Tabla de contenidos

Abstracto

Este trabajo presenta un estudio de regresión sobre datos de calidad del aire en la ciudad de Madrid, con el objetivo de modelar y predecir la concentración de benceno (C_6H_6) a partir de distintas variables ambientales y contaminantes medidas de forma horaria. El dataset utilizado contiene 9.354 registros e incluye, entre otros atributos, concentraciones de CO, NMHC, NOx y NO_2, respuestas de sensores específicos (serie PT08) y medidas de temperatura, humedad relativa y humedad absoluta. Todo el análisis práctico se ha desarrollado en notebooks de Python, en los que se documentan de forma detallada las fases de exploración, limpieza, modelado y evaluación.

En una primera parte se lleva a cabo una exploración sistemática del conjunto de datos y un preprocesamiento cuidadoso: se eliminan columnas sin contenido informativo, se identifican y sustituyen valores inválidos codificados como –200, y se descartan filas con información incompleta para trabajar con un conjunto consistente de observaciones. A continuación, en el Ejercicio 1, se construyen varios modelos de regresión lineal simple utilizando como predictores individuales las variables con mayor correlación con C6H6(GT). Dichos modelos se evalúan con métricas de regresión estándar (R^2, MAE, MSE y RMSE) y se apoyan en representaciones gráficas para interpretar el ajuste. Los resultados muestran que la señal del sensor PT08.S2(NMHC) ofrece la mejor capacidad predictiva como atributo aislado, llegando a explicar un porcentaje muy elevado de la variabilidad del benceno.

El Ejercicio 2 amplía el enfoque anterior incorporando múltiples variables de forma simultánea y entrenando distintos modelos de regresión, tanto lineales como basados en Random Forests. Se comparan varias configuraciones mediante las mismas métricas de error utilizadas en el ejercicio previo y se estudia la contribución relativa de cada atributo a través de medidas de importancia de variables y del análisis de los coeficientes del modelo. Este segundo bloque permite valorar hasta qué punto los modelos multivariantes capturan la variabilidad de la concentración de C_6H_6, identificar las características que más influyen en su predicción y discutir las principales limitaciones del estudio, así como posibles extensiones para mejorar la capacidad explicativa y predictiva en trabajos futuros.

1 Introducción

En este trabajo práctico se analizan datos de calidad del aire registrados en una ciudad mediante diversos sensores y equipos de medida, con el objetivo de estudiar y modelar la concentración de benceno en el aire urbano (C6H6(GT)). A partir de un conjunto de observaciones horarias que incluye tanto contaminantes (CO, NMHC, NOx, NO_2) como respuestas de sensores específicos y variables meteorológicas (temperatura, humedad relativa y absoluta), se investiga en qué medida estas variables permiten explicar y predecir el comportamiento del benceno, un compuesto de especial relevancia por su impacto en la salud.

El trabajo se organiza en dos ejercicios complementarios. En el Ejercicio 1 se aborda el problema desde la perspectiva de la regresión simple, construyendo modelos de regresión lineal basados en un único predictor. Para ello se seleccionan las variables con mayor correlación con C6H6(GT) y se ajusta un modelo por cada una de ellas, comparando su capacidad predictiva mediante métricas de error y visualizaciones. Este primer bloque sirve como línea base: permite identificar qué atributo individual (en particular la respuesta del sensor PT08.S2(NMHC)) ofrece la mejor aproximación a la concentración de benceno cuando se trabaja con un solo factor explicativo.

En el Ejercicio 2 se extiende el análisis al marco de la regresión múltiple, incorporando simultáneamente varios atributos como predictores. Se sigue un flujo de trabajo típico de aprendizaje automático: limpieza y preprocesamiento del dataset, análisis exploratorio de las variables, separación en conjuntos de entrenamiento y prueba, entrenamiento de distintos modelos de regresión y evaluación cuantitativa mediante métricas estándar. En esta fase se emplean, entre otros, modelos de regresión lineal, regresión Ridge y Random Forest Regressor, comparando sus resultados y analizando la importancia relativa de las variables más influyentes. Finalmente, se discuten las conclusiones globales del estudio: qué modelos ofrecen el mejor rendimiento, qué atributos tienen mayor impacto en la concentración de benceno y cuáles son las principales limitaciones del enfoque, proponiendo posibles mejoras y extensiones para trabajos futuros.

2 Ejercicio 1

2.1 Descripción del ejercicio y sus objetivos.

El objetivo del Ejercicio 1 es construir y analizar modelos de regresión lineal simple capaces de predecir la concentración de benceno en el aire urbano, representada por la variable objetivo C6H6(GT), a partir de un único atributo predictor. La idea central es estudiar qué variables del conjunto de datos se relacionan de forma más directa y lineal con el benceno y hasta qué punto, por sí solas, pueden explicar su variabilidad.

Para ello, se sigue un flujo de trabajo que incluye: (i) un análisis exploratorio inicial del conjunto de datos, (ii) un preprocesamiento orientado a la limpieza y preparación de las variables, (iii) la selección de los atributos candidatos con mayor correlación con C6H6(GT), (iv) el ajuste de varios modelos de regresión lineal simple —cada uno basado en un único predictor— y (v) la evaluación comparativa de dichos modelos mediante métricas de evaluación como R^2, MSE, MAE y RMSE. Este ejercicio proporciona una primera aproximación al problema y sirve como base de comparación frente al ejercicio posterior de regresión múltiple.

2.2 Descripción del conjunto de datos

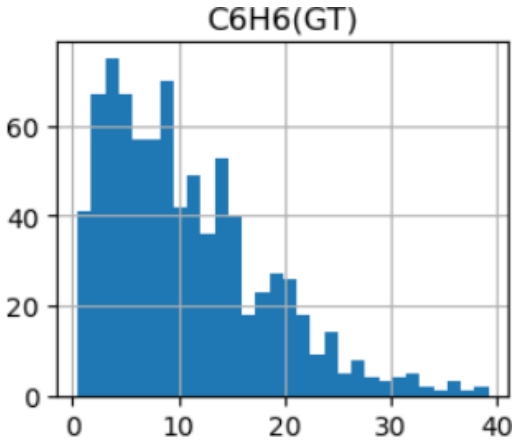
El conjunto de datos utilizado recoge mediciones horarias de calidad del aire en Madrid, con un total de 9.471 observaciones y 17 variables iniciales. Entre los atributos disponibles se incluyen:

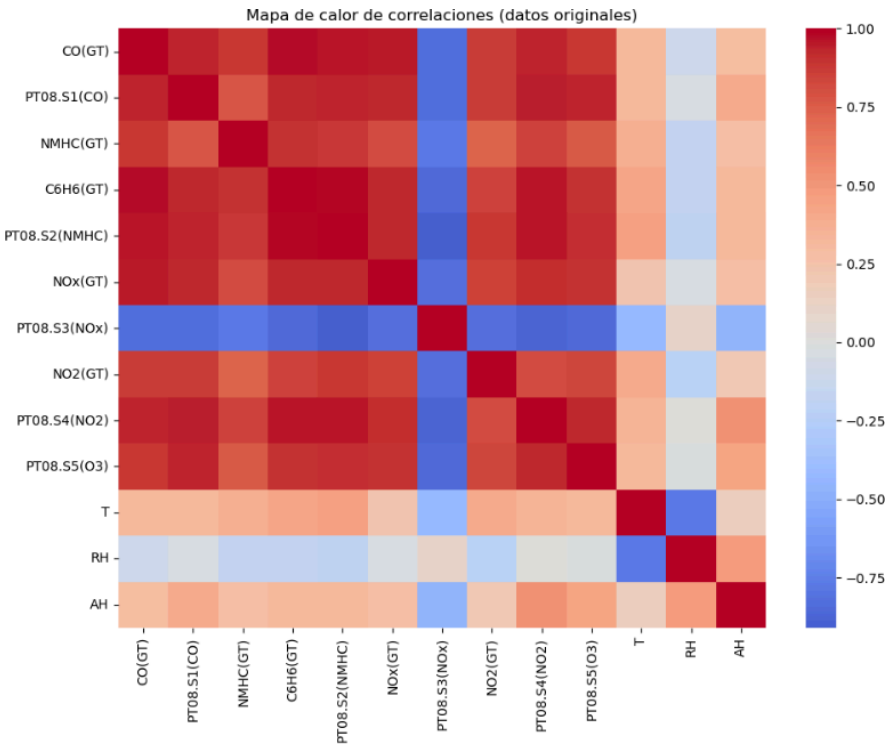
- Información temporal (Date, Time)
- Concentraciones de contaminantes medidas por analizadores de referencia (CO(GT), NMHC(GT), C6H6(GT), NOx(GT), NO2(GT))
- Respuestas de sensores específicos (PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3))
- Variables ambientales (T, RH, AH)

La variable objetivo del ejercicio es C6H6(GT), que representa la concentración horaria de benceno en µg/m³. El resto de variables numéricas se consideran potenciales predictores. Como punto de partida, se realiza un estudio descriptivo del conjunto de datos para examinar su estructura, tamaño, tipos de variables y las estadísticas descriptivas básicas (media, cuartiles y valores mínimo y máximo), con el fin de obtener una visión global inicial. Esta fase exploratoria permitió identificar diversas fuentes de error, como la presencia de un valor sistemático inválido (–200) utilizado como marcador de mediciones incorrectas y la existencia de columnas que únicamente contenían valores faltantes. Mediante las funciones df.head(), df.describe() y df.info(), se llevó a cabo esta inspección inicial que sirvió para detectar los

problemas que posteriormente serían abordados durante el preprocesamiento de los datos antes del entrenamiento de los modelos.

El análisis de distribuciones y correlaciones reveló que algunas variables muestran una relación especialmente fuerte con C6H6(GT), destacando entre ellas la respuesta del sensor PT08.S2(NMHC) y otros atributos asociados a contaminantes gaseosos. Estas variables se identificaron como las principales candidatas a actuar como predictores en los modelos de regresión lineal simple. Para documentar este proceso, se incluyen la distribución de la variable objetivo y la matriz de correlaciones del conjunto de variables numéricas, ilustrando así las técnicas de visualización utilizadas en el trabajo práctico. Cabe destacar que la matriz de correlación desempeñó un papel clave en la selección de las cinco variables con mayor correlación absoluta respecto a C6H6(GT), que posteriormente serían empleadas en la construcción y análisis de los modelos de regresión simple.





2.3 Preparación de datos

El preprocesamiento del conjunto de datos persigue eliminar fuentes de error y dejar el dataframe en un estado adecuado para el entrenamiento de los modelos de regresión. En primer lugar, se eliminan las columnas (variables) Unnamed: 15 y Unnamed: 16, ya que no únicamente presentan valores faltantes y no aportan valor estadístico. Asimismo, se descartan las variables Date y Time al no utilizarse como predictores en este ejercicio y no aportar información directamente interpretable dentro de un modelo de regresión lineal simple.

Durante el análisis descriptivo se detecta que todas las variables numéricas presentan el valor mínimo –200, que no tiene sentido físico dentro del contexto de las magnitudes medidas. Este valor se interpreta como un código de medición inválida o dato ausente. Para tratarlo adecuadamente, se reemplazan todos los valores –200 por NaN en el dataframe. Una vez realizados estos cambios, se procede a eliminar todas las filas que contienen valores faltantes en alguna de las columnas, con el objetivo de trabajar únicamente con observaciones completas y evitar problemas posteriores durante el entrenamiento de los modelos. Aunque este filtrado reduce ligeramente el número de registros disponibles, el tamaño final del conjunto de datos sigue siendo suficientemente grande para el análisis.

En cuanto al tratamiento de variables, no se realiza codificación de variables categóricas ni escalado o normalización, ya que el ejercicio se centra exclusivamente en variables numéricas y el modelo utilizado es una regresión lineal simple. Mantener las variables en su escala original facilita además la interpretación de los coeficientes. Como último paso en la sección de preprocesamiento de datos, para preparar la siguiente sección donde definimos los conjuntos de entrenamiento y testing y entrenamos los modelos, calculamos las correlaciones de todas las variables numéricas con la variable objetivo y elegimos las 5 variables con mayor valor de coeficiente de correlación ('PT08.S2(NMHC)', 'CO(GT)', 'PT08.S4(NO2)', 'PT08.S1(CO)', 'NOx(GT)') para realizar el estudio posterior de estas variables con la variable objetivo mediante regresión simple.

2.4 Experimentos

2.4.1 Definición del experimento

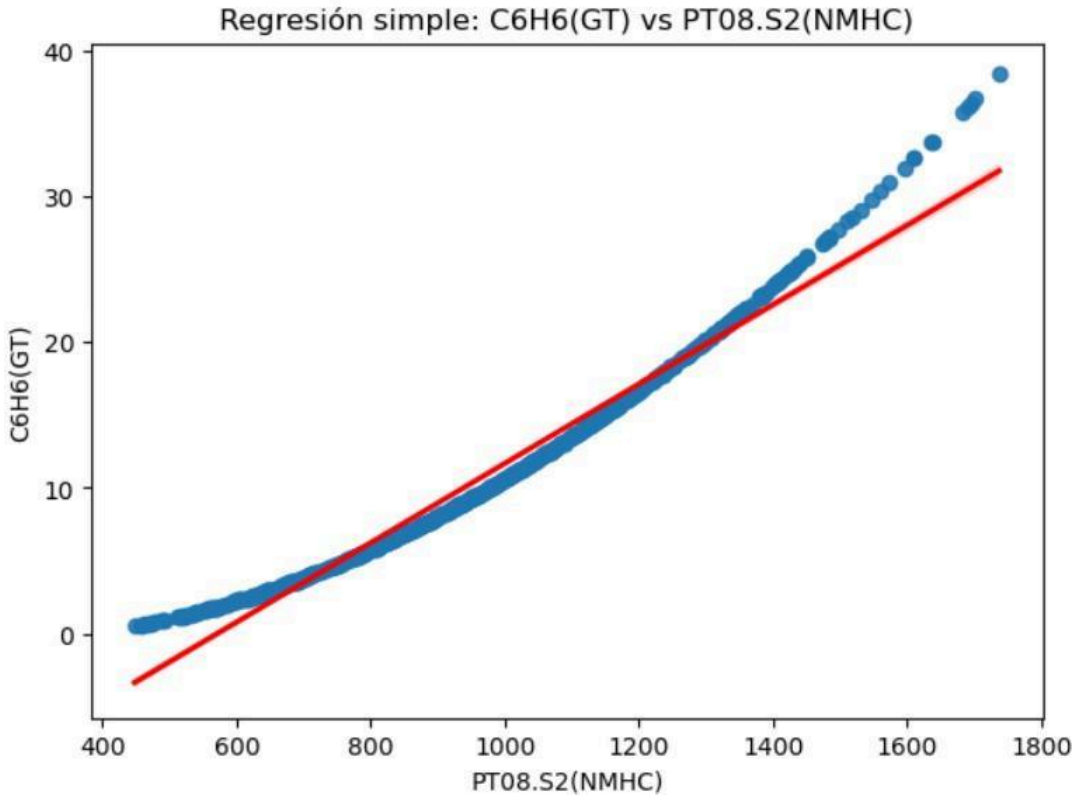
El experimento del Ejercicio 1 consiste en comparar varios modelos de regresión lineal simple, cada uno de ellos basado en un único atributo predictor, con el fin de determinar qué variable es más eficaz para predecir la concentración de benceno. Para ello, se seleccionaron como candidatos los cinco atributos con mayor coeficiente de correlación con la variable objetivo:

- PT08.S2(NMHC),
- CO(GT),
- PT08.S4(NO2),
- PT08.S1(CO),
- NOx(GT).

El conjunto de datos se divide en entrenamiento y prueba mediante un particionado 80%–20%, fijando una semilla aleatoria para garantizar la reproducibilidad. Sobre el conjunto de entrenamiento se ajustan cinco modelos de regresión lineal simple independientes, cada uno utilizando únicamente una de las variables candidatas como predictor.

Para documentar esta fase, se incluyen gráficos de dispersión con la recta de regresión superpuesta para cada uno de los modelos (por ejemplo, C6H6(GT) frente a PT08.S2(NMHC), CO(GT), etc.), lo que permite visualizar la dirección y fuerza de la relación lineal entre cada atributo y la variable objetivo. Como referencia se incluye la gráfica del modelo de regresión simple de la variable PT08.S2(NMHC) contra la variable objetivo, C6H6(GT):

```
Modelo 1: C6H6(GT) ~ PT08.S2(NMHC)
Coeficiente: 0.02724514887468094
Intercepto: -15.56369966593992
```



2.4.2 Evaluación del experimento

Una vez entrenados los modelos, se evalúa su desempeño sobre el conjunto de prueba mediante cuatro métricas estándar de regresión: el coeficiente de determinación R^2, el MAE (Mean Absolute Error), el MSE (Mean Squared Error) y el RMSE (Root Mean Squared Error).

Los resultados muestran diferencias claras entre los cinco modelos. El modelo basado en la variable PT08.S2(NMHC) obtiene el mejor rendimiento, con un valor de R^2 cercano a 0.97 y los valores más bajos de MAE, MSE y RMSE, lo que indica que este predictor es capaz de explicar una fracción muy elevada de la variabilidad de C6H6(GT) y produce errores medios reducidos. Le sigue en desempeño el modelo con CO(GT), que también presenta un ajuste bastante bueno aunque con errores algo mayores y menor capacidad explicativa que PT08.S2(NMHC). Los

modelos construidos a partir de PT08.S4(NO2), PT08.S1(CO) y NOx(GT) muestran un rendimiento inferior, con R^2 más bajos y errores más elevados, lo que refleja una relación lineal menos fuerte con la variable objetivo.

Para facilitar la comparación, se construye una tabla resumen con las métricas de evaluación de cada modelo:

	Predictor	R2	MAE	MSE	RMSE
0	PT08.S2(NMHC)	0.968561	0.990986	1.870566	1.367686
1	CO(GT)	0.948962	1.244345	3.036689	1.742610
2	PT08.S4(NO2)	0.926955	1.672284	4.346105	2.084731
3	PT08.S1(CO)	0.872490	2.155627	7.586674	2.754392
4	NOx(GT)	0.859624	2.091212	8.352230	2.890022

2.5 Preguntas

En relación con la pregunta planteada en el enunciado:

¿Hay alguna variable que sea capaz de predecir de forma precisa la concentración de benzeno?

Los resultados obtenidos en el Ejercicio 1 indican que, entre los atributos analizados, la respuesta del sensor PT08.S2(NMHC) es la variable que mejor predice la concentración de benceno cuando se utiliza de forma aislada en un modelo de regresión lineal simple. El valor de R^2 cercano a 0.97 y los errores MAE y RMSE significativamente bajos muestran que este predictor es capaz de reproducir con gran precisión el comportamiento de C6H6(GT) en el conjunto de prueba.

Aunque otros atributos como CO(GT) también presentan un rendimiento razonable, ninguno alcanza el nivel de precisión obtenido con PT08.S2(NMHC). Por tanto, dentro del marco de la regresión simple y para el conjunto de datos analizado, puede afirmarse que PT08.S2(NMHC) es una variable capaz de predecir de forma precisa la concentración de benceno, mientras que el resto de predictores individuales tienen una capacidad explicativa sensiblemente menor. Esta conclusión se apoya tanto en las métricas cuantitativas como en las visualizaciones obtenidas durante la evaluación de los modelos.

3 Ejercicio 2

3.1 Descripción del ejercicio y sus objetivos.

En el Ejercicio 2 trabajamos con un problema de regresión múltiple sobre datos de calidad del aire. El objetivo principal es predecir la concentración de benceno en el aire, representada por la variable C6H6(GT), utilizando como entrada varias variables medidas por la estación: otros contaminantes, respuestas de sensores y condiciones meteorológicas simples.

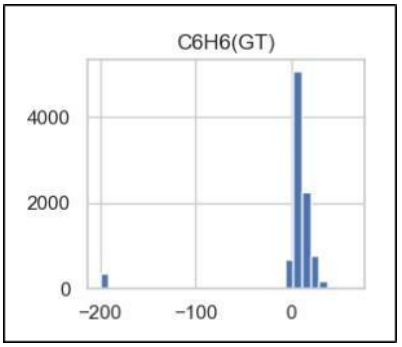
Para conseguirlo, se diseña un flujo de trabajo completo de Machine Learning: primero se limpia y prepara el conjunto de datos, después se entrenan distintos modelos de regresión y finalmente se comparan sus resultados. En concreto, se usan modelos de regresión lineal/Ridge y un modelo de Random Forest Regressor. El interés no es solo obtener buenas predicciones, sino también entender qué variables son más importantes para explicar el comportamiento de C6H6(GT) y qué capacidad real tienen estos modelos para describir la variabilidad del benceno en el aire urbano.

3.2 Descripción del conjunto de datos

El conjunto de datos contiene mediciones horarias de calidad del aire registradas por una estación urbana. Cada fila corresponde a una hora concreta y recoge, entre otros valores, la concentración de varios contaminantes y la respuesta de diferentes sensores electrónicos. En total el dataset original tiene alrededor de 9.500 observaciones y unas 17 columnas, aunque no todas ellas se utilizan en el modelo final.

Entre los atributos disponibles encontramos concentraciones de monóxido de carbono (CO(GT)), óxidos de nitrógeno (NOx(GT), NO2(GT)), así como respuestas de sensores específicos para distintos gases (PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3)).
También se incluyen variables meteorológicas como la temperatura (T), la humedad relativa (RH) y la humedad absoluta (AH). La variable objetivo que queremos predecir es C6H6(GT), que representa la concentración de benceno medida por el equipo de referencia.

En el análisis exploratorio se ha observado la distribución de C6H6(GT) mediante histogramas y estadísticos básicos. Esto permite comprobar el rango de valores más frecuentes y detectar posibles valores atípicos o medidas anómalas.

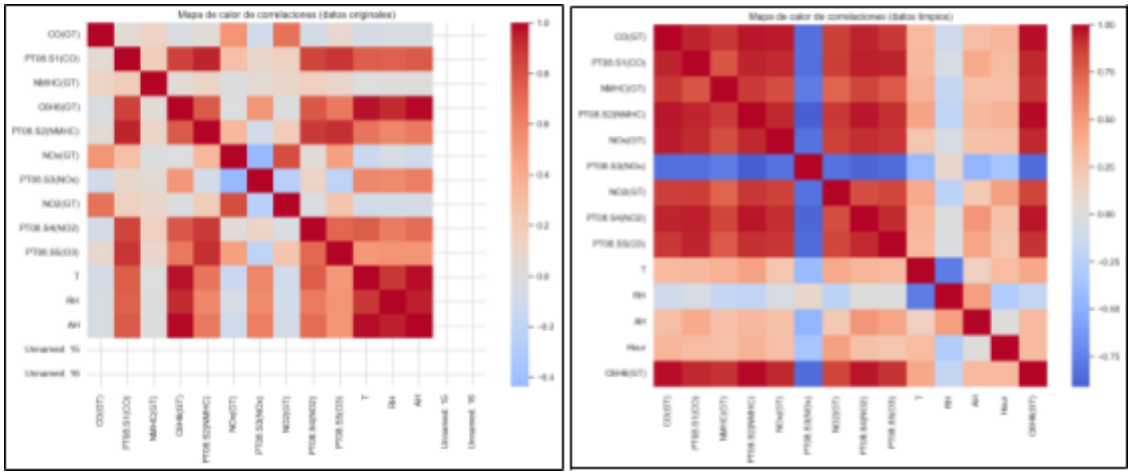


3.3 Preparación de datos

Antes de entrenar los modelos ha sido necesario limpiar y transformar el dataset. En este problema algunos sensores utilizan el valor –200 como código de error, por lo que en primer lugar se han sustituido todos esos valores por nulos (NaN) en las columnas de gases y sensores. También se han identificado columnas completamente vacías (por ejemplo, columnas auxiliares sin datos reales) que se han eliminado porque no aportaban información útil al análisis.

Una vez marcados los valores erróneos, se ha creado un subconjunto de columnas con las variables más relevantes: concentraciones de contaminantes (CO(GT), NOx(GT), NO2(GT)), respuestas de sensores (PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3)), variables meteorológicas (T, RH, AH) y una nueva variable Hour obtenida a partir de la hora del día. La variable C6H6(GT) se ha mantenido como objetivo. Sobre este subconjunto se han eliminado todas las filas que todavía contenían valores nulos en alguna de las columnas seleccionadas, quedándonos con un número más reducido de observaciones, pero completamente limpias.

Para facilitar el análisis de relaciones entre variables se ha calculado la matriz de correlación de este dataset ya limpio. El mapa de calor asociado permite ver qué variables están más correlacionadas entre sí y, en particular, cómo se relaciona C6H6(GT) con el resto de atributos. No se ha aplicado codificación de variables categóricas porque todas las variables usadas son numéricas. Tampoco se ha realizado un escalado específico, ya que los modelos elegidos (regresión lineal/Ridge y Random Forest) pueden trabajar razonablemente bien con las escalas originales en este contexto.

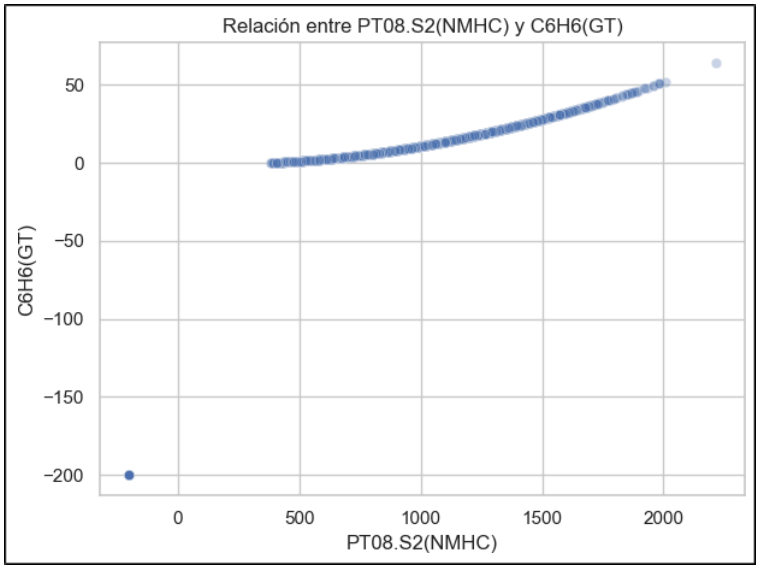


3.4 Experimentos

3.4.1 Definición del experimento

Una vez preparado el conjunto de datos se ha dividido en dos partes: un conjunto de entrenamiento con aproximadamente el 80 % de las observaciones y un conjunto de prueba con el 20 % restante. El conjunto de entrenamiento se utiliza para ajustar los modelos, mientras que el de prueba sirve para evaluar su rendimiento con datos que el modelo no ha visto durante el aprendizaje.

En este ejercicio se han probado dos tipos principales de modelos. El primer grupo lo forman los modelos de regresión basados en optimización: una regresión lineal múltiple estándar y una regresión Ridge. Ambos buscan una relación lineal entre las variables de entrada y C6H6(GT), pero Ridge añade un término de penalización sobre los coeficientes para evitar que se vuelvan demasiado grandes cuando hay mucha correlación entre las variables. El segundo tipo de modelo es un Random Forest Regressor, que combina muchos árboles de decisión y permite capturar relaciones no lineales y posibles interacciones entre variables. Para el Random Forest se han probado varias configuraciones sencillas de hiperparámetros (por ejemplo, distinto número de árboles y profundidad máxima) y se ha elegido la que funciona mejor en el conjunto de prueba.



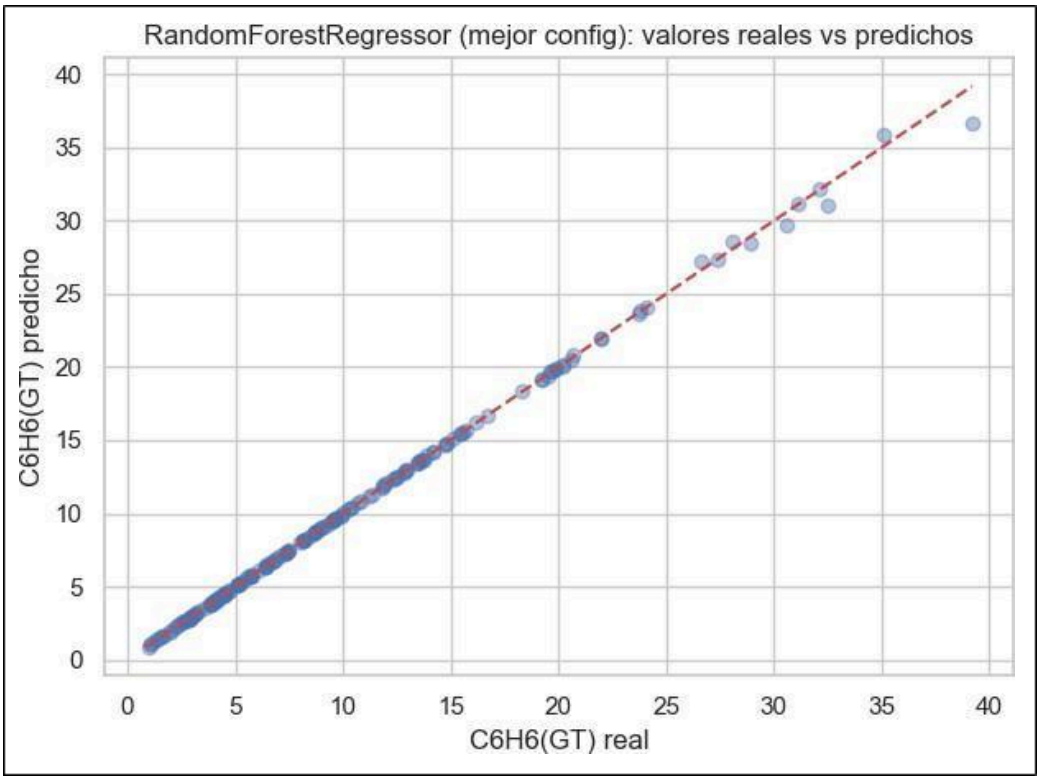
3.4.2 Evaluación del experimento

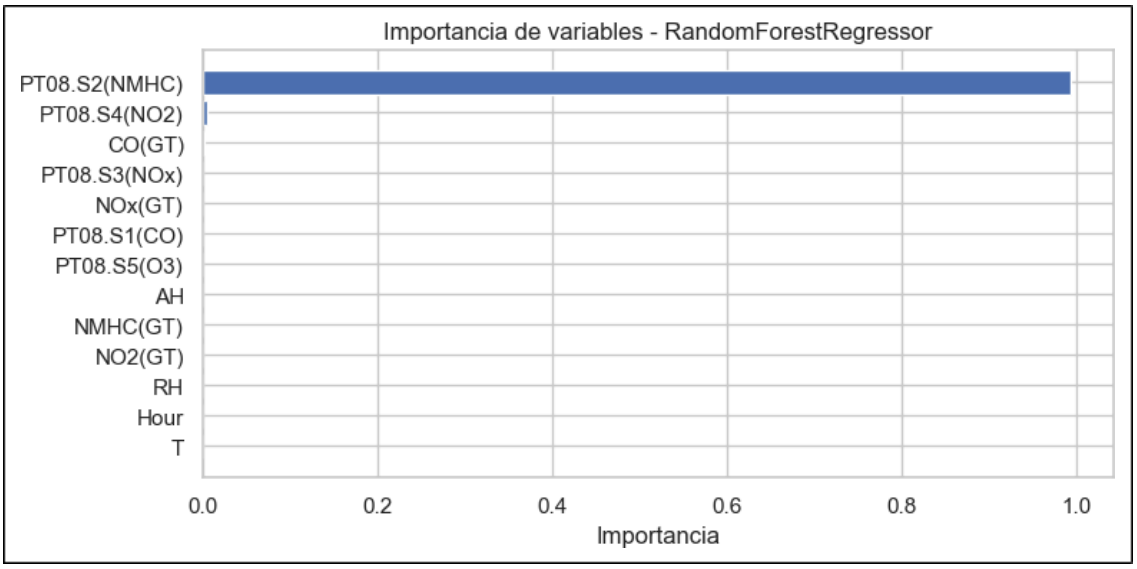
Para comparar los modelos se han utilizado varias métricas de regresión. El MAE (error absoluto medio) mide, en promedio, cuántas unidades se equivoca el modelo en sus predicciones. El MSE (error cuadrático medio) y su raíz cuadrada, el RMSE, penalizan más los errores grandes y permiten ver si hay predicciones muy alejadas del valor real. Por último, el coeficiente de determinación R^2 indica qué porcentaje de la variabilidad de C6H6(GT) logra explicar el modelo a partir de las variables de entrada: un R^2 cercano a 1 significa un ajuste muy bueno, mientras que valores más bajos indican que una parte importante de la variación no se está capturando.

Los resultados muestran que la regresión lineal y la regresión Ridge obtienen ya un rendimiento muy alto, con errores moderados y un R^2 alrededor de 0,99. La regresión Ridge ofrece valores prácticamente idénticos a la regresión lineal, lo que indica que la regularización no cambia mucho el ajuste en este caso. El Random Forest Regressor mejora aún más el rendimiento: consigue un MAE y un RMSE más pequeños y un R^2 cercano a 0,999, es decir, explica aproximadamente el 99,9 % de la variabilidad de C6H6(GT) en el conjunto de prueba. Por este motivo se considera que el Random Forest es el mejor modelo entre los probados.

Para entender mejor el comportamiento del Random Forest se ha analizado también el gráfico de valores reales frente a valores predichos y la distribución de los residuos (la diferencia entre lo real y lo predicho). En estos gráficos se observa que la mayoría de puntos se sitúan muy cerca de la diagonal ideal y que los residuos se concentran alrededor de cero, lo que confirma

el buen ajuste del modelo. Además, las importancias de las variables muestran claramente qué atributos están contribuyendo más a las predicciones.





3.5 Preguntas

En cuanto a los atributos que influyen más en C6H6(GT), el análisis de importancia del Random Forest indica que la variable clave es la respuesta del sensor PT08.S2(NMHC), asociado a los hidrocarburos no metánicos. Esta variable concentra la mayor parte de la importancia del modelo, lo que significa que pequeñas variaciones en esta medida tienen un impacto directo en la predicción de la concentración de benceno. En segundo lugar, aunque con un peso mucho menor, aparecen otras variables como PT08.S4(NO2) y CO(GT), junto con algunas respuestas adicionales de sensores y concentraciones de óxidos de nitrógeno. Las variables meteorológicas (T, RH, AH) o la hora del día (Hour) apenas aportan información al modelo, según las importancias obtenidas.

Respecto al porcentaje de variabilidad explicada, el mejor modelo (Random Forest Regressor) alcanza un R^2 aproximado de 0,999 sobre el conjunto de prueba. Esto significa que el modelo es capaz de explicar alrededor del 99,9 % de la variación observada en C6H6(GT) a partir de las variables de entrada utilizadas. Los modelos lineales también presentan un R^2 muy alto, cercano a 0,993, pero se quedan ligeramente por detrás del Random Forest.

A pesar de estos buenos resultados, el modelo tiene varias limitaciones. La más importante es que, tras la limpieza de datos y la eliminación de filas con valores nulos o código de error, solo se utiliza una fracción del dataset original. Esto reduce el tamaño de la muestra y puede hacer que el modelo se ajuste demasiado a ese subconjunto concreto. Además, los datos proceden de una sola ciudad y de un periodo determinado, por lo que no está claro si el modelo generalizaría bien a otras zonas o épocas del año. También es probable que haya factores externos importantes (como detalles del tráfico, condiciones de viento o fuentes concretas de emisión) que no están presentes en el dataset. Como posibles mejoras se podrían aplicar técnicas de imputación de datos para no perder tantas filas, recoger más variables relevantes y utilizar validación cruzada para obtener una estimación más robusta del rendimiento.

3 Conclusiones

3.1 Conclusiones generales

El trabajo realizado ha permitido recorrer de manera completa el ciclo de un proyecto de análisis de datos aplicado a un problema real de calidad del aire, desde la exploración inicial hasta la construcción y evaluación de modelos de regresión. A partir de un dataset con mediciones horarias de contaminantes, respuestas de sensores y variables ambientales, se ha visto la importancia de dedicar una parte significativa del esfuerzo a la comprensión y limpieza de los datos: identificar códigos de error sistemáticos como el valor –200, detectar columnas sin información útil y trabajar únicamente con observaciones consistentes ha sido fundamental para garantizar la validez de los resultados obtenidos en las fases posteriores de modelado.

En el Ejercicio 1, centrado en la regresión simple, se ha utilizado este conjunto de datos ya limpio para analizar la relación entre la concentración de benceno y distintos atributos individuales. La selección de las variables con mayor correlación con C6H6(GT) y el ajuste de varios modelos de regresión lineal simple han permitido cuantificar hasta qué punto un único predictor puede aproximar el comportamiento del benceno. Los resultados han mostrado que la respuesta del sensor PT08.S2(NMHC) destaca claramente sobre el resto, explicando una proporción muy elevada de la variabilidad de C6H6(GT) y proporcionando errores de predicción relativamente bajos. Este ejercicio ha servido como línea base, demostrando que ciertos sensores son extremadamente informativos por sí solos, pero también que otros atributos tienen una capacidad explicativa más limitada cuando se consideran de manera aislada.

El Ejercicio 2 ha ampliado esta perspectiva incorporando múltiples atributos simultáneamente y aplicando técnicas de regresión múltiple. Tras la preparación del conjunto de datos para este nuevo escenario, se han probado tanto enfoques lineales (regresión lineal y Ridge) como un modelo basado en árboles (Random Forest Regressor). La comparación de las métricas de evaluación ha puesto de manifiesto que, aunque los modelos lineales ya alcanzan un rendimiento muy alto, los modelos de tipo ensemble son capaces de capturar relaciones más complejas entre las variables y mejorar todavía más la precisión de las predicciones. Además, el análisis de la importancia de las variables ha permitido interpretar los modelos y conectar los resultados con el contexto del problema, identificando qué sensores y contaminantes están más estrechamente relacionados con la presencia de benceno en el aire. En conjunto, ambos ejercicios muestran cómo, partiendo de modelos simples y avanzando hacia enfoques más complejos, es posible obtener una visión progresivamente más rica y matizada del fenómeno estudiado, así como detectar las limitaciones actuales y posibles líneas de mejora para trabajos futuros.

3.2 Conclusiones específicas sobre los ejercicios

En el Ejercicio 1 se ha comprobado que, incluso trabajando con modelos de regresión lineal simple y un único predictor, es posible obtener una capacidad predictiva muy elevada para la concentración de benceno C6H6(GT). Entre las variables analizadas, la respuesta del sensor PT08.S2(NMHC) destaca claramente como mejor atributo individual: su modelo asociado alcanza un R^2 cercano a 0,97 y presenta los errores MAE y RMSE más bajos, lo que indica que este sensor recoge información fuertemente relacionada con la presencia de benceno en el aire urbano. Otros predictores, como CO(GT) o PT08.S4(NO2), también muestran cierta capacidad explicativa, pero con un rendimiento sensiblemente inferior al obtenido con PT08.S2(NMHC).

No obstante, este ejercicio también pone de relieve las limitaciones de los modelos de regresión simple. Trabajar con un único predictor implica que cualquier componente de la variabilidad de C6H6(GT) que dependa de la interacción entre varias variables o de relaciones no lineales no puede ser capturada por estos modelos. Además, el buen ajuste obtenido está condicionado al contexto específico del dataset utilizado (una única fuente de datos, periodo y localización concretos). Por tanto, aunque el Ejercicio 1 demuestra que ciertos sensores, como PT08.S2(NMHC), son muy informativos por sí solos, también motiva la necesidad de recurrir a enfoques de regresión múltiple y modelos más flexibles, como los explorados en el Ejercicio 2, para obtener una visión más completa.

Centrándonos en el Ejercicio 2, se ha demostrado que es posible predecir la concentración de benceno C6H6(GT) con gran precisión usando únicamente las medidas de otros contaminantes, las respuestas de los sensores y algunas variables meteorológicas simples. Los modelos lineales han logrado un R^2 muy alto, y el Random Forest Regressor ha mejorado aún más estos resultados, alcanzando un R^2 cercano a 0,999 y errores de predicción muy pequeños. Esto indica que la información recogida por los sensores, especialmente la asociada a hidrocarburos no metánicos, contiene prácticamente todo lo necesario para describir el comportamiento de C6H6(GT) en el conjunto de datos analizado.

Sin embargo, el ejercicio también pone de manifiesto aspectos que se podrían investigar más. El hecho de trabajar con un número relativamente pequeño de observaciones limpias y con datos de una sola estación hace que el modelo pueda estar muy adaptado a ese escenario concreto. Sería interesante repetir el estudio con más datos, de diferentes localizaciones y periodos, para comprobar hasta qué punto el modelo y las conclusiones se mantienen. También se podrían explorar otras técnicas de regresión y otros esquemas de validación (por ejemplo, validación cruzada o series temporales) para obtener una evaluación más robusta. Por último, incorporar nuevas variables externas relacionadas con el tráfico, la meteorología avanzada o la geografía urbana podría ayudar a entender mejor las causas de las variaciones en la concentración de benceno y mejorar aún más la capacidad explicativa del modelo.

Bibliografía

[1]
G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning with Applications in R (2.ª edición), Springer, 2021.

[2]
A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd Edition), O'Reilly Media, 2019.

[3]
W. McKinney, Python for Data Analysis: Data Wrangling with pandas, NumPy, and IPython (2nd Edition), O'Reilly Media, 2017.

[4]
J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media, 2016.

[5]
Canvas de Aprendizaje Estadístico y Data Mining, Apuntes de la asignatura, Universidad Francisco de Vitoria, plataforma Canvas, 2025.