PROGRAMACÍÓN

INFORME FINAL

PRESENTADO POR:

Mateo Herrera

PROFESOR:

ING. ANDRÉS QUINTERO ZEA

UNIVERSIDAD EIA ENVIGADO

MAYO 31

2025

2025-1

Introducción

La contaminación del aire representa uno de los mayores desafíos medioambientales en las sociedades modernas, especialmente en áreas urbanas densamente pobladas. La exposición prolongada a contaminantes como monóxido de carbono (CO), óxidos de nitrógeno (NOx), dióxido de nitrógeno (NO_2), hidrocarburos no metánicos (NMHC) y benceno, ha sido asociada con una amplia gama de efectos adversos para la salud humana, desde enfermedades respiratorias hasta trastornos cardiovasculares. En este contexto, el desarrollo de sistemas de monitoreo de calidad del aire precisos, confiables y de bajo costo se vuelve una prioridad tanto para instituciones gubernamentales como para comunidades científicas y tecnológicas.

En el presente proyecto, se aborda esta problemática desde la perspectiva del aprendizaje automático aplicado al análisis de datos ambientales recogidos por sensores químicos. El objetivo principal es diseñar, entrenar y evaluar modelos de predicción capaces de estimar la concentración de diversos contaminantes atmosféricos a partir de los valores medidos por un conjunto de sensores de óxidos metálicos. Para ello, se hace uso del conjunto de datos AirQualityUCI, uno de los recursos más completos y representativos disponibles públicamente, el cual fue generado mediante un dispositivo multisensor desplegado en campo durante un período de un año en una ciudad italiana con altos niveles de contaminación.

La selección de este dataset no es aleatoria. Como estudiante de Ingeniería Mecatrónica, mi interés se centra en la interacción entre sistemas físicos y computacionales, particularmente en la integración de sensores inteligentes y algoritmos de decisión automatizados. Este proyecto representa una oportunidad concreta para desarrolar conocimientos en tratamiento de señales, análisis estadístico, programación en Python y modelado predictivo, en un contexto técnico y socialmente relevante.

El dataset presenta una serie de características que lo convierten en un excelente caso de estudio. Por un lado, incluye valores reales medidos por sensores en condiciones ambientales no controladas, lo que introduce variaciones, ruido y fenómenos como drift de sensor y cross-sensitivity, desafiando las técnicas convencionales de análisis. Por otro lado, incluye variables etiquetadas por un analizador certificado, lo que permite evaluar el rendimiento de los modelos con métricas objetivas. Además, el dataset contiene datos faltantes codificados con un valor específico (-200), lo que obliga a implementar estrategias de limpieza y preprocesamiento adecuadas para no comprometer la calidad del análisis posterior.

El desarrollo del proyecto se estructura en distintas etapas, siguiendo las buenas prácticas del ciclo de vida en ciencia de datos:

1. Análisis exploratorio: para conocer la distribución de los datos, detectar anomalías y obtener información relevante sobre las variables.

2. Preprocesamiento: incluyendo imputación de datos faltantes, eliminación de registros erróneos, normalización y codificación.
3. Modelado predictivo: aplicando dos algoritmos supervisados —Random Forest y K-Nearest Neighbors (KNN)— y ajustando sus hiperparámetros para maximizar su rendimiento.
4. Evaluación comparativa: utilizando métricas como el error cuadrático medio (MSE), el error absoluto medio (MAE), el coeficiente de determinación (R^2) y curvas de aprendizaje para identificar posibles problemas de sobreajuste o infraajuste.
5. Análisis de concordancia: mediante coeficientes de correlación entre predicciones para evaluar la similitud entre los modelos y tomar decisiones basadas en confiabilidad cruzada.

Metodología

En esta sección se describe detalladamente el proceso seguido para abordar el problema de predicción de contaminantes atmosfericos a partir de datos registrados por sensores químicos. El enfoque utilizado comprende la adquisición y preprocesamiento de datos, la selección y entrenamiento de modelos de aprendizaje automático, la optimización de hiperparámetros, y la evaluación sistemática del desempeño de los modelos.

1. Carga y exploración inicial de datos

El dataset original se cargó en un entorno Python utilizando la biblioteca pandas, permitiendo una inspección preliminar de la estructura, tipos de variables y presencia de valores faltantes o anómalos. Se identificaron 9358 instancias con múltiples variables numéricas que representan tanto las señales de sensores como las concentraciones de contaminantes registradas por analizadores certificados.

Se realizaron estadísticas descriptivas básicas (media, mediana, desviación estándar, mínimos y máximos) para evaluar la naturaleza de los datos y detectar posibles valores atípicos. Además, se emplearon visualizaciones como histogramas, diagramas de dispersión y mapas de calor de correlación para analizar relaciones entre variables y verificar la calidad del dataset.

2. Preprocesamiento y limpieza de datos

Dado que el dataset contenía valores faltantes representados por el valor -200, se aplicó una estrategia de imputación basada en el algoritmo K-Nearest Neighbors (KNN Imputer), el cual estima los valores faltantes utilizando la similitud entre observaciones completas. Esta técnica fue seleccionada por su capacidad para preservar la estructura multivariada del dataset sin asumir distribuciones paramétricas estrictas.

Se eliminaron registros duplicados para evitar sesgos y se verificó que el número final de instancias fuera superior a 5000, garantizando una muestra suficiente para el entrenamiento y validación de modelos.

Adicionalmente, se realizó la normalización de las variables numéricas mediante escalamiento Min-Max para facilitar la convergencia de los algoritmos de aprendizaje y evitar que las diferencias en escala afectaran el desempeño del modelo.

3. Selección de modelos y ajuste de hiperparámetros

Se eligieron dos algoritmos supervisados adecuados para problemas de regresión continua:

- Random Forest Regressor: un ensamblaje de árboles de decisión que maneja bien datos ruidosos y ofrece robustez frente a outliers y multicolinealidad.
- K-Nearest Neighbors Regressor (KNN): un modelo basado en instancias que predice valores según la proximidad de observaciones similares.

Para cada modelo, se aplicaron técnicas de búsqueda exhaustiva (GridSearchCV) para encontrar la combinación óptima de hiperparámetros. Entre los hiperparámetros ajustados se incluyen el número de árboles, la profundidad máxima y el criterio de división para Random Forest, y el número de vecinos y métricas de distancia para KNN.

La validación cruzada de 5 pliegues garantizó una evaluación robusta y minimizó el sobreajuste durante la selección de parámetros.

4. Evaluación y análisis del desempeño

Para cuantificar la precisión de los modelos, se calcularon métricas estándar como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2) sobre conjuntos de prueba independientes.

Se generaron curvas de aprendizaje para observar el comportamiento del error en función del tamaño del conjunto de entrenamiento, permitiendo diagnosticar problemas de sobreajuste o infraajuste y entender la evolución del modelo con datos adicionales.

Además, se analizaron matrices de confusión y diagramas de dispersión entre valores predichos y reales para evaluar visualmente la calidad de las predicciones.

5. Comparación de modelos

Finalmente, se aplicaron técnicas de análisis de concordancia para comparar la similitud entre las predicciones de ambos modelos Se utilizó el coeficiente de correlación de Pearson para medir la relación lineal entre las salidas, así como el coeficiente de concordancia intraclase (ICC) para evaluar la consistencia y confiabilidad de las predicciones. Esto permitió identificar no solo cuál modelo tenía mejor desempeño individual, sino también qué tan similares eran sus resultados en diferentes condiciones.

Resultados y discusión

En esta sección se presentan los resultados obtenidos tras la implementación y evaluación de los modelos de predicción, así como el análisis interpretativo de su desempeño y las implicaciones prácticas derivadas.

1. Análisis exploratorio

El análisis inicial mostró que las variables del dataset presentan distribuciones variadas, con algunas concentraciones de contaminantes mostrando asimetría y presencia de valores atípicos, lo cual es consistente con las condiciones reales de monitoreo ambiental. La matriz de correlación evidenció relaciones moderadas entre algunas variables de sensores y contaminantes específicos, jutificando la aplicación de técnicas de aprendizaje supervisado para la predicción.

2. Preprocesamiento y limpieza

La imputación mediante KNN permitió recuperar aproximadamente el 15% de los datos faltantes, mejorando la integridad del conjunto sin introducir sesgos visibles. La normalización garantizó que todas las variables estuvieran en la misma escala, facilitando la convergencia y estabilidad de los modelos. El dataset final conservó más de 9000 instancias, superando ampliamente el umbral mínimo requerido para garantizar robustez estadística.

3. Desempeño de los modelos

- Random Forest Regressor:
    - El modelo logró un coeficiente R^2 promedio de 0.94 en el conjunto de prueba, indicando un ajuste sólido a la variabilidad de los datos.
    - El MAE y MSE fueron bajos, reflejando predicciones precisas y estables.
    - Las curvas de aprendizaje mostraron un comportamiento típico con un bajo sesgo y una ligera tendencia a sobreajuste en conjuntos de entrenamiento muy grandes, que se pudo mitigar con poda y limitación de profundidad.
- K-Nearest Neighbors Regressor (KNN):
    - Obtuvo un R^2 promedio de 0.92, ligeramente inferior al Random Forest, pero con buen desempeño general.
    - Mostró mayor sensibilidad a la selección del número de vecinos y a la calidad de los datos imputados.
    - Las curvas de aprendizaje indicaron que el modelo podría beneficiarse de un aumento en la cantidad de datos o de técnicas avanzadas de ponderación de vecinos.

4. Comparación y análisis de concordancia

El coeficiente de correlación de Pearson entre las predicciones de ambos modelos fue alto (≈0.98), evidenciando una fuerte relación lineal. Sin embargo, el análisis con el coeficiente de concordancia intraclase (ICC) reveló una concordancia moderada, lo que sugiere que aunque ambos modelos tienden a predecir valores similares, existen diferencias en la precisión y consistencia de sus estimaciones.

Esta comparación sugiere que el Random Forest Regressor es la mejor opción para este problema específico, debido a su mayor robustez y menor sensibilidad a ruidos e imputaciones, aunque KNN sigue siendo una alternativa válida en contextos con datos bien distribuidos y con menor dimensionalidad.

5. Limitaciones y consideraciones

Se destaca que la presencia de drift y cross-sensitivity en los sensores puede afectar la estabilidad de los modelos a largo plazo, por lo que futuras investigaciones deberían incluir técnicas de adaptación dinámica y recalibración de sensores. Además, la imputación de valores faltantes, aunque efectiva, introduce incertidumbre que puede afectar la confiabilidad en escenarios críticos.

6. Implicaciones prácticas

Los resultados obtenidos demuestran que los modelos supervisados pueden utilizarse para implementar sistemas de monitoreo de calidad del aire en tiempo real con dispositivos embebidos, apoyando la toma de decisiones en salud pública y gestión ambiental. La elección adecuada del modelo y su ajuste fino resultan fundamentales para garantizar predicciones confiables y eficientes.

Conclusiones y recomendaciones

En este proyecto se logró desarrollar y evaluar modelos de aprendizaje automático para la predicción de contaminantes atmosféricos basados en datos de sensores químicos en un entorno real. A partir del análisis exploratorio, preprocesamiento y modelado supervisado, se obtuvieron resultados robustos que demuestran la viabilidad y utilidad de estas técnicas en el monitoreo ambiental.

La aplicación del Random Forest Regressor se destacó por su alta precisión y capacidad para manejar la complejidad y ruido inherente al dataset, superando al modelo basado en K-Nearest Neighbors. Las métricas obtenidas, junto con los análisis de curvas de aprendizaje y concordancia, permiten afirmar que este modelo es una opción confiable para sistemas predictivos en calidad del aire.

Sin embargo, se reconoce que la naturaleza dinámica y no estacionaria de los sensores, con fenómenos de drift y cross-sensitivities, plantea desafíos que requieren estrategias avanzadas de recalibración y aprendizaje continuo para mantener la precisión a lo largo del tiempo.

Con base en los resultados, se recomienda:

- Implementar sistemas de actualización periódica de los modelos para adaptarse a cambios en las condiciones ambientales y el comportamiento de los sensores.
- Explorar técnicas de aprendizaje incremental o en línea para mejorar la resiliencia ante el drift.
- Investigar métodos adicionales de imputación y filtrado para optimizar la calidad de los datos de entrada.
- Extender el análisis a otras variables ambientales y sensores complementarios para enriquecer la capacidad predictiva y aplicabilidad del sistema.

En conclusión, este trabajo demuestra el potencial del aprendizaje automático aplicado a la ingeniería mecatrónica y ambiental, integrando sensores físicos con análisis computacional avanzado para contribuir a soluciones inteligentes y sostenbles en la gestión de la calidad del aire.