611	0.769	0.681	0.660	0.473	0.507	0.400
GB	0.772	0.684	0.664	0.680	0.738	0.827	0.695	0.690	0.677	0.715	0.600
BR	0.756	0.713	0.659	0.668	0.764	0.825	0.707	0.710	0.652	0.594	0.603
RF	0.778	0.721	0.692	0.694	0.771	0.838	0.736	0.716	0.689	0.655	0.623
KNR	0.515	0.651	0.482	0.420	0.597	0.713	0.566	0.595	0.537	0.595	0.410
SVR LINEAR	0.550	0.379	-0.422	0.313	0.417	0.677	0.397	0.454	0.334	0.396	0.253
SVR RBF	0.630	0.585	0.532	0.341	0.462	0.669	0.429	0.412	0.470	0.415	0.317

Tabla 4.2. Resultados de los modelos

4.7 Búsqueda de hiperparámetros

Cada modelo de algoritmo implementado requiere cierto número de hiperparametros definidos por la persona que crea el modelo, para esto, existen distintas técnicas que permiten optimizar dicho proceso a través de la prueba de distintas combinaciones de parámetros hasta conseguir la combinación con mejores resultados. En el caso de este trabajo, se utilizó la técnica conocida como *random search,* en donde se definen rangos de valores para la búsqueda y los mismos se escogen de manera aleatoria hasta elegir la combinación con las mejores métricas.

En el caso de este trabajo se realizó búsqueda de hiperparametros para los tres modelos con los cuales se obtuvieron los mejores resultados, es decir, *Gradient Boosting, Bagging Regressor y Random Forest.* A continuación, se muestran los resultados obtenidos luego de la búsqueda de hiperparametros.

Random Forest Regressor			
Contaminante	Original	Mejorado	Incremento
NO2	0.777687	0.783554	0.75%
SO2	0.720773	0.698193	-3.13%
CO	0.692369	0.708173	2.28%
PM_10	0.693741	0.705479	1.69%
PM_25	0.770583	0.776815	0.81%
O3	0.837531	0.833175	-0.52%
NO	0.735967	0.759127	3.15%
NOx	0.716189	0.737684	3.00%
TOL	0.688739	0.723602	5.06%
BEN	0.654783	0.744388	13.68%
EBE	0.622913	0.650289	4.39%

Tabla 4.3. Resultados del modelo *Random Forest Regressor*

Gradient Boosting Regressor			
Contaminante	Original	Mejorado	Incremento
NO2	0.77781	0.772382	-0.70%
SO2	0.671426	0.683709	1.83%
CO	0.677761	0.664058	-2.02%
PM_10	0.637441	0.679907	6.66%
PM_25	0.754149	0.737958	-2.15%
O3	0.809101	0.827429	2.27%
NO	0.664305	0.69468	4.57%
NOx	0.633829	0.689825	8.83%
TOL	0.691308	0.676663	-2.12%
BEN	0.669747	0.714769	6.72%
EBE	0.554914	0.599767	8.08%

Tabla 4.4. Resultados del modelo *Gradient Boosting Regressor*

Bagging Regressor			
Contaminante	Original	Mejorado	Incremento
NO2	0.756205	0.785331	3.85%
SO2	0.713029	0.674951	-5.34%
CO	0.659057	0.696154	5.63%
PM_10	0.668021	0.695505	4.11%
PM_25	0.764358	0.775175	1.42%
O3	0.824861	0.816165	-1.05%
NO	0.707282	0.774085	9.45%
NOx	0.710118	0.757344	6.65%
TOL	0.651796	0.72826	11.73%
BEN	0.594437	0.740656	24.60%
EBE	0.602509	0.664098	10.22%

Tabla 4.5. Resultados del modelo *Bagging Regressor*

	RF	GB	BR
NO2	0.783554	0.772382	0.785331
SO2	0.698193	0.683709	0.674951
CO	0.708173	0.664058	0.696154
PM_10	0.705479	0.679907	0.695505
PM_25	0.776815	0.737958	0.775175
O3	0.833175	0.827429	0.816165
NO	0.759127	0.69468	0.774085
NOx	0.737684	0.689825	0.757344
TOL	0.723602	0.676663	0.72826
BEN	0.744388	0.714769	0.740656
EBE	0.650289	0.599767	0.664098

Tabla 4.6. Resultado de los modelos Random Forest, Gradient Boosting Regressor y Bagging Regressor

Podemos observar cómo en algunos casos notamos una desmejora en los resultados obtenidos, esto se puede deber al hecho de la técnica de búsqueda de hiperparametros hace una búsqueda tomando valores aleatorios de los rangos especificados más no toma todos valores posibles.

4.8 Interpretabilidad de los resultados

Se analizaron algunos de los árboles de decisiones creados para el modelo con los mejores resultados, en este caso corresponde al modelo *random forest* cuando predice el ozono en el aire. La visualización de los árboles de decisiones creados por los modelos permite analizar cuáles son las variables con mayor influencia en nuestro modelo. En este caso podemos ver como en todos los árboles la variable al tope del árbol es velocidad del viento. Esto se debe a que la velocidad del viento tiene una gran influencia en la capacidad de esparcirse que tienen las moléculas en la atmosfera.

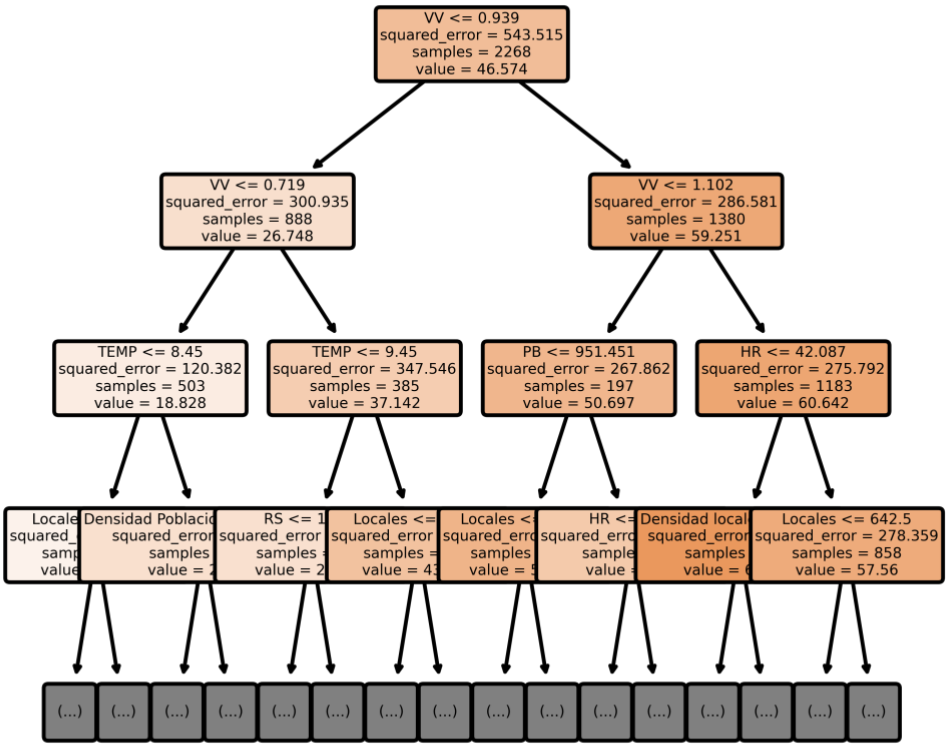

Ilustración 34. Árbol de decisión al predecir O3

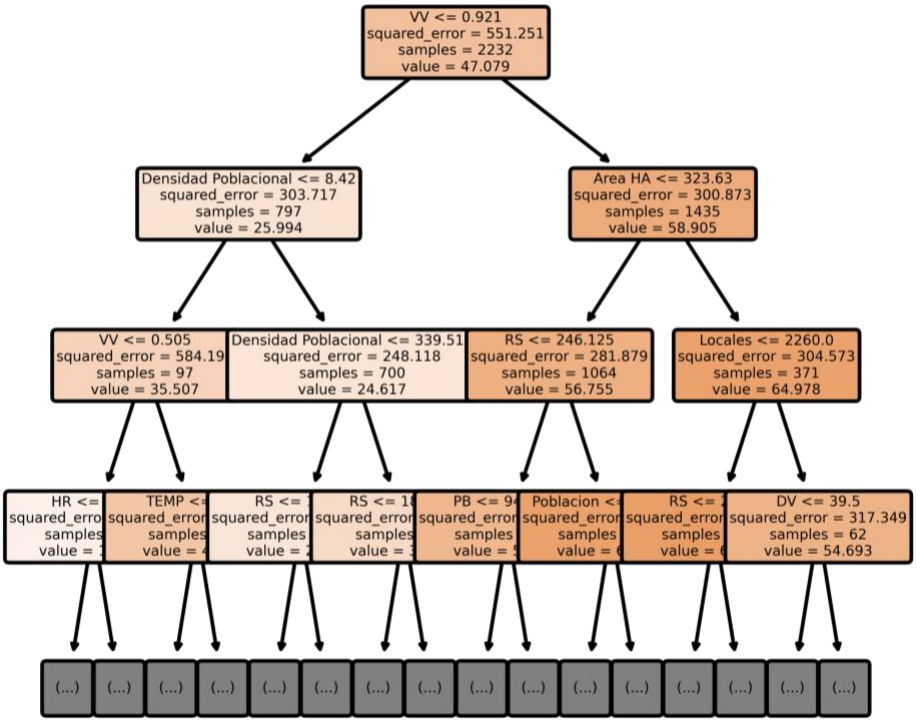

Ilustración 36. Árbol de decisión al predecir O3

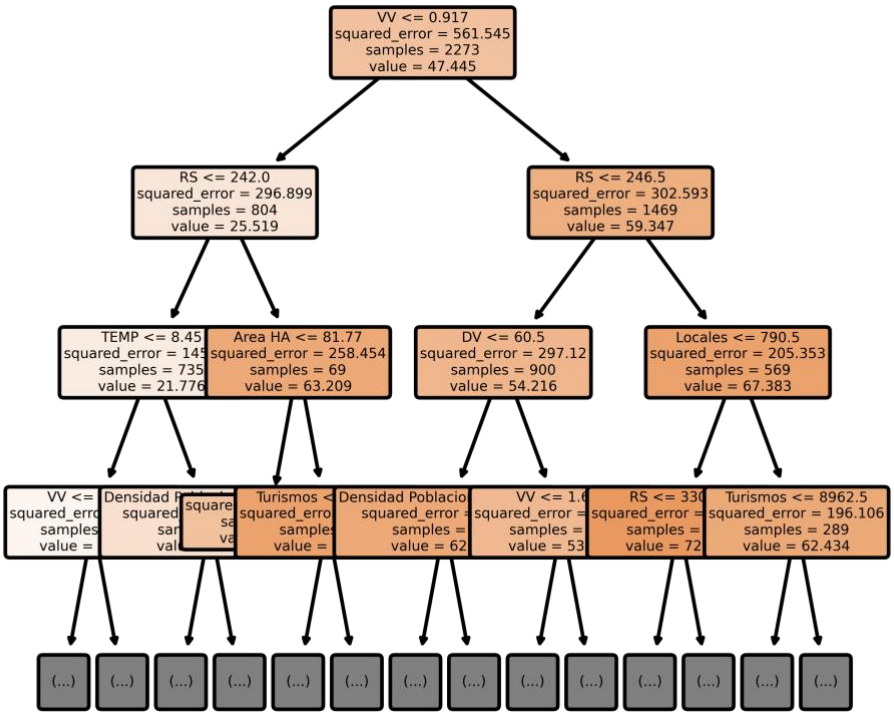

Ilustración 35. Árbol de decisión al predecir O3

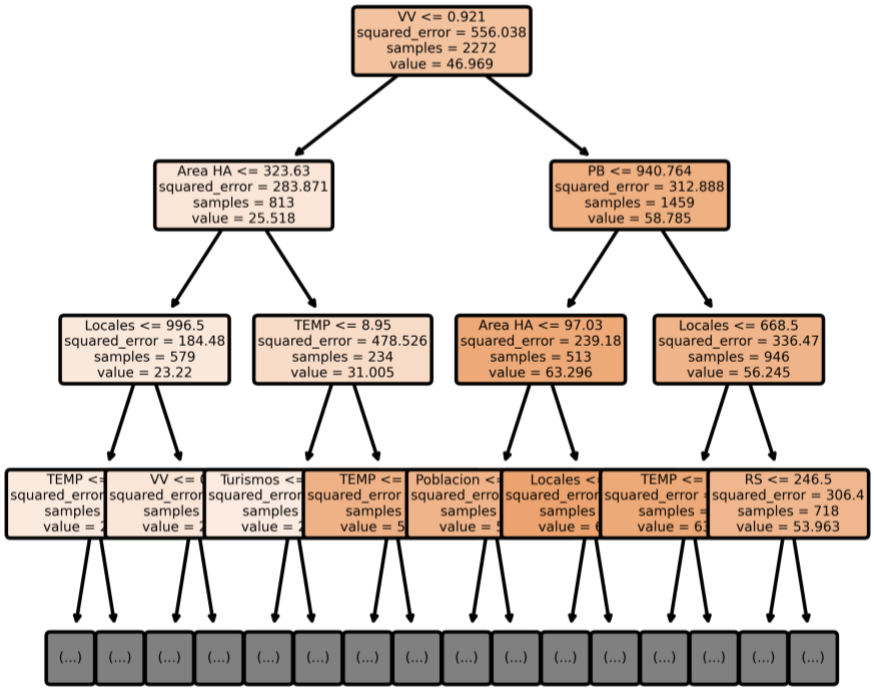

Ilustración 38. Árbol de decisión al predecir O3

De igual modo, existen múltiples técnicas que nos permiten analizar la importancia que tienen cada una de las variables dentro del modelo, como *permutation feature impotance*, en donde se reentrena el modelo realizando una permutación en una de las variables en cada entrenamiento, la mejora o desmejora del modelo aplicando dicha permutación permite conocer si dicha variable tiene una gran influencia en el modelo o no.

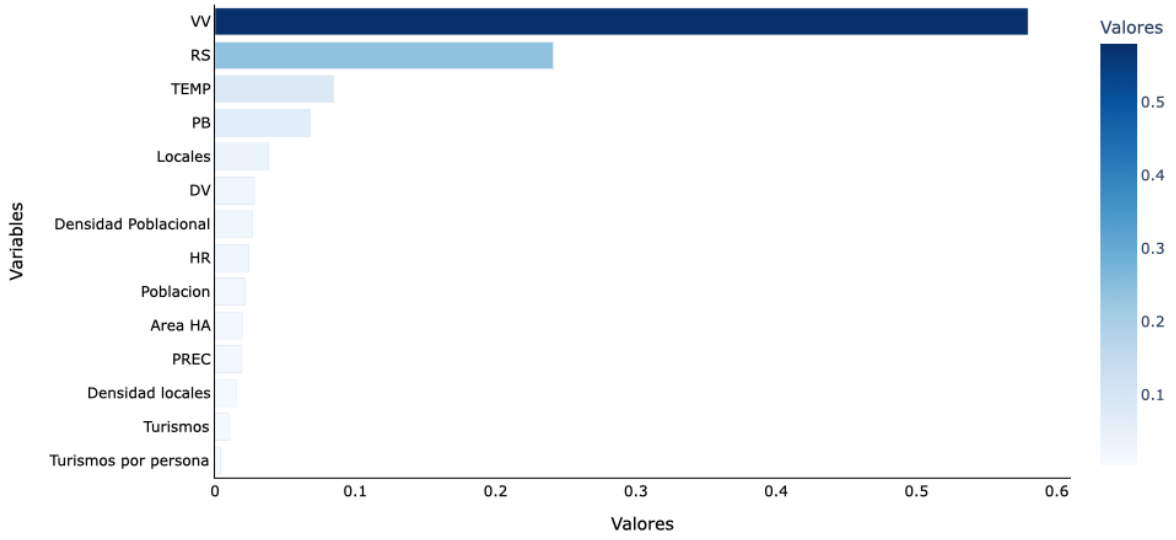

Ilustración 37. Importancia de las variables

A través, de la gráfica podemos observar resultados similares a los observados en el árbol de decisión, en donde la variable con mayor influencia en el modelo fue la velocidad del viento. Otro aspecto por observar a través de la gráfica de barras es el hecho de que las primeras 3 variables son variables meteorológicas, lo que nos permita concluir que las variables meteorológicas tienen una mayor relación con respecto a los valores del ozono en el ambiente que las variables con información demográfica.

Capítulo 5. CONCLUSIONES Y FUTURAS LÍNEAS DE TRABAJO

A partir del análisis y la investigación realizada se pudieron obtener una serie de conclusiones a partir de los objetivos planteados al comienzo del trabajo. En primer lugar, los datos obtenidos con respecto a la calidad del aire de Madrid y sus datos climatológicos históricos extraídos del portal de datos abiertos de ayuntamiento de Madrid mostraron una alta calidad, tomando en cuenta diversos aspectos como completitud, integridad y coherencia, acompañados de la documentación necesaria para un buen entendimiento de los datos.

Partiendo del procesamiento y análisis de los datos fuimos capaces de observar cuales eran los barrios de Madrid (con estaciones meteorológicas o de calidad del aire) con mayor densidad poblacional, lo cuales fueron barrio El pilar, comillas y Argüelles. Por otro lado, el barrio con la mayor densidad de locales se determinó que era sol, y lo barrios con mayor relación entre turismos y personas fueron el barrio recoletos y el viso. De igual manera, al estudiar las relaciones lineales que existían entre nuestras variables, se determinó que existe una mayor relación lineal entre las variables meteorológicas y el *target* (contaminantes del aire) que entre las variables demográficas y las variables de salida.

Al momento de implementar distintos algoritmos comúnmente utilizados para resolver problemas de regresión como es el caso de este trabajo, se obtuvieron las mejores métricas utilizando algoritmos basados en arboles de decisión, en contrariedad a algoritmos más sencillos como lo es la regresión lineal. Estos resultados se podrían deber al hecho que la relación que existe entre nuestras variables de entrada y nuestras variables de salida no consiste simplemente en una relación lineal, y los árboles de decisión son capaces de sobrellevar esta relación no lineal. El contaminante con el cual se obtuvieron los mejores resultados fue el ozono, lo que nos puede llevar a la conclusión que es un contaminante cuya presencia en el aire se puede ver altamente influenciada por las condiciones climatológicas.

Como último aspecto fuimos capaces de determinar la variable más influyente en nuestro modelo de predicción, la velocidad del aire. Este dato meteorológico tiene una gran influencia en la contaminación del aire y esto se debe a que ciertos contaminantes, como es el caso del ozono, el cual se ve beneficiado (a aumentar) cuando existe una disminución de la velocidad del viento, ya que dificulta la dispersión del contaminante en la atmosfera.

Existen múltiples análisis e investigaciones que podrían proceder de esta investigación. Como futuras líneas de trabajo que derivan de este trabajo se plantean los siguientes puntos:

- La alimentación diaria al modelo de datos provenientes de la API del portal de datos abiertos de ayuntamiento de Madrid que permita un reentrenamiento con cierta frecuencia del modelo como también la obtención de datos de la contaminación del aire futuros a través del pronóstico del clima.
- La creación de una plataforma que permita visualizar y conocer los valores históricos y futuros de la contaminación del aire en Madrid, siendo los datos futuros los obtenidos por el modelo de predicción.
- La mejora de los modelos planteados en este trabajo y la inclusión de nuevas variables capaces de aportar información relevante para la predicción.
- Un análisis extensivo que permita conocer a profundidad la relación existente entre el clima y la contaminación del aire.

BIBLIOGRAFÍA

Ayuntamiento de Madrid. (25 de Julio de 2022). *datos.madrid.es*. Obtenido de Portal de datos abiertos del Ayuntamiento de Madrid: https://datos.madrid.es/portal/site/egob

Pedhazur, E. J. (1982). *Multiple regression in behavioral research. Explanation and prediction.* New York: Holt, Rinehart and winston.

Pedregosa, F. a. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.