Primera Entrega Corregida (Comentarios):

La profesora comentó que el trabajo presentaba varias deficiencias formales y conceptuales que debían de ser corregidas. Comentaba que en el archivo *main* no se encontraba actualizado y la estructura de carpetas no seguía las especificaciones establecidas, lo que dificulta la comprensión y evaluación del proyecto. Asimismo, se observan numerosas erratas que afectan a la claridad y al rigor académico del documento. Desde un punto de vista metodológico, no quedaba claramente definido el horizonte temporal con el que se pretende predecir los valores, un aspecto clave que condiciona de manera significativa el diseño y la construcción del modelo. En esta misma línea, tampoco se ha definido adecuadamente la variable *target*, lo cual supone una carencia fundamental, ya que la correcta formulación del problema de predicción es un requisito imprescindible para el desarrollo coherente y válido del modelo.

Link repositorio

https://github.com/sebastiangs1001-python/MachineL

Participantes	Link Repositorio	Correo Institucional:
Sebastian Gutierrez Sanchez	github.com/sebastiangs1001-python	sebastian.gutierrez@cunef.edu
Pablo Noe Fernandez Gallegos	github.com/pablofdzg0-cmyk	p.fernandezgallegos@cunef.edu

Descripción del problema del negocio:

La contaminación del aire en las ciudades es un problema importante porque afecta a la salud (más problemas respiratorios y cardíacos) y obliga a tomar decisiones como restricciones de tráfico o avisos a la población. Para eso hacen falta datos fiables de contaminantes como CO y NO_2. Las estaciones oficiales miden muy bien, pero son pocas y son muy caras, así que la información es limitada. Han aparecido sensores “low-cost” que se pueden poner en muchos sitios, pero sus lecturas son ruidosas y dependen mucho de la temperatura y humedad, por lo que no se pueden usar directamente.

En este trabajo usamos modelos de Machine Learning para calibrar estos sensores y transformar sus señales en estimaciones más fiables de contaminación, además de predecir niveles de calidad del aire y poder anticipar posibles episodios de mayor contaminación.

Origen del dataset

Dataset utilizado: Air Quality Data Set
(https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set/data)

El dataset empleado en este proyecto procede de la plataforma Kaggle, donde se encuentra publicado bajo el nombre Air Quality Data Set por el usuario fedesoriano. Este dataset Recoge mediciones horarias realizadas por una estación de monitorización urbana y una serie de sensores químicos (PT08) y analizadores de referencia, que registran concentraciones de distintos contaminantes (como CO, NO_2, benceno, etc.), junto con variables meteorológicas (temperatura, humedad relativa y absoluta) y la información de fecha y hora.

Ventajas del dataset:

1. El dataset presenta un alto volumen de observaciones, lo que incrementa la robustez estadística de los análisis y permite capturar variaciones temporales con un elevado nivel de detalle.
2. Incluye múltiples variables ambientales y de sensores, lo que posibilita el estudio conjunto de distintos factores que influyen en la calidad del aire desde una perspectiva multivariante.
3. Los datos reflejan condiciones reales de medición en entornos urbanos, lo que aporta relevancia práctica a las conclusiones obtenidas.

Desventajas del dataset:

1. Al tratarse de datos procedentes de sensores, el dataset puede contener ruido, valores atípicos o mediciones erróneas, lo que hace necesario aplicar técnicas de limpieza y preprocesamiento antes del análisis.
2. La ausencia de una vinculación directa con estaciones oficiales de control ambiental limita la validación externa de la precisión absoluta de las mediciones.
3. La presencia de valores faltantes en algunas variables puede afectar a determinados análisis, obligando a tomar decisiones metodológicas sobre imputación o exclusión de datos.

Planteamiento del problema:

Tipos de Modelos a utilizar:

En este contexto académico, nos percatamos que el uso de modelos supervisados era la mejor opción, trabajaremos con modelos de:

1. Regresión Lineal (BaseLine Model)
2. Stochastic Gradient Descent
3. Random Forest

Explicación de las variables:

1. **Date**: Fecha de la medición, en formato día/mes/año.
2. **Time**: Promedio de la hora de la medición, en formato hora.minuto.segundo.
3. **CO(GT):** Concentración horaria “verdadera” de monóxido de carbono (CO) medida por un analizador de referencia. Unidad: mg/m³.
4. **PT08.S1(CO):** Señal promedio horaria del sensor S1 (basado en óxido de estaño), diseñado principalmente para detectar CO. Es una lectura del sensor en unidades instrumentales.
5. **NMHC(GT):** Concentración horaria “verdadera” de hidrocarburos no metánicos (NMHC) medida por analizador de referencia. Unidad: µg/m³.
6. **C6H6(GT):** Concentración horaria “verdadera” de benceno (C_6H_6), un contaminante orgánico, medida por analizador de referencia. Unidad: µg/m³.
7. **PT08.S2(NMHC):** Señal promedio horaria del sensor S2 (basado en titania), nominalmente orientado a la detección de NMHC. Lectura en unidades instrumentales.
8. **NOx(GT):** Concentración horaria “verdadera” de óxidos de nitrógeno totales (NOx) medida por analizador de referencia. Unidad: ppb.
9. **PT08.S3(NOx):** Señal promedio horaria del sensor S3 (óxido de tungsteno), diseñado para ser sensible a NOx.
10. **NO2(GT):** Concentración horaria “verdadera” de dióxido de nitrógeno (NO_2) medida por analizador de referencia. Unidad: µg/m³.
11. **PT08.S4(NO2):** Señal promedio horaria del sensor S4 (óxido de tungsteno), nominalmente orientado a la detección de NO_2.
12. **PT08.S5(O3):** Señal promedio horaria del sensor S5 (óxido de indio), diseñado para ser sensible al ozono (O_3).
13. **T:** Temperatura del aire en grados Celsius (°C), promedio horario.
14. **RH:** Humedad relativa del aire en porcentaje (%), promedio horario.
15. **AH:** Humedad absoluta del aire, es decir, cantidad de vapor de agua por volumen de aire. Unidad habitual: g/m³.

Variables Excluidas y Tratamiento de Datos :

Vamos a tratar los valores nulos de la siguiente forma: en primer lugar, eliminaremos las columnas Unnamed: 15 y Unnamed: 16, ya que presentan un 100 % de valores nulos y no aportan información útil al análisis. Para el resto de variables, el porcentaje de valores faltantes es muy bajo (se tiene 114 de NaaN lo que implica un ≈1,2 % en cada columna de 100%), por lo que eliminaremos las filas que contengan NaN. Permitiéndonos mantener prácticamente el mismo tamaño muestral.

Los datos nulos son indicados con el valor -200. Por Lo que lo eliminaremos para evitar añadir ruido. Pues en el dataset se utiliza como código para “medición ausente o inválida” y no como valores reales propios de la variable.
La variable NMHC(GT) tiene alrededor de un 90 % de valores ausentes, por lo que se considerará inutilizable para el modelado y se eliminará del conjunto de variables.

Para el modelo de regresión con CO(GT) como variable objetivo, eliminaremos las filas que tengan CO(GT) vacío. Las variables predictoras seleccionadas (sensores químicos (PT08) variables meteorológicas y temporales) tienen en torno a un 4 % de valores faltantes; dado que es un porcentaje bajo, eliminaremos también las filas que tengan NaN en cualquiera de estas columnas, de forma que el modelo se entrene solo con observaciones completas.

Se creó una tabla para mostrar la distribución de todas las variables numéricas del dataset mediante histogramas con estimación de densidad (KDE), lo que permite identificar la forma de las distribuciones, la presencia de valores atípicos y posibles asimetrías.
Tras realizar la limpieza del dataset disponemos ya de un conjunto de datos consistente listo para el análisis. A partir de este punto podemos comenzar a plantear y entrenar los distintos modelos de machine learning (específicamente Regresión Lineal, Random Forest y XG Boost), definiendo claramente la variable objetivo y las características (sensores, variables meteorológicas y temporales) que utilizaremos como entrada en cada uno de ellos.

Variable Target:

La variable *CO(GT)* se establece como variable objetivo, al representar el contaminante cuya predicción constituye el propósito principal del modelo. Las variables explicativas seleccionadas incluyen las respuestas de los sensores químicos (PT08), junto con variables meteorológicas y temporales, las cuales aportan información contextual relevante para la modelización de la concentración de CO.

Con el fin de evitar problemas de *data leakage* y garantizar un planteamiento realista del modelo, no se incorporan como variables predictoras otras mediciones directas de contaminantes tales como *NMHC(GT)*, *C6H6(GT)*, *NOx(GT)* y *NO2(GT)*, dado que su inclusión podría introducir información que no estaría disponible en un escenario de predicción real.

Ingeniería y selección de variables (Feature Engineering)

Como parte del proceso de modelización, se llevó a cabo una fase de ingeniería de variables con el objetivo de adaptar las características originales del dataset al marco de la regresión lineal y mejorar la capacidad explicativa del modelo.

Las variables utilizadas incluyen principalmente las respuestas de los sensores químicos PT08 (PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2) y PT08.S5(O3)), junto con variables meteorológicas (T, RH, AH) y componentes temporales derivados de la fecha y la hora de medición. Estas variables se mantienen en su forma continua original, ya que representan magnitudes físicas cuya relación con la concentración de CO puede aproximarse de manera lineal.

Dado que las variables presentan diferentes escalas y órdenes de magnitud, se aplicó un proceso de escalado para homogeneizar los rangos de los predictores. Esta transformación

permite evitar que variables con valores numéricos más elevados tengan un peso desproporcionado en el ajuste del modelo y facilita una interpretación más equilibrada de los coeficientes de la regresión.

En el caso de las variables temporales, se realizó su transformación a formato numérico, permitiendo al modelo capturar patrones asociados a la variabilidad horaria y diaria de la concentración de monóxido de carbono. Estas transformaciones resultan especialmente relevantes en un contexto de datos ambientales con estructura temporal.

No se aplicaron transformaciones logarítmicas ni agregaciones adicionales, con el fin de mantener la interpretabilidad del modelo y evaluar en primera instancia la capacidad de un enfoque lineal sobre las variables originales del sistema de sensores.

Modelado y Comparación de Modelos

Modelo A: Regresión Lineal con Predicción de Variable CO(GT)

Para abordar el problema de estimación de la concentración de monóxido de carbono (*CO(GT)*), se ha seleccionado como modelo base un enfoque de regresión lineal, debido a su simplicidad, interpretabilidad y adecuación inicial al comportamiento observado en los datos.

El análisis exploratorio previo muestra que la variable *CO(GT)* presenta una distribución continua y asimétrica, con una concentración elevada de valores bajos y una cola hacia valores más altos. Este comportamiento es característico de variables ambientales y resulta compatible con un enfoque de regresión, en el que se modeliza una variable continua a partir de un conjunto de predictores.

Asimismo, a partir de la matriz de gráficos bivariantes (*pairplot*) entre *CO(GT)* y las señales de los sensores químicos *PT08*, se observan relaciones claras y consistentes entre la variable objetivo y algunas de las respuestas de los sensores, especialmente *PT08.S1(CO)*. Aunque estas relaciones no son perfectamente lineales y presentan dispersión —esperable en datos procedentes de sensores—, los patrones observados sugieren que un modelo lineal puede capturar una parte significativa de la relación entre las señales de los sensores y la concentración real de CO.

La regresión lineal se utiliza, por tanto, como modelo de referencia (baseline) que permite:

1. Establecer un punto de comparación frente a modelos más complejos.
2. Evaluar hasta qué punto las relaciones observadas pueden explicarse mediante combinaciones lineales de las variables explicativas.
3. Facilitar la interpretación de los coeficientes y la influencia relativa de cada sensor y variable ambiental.

Este enfoque resulta especialmente adecuado en una primera fase del análisis, ya que permite validar la coherencia del planteamiento del problema antes de introducir modelos no

lineales o de mayor complejidad. Entonces para el Modelo B de Regresión Lineal que se presenta a continuación, decidimos eliminar la variable PT08.S1(CO).

Modelo B: Regresión Lineal con Predicción de Variable CO(GT) sin PT08.S1(CO).

Además del modelo base, se plantea un segundo modelo de regresión lineal (Modelo B) en el que se excluye la variable PT08.S1(CO) del conjunto de predictores. El objetivo de este experimento es evaluar la dependencia del modelo respecto al sensor más directamente asociado al monóxido de carbono y analizar la capacidad predictiva del resto de señales disponibles.

Durante el análisis exploratorio se observó que PT08.S1(CO) presenta una relación especialmente fuerte con la variable objetivo *CO(GT)*, lo que sugiere que este sensor actúa como un predictor dominante dentro del modelo. Si bien esta característica puede mejorar el rendimiento predictivo, también puede ocultar la contribución del resto de sensores y variables ambientales.

Al excluir PT08.S1(CO), el modelo se entrena únicamente con:

- Las señales restantes de los sensores químicos (*PT08.S2(NMHC)*, *PT08.S3(NOx)*, *PT08.S4(NO2)* y *PT08.S5(O3)*).
- Las variables meteorológicas (*T*, *RH*, *AH*).

Este planteamiento permite analizar si el modelo es capaz de inferir la concentración de CO a partir de señales indirectas, simulando un escenario más restrictivo en el que el sensor específicamente diseñado para CO no está disponible o presenta fallos. De este modo, se evalúa la robustez del enfoque de regresión lineal y la redundancia informativa existente entre los distintos sensores del sistema.

La comparación entre el modelo base y el Modelo B permite cuantificar el impacto de PT08.S1(CO) sobre el rendimiento del modelo y aporta información relevante para la selección final de variables y la interpretación de los resultados.

Conclusión comparativa entre el Modelo A y el Modelo B

Al comparar el Modelo A (que incluye el sensor PT08.S1(CO)) con el Modelo B (sin este sensor), se observa que el primero ofrece mejores resultados en la predicción de la concentración de monóxido de carbono (*CO(GT)*).

El Modelo A alcanza un valor de R^2 de 0.86, lo que indica que explica una gran parte de la variabilidad de CO(GT). Además, presenta errores más bajos tanto en MAE como en RMSE, lo que se traduce en predicciones más ajustadas a los valores reales.

En el caso del Modelo B, al eliminar el sensor PT08.S1(CO), el rendimiento disminuye ligeramente (R^2 = 0.83, con errores algo mayores). Esto sugiere que este sensor aporta información importante para estimar CO(GT). Sin embargo, el descenso no es drástico, ya que el modelo sigue siendo capaz de explicar más del 80 % de la variabilidad del contaminante utilizando únicamente el resto de sensores y las variables meteorológicas.

En conjunto, estos resultados indican que, aunque PT08.S1(CO) mejora la precisión del modelo, no es el único sensor relevante. El resto de variables también contribuyen de forma significativa a la predicción, lo que hace que el enfoque sea razonablemente robusto e útil incluso en escenarios en los que este sensor no esté disponible o presente fallos.

Modelo C: Stochastic Gradient Descent con Modelo A (Sin Tuning).

Además del modelo de regresión lineal clásico, se entrenó un modelo de regresión lineal basado en Stochastic Gradient Descent (SGD) utilizando el mismo conjunto de variables que en el Modelo A, es decir, incluyendo el sensor PT08.S1(CO) junto con el resto de sensores químicos y temporales.

El uso de SGD permite entrenar el modelo de forma iterativa, ajustando los coeficientes a partir de pequeños subconjuntos de datos en lugar de utilizar todo el dataset en cada iteración. Este enfoque resulta especialmente útil en contextos con un número elevado de observaciones y permite evaluar si una optimización incremental es capaz de alcanzar resultados comparables al modelo lineal tradicional.

Desde el punto de vista conceptual, este modelo mantiene la misma hipótesis de linealidad que la regresión lineal base, pero introduce un método de entrenamiento diferente, lo que permite analizar la estabilidad del modelo y su sensibilidad al proceso de optimización. De este modo, el modelo SGD se plantea como una alternativa al modelo base, no tanto por introducir mayor complejidad, sino por ofrecer una forma distinta de ajuste que puede resultar más eficiente o flexible en determinados escenarios.

La comparación de los resultados obtenidos con SGD frente a la regresión lineal clásica permite evaluar si el rendimiento del modelo depende principalmente de la formulación lineal del problema o del método utilizado para estimar los coeficientes, aportando así una visión más completa del proceso de modelización.

Modelo D: Stochastic Gradient Descent con Modelo B (Sin Tuning).

Como extensión del modelo basado en Stochastic Gradient Descent, se entrenó un Modelo D utilizando el mismo enfoque de optimización incremental, pero excluyendo el sensor PT08.S1(CO) del conjunto de variables explicativas, siguiendo el planteamiento del Modelo B.

Este modelo tiene como objetivo evaluar si el método de entrenamiento mediante SGD es capaz de mantener un rendimiento aceptable cuando se elimina el sensor más directamente relacionado con la variable objetivo. De este modo, se analiza tanto la influencia del conjunto de variables seleccionado como la robustez del algoritmo de optimización frente a un escenario más restrictivo.

El Modelo D se entrena utilizando únicamente las señales restantes de los sensores químicos (*PT08.S2(NMHC)*, *PT08.S3(NOx)*, *PT08.S4(NO2)* y *PT08.S5(O3)*), junto con las variables meteorológicas (*T*, *RH*, *AH*) y las variables temporales. Al igual que en el resto de

modelos, se mantiene la hipótesis de una relación aproximadamente lineal entre las variables explicativas y la concentración de monóxido de carbono.

La comparación entre este modelo y su equivalente con PT08.S1(CO) permite analizar hasta qué punto la pérdida de información derivada de la exclusión de dicho sensor puede compensarse mediante el resto de variables y si el enfoque basado en SGD ofrece ventajas adicionales frente a la regresión lineal clásica en este contexto.

Conclusión comparativa entre el Modelo C y el Modelo D

Los resultados obtenidos con los modelos basados en Stochastic Gradient Descent refuerzan las conclusiones extraídas previamente con la regresión lineal clásica. El Modelo C, que incluye el sensor *PT08.S1(CO)*, presenta un mejor rendimiento, con valores de MAE ≈ 0.34 y RMSE ≈ 0.49, lo que indica predicciones más cercanas a los valores reales de CO(GT).

En el Modelo D, donde se excluye este sensor, los errores aumentan (MAE ≈ 0.40, RMSE ≈ 0.53), lo que confirma que *PT08.S1(CO)* aporta información relevante para la predicción de la concentración de monóxido de carbono. Aun así, el modelo sigue ofreciendo resultados razonables, lo que indica que el resto de sensores y las variables meteorológicas conservan una capacidad explicativa significativa.

La comparación entre ambos modelos sugiere que el método de entrenamiento mediante SGD es capaz de capturar adecuadamente las relaciones presentes en los datos, obteniendo resultados muy similares a los del modelo lineal tradicional. La principal diferencia en el rendimiento no proviene del algoritmo de optimización, sino de la selección de variables, y en particular de la inclusión o exclusión del sensor PT08.S1(CO).

En conjunto, estos resultados confirman que el enfoque lineal es estable frente al método de entrenamiento utilizado y que la pérdida de rendimiento al eliminar PT08.S1(CO) es consistente tanto en la regresión lineal clásica como en el modelo basado en SGD.

Modelo E: SGD con Modelo A-Tuneado:

Como extensión del modelo basado en Stochastic Gradient Descent, se desarrolló una versión ajustada (tuneada) del Modelo A, manteniendo el mismo conjunto de variables explicativas que incluye el sensor PT08.S1(CO) junto con el resto de sensores químicos, variables meteorológicas y componentes temporales.

El objetivo de este ajuste fue mejorar el rendimiento del modelo mediante la optimización de sus hiper parámetros, analizando si un entrenamiento más controlado permite obtener predicciones más precisas sin modificar la estructura del modelo ni el conjunto de variables utilizadas.

Para ello, se aplicó un proceso de búsqueda sistemática de hiper parámetros, evaluando distintas configuraciones relacionadas con la velocidad de aprendizaje y el número de iteraciones del algoritmo. Este proceso se realizó utilizando un esquema de validación que

respeta el orden temporal de los datos, garantizando que el modelo se entrena únicamente con observaciones anteriores a las utilizadas para su evaluación, lo que resulta especialmente relevante en un contexto de datos ambientales con dependencia temporal.

El ajuste de hiper parámetros permitió identificar una configuración más estable del modelo, mejorando el equilibrio entre velocidad de convergencia y precisión. De este modo, el modelo tuneado sirve para evaluar hasta qué punto el rendimiento del enfoque basado en SGD puede incrementarse mediante una optimización adecuada, sin necesidad de recurrir a modelos más complejos.

En conjunto, este modelo ajustado permite comparar de forma directa el impacto del *tuning* frente a las versiones no ajustadas del modelo, aportando una visión más completa sobre el papel del método de entrenamiento en la calidad de las predicciones de *CO(GT)*.

Modelo F: SGD con Modelo B-Tuneado:

Se entrenó una versión ajustada del modelo basado en Stochastic Gradient Descent siguiendo el planteamiento del Modelo B, en el que se excluye el sensor PT08.S1(CO) del conjunto de variables explicativas. El objetivo principal de este modelo es analizar si el ajuste de hiper parámetros puede mitigar la pérdida de rendimiento asociada a la eliminación del sensor más directamente relacionado con la variable objetivo.

El proceso de *tuning* se centró en optimizar el comportamiento del algoritmo durante el entrenamiento, ajustando parámetros como la velocidad de aprendizaje y el número de iteraciones. Este ajuste se realizó respetando la estructura temporal de los datos, de modo que el modelo se entrena siempre con observaciones anteriores a las utilizadas para su evaluación, garantizando así un escenario realista y evitando fugas de información temporal.

Este modelo resulta especialmente útil para evaluar la capacidad del enfoque lineal entrenado con SGD cuando se dispone de información más limitada, simulando situaciones en las que el sensor específico de CO no está disponible o presenta fallos. La comparación de sus resultados con los modelos equivalentes que incluyen PT08.S1(CO) permite analizar hasta qué punto el rendimiento del modelo depende de la selección de variables frente al método de entrenamiento empleado.

En conjunto, el SGD con Modelo B tuneado aporta una visión más completa sobre la robustez del sistema de modelización y sobre el papel del *tuning* en escenarios con restricciones en el conjunto de sensores.

Conclusión comparativa entre el Modelo E y el Modelo F

La comparación entre los modelos basados en Stochastic Gradient Descent con ajuste de hiper parámetros confirma las tendencias observadas en las versiones no tuneadas y en los modelos de regresión lineal clásica.

El SGD con Modelo A, que incluye el sensor *PT08.S1(CO)*, obtiene mejores resultados, con un MAE ≈ 0.35 y un RMSE ≈ 0.49, lo que indica predicciones más próximas a los valores

reales de la concentración de monóxido de carbono. El proceso de *tuning* permite estabilizar el entrenamiento y mantener un buen equilibrio entre precisión y generalización, aunque la mejora respecto a versiones no ajustadas es moderada.

Por el contrario, el SGD con Modelo B, en el que se excluye el sensor *PT08.S1(CO)*, presenta errores más elevados (MAE ≈ 0.40, RMSE ≈ 0.53). Estos resultados muestran que, incluso tras el ajuste de hiper parámetros, la ausencia de este sensor no puede compensarse completamente mediante el resto de variables disponibles, lo que confirma su importancia dentro del sistema de predicción.

En conjunto, esta comparación indica que la selección de variables tiene un impacto mayor en el rendimiento del modelo que el propio ajuste del algoritmo, y que el *tuning* mejora la estabilidad del SGD pero no sustituye la información aportada por los predictores más relevantes. Aun así, el modelo B tuneado sigue ofreciendo resultados razonables, lo que refuerza la idea de que el enfoque es relativamente robusto ante la pérdida de un sensor clave.

Evaluación de los Modelos Random Forest

Como parte del análisis, se entrenaron distintos modelos basados en Random Forest, tanto con el conjunto de variables del Modelo A (incluyendo *PT08.S1(CO)*) como del Modelo B (sin dicho sensor), considerando versiones con y sin ajuste de hiper parámetros. Estos modelos se incluyeron con un carácter comparativo y exploratorio, con el fin de contextualizar los resultados obtenidos por los modelos principales, sin que constituyen el eje central del trabajo.

Los resultados muestran que los modelos Random Forest presentan errores similares o superiores a los obtenidos con la regresión lineal y los modelos basados en SGD. Además, el proceso de *tuning* no aporta mejoras relevantes en términos de MAE y RMSE. En este contexto, la mayor complejidad del modelo no se traduce en un aumento de la precisión, por lo que su uso no aporta ventajas claras frente a enfoques más simples.

En consecuencia, los modelos Random Forest se mantienen como referencia comparativa, pero no se consideran candidatos prioritarios para la solución final. Esta decisión permite mantener el foco del trabajo en modelos más interpretables y alineados con el objetivo principal del proyecto.

Modelo	PT08.S1	R^2	MAE	RMSE
Regresión Lineal – Modelo A	Sí	8.605	3.455	4.886
Regresión Lineal – Modelo B	No	8.339	4.007	5.331
SGDRegressor – Modelo A (sin tuning)	Sí	—	3.447	4.860

SGDRegressor – Modelo A (tuneado)	Sí	—	3.462	4.891
SGDRegressor – Modelo B (sin tuning)	No	—	3.990	5.304
SGDRegressor – Modelo B (tuneado)	No	—	4.008	5.332
Random Forest – Modelo A (sin tuning)	Sí	—	3.333	4.968
Random Forest – Modelo A (tuneado)	Sí	—	3.442	5.074
Random Forest – Modelo B (sin tuning)	No	—	3.473	5.063
Random Forest – Modelo B (tuneado)	No	—	3.620	5.212

Entrenamientos y Evaluaciones Finales:

La comparación entre los distintos modelos entrenados muestra que los enfoques lineales, tanto en su versión clásica como mediante Stochastic Gradient Descent, ofrecen los mejores resultados para la predicción de la concentración de monóxido de carbono (*CO(GT)*). En todos los casos, los modelos que incluyen el sensor PT08.S1(CO) presentan un rendimiento superior, reflejado en valores más bajos de MAE y RMSE, lo que confirma la relevancia de este sensor en la estimación de CO.

La exclusión de PT08.S1(CO) conlleva una disminución moderada del rendimiento, aunque los modelos siguen siendo capaces de explicar una parte significativa de la variabilidad de la variable objetivo. Esto indica que el resto de sensores químicos y las variables meteorológicas aportan información complementaria útil, lo que refuerza la robustez del enfoque propuesto.

En cuanto a los modelos basados en Random Forest, su inclusión se realizó con un carácter comparativo. Los resultados obtenidos no mejoran los alcanzados por los modelos lineales y presentan errores similares o superiores, incluso tras el ajuste de hiper parámetros. Por ello, no se consideran candidatos prioritarios para la solución final, ya que su mayor complejidad no se traduce en una mejora clara del rendimiento.

En conjunto, los resultados ponen de manifiesto que la selección de variables tiene un impacto más relevante que la elección de algoritmos más complejos, y que los modelos lineales constituyen una solución adecuada y eficiente para este problema.

Conclusión Final del Proyecto

En este proyecto se ha abordado el problema de predecir la concentración de monóxido de carbono en entornos urbanos a partir de señales de sensores de bajo coste, variables meteorológicas y componentes temporales, formulando en términos de un problema de aprendizaje supervisado de regresión.

A lo largo del trabajo se ha seguido un proceso completo de análisis de datos y modelización, comenzando con un análisis exploratorio que permitió comprender el comportamiento de las variables y justificar la definición de *CO(GT)* como variable objetivo. Posteriormente, se llevó a cabo un proceso de ingeniería y selección de variables orientado a evitar fugas de información y a construir un conjunto de predictores coherente con un escenario real de predicción.

Los resultados obtenidos muestran que es posible predecir la concentración de CO con un nivel de precisión razonable utilizando modelos relativamente simples e interpretables, sin necesidad de recurrir a algoritmos altamente complejos. Este aspecto resulta especialmente relevante desde el punto de vista práctico, ya que facilita la comprensión del modelo y su posible aplicación en sistemas de monitorización de la calidad del aire basados en sensores *low-cost*.

Como limitaciones del estudio, cabe destacar la dependencia del rendimiento respecto a determinados sensores clave y la posible presencia de ruido inherente a los datos de sensores químicos. Como líneas futuras de trabajo, se podría explorar la incorporación de modelos específicos de series temporales, técnicas de regularización más avanzadas o la validación del enfoque en otros entornos urbanos.

En conclusión, el proyecto demuestra que un enfoque bien planteado de aprendizaje automático puede contribuir de forma efectiva a la predicción de la calidad del aire, aportando una herramienta útil para apoyar la toma de decisiones en contextos urbanos y ambientales.