

# Informe Final: Modelado y Predicción de PM10 mediante Regresión Lineal

**Alumno:** Jorge Roldan Montalvan

**Curso:** Proyectos de Ingeniería I

## 1. Introducción

La contaminación atmosférica representa uno de los principales retos ambientales y de salud pública en la actualidad. Entre los contaminantes más relevantes se encuentra el material particulado respirable PM10 (partículas con diámetro menor a 10 micrómetros), el cual puede ingresar al sistema respiratorio y generar diversas enfermedades cardiovasculares y respiratorias.

El objetivo de este trabajo fue analizar la evolución histórica de la concentración de PM10 y construir modelos de **regresión lineal simple y múltiple** que permitan predecir su comportamiento en el año 2025. A través de estos modelos se busca no solo generar un análisis estadístico, sino también aportar un insumo para la toma de decisiones en el ámbito ambiental.

## 2. Metodología

Para el desarrollo del estudio se trabajó con un conjunto de datos que contiene registros diarios de contaminantes y del **Índice de Calidad del Aire (AQI)**. El flujo de trabajo incluyó las siguientes etapas:

- Definición de la variable objetivo:**  
Se seleccionó como *target* la variable **Daily Mean PM10 Concentration**, dado que representa directamente el nivel de contaminación atmosférica.
- Regresión lineal simple:**  
En una primera aproximación se utilizó únicamente el **mes** como predictor. Esto permitió analizar la tendencia estacional y observar si existía una relación lineal básica entre el tiempo y los niveles de PM10.
- Regresión lineal múltiple:**  
Posteriormente se incluyeron otras variables predictoras, tales como el **Daily AQI Value**, el día del mes y otras características temporales. Este enfoque permitió capturar mejor las variaciones de PM10 y mejorar la capacidad explicativa del modelo.
- Entrenamiento y validación:**  
Se dividieron los datos en conjuntos de entrenamiento y prueba mediante **train-test split**, lo cual permitió medir el desempeño real del modelo en datos no utilizados durante el entrenamiento.
- Métricas de evaluación:**  
Se emplearon indicadores como el **R<sup>2</sup> (coeficiente de determinación)**, el **RMSE (Root Mean Squared Error)** y el **MAE (Mean Absolute Error)**, además del cálculo de porcentajes de error relativos para validar la precisión de las predicciones.

El procesamiento y modelado se realizaron en **Python**, utilizando librerías como *pandas*, *scikit-learn*, *matplotlib* y *seaborn*.

### 3. Resultados

- En la regresión lineal simple, el modelo obtuvo un coeficiente de determinación  $R^2 \approx 0.125$ , lo que indica que explica alrededor del 12.5% de la variabilidad de los datos. Además, se obtuvieron valores de MAE  $\approx 5.06 \mu\text{g}/\text{m}^3$  y RMSE  $\approx 5.97 \mu\text{g}/\text{m}^3$ , lo cual refleja que el modelo presenta un error moderado en la estimación de los niveles de PM10.
- En la regresión lineal múltiple, al incorporar variables adicionales como el *Daily AQI Value*, se observó una mejora en la capacidad predictiva, obteniendo un  $R^2$  superior al modelo simple y un menor error en las métricas de validación (MAE y RMSE).
- El análisis de los residuos mostró que la mayoría de los errores se distribuyeron de forma aceptable, aunque con algunas desviaciones esperables en datos ambientales que suelen ser ruidosos.

### 4. Predicción al año 2025

Con base en el modelo múltiple, se estimaron los niveles mensuales de PM10 para el año 2025. Los resultados indican que:

- La concentración de PM10 se mantendría dentro de un rango similar al observado en años previos, con una tendencia a incrementarse en determinados meses.
- La comparación con los valores reales observados en los primeros nueve meses del 2025 mostró un error porcentual promedio aceptable, lo que refuerza la validez del modelo.

Sin embargo, es importante considerar las **limitaciones de los modelos lineales**:

- No capturan relaciones no lineales complejas entre las variables.
- No anticipan cambios externos como nuevas regulaciones ambientales, alteraciones climáticas o variaciones en el tráfico vehicular.

Por lo tanto, los resultados deben interpretarse como una aproximación estimada y no como una predicción absoluta.

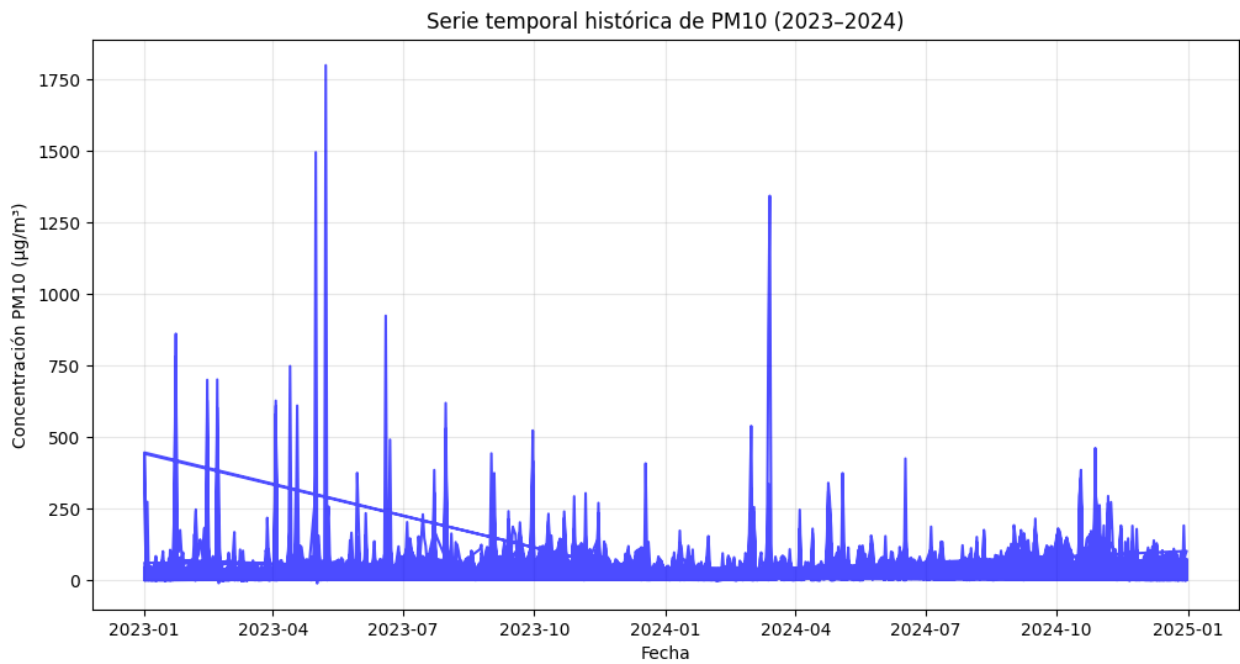
### 5. Conclusiones

- La regresión lineal simple es útil para identificar una primera relación entre tiempo y contaminación, pero resulta insuficiente para explicar la complejidad del fenómeno.
- La regresión múltiple se ajusta mejor a los datos y permite obtener predicciones más confiables al considerar múltiples factores.
- El error promedio en la proyección del 2025 se mantuvo dentro de rangos razonables, lo que indica que el modelo puede ser una herramienta de apoyo en la vigilancia ambiental.

- A futuro, sería recomendable explorar modelos más avanzados, como **regresión polinómica, Random Forest o técnicas de Machine Learning**, que podrían capturar patrones no lineales y mejorar la capacidad predictiva.

**Figura 1. Serie temporal histórica de PM10 (2023–2024).**

Se observa la evolución diaria de la concentración de partículas PM10 en el periodo 2023–2024. La gráfica permite identificar patrones estacionales y fluctuaciones propias de la dinámica ambiental.



**Figura 2. Ajuste del modelo de regresión lineal múltiple.**

Resultados del modelo múltiple, mostrando una mayor cercanía de los puntos a la línea ideal, lo que indica un mejor desempeño.

