“Año de la recuperación y consolidación de la economía peruana”

UNIVERSIDAD PERUANA CAYETANO HEREDIA

FACULTAD DE CIENCIAS E INGENIERÍA

TÍTULO:

“Predicción Temporal de CO en Mississippi utilizando Técnicas de Machine Learning”

Autor:

Rioja Cruz Vanesa Doris

CURSO:

Proyectos de Ingeniería I

GRUPO:

AG2 - Subgrupo 3

FECHA:

20 de setiembre del 2025

2025 - II

Introducción

La predicción de contaminantes atmosféricos constituye un desafío relevante para la gestión de la calidad del aire. En este trabajo se aborda la estimación de concentraciones diarias de monóxido de carbono (CO) para el estado de Mississippi durante el año 2025. Se utilizan datos históricos (2022–2024) proporcionados por la EPA - Air Data, los cuales incluyen tanto el valor diario del CO (Daily Max 8-hour CO Concentration) como el índice de calidad del aire (Daily AQI Value).

El objetivo es evaluar distintos enfoques de modelado: un modelo de regresión base utilizando exclusivamente variables temporales y un modelo de regresión jerárquico que incorpora la relación entre CO y AQI mediante predicciones intermedias (AQI_pred → CO_pred). Se comparan los resultados en términos de métricas estadísticas y capacidad de generalización.

Metodología

Paso 1: Recolección y preparación de datos

- **Fuente:** Los datos provienen de la *Environmental Protection Agency (EPA)*.
- **Cobertura temporal:** 2022–2025.
- **Área geográfica:** Mississippi
- **Variables principales:**
    - **Date:** fecha de la observación (diaria).
    - **Daily Max 8-hour CO Concentration (CO):** concentración máxima de CO en ppm.
    - **Daily AQI Value (AQI):** índice de calidad del aire correspondiente.
- **Datos en bruto:**
    - CO_Mississippi_2022.csv
    - CO_Mississippi_2023.csv
    - CO_Mississippi_2024.csv
    - CO_Mississippi_2025.csv (únicamente para comparación)

- **Procesamiento previo:**

- Se concatenaron los dataframe de 2022-2024 para lograr un solo dataframe a partir del cual se realizará el entrenamiento y prueba.
- Se unificó la frecuencia a diaria continua usando *asfreq("D")*.
- Se aplicó interpolación lineal para cubrir valores faltantes de CO y AQI.
- Se renombraron columnas a un formato estándar (Date, CO, AQI).
- Se generaron nuevas variables: lags y medias móviles (MAs) para capturar memoria temporal, y funciones estacionales (seno/coseno) para representar periodicidad semanal y anual.
- En el modelo basado en fecha, se añadieron variables categóricas (dummies de día de semana y mes).

Se aplicaron dos estrategias principales:

1. **Modelo base (features de fecha):**

- Variables de entrada: tendencia temporal, componentes estacionales (sin/cos de 7 y 365 días), dummies de día de la semana y mes.
- No utiliza información de AQI.
- Sirve como línea base para evaluar si los patrones de fecha y estacionalidad son suficientes para predecir CO.
- Entrenamiento: enero 2022 – marzo 2024. Validación: abril - diciembre 2024.
- Predicción: se construyó un forecast completo para todo el año 2025.
- Algoritmo: XGBoost (reg:squarederror).

2. **Modelo jerárquico (relación AQI–CO):**

- Motivación: existe una correlación directa entre AQI y CO, lo que justifica un esquema jerárquico en dos etapas.
- Primera etapa (AQI):
  - Entrenamiento de un modelo de AQI utilizando lags (1, 2, 3, 7, 14, 28, 56 días), medias móviles (7, 14, 28, 30 días) y estacionalidad (sin/cos de 7 y 365 días).
  - Se generan predicciones *out-of-fold (OOF)* para 2022–2024 y un forecast recursivo para 2025.
- Segunda etapa (CO):
  - El CO se modela con sus propios lags/MAs, más los de AQI_pred.
  - Se entrenan dos variantes de XGBoost:
    - **Central:** reg:squarederror.

        - **Cuantil (q95):** reg:quantileerror (para capturar picos).
    - Se combinan en un ensemble (0.85 central / 0.15 cuantil).
- Entrenamiento: enero 2022 – marzo 2024. Validación: abril - diciembre 2024.
- Predicción: recursiva, día a día, sin usar datos reales de 2025.

Evaluación de resultados:

- Métricas: RMSE, MAE y R^2.
- Tablas comparativas en escalas diaria, semanal y mensual para 2025.
- Visualización de curvas reales vs. predichas.

Resultados

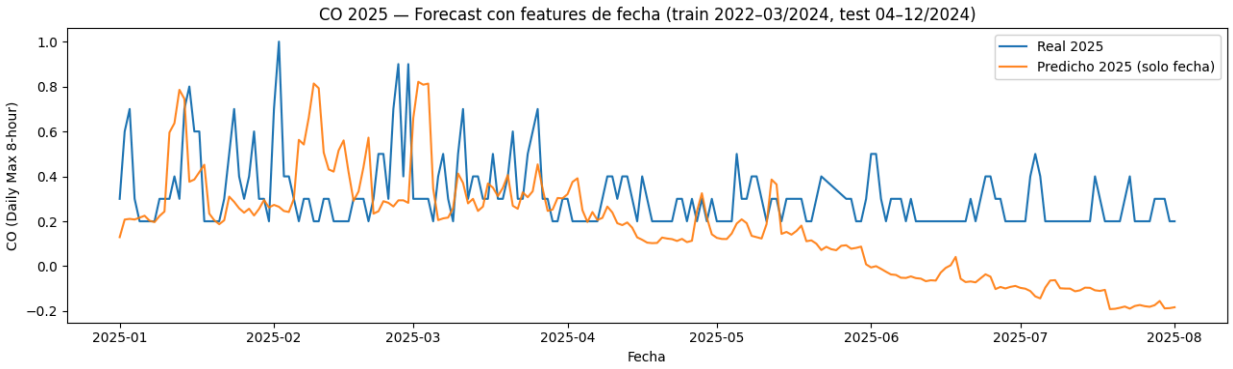
a. Modelo base (features de fecha)

En la validación de abril a diciembre de 2024, el modelo obtuvo métricas de **RMSE ≈ 0.186, MAE ≈ 0.151 y R^2 ≈ -0.57**, lo que indica un bajo poder explicativo: aunque el error absoluto medio es relativamente bajo, el valor negativo del R^2 muestra que el modelo no logra superar una predicción ingenua basada en la media.

En la comparación con los datos reales de 2025 (enero–agosto), los resultados fueron aún más limitados (**RMSE ≈ 0.277, MAE ≈ 0.226 y R^2 ≈ -2.54**), reflejando que las predicciones se alejan considerablemente de la dinámica real.

Estos valores muestran que el modelo, basado únicamente en información de fecha (tendencia, estacionalidad semanal y mensual), captura ciertos patrones globales, pero no logra representar adecuadamente la variabilidad del CO para los meses de 2025.

En el gráfico de comparación, se observa que las predicciones siguen una trayectoria más suavizada y decreciente, sin reflejar los picos abruptos ni la amplitud de las variaciones diarias reales. Esto confirma que los atributos temporales por sí solos no son suficientes para capturar la complejidad del contaminante.



b. Modelo jerárquico con AQI_pred + ensemble

En la validación del AQI durante 2024, el desempeño fue moderado (**RMSE ≈ 1.55, MAE ≈ 1.21 y R^2 ≈ 0.32**), mostrando que la predicción del índice captura parcialmente su dinámica, aunque con dificultades para reproducir los picos abruptos.

Para la validación del CO (abril–diciembre 2024), el modelo logró un mejor ajuste:

- Modelo central: **RMSE ≈ 0.124, R^2 ≈ 0.30**.
- Modelo cuantil q95: **RMSE ≈ 0.134, R^2 ≈ 0.19**.
- Ensemble 0.85/0.15: **RMSE ≈ 0.124, R^2 ≈ 0.31**.

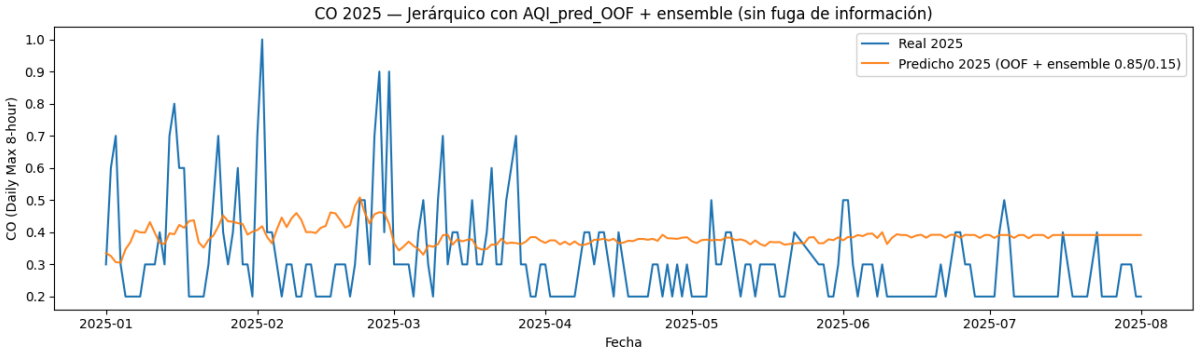
Estos valores muestran que el modelo jerárquico, que combina la predicción del AQI con los rezagos y medias móviles del CO, logró mejorar la capacidad de explicación respecto al Modelo 1, aunque aún con limitaciones en la precisión de los picos.

En la comparación con 2025 (enero–agosto), las métricas fueron **RMSE ≈ 0.163, MAE ≈ 0.138 y R^2 ≈ -0.23**. Esto refleja que las predicciones replican el nivel medio de la serie, pero todavía presentan dificultades para seguir la variabilidad extrema.

Las tablas comparativas de 2025 respaldan esta conclusión:

- A nivel diario, los errores son más altos en días con incrementos bruscos.
- En escalas semanales y mensuales, los promedios suavizan la discrepancia, con errores porcentuales absolutos (APE%) en el rango de 30–60%.

En el gráfico de comparación, se aprecia que el modelo reproduce de forma estable el nivel medio de las concentraciones de CO, pero no logra seguir los picos más pronunciados, lo que limita su capacidad predictiva en episodios críticos.



Discusión

El contraste entre ambos enfoques pone en evidencia las limitaciones de los modelos en función de la información disponible:

Modelo basado sólo en fecha:

- Insuficiente para capturar la variabilidad real del CO, ya que depende únicamente de atributos temporales (tendencia, estacionalidad semanal y mensual), el modelo ignora relaciones ambientales clave que influyen directamente en la concentración de contaminantes.
- Sus predicciones se limitaron a reproducir una tendencia suavizada y decreciente.
- No logró seguir los picos abruptos de 2025, lo que se reflejó en un R^2 negativo.
- En la práctica, su desempeño fue similar al de una media histórica.

Modelo jerárquico AQI → CO con ensemble:

- Mostró métricas más favorables en la validación 2024, indicando que el modelo aprendió la relación entre AQI y CO.
- Sin embargo, en 2025 volvió a mostrar R^2 negativo por dos factores principales:
    - Falta de variables exógenas (meteorología, tráfico, emisiones), que explicarían los cambios de dinámica.
    - Los datos disponibles no siguen una secuencia estable en el tiempo, lo que afecta la extrapolación.
- El ensemble 0.85/0.15 equilibró la predicción central con un modelo más sensible a extremos, reduciendo la subestimación de picos, aunque sin resolverla por completo.

Limitaciones generales:

- Errores más altos en días con picos abruptos; los promedios semanales y mensuales suavizan la discrepancia.
- El nivel medio del CO fue mejor replicado que su variabilidad diaria.
- El R^2 negativo en 2025 es un fenómeno esperado en extrapolación temporal sin suficientes señales exógenas.

Conclusión:

- El modelo jerárquico con ensemble es el más apropiado entre los probados, mostrando consistencia en validación y error moderado en 2025.
- Para futuras mejoras, sería necesario incluir variables meteorológicas (temperatura, viento, humedad) y otras exógenas, que permitan capturar la dinámica real de los contaminantes y mejorar la robustez de las predicciones.

Referencias

- T. Chen y C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
- U.S. Environmental Protection Agency (EPA), “Air Quality System (AQS) Data Mart,” 2024. [Online]. Available: https://www.epa.gov/airdata
- J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.