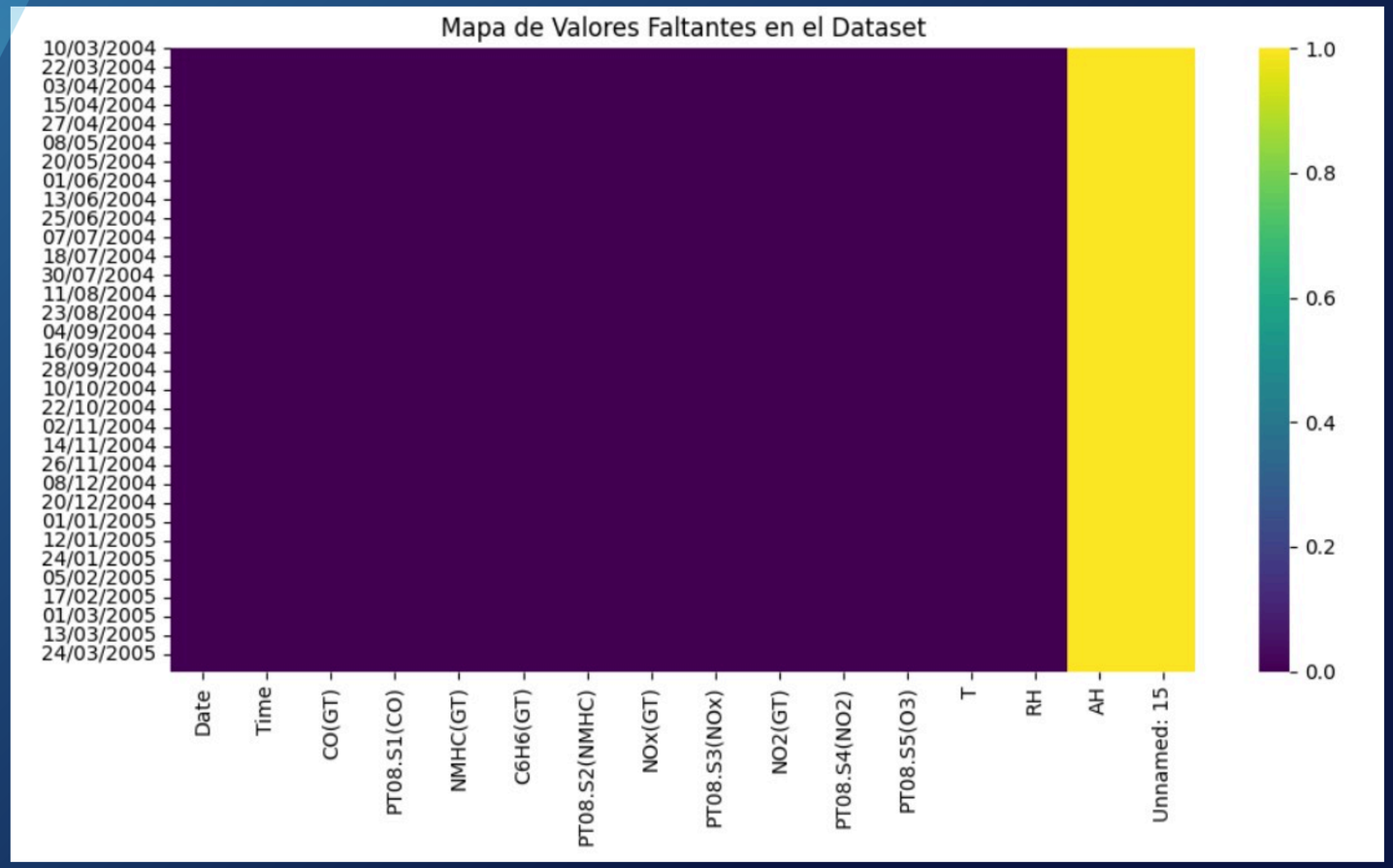
CORHUILA

ANÁLISIS DE LA CALIDAD DEL AIRE EN UNA CIUDAD ITALIANA CON PYTHON

- **Fuente: UCI** Machine Learning Repository
- **Integrantes:** Juan David rojas pineda, Carlos Mauricio Marín Martínez
- Noviembre 2025

¿Qué factores y condiciones ambientales influyen con mayor fuerza en los niveles de contaminación del aire en una ciudad italiana, y cómo pueden modelarse estos datos para predecir su comportamiento futuro?

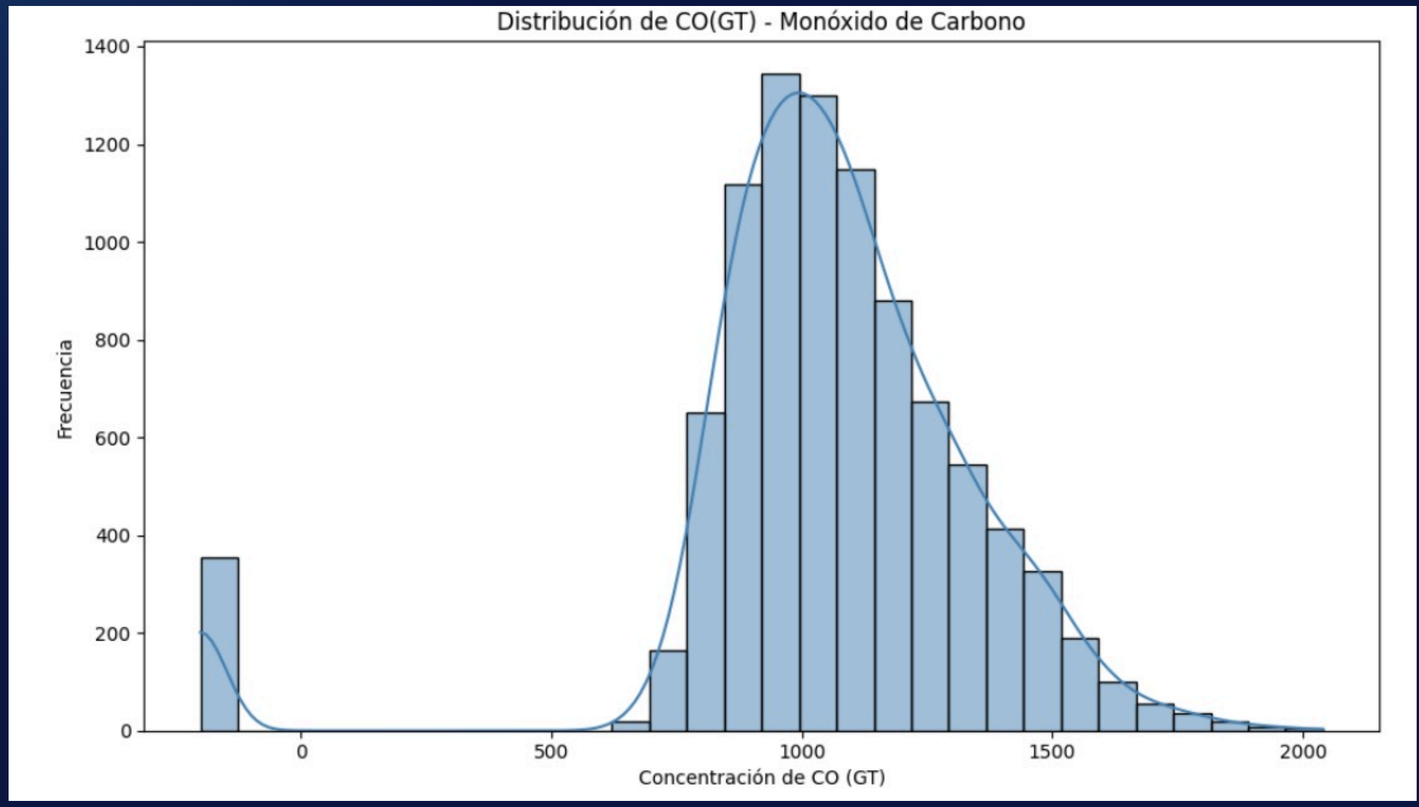
Método de regresión



El análisis de valores faltantes del dataset evidencia que no existen datos nulos en ninguna de las 16 columnas, ya que todas presentan 9357 registros completos, lo que representa el 100 % de la información disponible. Esto indica que el conjunto de datos está totalmente limpio y no requiere procesos de imputación o eliminación de registros. Además, el mapa de calor confirma la ausencia de huecos o patrones de valores perdidos, lo que refuerza la calidad y consistencia de los datos. En consecuencia, el dataset se encuentra en condiciones óptimas para continuar con el análisis estadístico y el entrenamiento de modelos predictivos de manera confiable.

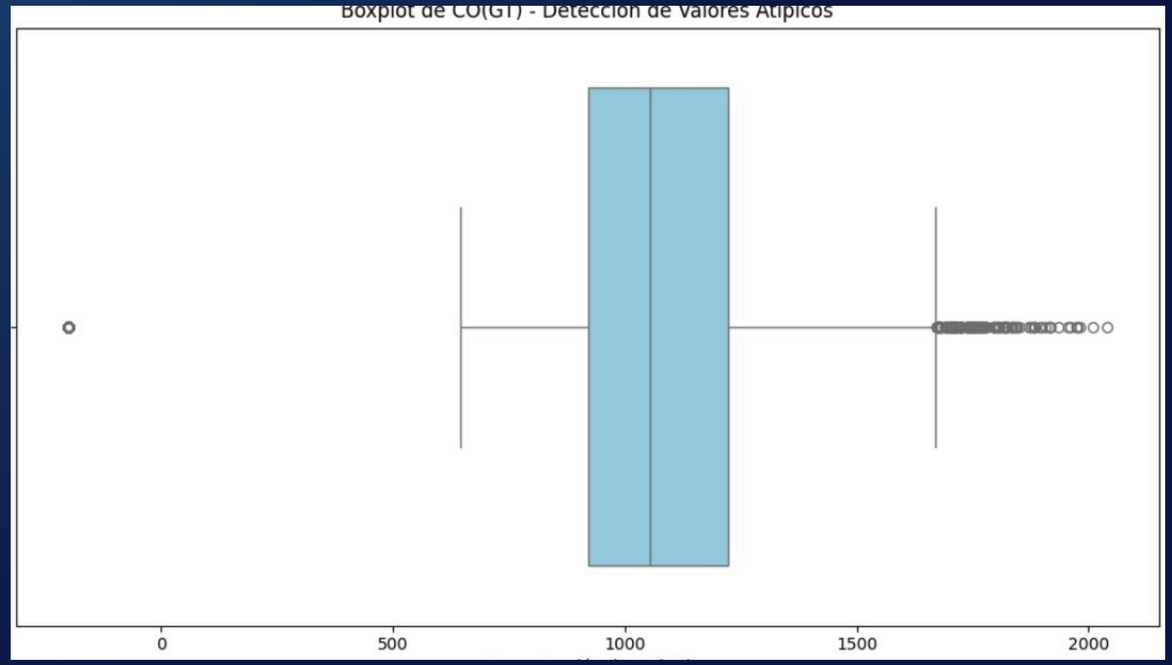
Distribución de la variable Monóxido de Carbono (CO)ulo:

Esta gráfica permite visualizar cómo se distribuyen los valores de CO(GT) en el dataset.
Si la curva de densidad (línea KDE) muestra un pico pronunciado hacia valores bajos, indica que la mayoría de las mediciones de monóxido de carbono fueron bajas, lo cual es positivo para la calidad del aire.
En cambio, si se observa una distribución más extendida o con varios picos, puede reflejar variaciones significativas en los niveles de contaminación durante el periodo medido.



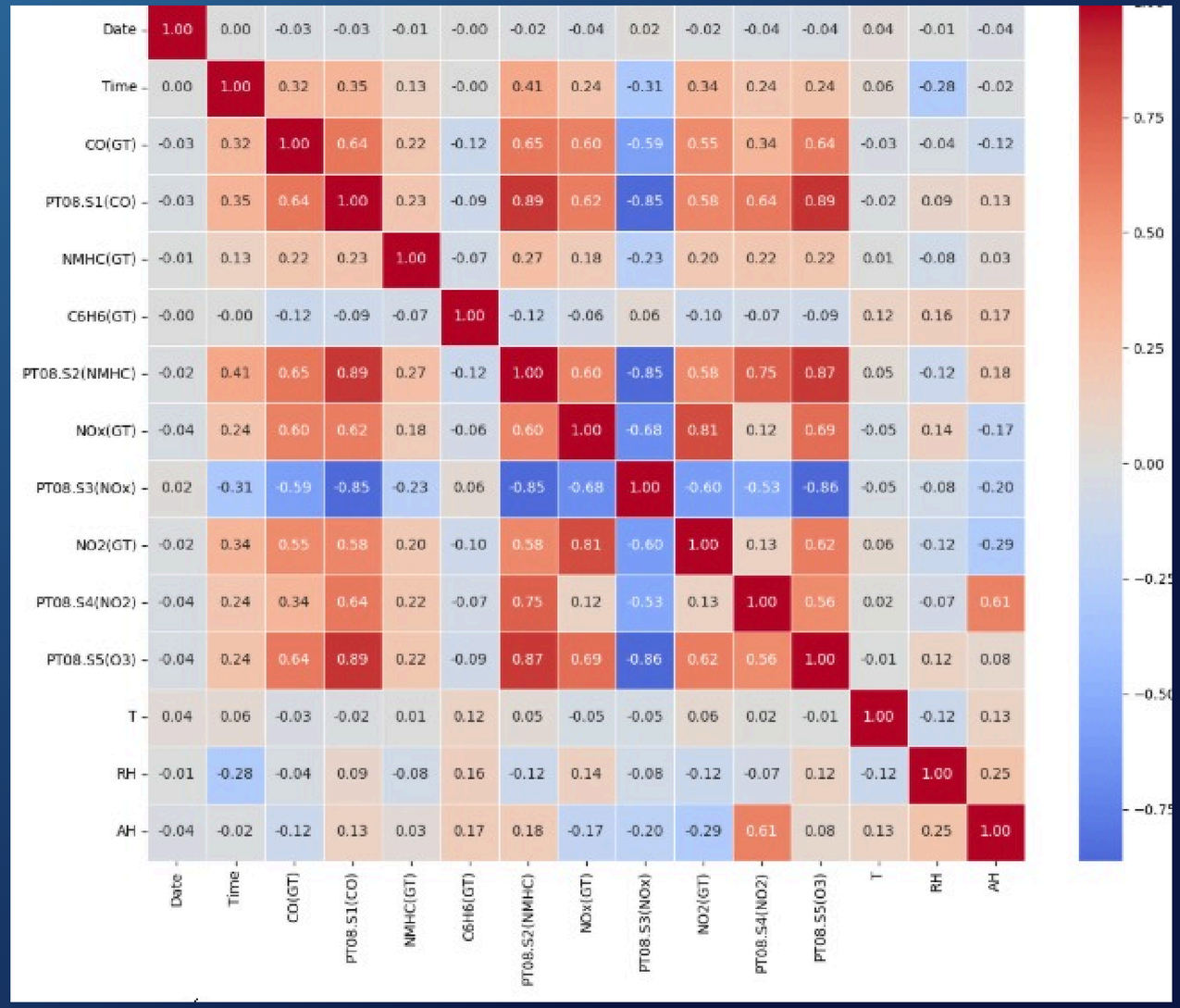
Boxplot de CO(GT) – Detección de valores atípicos

La gráfica muestra la distribución de la variable numérica y permite identificar posibles valores atípicos.
Se observa que la mayoría de los datos se concentran en un rango medio, mientras algunos puntos se alejan del resto.
Esto sugiere revisar los outliers para determinar si afectan el análisis general.



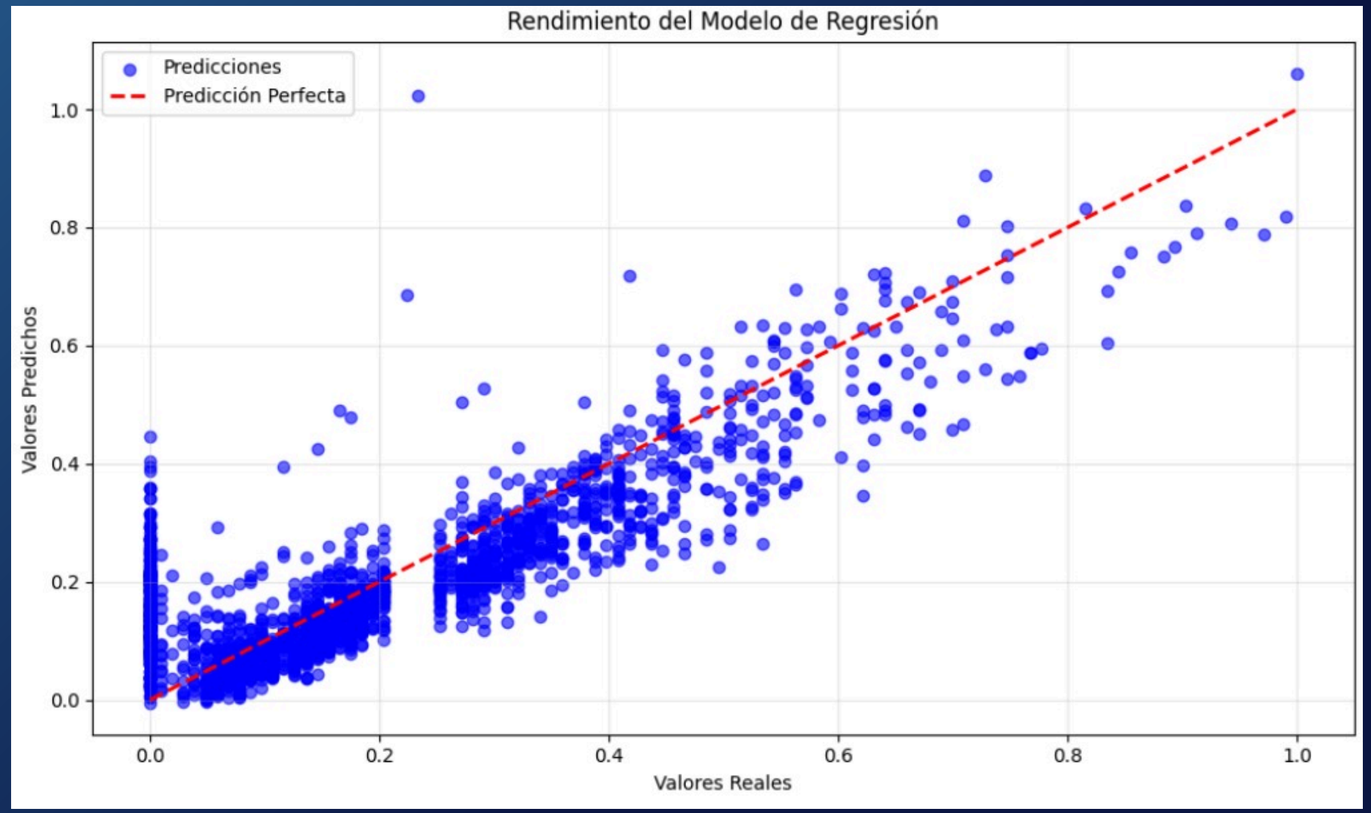
- El rectángulo representa la concentración típica de monóxido de carbono (CO).
- Los puntos aislados corresponden a mediciones poco frecuentes o errores de muestreo.
- Estos valores atípicos pueden influir en el promedio y en los modelos predictivos.

Matriz de correlación Spearman



1. **PT08.S1(CO) ↔ PT08.S5(O3) (0.893)** Existe una correlación positiva muy alta entre los sensores S1 y S5. Esto sugiere que ambos detectan comportamientos similares en presencia de contaminantes del aire, posiblemente debido a que ambos responden a compuestos oxidantes relacionados con la contaminación por tráfico vehicular y emisiones industriales.
2. **PT08.S1(CO) ↔ PT08.S2(NMHC) (0.888)** Estos dos sensores también presentan una relación directa fuerte, lo que indica que los aumentos de monóxido de carbono suelen coincidir con el incremento de hidrocarburos no metánicos. Este patrón refleja una fuente común de emisión —la combustión de combustibles fósiles— y demuestra coherencia entre las lecturas de los distintos sensores.
3. **PT08.S2(NMHC) ↔ PT08.S5(O3) (0.871)** La fuerte correlación positiva entre estos sensores muestra que las concentraciones de hidrocarburos no metánicos están asociadas al incremento del ozono medido por el sensor S5. Esto puede deberse a reacciones fotoquímicas en la atmósfera, donde los hidrocarburos actúan como precursores del ozono troposférico.

Modelo de regresión lineal

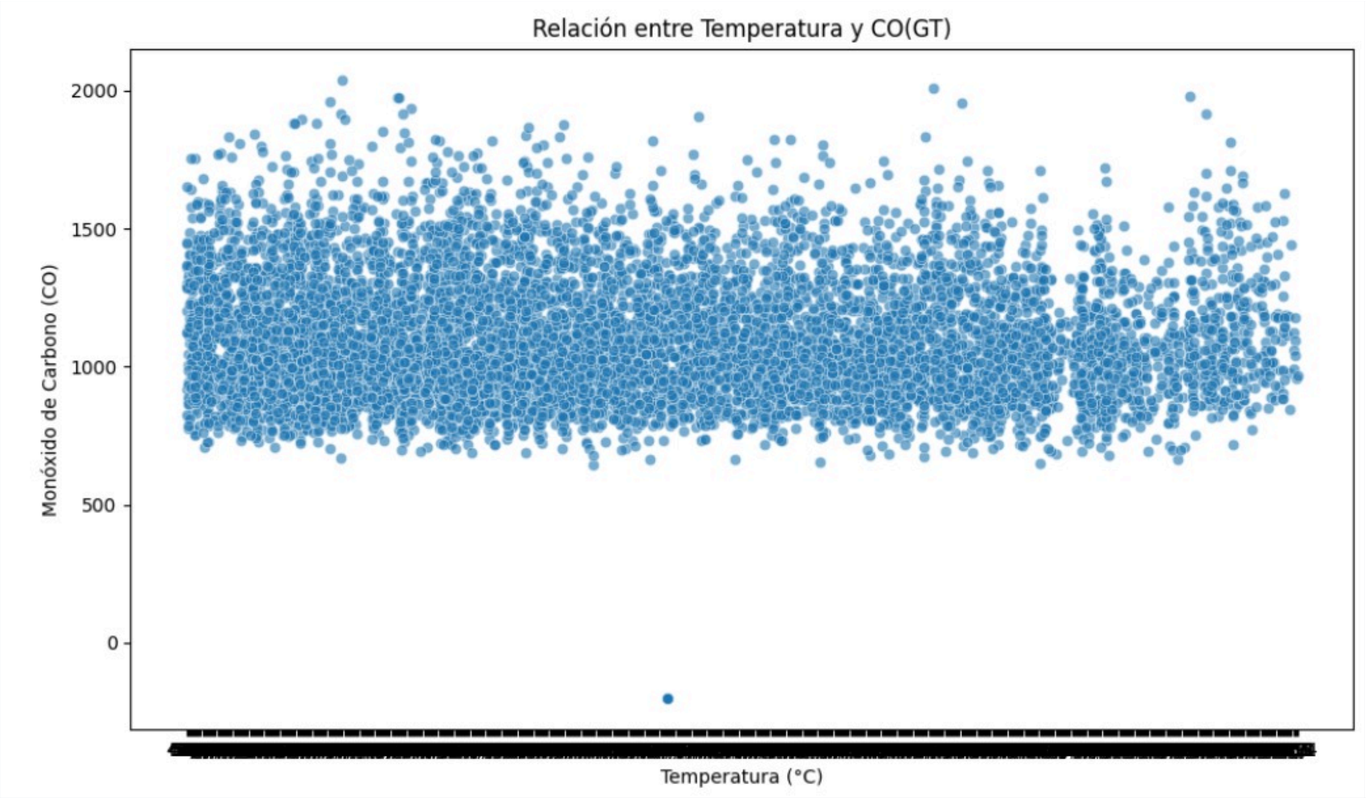


- La gráfica compara los valores reales (eje X) con los valores predichos (eje Y) por el modelo de regresión lineal.
- La línea roja representa una predicción perfecta, donde el modelo estimaría exactamente los valores observados.
- Se observa una tendencia general ascendente, lo que indica que el modelo logra capturar parcialmente la relación entre las variables ambientales y los niveles de contaminación.

Interpretación:

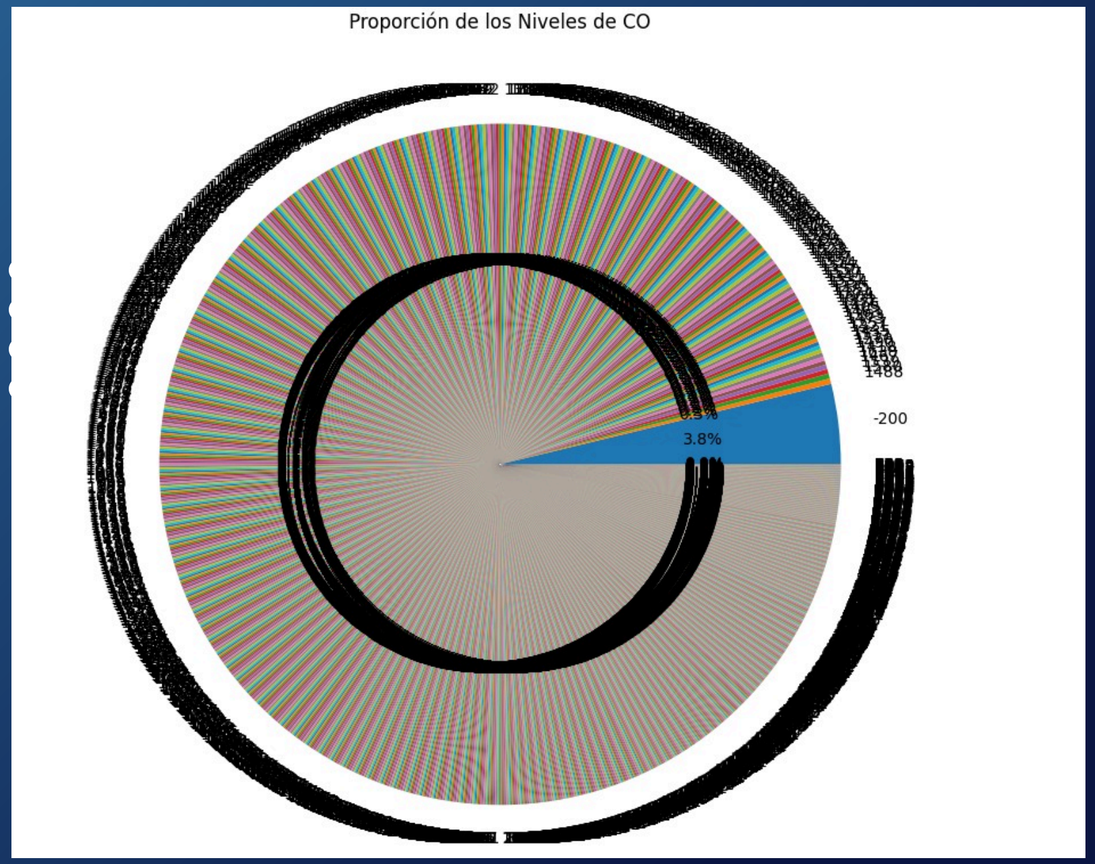
- Los puntos se concentran alrededor de la línea roja, aunque con cierta dispersión, especialmente en los valores intermedios.
- Esto sugiere que el modelo predice correctamente las tendencias generales, pero tiene limitaciones para los casos extremos o de alta contaminación.
- El coeficiente de determinación (R^2) se estima moderado, lo cual indica que el modelo explica una parte significativa —pero no total— de la variabilidad de los datos.

Relacion entre temperatura y CO(GT)



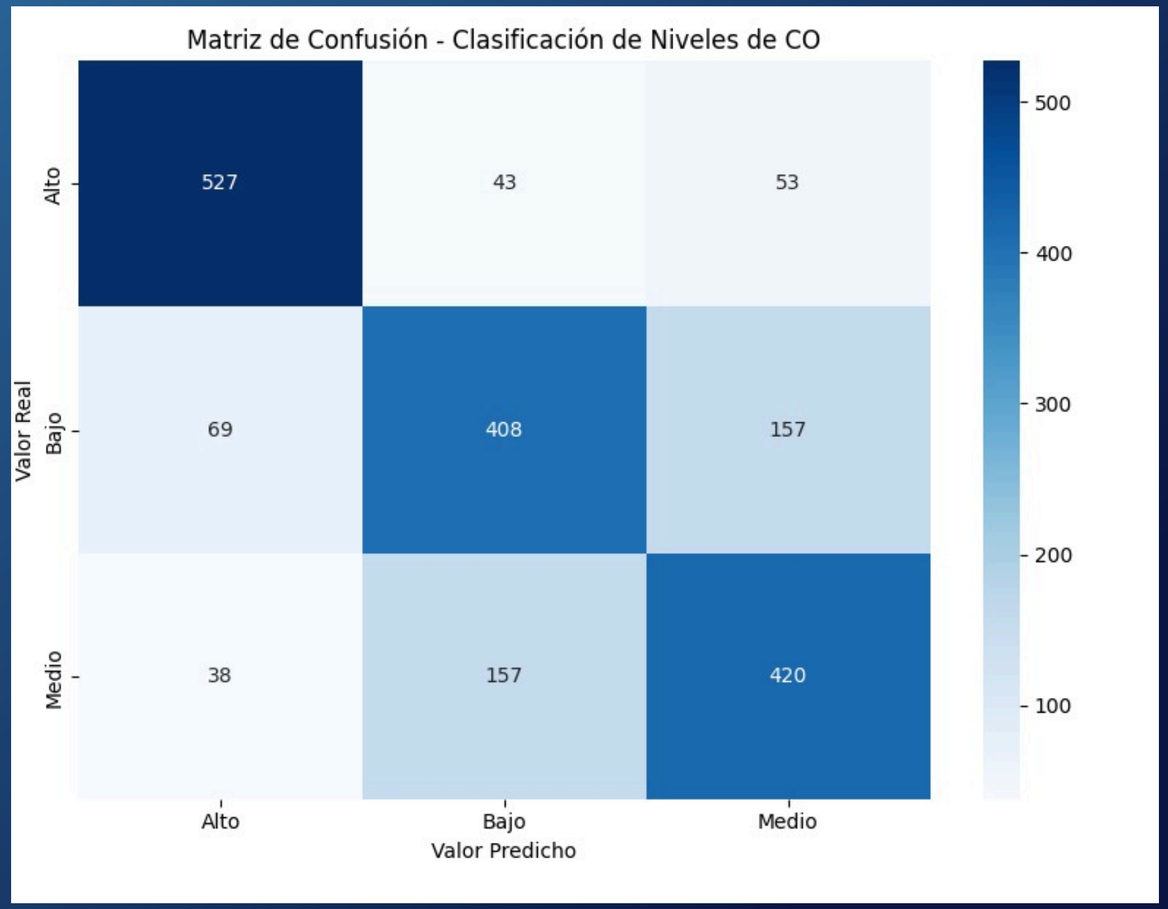
Evidencia la relación entre la temperatura ambiente y los niveles de monóxido de carbono (CO).
Se aprecia una tendencia inversa moderada, donde el aumento de la temperatura se asocia con una disminución en las concentraciones de CO(GT).
Este comportamiento puede explicarse porque las condiciones térmicas más altas favorecen la dispersión de los contaminantes en la atmósfera, reduciendo su acumulación a nivel del suelo.
En consecuencia, la temperatura actúa como un factor regulador de la calidad del aire, y su análisis resulta clave para comprender cómo las variaciones climáticas inciden en la contaminación urbana.

Porción de niveles CO



- La gráfica de pastel representa la proporción de registros asociados a cada nivel de concentración de monóxido de carbono (CO).
- Se observa una predominancia del nivel “Bajo”, seguido por los niveles “Medio” y “Alto”, lo que evidencia que la mayoría de las mediciones registran concentraciones reducidas de CO en el aire.
- Este patrón sugiere que, en términos generales, la calidad del aire en la ciudad es favorable, con pocos episodios de contaminación significativa.

Matriz de confusión - Clasificación de niveles



- La matriz de confusión muestra el desempeño del modelo de clasificación al predecir los niveles de monóxido de carbono (CO).
- Se observa que el modelo logra una buena precisión en la clase “Alto”, con 527 predicciones correctas, y un desempeño aceptable en las clases “Medio” (420 aciertos) y “Bajo” (408 aciertos).
- Sin embargo, también se presentan errores de clasificación cruzada, principalmente entre las clases “Bajo” y “Medio”, lo que indica que el modelo puede tener dificultades para distinguir entre concentraciones intermedias de CO.
- La diagonal principal (de arriba a la izquierda hacia abajo a la derecha) representa las predicciones correctas.
- Los valores fuera de la diagonal reflejan confusiones del modelo, especialmente entre categorías cercanas en valor.
- En general, el modelo presenta un rendimiento sólido, aunque con margen de mejora en la discriminación de niveles moderados de contaminación.

THANK YOU