




Proyecto de Análisis de Datos: Calidad del Aire (Preguntas)

 Profesora: Sol Del Valle FIGUEROA

 Barozzi Eugenia

 ISPC | Primer evidencia de Desarrollo de Modelo de IA machine (aprendizaje automático)

 Colab:  Ev1-Calidad-del-Aire.pynb | GitHub:  [Proyecto](#)

 Proyecto completo: [Pdf Completo](#)

Preguntas antes del EDA

- 1. ¿Los datos son estructurados o no estructurados?** Los datos son estructurados, organizados en filas y columnas dentro de un archivo CSV. Es tabular y facilita su procesamiento y análisis mediante herramientas como Pandas en Python.
- 2. ¿Cómo es el volumen de los datos?** El dataset contiene aproximadamente 9,358 filas y 15 columnas. Este volumen de datos se considera adecuado para la aplicación de técnicas de regresión supervisada sin necesidad de recurrir a metodologías de Big Data.
- 3. ¿Qué tipo de problema se resolverá?** El objetivo principal es abordar un problema de regresión. Buscamos predecir las concentraciones de gases contaminantes en el aire, siendo el Monóxido de Carbono (CO) un ejemplo clave. Estas concentraciones son valores numéricos continuos, l
- 4. ¿Las variables objetivo son continuas o categóricas?** Estas concentraciones son valores numéricos continuos. Representan mediciones en unidades físicas específicas y pueden tomar cualquier valor dentro de un rango.
- 5. ¿Qué tipo de datos contiene cada columna?** representan concentraciones de gases (como CO(GT), NOx(GT), C6H6(GT)), señales de sensores electrónicos (PT08.S1(CO) hasta PT08.S5(O3)), y condiciones meteorológicas (Temperatura, Humedad Relativa, Humedad Absoluta). También hay columnas que representan **información temporal** (Fecha y Hora), que posteriormente se combinaron en un formato datetime.
- 6. Existe una dimensión temporal en los datos? ¿Se trata de una serie de tiempo?** Sí, existe una clara dimensión temporal. Los datos incluyen columnas de Fecha y Hora, lo que indica que las observaciones fueron registradas en momentos específicos. Al combinar estas

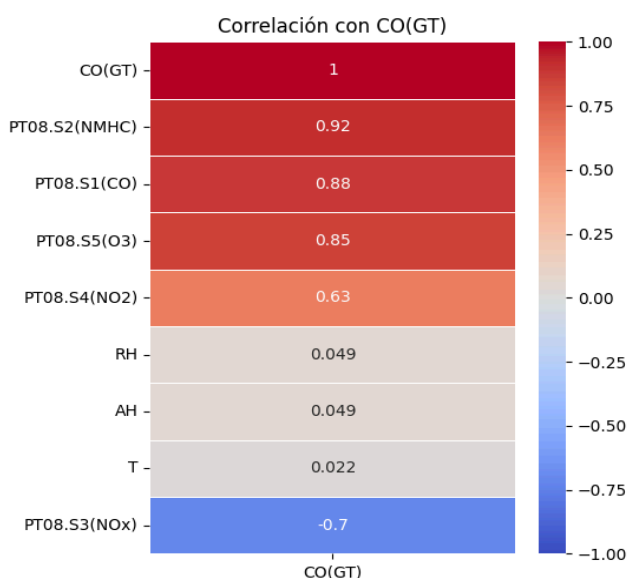
columnas en 'Datetime' y establecerla como índice, confirmamos que **se trata de una serie de tiempo**, donde cada fila representa una medición en un punto específico en el tiempo.

7. ¿Cuál es la granularidad o frecuencia temporal de las mediciones? La frecuencia temporal de las mediciones es horaria. Cada registro en el dataset corresponde a una lectura tomada en un intervalo de una hora.

8. ¿Se conocen valores especiales o códigos que necesitan ser interpretados? Hasta el momento, el valor especial identificado es -200, que se interpreta como un indicador de datos faltantes.

9. ¿Qué herramientas y librerías son apropiadas para trabajar estos datos? Las herramientas y librerías apropiadas incluyen pandas para la manipulación y análisis de datos tabulares, seaborn y matplotlib para la visualización de datos, scikit-learn para la implementación de modelos de aprendizaje automático (regresión), statsmodels para el análisis estadístico y numpy para operaciones numéricas.

2. Preguntas sobre el análisis



1. ¿Qué distribución estadística tienen los datos? En el análisis exploratorio (de resumen estadístico y histograma) reveló que la mayoría de las variables presentaban distribuciones sesgadas hacia la derecha. La variable Humedad Relativa (RH) mostró una distribución más cercana a la normal. Se identificó una dispersión variable entre las columnas y la presencia de valores atípicos en contaminantes como “CO(GT)” Monóxido de carbono real y “NO2(GT)” Dióxido de nitrógeno, así como en la temperatura “(T)”.

2. ¿Qué columnas son más importantes?

Considerando el objetivo de predecir la contaminación del aire, el objetivo principal en

este proyecto se centró en predecir la concentración de Monóxido de Carbono (CO(GT)). Por lo tanto, priorizamos aquellas variables que mostraban una relación más directa y fuerte con el CO(GT) en nuestro análisis exploratorio inicial (especialmente en los scatter plots y la matriz de correlación).

'C6H6(GT)', # Benceno (correlación más alta)

'PT08.S2(NMHC)', # Señal del sensor de NMHC (segunda más alta)

'PT08.S1(CO)', # Señal del sensor de CO (tercera más alta)

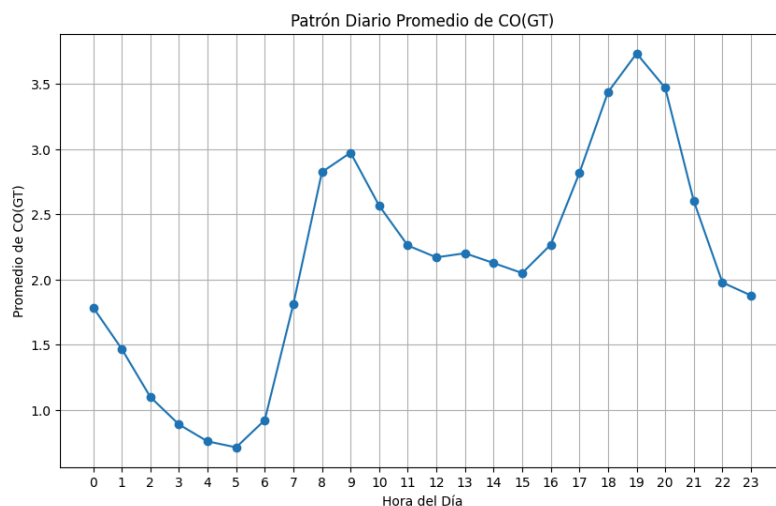
'PT08.S5(O3)', # Señal del sensor de ozono (cuarta más alta)

'NOx(GT)', # Óxidos de nitrógeno (quinta más alta)

'PT08.S3(NOx)' # Señal del sensor de NOx (correlación negativa relevante)

3. ¿Qué tipo de limpieza se realizó? La limpieza principal identificada fue el tratamiento de los valores faltantes, representados por el valor -200 en las columnas numéricas. Estos valores fueron reemplazados por NaN para facilitar su manejo en el análisis. Además, las columnas 'Date' y 'Time' requirieron ser combinadas y convertidas al formato datetime en una nueva columna 'Datetime' para permitir el análisis de series temporales y la extracción de características basadas en el tiempo.

4 ¿Las lecturas de los sensores son más útiles que las condiciones meteorológicas para predecir linealmente la concentración de CO? Respuesta: Sí, las lecturas de los cinco sensores químicos (PT08.S1 a PT08.S5) muestran correlaciones mucho más fuertes con la concentración de Monóxido de Carbono (CO(GT)) en comparación con las variables meteorológicas como la temperatura y la humedad.



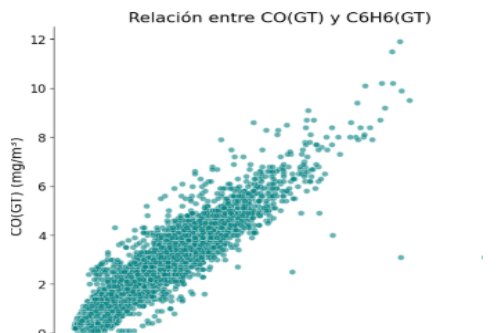
5. ¿Existen patrones cíclicos?:

El gráfico del "Patrón Diario Promedio de CO(GT)" explica los patrones cíclicos diarios. Muestra cómo varía la concentración promedio de Monóxido de Carbono a lo largo de las diferentes horas del día, revelando ese "sube y baja" que se repite diariamente.

6. ¿Hay estacionalidad y estacionariedad?

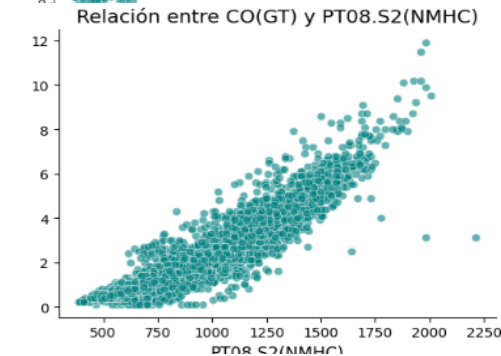
1. Estacionalidad: Sí, observamos fluctuaciones mensuales en el promedio de CO(GT) dentro del año. Sin embargo, con datos de un solo año, no podemos confirmar una estacionalidad anual consistente que se repita año tras año. Necesitaríamos datos de varios años para eso. Por ahora, vemos variaciones intra-anales.

2. Estacionariedad: Según la prueba ADF, la serie de CO(GT) es estacionaria. Esto significa que, estadísticamente, su media y varianza parecen ser constantes a lo largo del tiempo (después de tener en cuenta las fluctuaciones a corto plazo). Sin embargo, esta conclusión no niega la presencia de patrones cíclicos diarios y las variaciones mensuales. La estacionariedad en el sentido del ADF se refiere más a la ausencia de tendencias estocásticas o raíces unitarias.

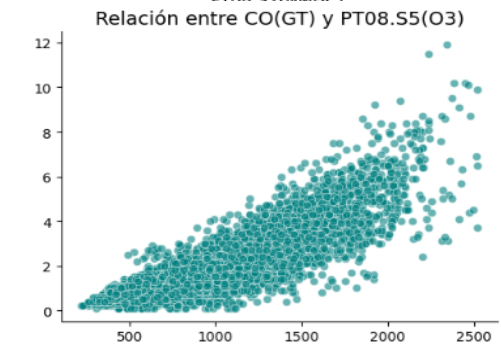


7. ¿Qué tipo de contaminante parece estar más directamente relacionado con la presencia de Monóxido de Carbono (CO)? Respuesta:

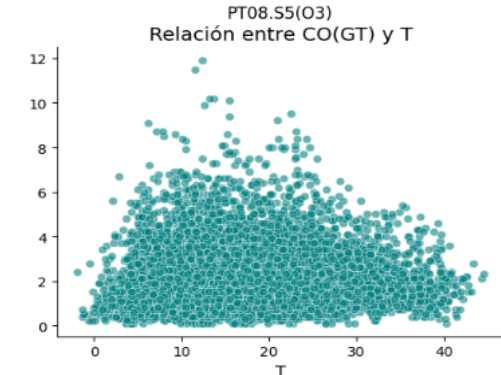
La concentración real de Benceno tiene una correlación positiva muy alta con el CO. Esto sugiere que sus fuentes de emisión están fuertemente relacionadas (principalmente tráfico vehicular y combustión).



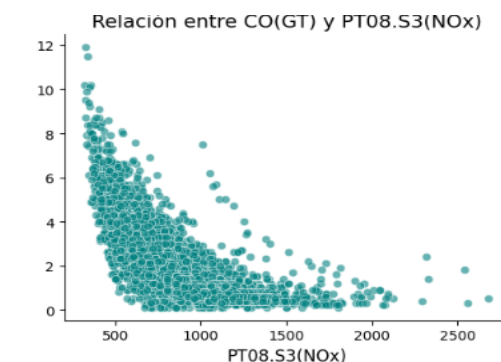
8. ¿Qué otro tipo de contaminante parece estar más directamente relacionado con la presencia de Monóxido de Carbono (CO)? Los hidrocarburos no metánicos (NMHC) muestran una fuerte correlación positiva con el Monóxido de Carbono (CO), sugiriendo una posible relación en sus fuentes de emisión o procesos atmosféricos.



9. ¿El ozono (O3) y el Monóxido de Carbono (CO) tienden a aumentar o disminuir juntos en este entorno? Según los datos, existe una correlación positiva entre la señal del sensor de ozono (PT08.S5(O3)) y la concentración de CO(GT), sugiriendo que tienden a aumentar o disminuir de manera similar.



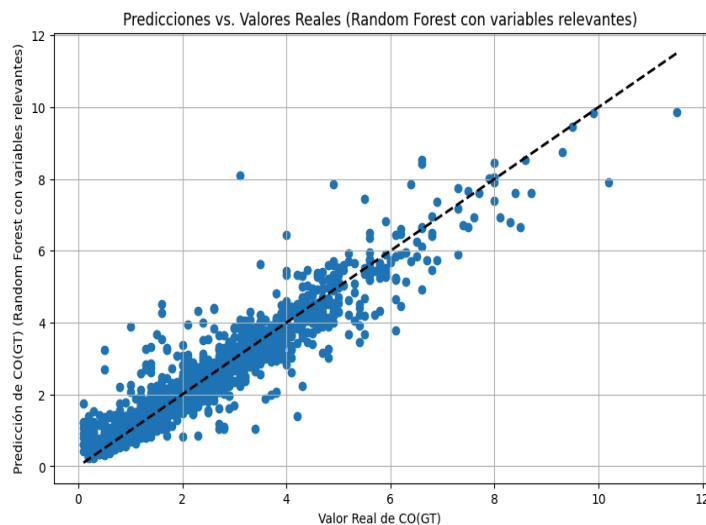
10. ¿Qué factor ambiental (de los medidos) parece tener la menor influencia lineal en la concentración de CO?: La temperatura (T) muestra la correlación lineal más débil con la concentración de Monóxido de Carbono (CO(GT)) entre las variables meteorológicas medidas en este dataset.



11. ¿Qué contaminante medido por un sensor muestra una relación inversa con el Monóxido de Carbono (CO)? La señal del sensor de Óxidos de Nitrógeno totales (PT08.S3(NOx)) presenta una correlación negativa con la concentración de Monóxido de Carbono (CO(GT)), indicando una relación inversa entre estas mediciones.

12. ¿Las condiciones de alta humedad favorecen o desfavorecen significativamente la presencia de CO según los datos? La Humedad Relativa (RH) y la Humedad Absoluta (AH) muestran correlaciones muy débiles con la concentración de CO(GT), sugiriendo que la humedad no tiene una influencia lineal significativa en los niveles de este contaminante.

13. ¿La relación entre el sensor de NOx y el CO sugiere que reducir las emisiones de NOx necesariamente aumentaría el CO? **Respuesta:** No necesariamente. La relación inversa observada entre el sensor de NOx y el CO podría ser resultado de procesos atmosféricos complejos o diferentes fuentes de emisión, por lo que una reducción de NOx no implicaría automáticamente un aumento de CO.



14. ¿Cuál es el resultado del modelo de regresión?

El modelo de regresión resultó con un **Coefficiente de Determinación (R^2):** El valor de **0.9020** y con un **Error Cuadrático Medio (MSE):** El valor de **0.2192**. El modelo de Random Forest es capaz de predecir la concentración de Monóxido de Carbono con una **precisión considerable**, explicando una alta proporción de la varianza en los datos de prueba. La señal del sensor de hidrocarburos no metánicos (**PT08.S2(NMHC)**) emerge como el factor más importante para estas

predicciones, seguido por la concentración de Benceno y los Óxidos de Nitrógeno.