1. INTRODUCCIÓN

La calidad del aire al pasar los años se ha convertido en un problema en la sociedad, esto se debe a una rápida expansión urbana, el aumento de vehículos y a las actividades industriales. Esta contaminación atmosférica es crítica en la ciudad de Alabama, teniendo como emisión principal a partículas de CO (Monóxido de Carbono) donde las cuales superan los límites recomendados por la Organización Mundial de la Salud (OMS).

Párrafo 1: A nivel global (Hablar algo de la ONU y estadísticas o reportes de CO

Párrafo 2: Nivel América(Estados unidos tiene la mayor emisión de CO o algo similar)

Párrafo 3: Nivel estados unidos

Párrafo 4: Nivel Alabama

Alumna: Mercedes Castañeda Reátegui
DNI: 76880566
Correo: mercedes.castaneda@upch.pe
Descripción: Soy una estudiante de ingeniería ambiental de 6to ciclo, interesada en minería y nanomateriales. Me gustaría enfocarme en el uso de nuevas tecnologías para reducir el impacto ambiental y mejorar la sostenibilidad en la industria minera.

2. METODOLOGÍA

A Continuación se detalla la metodología utilizada para el modelo predictivo de concentración de CO, profundizando en cada una de las etapas:

1. Recolección de datos:

En primer lugar, se descargaron los datos de la página web U.S. EPA por sus siglas en inglés United States Environmental Protection Agency, específicamente de uno de los indicadores de la calidad de aire (Download Daily Data), los parámetros seleccionados fueron las concentraciones registradas de CO (monóxido de carbono) en el área geográfica de Alabama de los años 2022 y 2023. En segundo lugar, se combinaron o concatenaron ambas datasets (conjunto de datos) con la finalidad de ser explorados posteriormente.

2. Preparación de datos:

Dentro de los datos descargados existe la variable “Dates” (fechas), por lo que es importante asegurarnos que estén en su formato correspondiente, para ello, se ordenan dichos valores en orden creciente y posterior a ello se utiliza la función time spet, la cual nos va a permitir asignar valores numéricos en secuencia del orden asignado, por ejemplo, 0 -> 1 de enero del 2022, 1 -> 2 de enero del 2022, 2 -> 3 de enero del 2022 y así sucesivamente con los días restantes. Posterior a ello, se dividió el dataset de forma que el 80% de los primeros registros estén en el training y el 20% estén en el test.

Al obtener la longitud total de fechas para ambos años se esperaría un valor de 730 (365x2 = 730 días), pero, se obtienen 717, dando alusión que en algunos días no se registraron datos. Sin embargo, para mantener la tendencia en los datos se utilizó el método forward fill, lo cual nos va a permitir completar las fechas faltantes.

3. Extracción de características:

En este paso, se ha considerado la relevancia de las columnas en la variación de las concentraciones de CO. Aunque la mayoría de las columnas son significativas, se ha identificado que muchas de ellas no tienen influencia en la variación de las concentraciones de monóxido de carbono. Por esta razón, las columnas eliminadas son las siguientes:

- **POC(Parameter Ocurrence Code)**
- **Source**
- **Site ID**

- **Local Site Name**
- **Units**
- **AQS Parameter Code**
- **Method Code**
- **CBSA Code**
- **CBSA Name**
- **State FIPS Code**
- **County**
- **State**
- **County FIPS Code**
- **Site Latitude y Site Longitude**

4. **Selección de un modelo**

El modelo empleado es una regresión lineal, una técnica estadística que analiza la relación entre una variable dependiente y una o más variables independientes. Su objetivo es identificar la mejor línea recta que representa esta relación en un conjunto de datos, minimizando la diferencia entre los valores reales y los predichos. El proceso comienza con la carga y exploración de los datos, seguido por el ajuste del modelo de regresión lineal. Una vez entrenado, el modelo se puede usar para hacer predicciones con nuevos datos.

En este caso, se ha trabajado con el AQI (Índice de Calidad del Aire), una métrica que mide la calidad del aire basándose en la concentración de diversos contaminantes, incluido el monóxido de carbono (CO). El valor del AQI está diseñado para reflejar el impacto en la salud, niveles de contaminantes en el aire. Un valor más alto de AQI indica una peor calidad del aire y una mayor presencia de contaminantes, como el CO. Dado que el CO es uno de los contaminantes evaluados en el cálculo del AQI, existe una correlación directa entre el AQI y la concentración de CO: a medida que aumenta la concentración de CO, también lo hace el AQI, lo que señala un mayor riesgo para la salud. Por lo tanto, el valor del AQI puede servir como un indicador de la concentración de CO, dado que ambos aumentan en situaciones de alta contaminación.

5. **Entrenamiento y evaluación:**

Para el entrenamiento de este modelo se usó "Date", "Time", "Daily AQI Value", "Actual Concentration" y "Predicted Concentration" para predecir la Concentración de CO, previamente pasado por pipeline que en machine learning es una secuencia de pasos automatizados que transfor y esto facilita la organización y reutilización de procesos, asegurando que cada paso se ejecute en un flujo estructurado y eficiente. Además, automatiza el proceso, reduciendo errores y optimizando el rendimiento del modelo. Estos datos fueron exportados mediante un pkl, este es un formato utilizado para almacenar objetos serializados mediante el módulo pickle. Serializar un objeto convierte su estado en una secuencia de bytes que se puede guardar en un archivo

y posteriormente deserializar para restaurar el objeto original. Los archivos .pkl son útiles para guardar modelos de machine learning, estructuras de datos complejas y otros objetos, permitiendo su reutilización sin necesidad de recalcular o reconstruirlos.

6. **Predicción:**

Finalmente, nos queda usar el modelo entrenado con los parámetros correctos para predecir nuevos datos, por ejemplo:

	Date	Time	Daily AQI Value	Actual Concentration	Predicted Concentration
573	2023-08-07	583	3	0.30000000	0.27437275
574	2023-08-08	584	5	0.40000000	0.43643077
575	2023-08-09	585	5	0.40000000	0.43641840
576	2023-08-10	586	2	0.20000000	0.19330045
577	2023-08-11	587	3	0.30000000	0.27432327

Asimismo, en una regresión lineal, es importante considerar los valores de las métricas, los cuales nos van a servir para evaluar la calidad del modelo ajustado y su capacidad para hacer predicciones precisas sobre nuevos datos.

En este caso se obtiene que el Error Absoluto Medio (MAE) = 0.0231, esto significa que en promedio, las predicciones del modelo se desvían en aproximadamente 0.0231 unidades del valor real de la Concentración Máxima Diaria de CO en 8 horas; el Error Cuadrático Medio (MSE) = 0.00067, esto representa el promedio de las diferencias al cuadrado entre los valores predichos y los reales; la Raíz del Error Cuadrático Medio (RMSE) = 0.0259, esto indica que la magnitud típica del error es pequeña, lo que sugiere un buen rendimiento del modelo; y el Coeficiente de Determinación (R^2): = 0.959, esto significa que aproximadamente el 95.9% de la variabilidad en la concentración de CO es explicada por el modelo.En conclusión, el modelo presenta un ajuste fuerte con los datos.

3. **RESULTADOS:**
4. **DISCUSIONES: (OPCIONAL)**
5. **REFERENCIAS: FORMATO IEEE**