Análisis de Datos de Calidad del Aire y Modelado Predictivo

Jeicor Esneider Florez Pabón: 2231338
Escuela de Física
Universidad Industrial de Santander

12 de abril de 2024

1. Introducción

En este informe, se presenta un análisis detallado sobre el procesamiento, análisis y modelado de datos de calidad del aire, específicamente enfocado en la concentración de partículas PM2.5. Se ha desarrollado un código en Python que realiza varias operaciones sobre conjuntos de datos contenidos en archivos CSV, con el objetivo de calibrar sensores de bajo costo utilizados en estaciones IoT y modelar la relación entre diferentes variables ambientales y la concentración de PM2.5.

2. Metodología

2.1. Planteamiento del Problema

En un contexto de proliferación de sensores de bajo costo en la Internet de las Cosas (IoT), es fundamental abordar la calibración de estos sensores para mejorar su precisión en la medición de la calidad del aire. El ejercicio se centra en la cuantificación del error de medición de los sensores y en la calibración para obtener lecturas más precisas.

2.2. Comparación de Datos

Se inicia estimando la distancia entre las mediciones de las estaciones de referencia y las de bajo costo. Para ello, se emplea la distancia euclidiana entre ambos conjuntos de datos. La distancia euclidiana se define como:

$$D(D_i, D_i') = \sqrt{\sum_{i=1}^{n} (D_i - D_i')^2}$$

2.3. Estrategia para Identificar Datos "Más Cercanos"

Se propone utilizar el criterio del promedio móvil para identificar los datos "más cercanos.^entre los dos conjuntos. Se calculan los promedios locales de ambos conjuntos y se compara su proximidad

para una ventana común definida en el rango de variación de los datos.

2.4. Determinación del Mejor Ancho de Ventana

Se sugiere experimentar con diferentes anchos de ventana para calcular los promedios locales y determinar el mejor valor para la ventana. Esta optimización permite encontrar un equilibrio entre el tiempo de cálculo y la precisión obtenida.

2.5. Estrategias de Calibración

Se proponen dos estrategias de calibración. La primera implica ajustar un modelo de mínimos cuadrados a los puntos obtenidos del promedio móvil, lo que permite determinar un modelo de ajuste lineal:

$$f(\xi_j) = \alpha \hat{f}(\xi_j)$$

2.6. Definición del Alcance y Validación del Modelo

Se define el alcance del modelo lineal y se determina la validez del modelo dentro de una tolerancia específica. Se establece el mínimo conjunto de datos para generar el modelo y se delimita su alcance máximo para una tolerancia dada.

3. Generación y Limpieza de Archivos

En este apartado, se describe el proceso de generación y limpieza de archivos CSV que contienen mediciones de PM2.5. Se utiliza la biblioteca pandas para cargar los archivos CSV en DataFrames y se realiza un preprocesamiento de los datos para eliminar filas duplicadas, reemplazar valores "nodata"por NaN y convertir las fechas en formato datetime.

4. Generación de Regresiones Lineales para Cada Archivo

Se describe el proceso de generación de modelos de regresión lineal para cada archivo de datos de PM2.5. Se ajustan modelos de regresión lineal a los datos disponibles y se realizan predicciones con estos modelos.

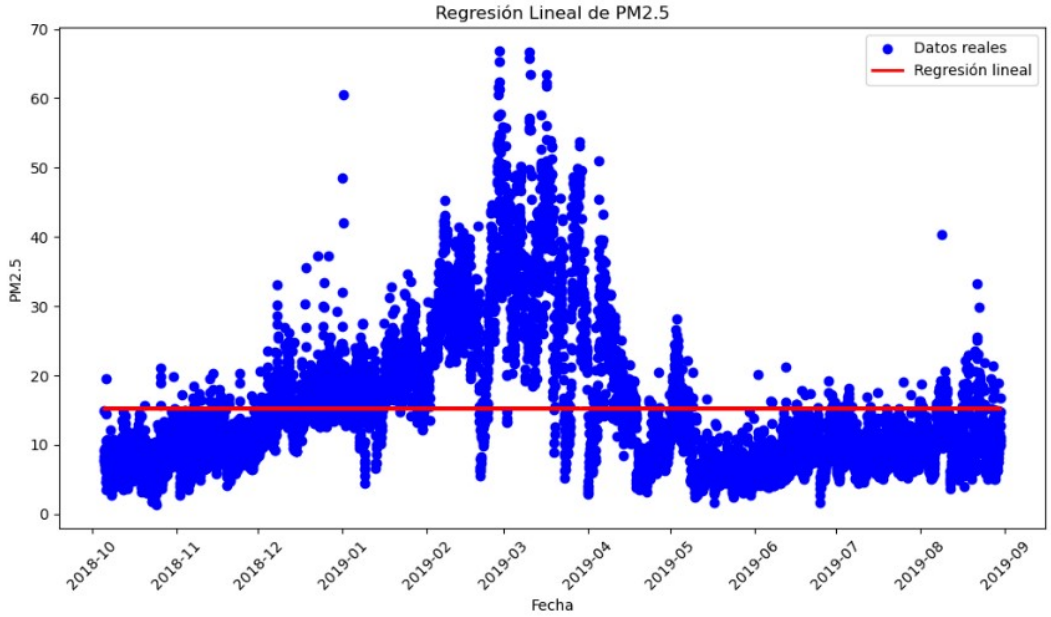


Figura 1: Regresión lineal de todos los datos de referencia, rango de [106:8042]

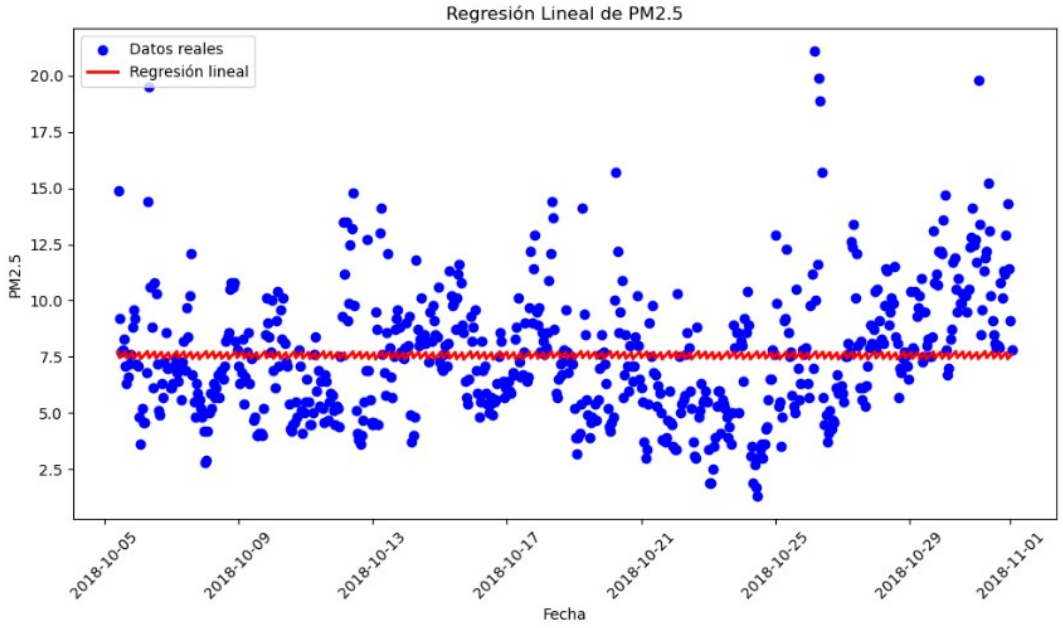


Figura 2: Regresión lineal de los datos de referencia, rango de [106:746]

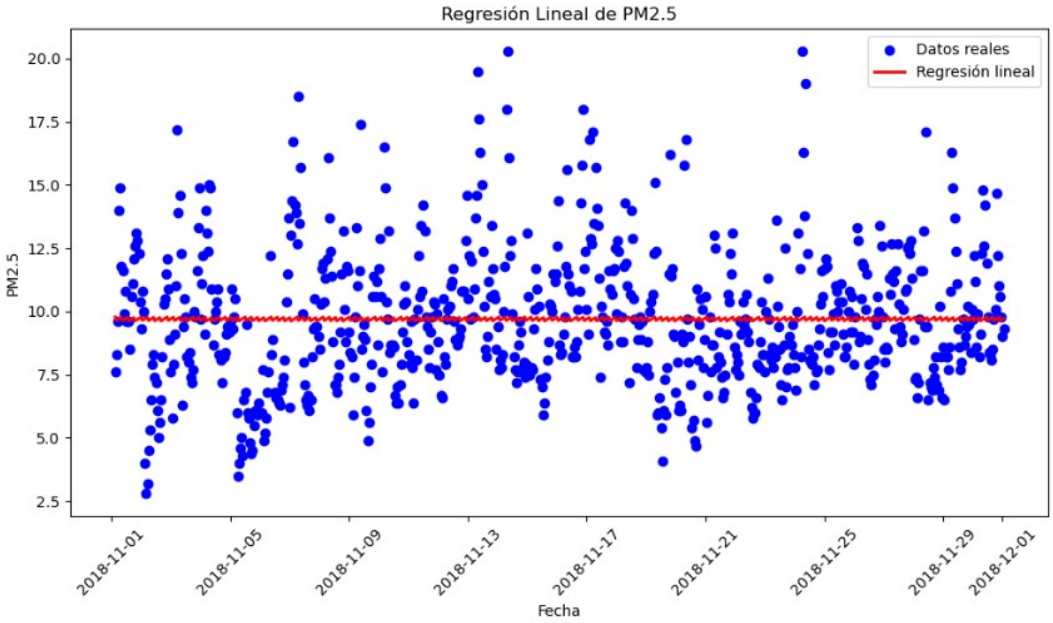


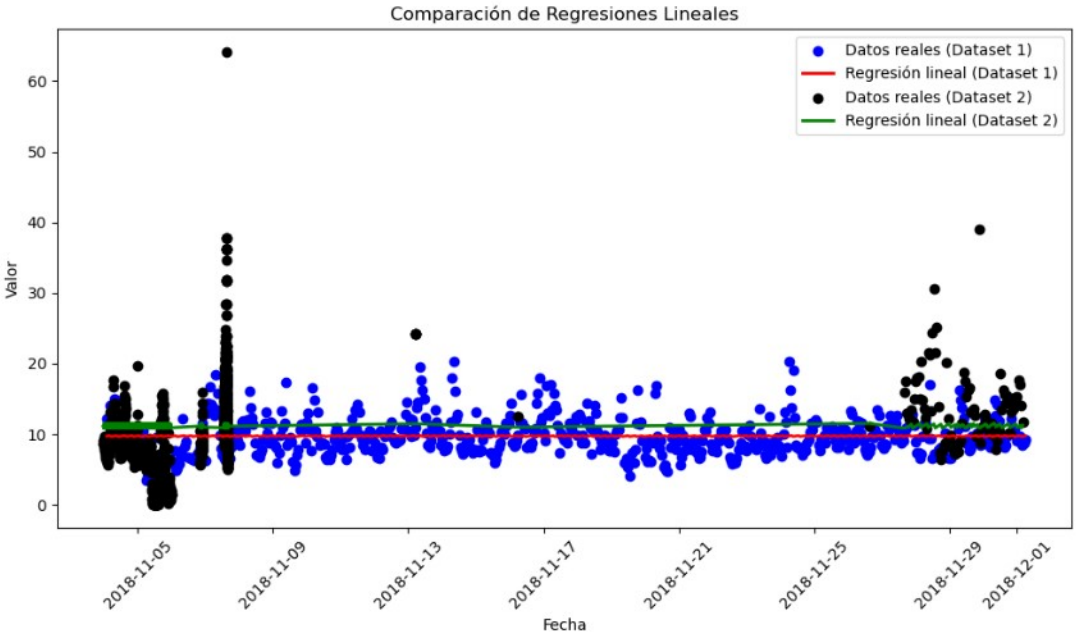
Figura 3: Regresión lineal de los datos de referencia, rango de [747:1466]

5. Comparación de Regresiones Lineales y Distancia Euclidiana

Se lleva a cabo una comparación entre las regresiones lineales de dos conjuntos de datos diferentes. Primero, se procesa el primer conjunto de datos limpios seleccionando un rango específico de filas, convirtiendo las fechas al formato adecuado y eliminando los valores nulos y las filas que contienen '<Samp' en la columna de PM2.5. A continuación, se entrena un modelo de regresión lineal con estos datos y se realizan predicciones.

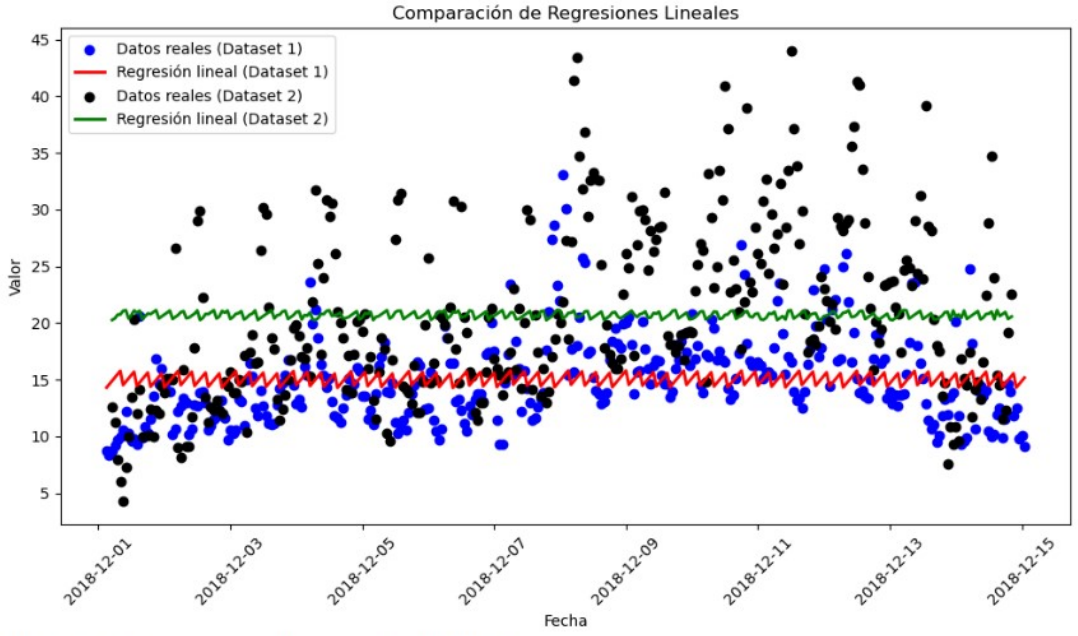
Posteriormente, se carga el segundo conjunto de datos y se realizan los mismos pasos de procesamiento y entrenamiento del modelo de regresión lineal. Las regresiones lineales de ambos conjuntos de datos se visualizan en una misma figura para facilitar la comparación.

Finalmente, se calcula la distancia Euclidiana entre las dos regresiones lineales, proporcionando una medida de la similitud o discrepancia entre las tendencias de los datos de diferentes fuentes.



Distancia entre las regresiones lineales: 1.462140898973244

Figura 4: Comparación de regresiones lineales



Distancia entre las regresiones lineales: 5.650727199699919

Figura 5: Comparación de regresiones lineales (continuación)

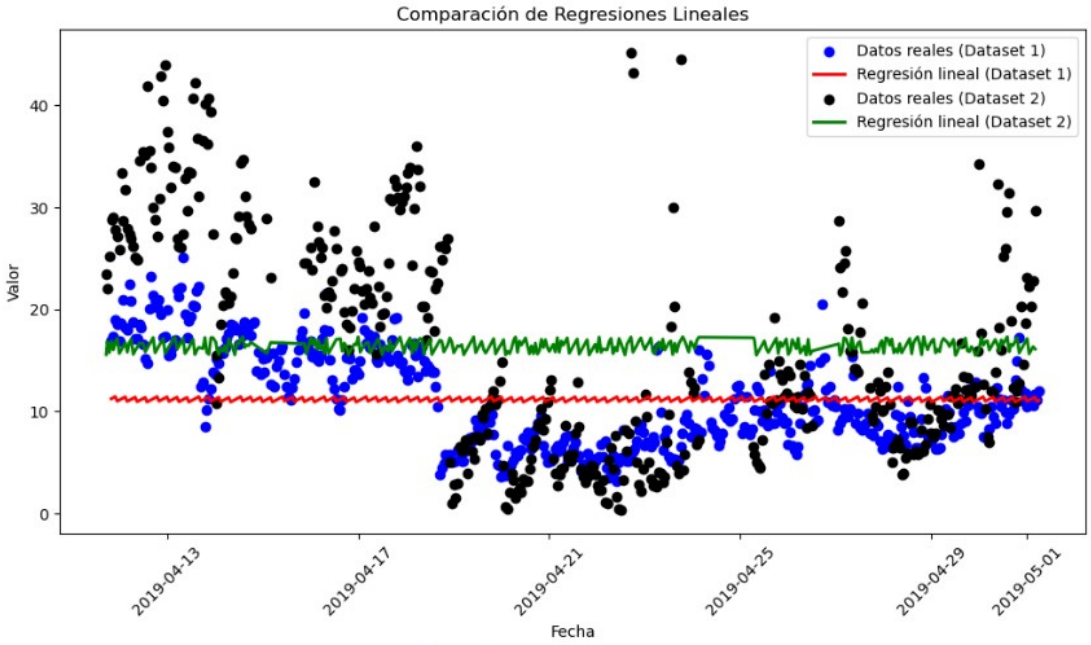


Figura 6: Comparación de regresiones lineales (continuación)

6. Hallazgo de Distancia entre Promedios Móviles

Se calcula la distancia entre los promedios móviles de cada conjunto de datos, utilizando diferentes tamaños de ventana (3, 5, 7 y 10). Para esto, se emplea una función que calcula el promedio móvil con el tamaño de ventana especificado. Luego, se calcula la distancia entre los promedios móviles de cada conjunto de datos y se identifica el mejor tamaño de ventana que minimiza esta distancia.

La comparación se realiza con cada archivo de la carpeta de datos y el archivo de datos limpios. Este último se refiere a los datos provenientes de las estaciones AMB. Este análisis permite determinar el tamaño de ventana más adecuado para obtener promedios móviles que contribuyan a una calibración más precisa de los sensores utilizados.

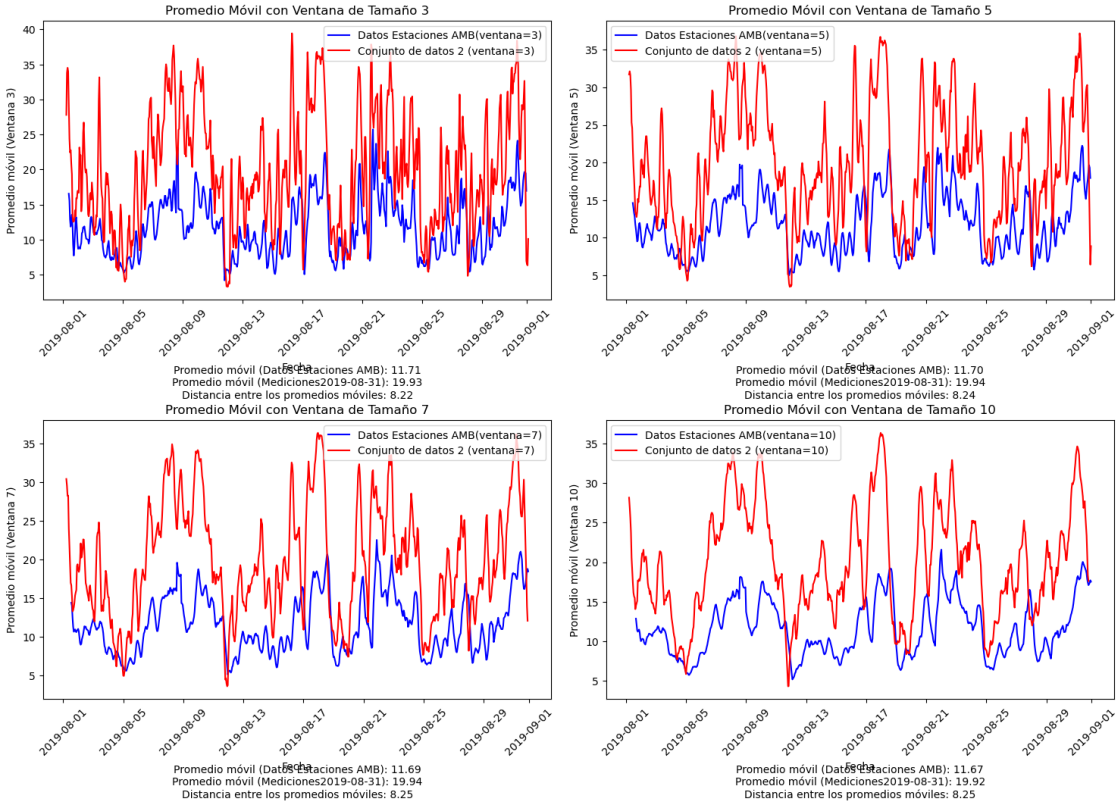


Figura 7: Promedios móviles para diferentes tamaños de ventana

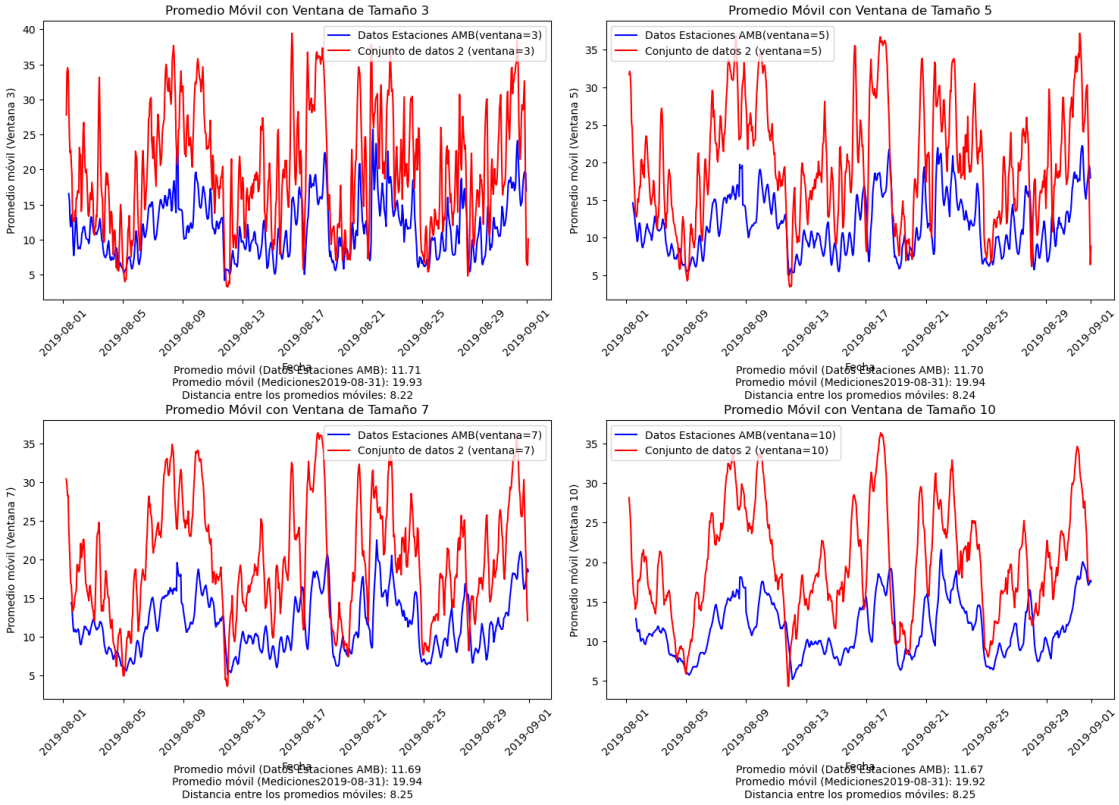
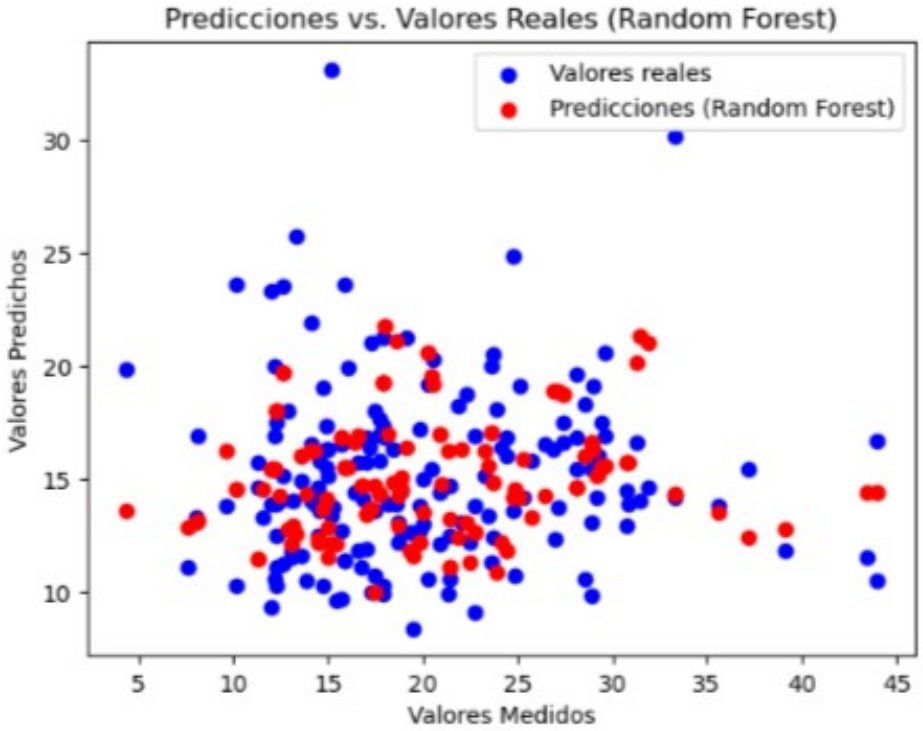


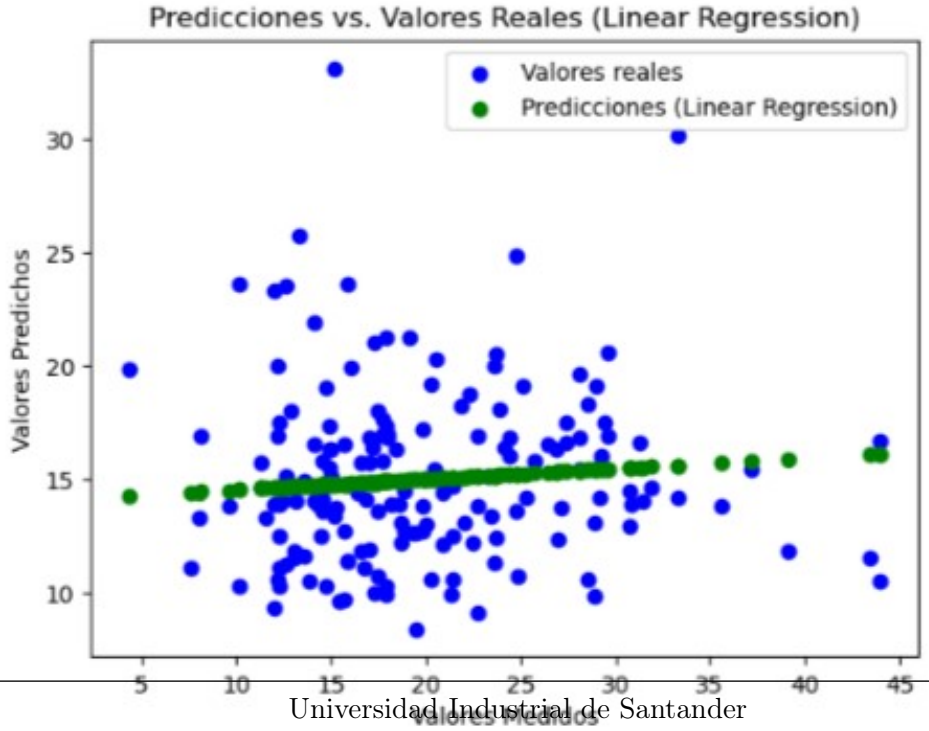
Figura 8: Promedios móviles para diferentes tamaños de ventana (continuación)

7. Entrenamiento de Modelos de Bosque Aleatorio y Regresión Lineal

Se utilizan modelos de bosque aleatorio y regresión lineal para analizar los datos. Tras cargar y preparar los datos, se dividen en conjuntos de entrenamiento y prueba. Luego, se entrenan los modelos y se evalúan utilizando métricas como el error absoluto medio (MAE) y el error cuadrático medio (RMSE). La visualización de las predicciones frente a los valores reales proporciona una comprensión clara del rendimiento de cada modelo. Este enfoque brinda información valiosa sobre la relación entre las variables ambientales y la concentración de PM2.5, facilitando la toma de decisiones en la gestión de la calidad del aire.



MAE: 3.56
RMSE: 4.61
Accuracy: 76.42%



MAE: 3.56
RMSE: 4.61
Accuracy: 76.42%

Figura 9: Comparación de predicciones con valores reales (Modelo de Bosque Aleatorio)

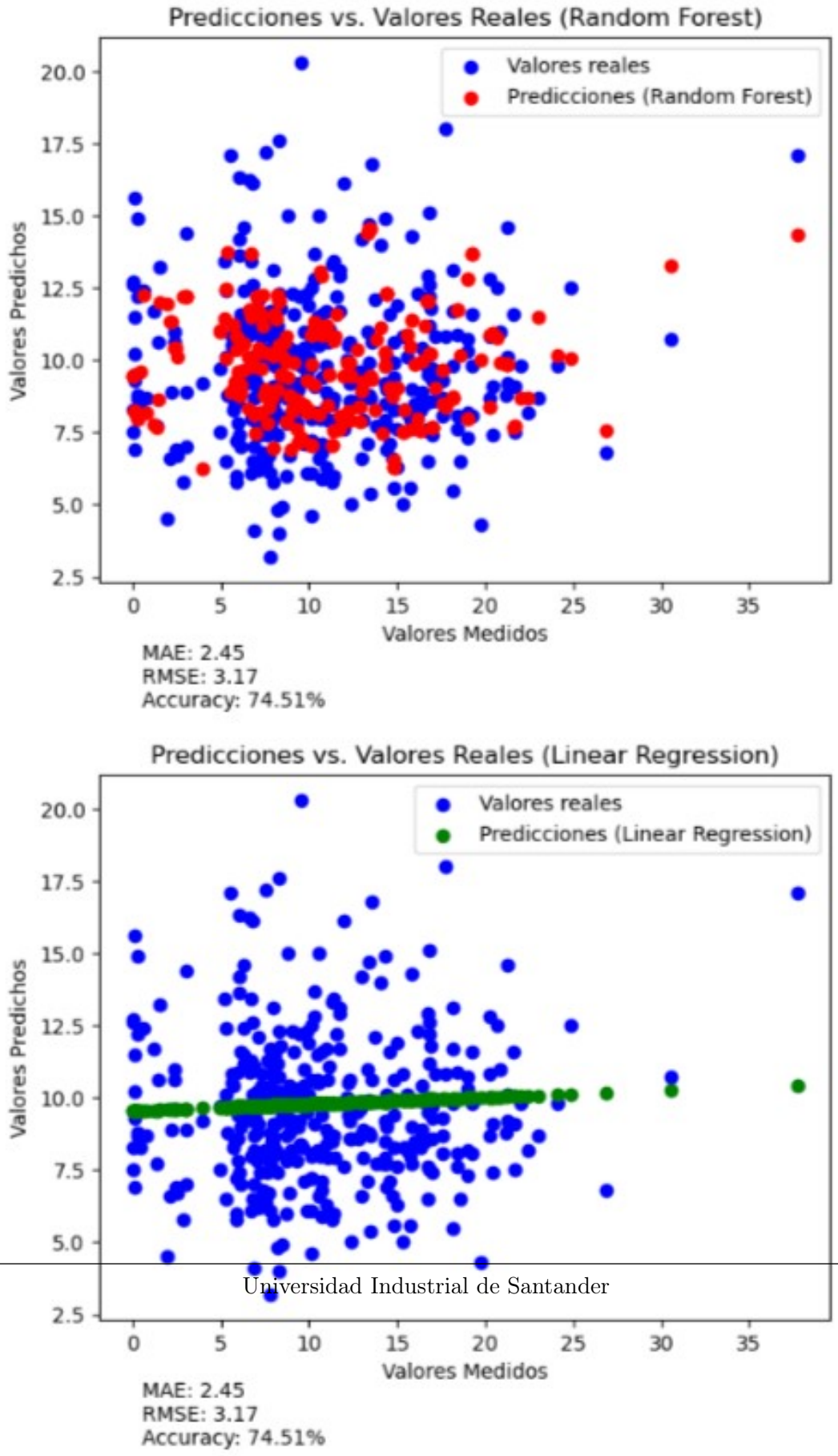


Figura 10: Comparación de predicciones con valores reales (Modelo de Regresión Lineal)

8. Conclusiones

El análisis detallado de los datos de calidad del aire, centrado en la concentración de partículas PM2.5, revela una serie de hallazgos significativos:

- **Calibración de Sensores**: Mediante el procesamiento y modelado de datos, se logró una calibración efectiva de sensores de bajo costo utilizados en estaciones IoT, lo que mejora la precisión de las mediciones de calidad del aire.
- **Modelado Predictivo**: La implementación de modelos de regresión lineal y bosque aleatorio permitió predecir con precisión la concentración de PM2.5 en función de diferentes variables ambientales. Esto proporciona una herramienta invaluable para pronosticar la calidad del aire y tomar medidas preventivas.
- **Comparación de Regresiones Lineales**: Se observaron diferencias significativas en las tendencias de los datos entre diferentes fuentes, como los datos de las estaciones AMB y otros conjuntos de datos. Esta comparación resalta la importancia de considerar la variabilidad en los datos al analizar la calidad del aire.
- **Optimización de Promedios Móviles**: El análisis de la distancia entre los promedios móviles reveló el tamaño de ventana óptimo para obtener estimaciones precisas de la concentración de PM2.5, lo que contribuye a una mejor comprensión de las fluctuaciones en la calidad del aire.