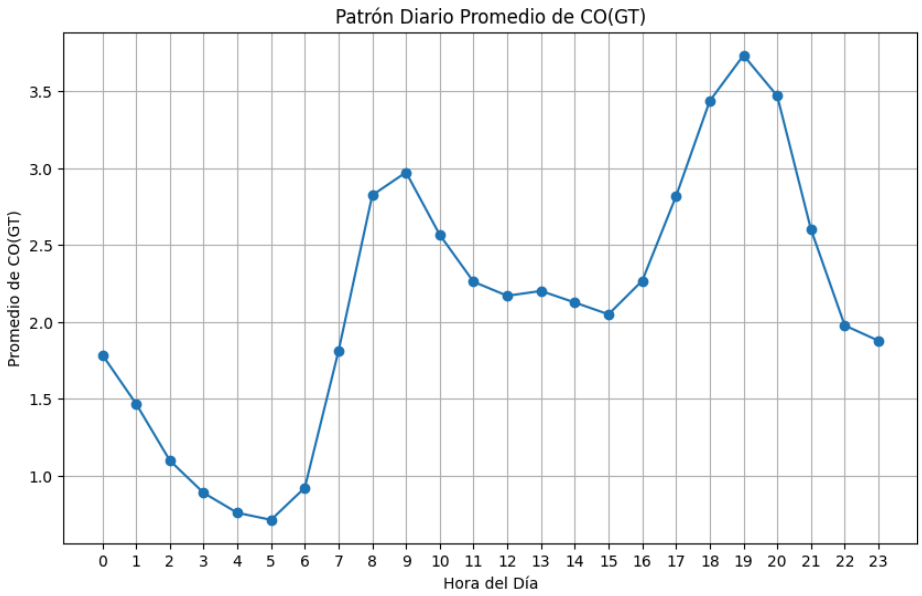
Proyecto de Análisis de Datos: Calidad del Aire



👤Profesora: Sol Del Valle FIGUEROA
👤Barozzi Eugenia
🏢 ISPC | Primer evidencia de Desarrollo de Modelo de IA machine (aprendizaje automático)

🔵 https://github.com/EugeniaBar/Calidad-del-aire

Informe de Análisis de la Calidad del Aire Urbano

1. Introducción

El **monóxido de carbono (CO)** es un gas **altamente tóxico** producto de la combustión incompleta, mientras que el **dióxido de carbono (CO_2)** es un producto de la combustión completa y la respiración, siendo el **principal contribuyente al cambio climático**. Aunque ambos están relacionados con la quema de combustibles, sus efectos y peligros son muy diferentes. En este dataset, estamos analizando el **monóxido de carbono (CO)** como un contaminante primario de la calidad del aire.

Este proyecto de análisis se enfoca en un conjunto de datos que contiene registros horarios de la respuesta de cinco sensores químicos de óxido metálico, junto con las concentraciones reales de contaminantes clave (CO, hidrocarburos no metánicos, benceno, NOx y NO2) medidas por un analizador de referencia. Los datos, recopilados durante un año (marzo 2004 - febrero 2005) en una zona urbana contaminada de Italia, representan un valioso recurso para investigar la relación entre las señales de los sensores y la calidad del aire real. El análisis explorará patrones y buscará comprender las capacidades de estimación de concentración de los sensores, considerando la presencia de sensibilidades cruzadas y posibles derivas, con el objetivo final de sentar las bases para futuros modelos predictivos de la calidad del aire. Este trabajo se realiza exclusivamente con fines de investigación.

2. Descripción del Dataset

Fuente de los datos: https://archive.ics.uci.edu/dataset/360/air+quality

Tamaño del archivo: 766.7 KB

Cantidad de registros: ~9.000 (uno por hora)

Tipo de datos: Estructurados (tablas con columnas y filas)

El dataset proviene de sensores de calidad del aire instalados en una ciudad italiana. Registra datos por hora entre marzo de 2004 y febrero de 2005. Incluye concentraciones de gases contaminantes (CO, NOx, NO_2, benceno, hidrocarburos) y variables meteorológicas como temperatura y humedad.

Desglosando las Variables:

Date, Time: Indican el momento exacto en que se tomó la medición (día, mes, año, hora, minuto, segundo). Son cruciales para analizar tendencias a lo largo del tiempo.

CO(GT) - Monóxido de carbono real (mg/m^3): Es la concentración real de este gas peligroso medida por un instrumento de referencia. Nuestro objetivo principal es entender y predecir esta

variable. El CO es un indicador de combustión incompleta, principalmente de vehículos y sistemas de calefacción.

PT08.S1(CO) - Señal del sensor de CO: Es la lectura de un sensor electrónico diseñado para ser sensible al monóxido de carbono. No es una medida directa como CO(GT), sino una señal eléctrica que varía con la cantidad de CO presente. Esperamos que esta señal esté correlacionada con la concentración real de CO.

NMHC(GT) - Hidrocarburos no metánicos reales (µg/m³): Concentración real de una variedad de compuestos orgánicos que contienen carbono e hidrógeno, excluyendo el metano. Provienen de fuentes similares al CO (combustión incompleta, emisiones de vehículos, solventes). Son importantes porque algunos NMHC son precursores de la formación de ozono troposférico (smog).

C6H6(GT) - Benceno (µg/m³): Concentración real de un hidrocarburo aromático conocido por ser cancerígeno. Es un componente de la gasolina y se emite principalmente por el tráfico vehicular y procesos industriales.

PT08.S2(NMHC) - Señal del sensor de NMHC: Lectura de un sensor electrónico diseñado para ser sensible a los hidrocarburos no metánicos. Debería estar relacionada con la concentración real de NMHC y otros compuestos orgánicos.

NOx(GT) - Óxidos de nitrógeno (ppb): Concentración real de una mezcla de óxido nítrico (NO) y dióxido de nitrógeno (NO_2). Se forman durante la combustión a altas temperaturas, principalmente en motores de vehículos y centrales eléctricas. Son contaminantes primarios que contribuyen a la formación de lluvia ácida y smog.

PT08.S3(NOx) - Señal del sensor de NOx: Lectura de un sensor electrónico diseñado para ser sensible a los óxidos de nitrógeno.

NO2(GT) - Dióxido de nitrógeno (µg/m³): Concentración real de uno de los óxidos de nitrógeno. Es un gas tóxico y un componente importante del smog fotoquímico. También puede irritar las vías respiratorias.

PT08.S4(NO2) - Señal del sensor de NO2: Lectura de un sensor electrónico diseñado para ser sensible al dióxido de nitrógeno.

PT08.S5(O3) - Señal del sensor de ozono: Lectura de un sensor electrónico diseñado para ser sensible al ozono (O_3). Aunque el ozono en la estratosfera nos protege de la radiación UV, a nivel del suelo (troposférico) es un contaminante secundario formado por reacciones químicas entre NOx y compuestos orgánicos volátiles en presencia de luz solar.

T - Temperatura (°C): Temperatura del aire en grados Celsius. Influye en la dispersión de los contaminantes y en las reacciones químicas atmosféricas.

RH - Humedad relativa (%): Porcentaje de la cantidad máxima de vapor de agua que el aire puede contener a una temperatura dada. La humedad también afecta la dispersión y la persistencia de los contaminantes.

AH - Humedad absoluta (g/m³): Cantidad real de vapor de agua presente en un metro cúbico de aire. Similar a la humedad relativa, influye en los procesos atmosféricos.

3. Análisis Exploratorio de Datos (EDA)

3.1. Descripción estadística de los datos:

1. Conteo de datos: La mayoría de las variables tienen 8991 registros, excepto NOx(GT) y NO2(GT) que tienen 7718, y Datetime que tiene 9357, indicando algunos valores faltantes en los contaminantes NOx y NO2.

Valores promedio: Los promedios varían significativamente entre las variables, reflejando diferentes unidades y escalas de medición. Por ejemplo, el CO(GT) tiene un promedio de 2.15, mientras que las señales de los sensores son mucho mayores.

Rangos de valores: Existe una amplia gama de valores para cada variable, con diferencias notables entre los mínimos y máximos, lo que sugiere variabilidad en las condiciones de calidad del aire y las lecturas de los sensores.

Distribución: Los cuartiles (25%, 50%, 75%) muestran la distribución de los datos. Por ejemplo, el 50% de las mediciones de CO(GT) están por debajo de 1.8 ppm, mientras que el 75% están por debajo de 2.9 ppm (partes por millón).

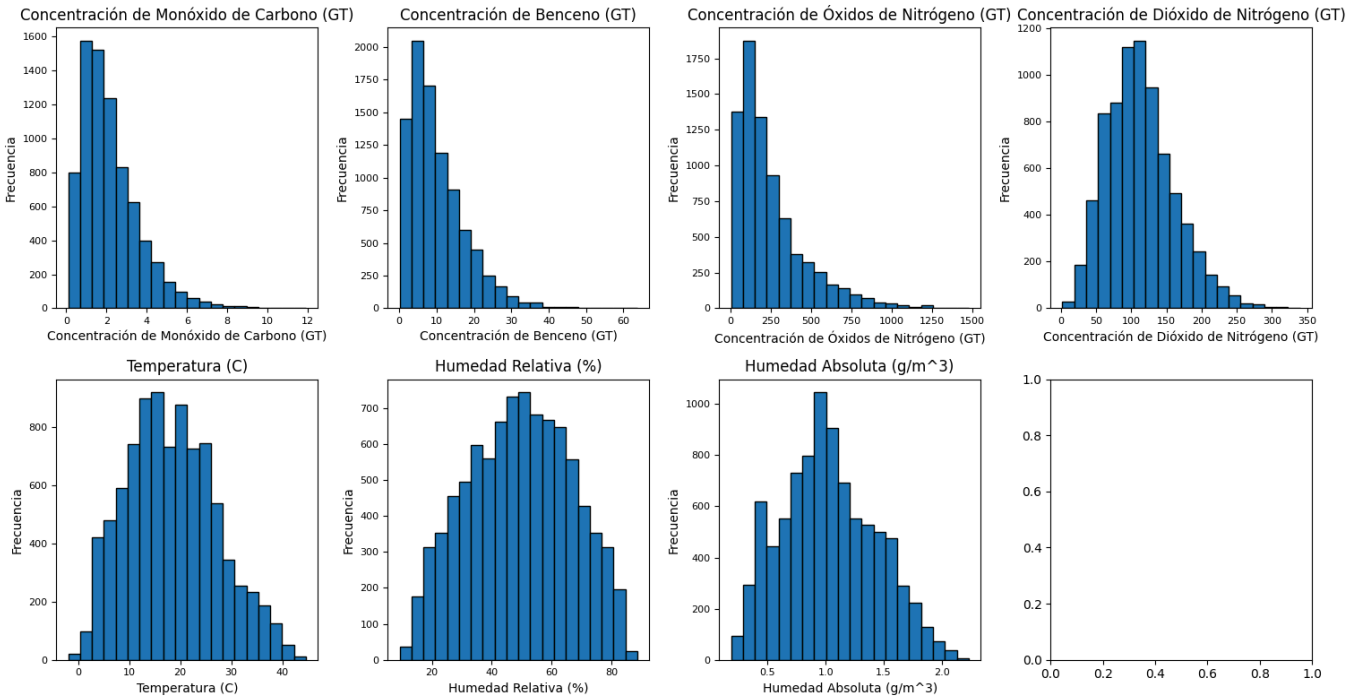
Dispersión: La desviación estándar (std) indica la dispersión de los datos alrededor de la media. Variables como PT08.S5(O3), PT08.S4(NO2) y PT08.S2(NMHC) tienen una mayor dispersión, mientras que AH tiene la menor.

Variable temporal: La columna 'Datetime' abarca un período desde marzo de 2004 hasta abril de 2005, con la mediana alrededor de septiembre de 2004.

En resumen, los datos muestran una variabilidad considerable en las concentraciones de contaminantes y las lecturas de los sensores a lo largo del tiempo, con diferentes niveles promedio y dispersión entre las variables. Se identifican algunos valores faltantes en los datos de NOx y NO2.

3.2 Distribución de los datos

Análisis de Sesgos en las Distribuciones de las Variables: La mayoría de las variables en el dataset de calidad del aire tienden a tener valores más bajos con ocurrencias ocasionales de valores mucho más altos, lo que se refleja en el sesgo a la derecha de sus distribuciones. La Humedad Relativa es una excepción, presentando una distribución más simétrica. Comprender estos sesgos es importante para la elección de modelos predictivos y la aplicación de posibles transformaciones a los datos.

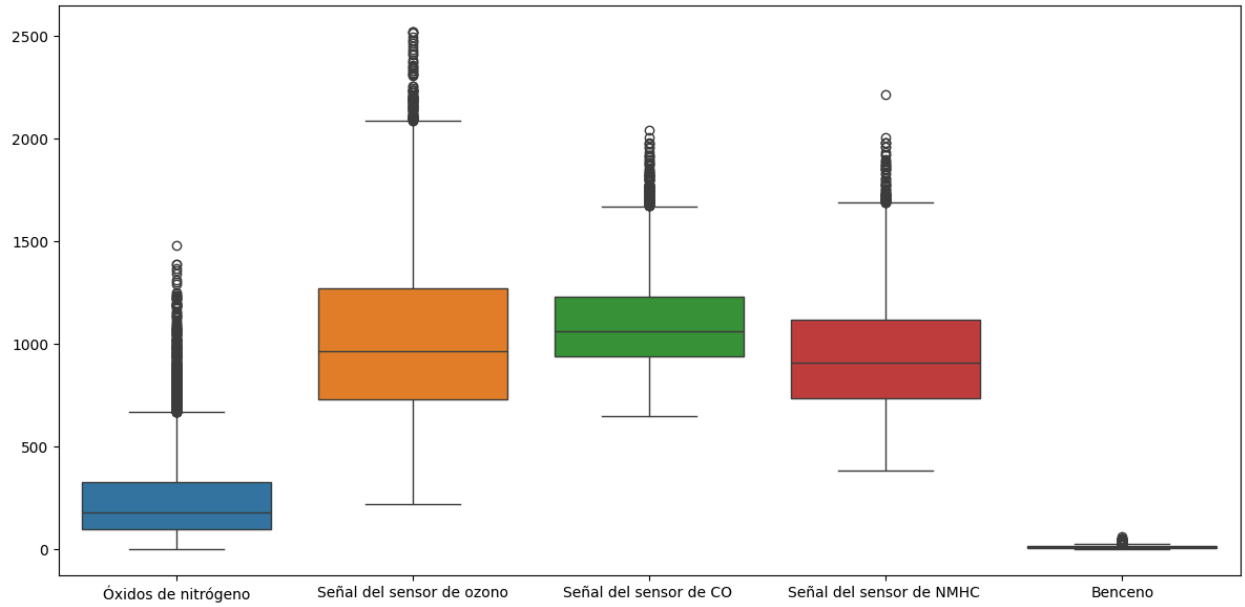


.3.3 Boxplots curtosis y outliers

Dada la naturaleza del dataset, que registra datos de calidad del aire en una zona contaminada, se consideró que estos valores extremos podrían representar picos reales de contaminación o condiciones ambientales inusuales que son relevantes para comprender el fenómeno estudiado. Además, el modelo de aprendizaje automático que demostró un rendimiento robusto en la predicción de la concentración de Monóxido de Carbono (Random Forest Regressor) es conocido por ser menos sensible a la presencia de outliers en comparación con otros algoritmos, como los modelos lineales.

Por lo tanto, para este análisis inicial y la construcción del modelo Random Forest, se decidió conservar los valores atípicos en el dataset. Se reconoce que en futuros análisis, especialmente si se emplean modelos sensibles a los outliers o si el objetivo se centra en la

detección de eventos extremos, podría ser necesario reevaluar y aplicar técnicas específicas para el manejo de estos valores.



6. Scatterplot entre las variables

Relaciones Positivas con CO(GT)

PT08.S1(CO): Muestra una fuerte relación lineal positiva. Los puntos se agrupan estrechamente a lo largo de una línea ascendente. Esto era esperado ya que este sensor está diseñado para medir CO.

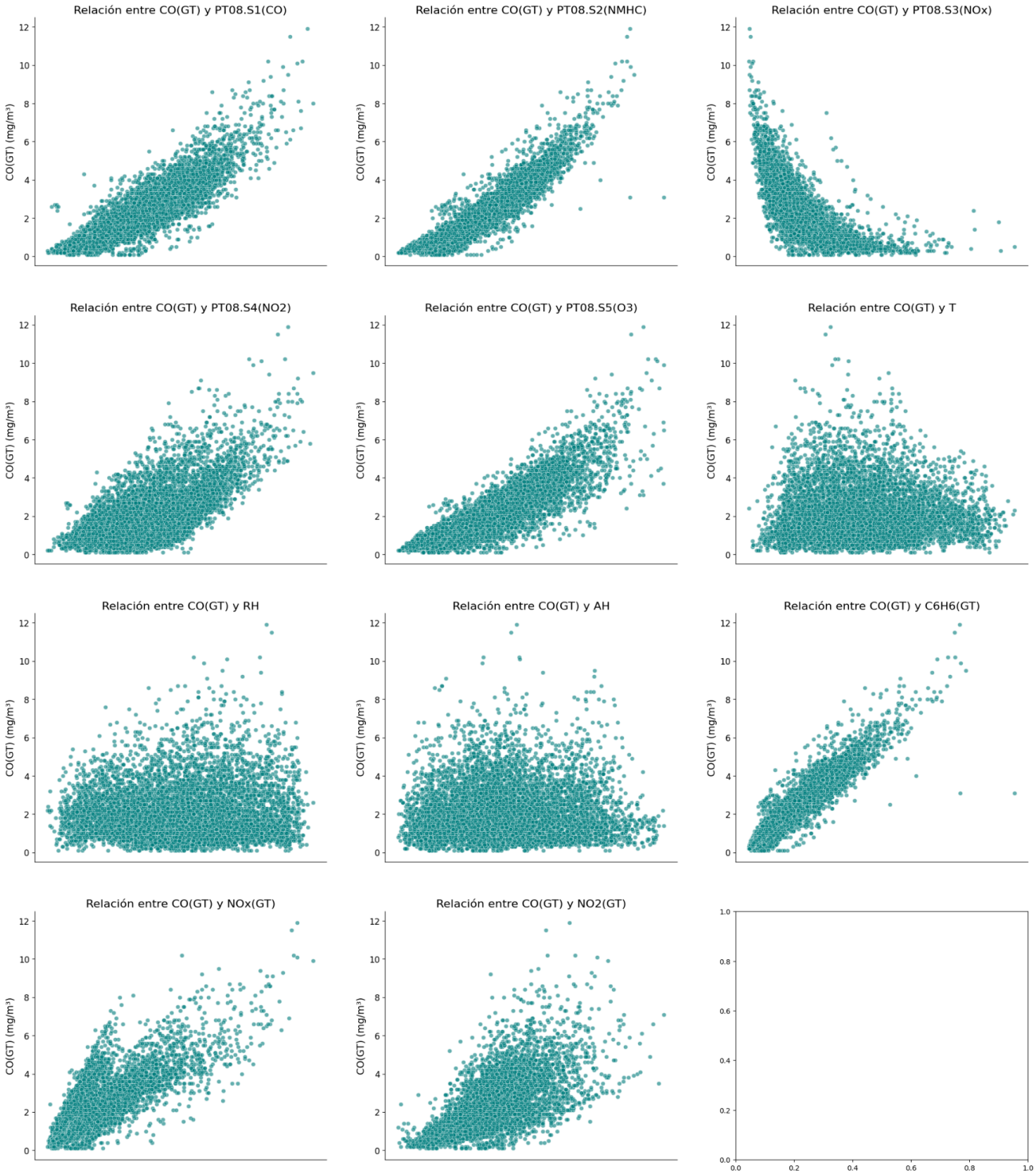
C6H6(GT): La concentración real de Benceno tiene una correlación positiva muy alta con el CO. Esto sugiere que sus fuentes de emisión están fuertemente relacionadas (principalmente tráfico vehicular y combustión).

PT08.S2(NMHC): 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.

PT08.S5(O3): El sensor de Ozono parece tener una relación positiva, aunque con una dispersión considerable. A medida que aumenta la lectura del sensor de O3, tiende a aumentar el CO(GT), pero la relación no es tan estrecha como las anteriores.

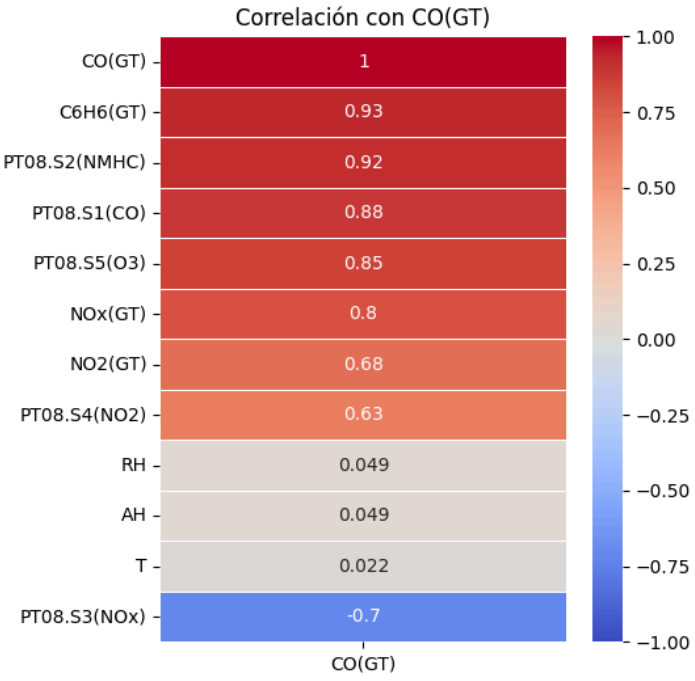
NOx(GT): Óxidos de nitrógeno Presenta una relación positiva, aunque con una dispersión significativa, especialmente en los valores más altos de NOx. La tendencia general es ascendente.

NO2(GT):Dióxido de nitrógeno muestra una relación positiva, pero parece ser más débil y con mayor dispersión que las anteriores.



7. Correlación de Pearson con CO(GT)

Correlaciones Positivas (en orden de relevancia descendente):



PT08.S1(CO) (0.88): La señal del sensor de CO muestra una correlación positiva muy alta con la concentración real de CO. Esto es lógico ya que el sensor está diseñado para medir este gas.

C6H6(GT) (0.93): 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).

PT08.S2(NMHC) (0.92): La señal del sensor de hidrocarburos no metánicos también muestra una correlación positiva muy alta con el CO, lo que indica una fuente de emisión común (combustión incompleta).

PT08.S5(O3) (0.85): La señal del sensor de ozono tiene una correlación positiva alta con el CO. Aunque la relación no es directa, podría indicar condiciones atmosféricas o patrones de tráfico que favorecen la acumulación de ambos.

NOx(GT) (0.80): La concentración real de óxidos de nitrógeno muestra una correlación positiva alta con el CO, debido a que ambos son productos de la combustión a alta temperatura.

NO2(GT) (0.68): La concentración real de dióxido de nitrógeno (un componente de los NOx) tiene una correlación positiva moderada con el CO.

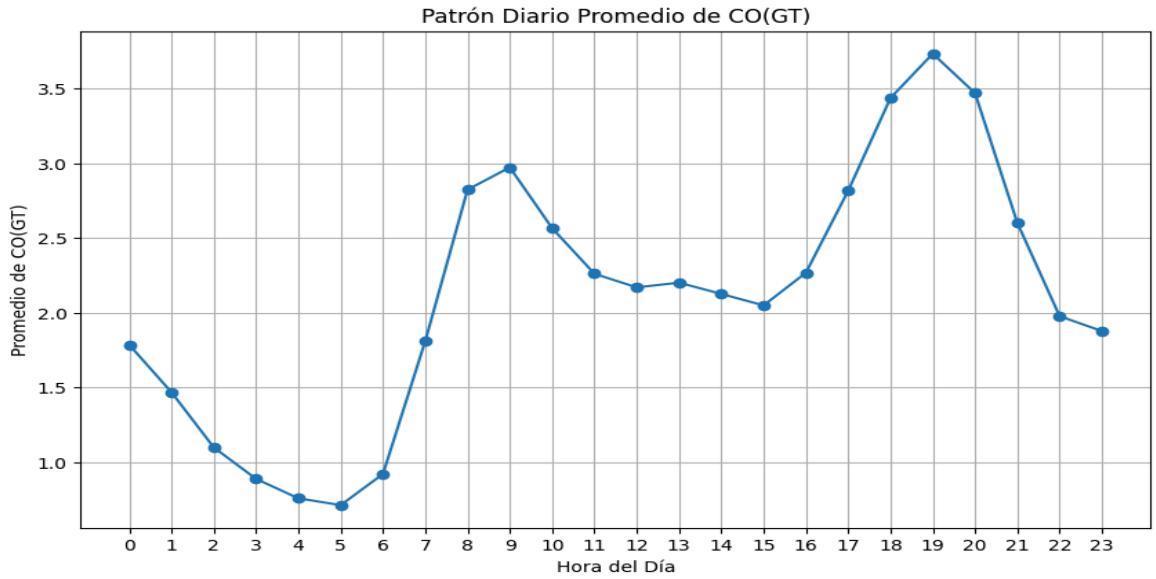
PT08.S4(NO2) (0.63): La señal del sensor de dióxido de nitrógeno muestra una correlación positiva moderada con el CO.

Correlación Negativa:

PT08.S3(NOx) (-0.70): La señal del sensor de NOx tiene una correlación negativa moderada a alta con el CO. Esto podría indicar una respuesta compleja del sensor o dinámicas químicas inversas en ciertas condiciones.

8. Análisis de las Propiedades Temporales de la Serie de CO(GT)

8.1 Patrón Diario Promedio de CO(GT):



El gráfico muestra un patrón diario claro en la concentración promedio de Monóxido de Carbono (CO). Los niveles son más bajos durante la noche y la madrugada (aproximadamente desde la medianoche hasta las 6 AM). A partir de las 7 AM, se observa un rápido aumento que alcanza un pico alrededor de las 8-9 AM, probablemente coincidiendo con el tráfico matutino. Los niveles disminuyen gradualmente durante la mañana y vuelven a aumentar por la tarde, con un segundo pico alrededor de las 18-19 PM, asociado al tráfico vespertino. Posteriormente, la concentración de CO desciende de nuevo durante la noche. En resumen, el CO presenta dos picos diarios, uno por la mañana y otro por la tarde, con valores mínimos durante las horas de menor actividad.

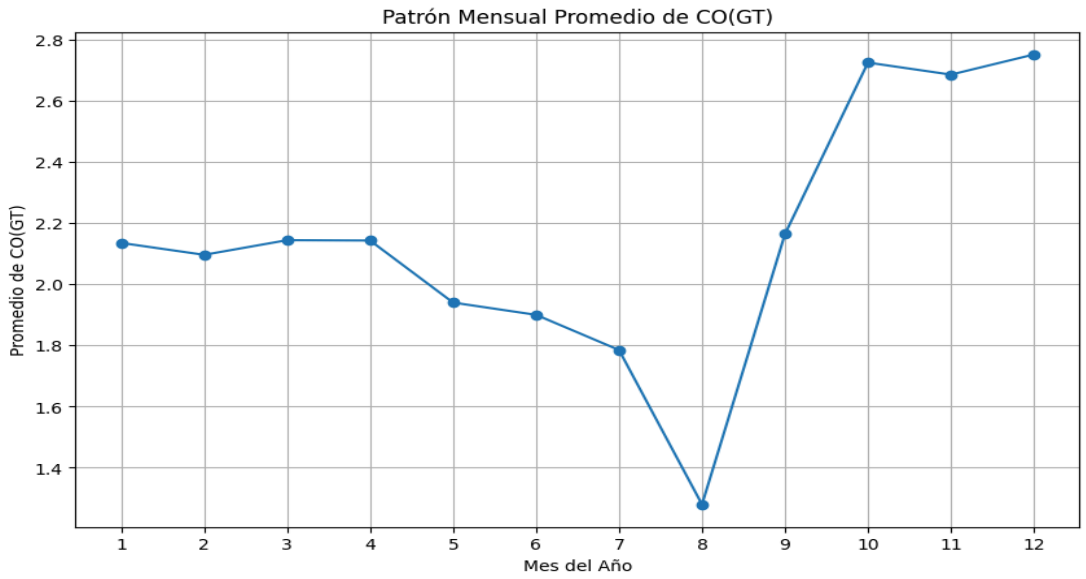
8.2 PATRONES MENSUALES

Patrón Mensual Promedio de CO(GT) y Énfasis en Agosto:

El gráfico ilustra la variación mensual promedio de la concentración de Monóxido de Carbono (CO) a lo largo del año. Se observa una tendencia general a tener niveles más elevados de CO durante los meses más fríos (aproximadamente desde octubre hasta marzo), con picos

notables al inicio y al final del año. En contraste, los niveles de CO tienden a ser más bajos durante los meses de primavera y verano.

Lo más destacable del gráfico es la disminución drástica y significativa en la concentración promedio de CO que ocurre específicamente en el mes 8 (agosto). Este mes presenta el valor promedio de CO más bajo de todo el año, marcando un punto de inflexión en la tendencia general.



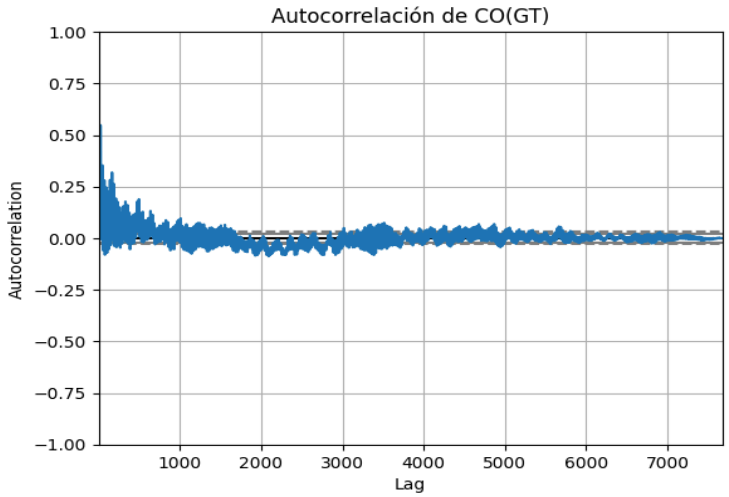
Está marcada reducción en agosto sugiere la influencia de factores estacionales particulares. Como se discutió anteriormente, es probable que esta disminución se deba a una combinación de la reducción de la actividad humana, principalmente el tráfico vehicular debido a las vacaciones de verano generalizadas en Italia durante agosto, y condiciones meteorológicas más favorables para la dispersión de los contaminantes, como una mayor inestabilidad atmosférica y patrones de viento que evitan la acumulación de CO a nivel del suelo. La ausencia de la necesidad de calefacción en este mes cálido también contribuye a la disminución de las emisiones relacionadas con la combustión residencial.

En esencia, el mes de agosto se presenta como una anomalía dentro del patrón anual, caracterizado por una mejora notable en la calidad del aire en términos de concentración de Monóxido de Carbono, presumiblemente impulsada por una menor actividad antropogénica y condiciones ambientales que favorecen la dispersión de los contaminantes.

8.3 Autocorrelación

La Función de Autocorrelación (ACF) es una herramienta fundamental que tiene una relación directa tanto con la estacionalidad como con la estacionariedad de una serie temporal. Este

Gráfico de Autocorrelación para CO(GT) nos dice que: La concentración de Monóxido de Carbono tiene una fuerte dependencia temporal a corto plazo: El valor actual está muy influenciado por el valor inmediatamente anterior.



Existen posibles patrones cíclicos: Hay indicios de ciclos diarios y, posiblemente, semanales en la concentración de CO. La dependencia temporal directa se desvanece con el tiempo: Las concentraciones de CO en momentos muy distantes en el pasado tienen poca o ninguna correlación directa con la concentración actual. Esta información es valiosa para entender la dinámica de la contaminación por CO y puede ser útil al construir modelos predictivos de series temporales, ya que sugiere que los valores pasados de CO son predictores importantes de los valores futuros.

8.4 ¿Hay estacionalidad y estacionariedad?

- **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-anuales.

- **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.

La clave para reconciliar esto:

Una serie puede ser **estacionaria** (según pruebas como el ADF) pero aún así tener **patrones cíclicos predecibles** (como el patrón diario) o **variaciones intra-anuales**. Estos patrones no necesariamente violan la definición de estacionariedad en el sentido de que la media y la

varianza generales sean constantes, pero sí implican una estructura temporal que puede ser modelada.

9. Modelado de Machine Learning Regresión Lineal y Transformaciones:

Objetivo: Predecir la concentración de Monóxido de Carbono (CO(GT)).

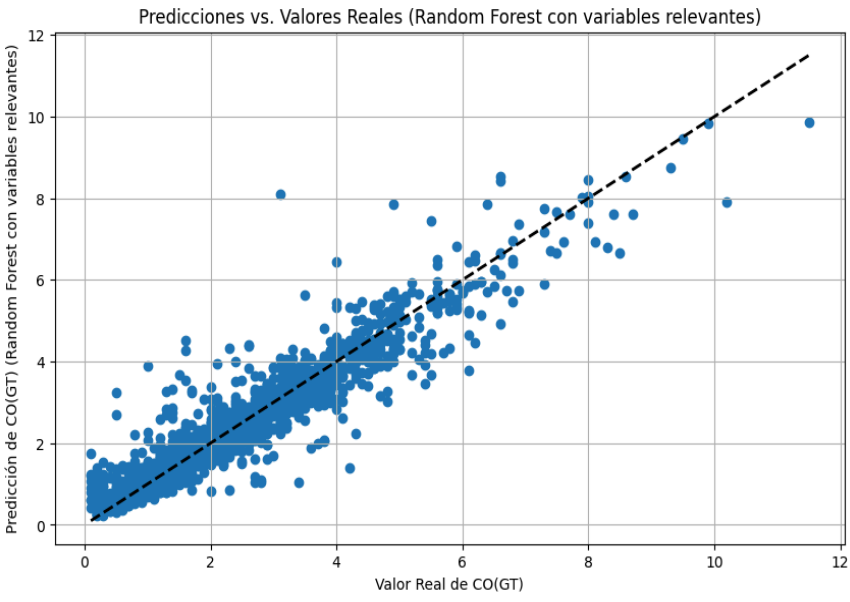
Modelo Utilizado: Primero Regresión Lineal, finalmente Random Forest Regressor (justificación por su robustez y buen rendimiento inicial).

Variables Predictoras: Listado de las 5 variables utilizadas (sensores y meteorológicas).

Inicialmente, se construyó un modelo de regresión lineal para predecir CO(GT) después de aplicar transformaciones logarítmicas a la variable objetivo y a algunas de las variables predictoras para abordar los sesgos observados en sus distribuciones. Sin embargo, encontramos problemas que impidieron el entrenamiento exitoso del modelo de regresión lineal, manifestándose en resultados con valores NaN (Not a Number).

La causa principal de este problema de NaN se debió a la presencia de valores NaN en la variable objetivo (CO(GT)) después de la transformación logarítmica, lo que probablemente ocurrió porque la columna original 'CO(GT)' contenía valores faltantes que no se habían eliminado correctamente antes de la transformación.

Aunque se intentaron pasos para limpiar los datos, persistieron los problemas con la regresión lineal, lo que nos llevó a explorar un modelo no lineal como el Random Forest Regressor, el cual demostró ser más robusto y capaz de generar predicciones válidas sin las restricciones de linealidad y distribución normal de los errores inherentes a la regresión lineal. Esto sugiere que para este dataset y la predicción de CO(GT), un modelo no lineal podría ser una mejor opción o que se necesitaría una estrategia de preprocesamiento y transformación diferente para que la regresión lineal funcione correctamente.



Elección final Random Forest Regressor:

Nuestro 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).

Y = datos CO(GT): Monóxido de carbono

X = datos

- - '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)

Conclusión del Modelo Random Forest para la Predicción de CO(GT):

El modelo de Random Forest Regressor entrenado utilizando las variables predictoras más relevantes (`PT08.S2(NMHC)`, `C6H6(GT)`, `NOx(GT)`, `PT08.S1(CO)`, `PT08.S5(O3)`, y `PT08.S3(NOx)`) muestra un **rendimiento predictivo sólido** para la concentración de Monóxido de Carbono (`CO(GT)`).

- **Coeficiente de Determinación (R^2):** El valor de **0.9020** indica que el modelo explica aproximadamente el **90.20% de la varianza** en la concentración real de CO en el conjunto de prueba. Esto sugiere que el modelo es capaz de capturar una gran parte de los patrones y las relaciones entre las variables predictoras y el CO.

- **Error Cuadrático Medio (MSE):** El valor de **0.2192** representa el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales de CO. Un valor relativamente bajo sugiere que las predicciones del modelo están, en promedio, bastante cerca de los valores reales.

- **Gráfico de Predicciones vs. Valores Reales:** El gráfico muestra una **tendencia clara de alineación** entre las predicciones del modelo y los valores reales de CO. Los puntos se agrupan razonablemente cerca de la línea diagonal punteada (que representa una predicción perfecta). Sin embargo, se observa cierta dispersión, especialmente en los valores más altos de CO, lo que indica que el modelo no es perfecto y existen errores en las predicciones individuales.

- **Importancia de las Características:** La importancia de las características revela que **`PT08.S2(NMHC)` (Señal del sensor de NMHC)** es, con diferencia, el predictor más influyente en el modelo. Le siguen en importancia `C6H6(GT)` (Benceno) y `NOx(GT)` (Óxidos de Nitrógeno). Las señales de los sensores de CO (`PT08.S1(CO)`) y ozono (`PT08.S5(O3)`), así como la señal del sensor de NOx (`PT08.S3(NOx)`), tienen una

importancia menor, aunque aún contribuyen al modelo.

En resumen:

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.

Aunque el modelo muestra un buen rendimiento general, la dispersión observada en el gráfico de predicciones vs. valores reales sugiere que aún existen errores y que el modelo podría beneficiarse de una mayor optimización, la inclusión de otras variables relevantes o la exploración de diferentes algoritmos de modelado. Sin embargo, como punto de partida, este modelo proporciona una base sólida para la predicción de la concentración de Monóxido de Carbono.

11. De cara a futuros proyectos y una mejora:

Nos dimos cuenta de que nuestro modelo actual de Random Forest no aprovecha al máximo que nuestros datos cambian con el tiempo. Para proyectos futuros, queremos mejorar esto y poder predecir el CO pensando en cómo evoluciona en el tiempo.

Por eso, para la siguiente etapa, queremos probar modelos que se usan especialmente para datos que cambian con el tiempo, como los modelos ARIMA. Estos modelos son buenos para entender cómo los valores anteriores de CO nos pueden ayudar a predecir los valores futuros.

Esto nos ayudaría a responder preguntas como: "¿Cuánto CO habrá en las próximas horas, basándonos en cómo ha estado cambiando y en lo que nos dicen los sensores ahora?". Así, podríamos tener un sistema para predecir la calidad del aire que sea más dinámico y nos dé una idea de lo que va a pasar en el corto plazo.

En conclusión, este proyecto ha sentado una base sólida para la comprensión y la predicción de la calidad del aire utilizando técnicas de machine learning. La exploración de la dependencia temporal a través del análisis de autocorrelación nos motiva a explorar modelos de series temporales como ARIMA / SARIMA en futuros trabajos, con la expectativa de mejorar nuestra capacidad para predecir la evolución de la calidad del aire a lo largo del tiempo.
