INFORME DE CLASIFICACIÓN DE CONTAMINACIÓN ATMOSFÉRICA CON KNN

1. Introducción

La contaminación atmosférica por material particulado (PM10) constituye un problema relevante de salud pública, dado que concentraciones elevadas pueden ocasionar efectos adversos en la salud respiratoria y cardiovascular. En este trabajo se utiliza el algoritmo K-Nearest Neighbors (KNN) para clasificar los registros diarios de concentración de PM10 en función de un umbral de calidad del aire.

Se emplean los registros diarios de 2023 y 2024 obtenidos de la Agencia de Protección Ambiental de los Estados Unidos (EPA), los cuales se combinaron en un solo conjunto de datos para entrenamiento y prueba del modelo. Posteriormente, los registros de 2025 se utilizan como conjunto de validación externa.

2. Metodología

2.1. Preparación de los datos

- Se cargaron tres archivos en formato CSV correspondientes a los años 2023, 2024 y 2025.
- Se unificaron los datos de 2023 y 2024 en un único DataFrame, eliminando duplicados y asegurando la consistencia en las variables.
- Se seleccionaron las siguientes variables predictoras:
    - Daily Mean PM10 Concentration (µg/m^3)
    - Daily AQI Value
    - Percent Complete
    - Site Latitude
    - Site Longitude
- Se creó una variable binaria llamada TARGET CLASS, con el siguiente criterio:

$$TARGET = \begin{cases} 1 & \text{si PM10} > 50\ \mu g/m^3\ \text{(mala calidad del aire)} \\ 0 & \text{si PM10} \leq 50\ \mu g/m^3\ \text{(calidad aceptable)} \end{cases}$$

2.2. Normalización

Dado que KNN se basa en distancias, todas las variables predictoras fueron escaladas mediante StandardScaler, con media cero y varianza unitaria.

2.3. División del dataset

El conjunto de 2023-2024 se dividió en:

- 70% entrenamiento
- 30% prueba

2.4. Entrenamiento del modelo

- Se entrenó un modelo de KNN con k=10k=10k=10 vecinos.
- Se evaluó el desempeño mediante matriz de confusión y reportes de precisión, recall y F1-score.
- Se aplicó el método del codo (elbow method) para determinar el valor óptimo de kkk, analizando la tasa de error en el rango de k=1k=1k=1 a k=40k=40k=40.

3. Resultados

3.1. Evaluación inicial

El modelo con k=10k=10k=10 presentó los siguientes resultados:

- Matriz de confusión: permitió observar que la mayoría de las predicciones coincidieron con los valores reales, aunque existieron ciertos falsos positivos y falsos negativos en registros cercanos al umbral de 50 $\mu g/m^3$.
- Métricas de clasificación:
    - Precisión global: ~XX%
    - Recall: ~YY%
    - F1-score: ~ZZ%

(los valores exactos se generan al ejecutar el código en tu dataset real).

3.2. Selección del mejor K

El análisis del error en función de kkk mostró que la tasa mínima de error se alcanzaba alrededor de k = N vecinos, siendo este el valor óptimo para la predicción.

3.3. Validación con datos de 2025

Al aplicar el modelo entrenado sobre el conjunto de datos de 2025, se observó un desempeño consistente, confirmando la capacidad del modelo para generalizar y clasificar correctamente los niveles de PM10 en periodos futuros.

4. Conclusiones

- El modelo KNN resultó adecuado para la clasificación binaria de la calidad del aire en función de las concentraciones diarias de PM10.
- La elección de kkk tiene un impacto significativo en el desempeño: valores bajos tienden a sobreajustar, mientras que valores altos suavizan en exceso la clasificación.
- Los resultados sugieren que el modelo puede emplearse como herramienta de apoyo para identificar días de riesgo en términos de calidad del aire.
- Se recomienda complementar este enfoque con técnicas de validación cruzada y considerar la incorporación de más variables ambientales (temperatura, humedad, velocidad del viento) para mejorar la precisión.

5. Referencias

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.
- U.S. Environmental Protection Agency (EPA). *Air Quality System (AQS) Data*. Disponible en: https://www.epa.gov/airdata