Tecnológico de Monterrey

Análisis de Regresión Lineal Aplicado a un Dataset de Calidad del Aire

Adrián Matute Beltrán A01703889

24 de agosto del 2024

El presente reporte tiene como objetivo analizar las relaciones entre diversas variables medioambientales registradas en un área urbana utilizando un modelo de regresión lineal. El dataset empleado en este análisis, obtenido del UCI Machine Learning Repository, contiene datos recopilados durante un año en una zona urbana con alta contaminación atmosférica. La variable objetivo de este estudio es la concentración de Benceno (C6H6(GT)), y se busca predecir su comportamiento en función de otras variables atmosféricas y respuestas de sensores químicos.

El propósito de este análisis es determinar la precisión de un modelo de regresión lineal al predecir la concentración de Benceno, identificar las variables más influyentes en su predicción y evaluar la efectividad del modelo.

Descripción del Dataset

El dataset utilizado para el análisis es “Air Quality Data ”, contiene 9358 registros utilizables de respuestas promediadas por hora de varios sensores químicos de óxido de metal, junto con datos de un analizador de referencia certificado. Las variables incluidas en el análisis son:

- CO(GT): Concentración horaria promedio de CO en mg/m³ (medida de referencia).
- PT08.S1(CO): Respuesta del sensor dirigido nominalmente a CO.
- NMHC(GT): Concentración horaria promedio de hidrocarburos no metánicos en microg/m³.
- C6H6(GT): Concentración horaria promedio de Benceno en microg/m³.
- PT08.S2(NMHC): Respuesta del sensor dirigido nominalmente a NMHC.
- NOx(GT): Concentración horaria promedio de NOx en ppb.
- PT08.S3(NOx): Respuesta del sensor dirigido nominalmente a NOx.
- NO2(GT): Concentración horaria promedio de NO2 en microg/m³.
- PT08.S4(NO2): Respuesta del sensor dirigido nominalmente a NO2.
- PT08.S5(O3): Respuesta del sensor dirigido nominalmente a O3.
- T: Temperatura en °C.
- RH: Humedad relativa en %.
- AH: Humedad absoluta.

Se realizó un proceso de preprocesamiento para manejar los valores faltantes y preparar los datos para el análisis.

Preprocesamiento de Datos

El preprocesamiento de los datos fue crucial para asegurar la calidad del análisis. A continuación se describen las etapas de este proceso:

1. Conversión de Tipos de Datos: Algunas columnas contenían datos numéricos representados como cadenas de texto. Se realizó la conversión de éstos valores reemplazando comas por puntos decimales y transformando los datos a tipo 'float64'.

<bound method NDFrame.describe of Date Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC) ... PT08.S3(NOx) NO2(GT) PT08.S4(NO2) PT08.S5(O3) T RH AH

0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0	...	1056.0	113.0	1692.0	1268.0	13.6	48.9	0.7578
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0	...	1174.0	92.0	1559.0	972.0	13.3	47.7	0.7255
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0	...	1140.0	114.0	1555.0	1074.0	11.9	54.0	0.7502
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0	...	1092.0	122.0	1584.0	1203.0	11.0	60.0	0.7867
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0	...	1205.0	116.0	1490.0	1110.0	11.2	59.6	0.7888
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1226	30/04/2004	20.00.00	4.4	1449.0	501.0	19.5	1282.0	...	625.0	133.0	2100.0	1569.0	19.1	61.1	1.3345
1227	30/04/2004	21.00.00	3.1	1363.0	234.0	15.1	1152.0	...	684.0	110.0	1951.0	1495.0	18.2	65.4	1.3529
1228	30/04/2004	22.00.00	3.0	1371.0	212.0	14.6	1136.0	...	689.0	102.0	1927.0	1471.0	18.1	66.1	1.3579
1229	30/04/2004	23.00.00	3.1	1406.0	275.0	13.7	1107.0	...	718.0	108.0	1872.0	1384.0	17.7	66.9	1.3422
1230	01/05/2004	00.00.00	3.5	1425.0	275.0	15.2	1155.0	...	709.0	110.0	1936.0	1789.0	17.8	66.8	1.3460

2. Manejo de Valores Faltantes: Los valores faltantes en las columnas de interés, representados por el valor -200, fueron reemplazados por NaN. Posteriormente, se eliminaron todas las filas que contenían estos valores faltantes para asegurar un dataset limpio y completo para el análisis.

3. Selección de Variables: Se seleccionaron las siguientes variables como predictoras: T, RH, AH, PT08.S1(CO), PT08.S2(NMHC), y PT08.S3(NOx). La variable objetivo fue la concentración de Benceno (C6H6(GT)).

Metodología

Para modelar la relación entre las variables predictoras y la concentración de Benceno, se implementó un modelo de regresión lineal desde cero, sin el uso de bibliotecas de machine learning preconfiguradas.

Implementación del Modelo

El modelo de regresión lineal fue implementado utilizando la ecuación normal:

$$\theta = (X^T X)^{-1} X^T y$$

Donde X es la matriz de variables predictoras y y es el vector de la variable objetivo. El vector de coeficientes β fue calculado a partir de los datos y luego utilizado para realizar predicciones sobre el conjunto de datos.

Evaluación del Modelo

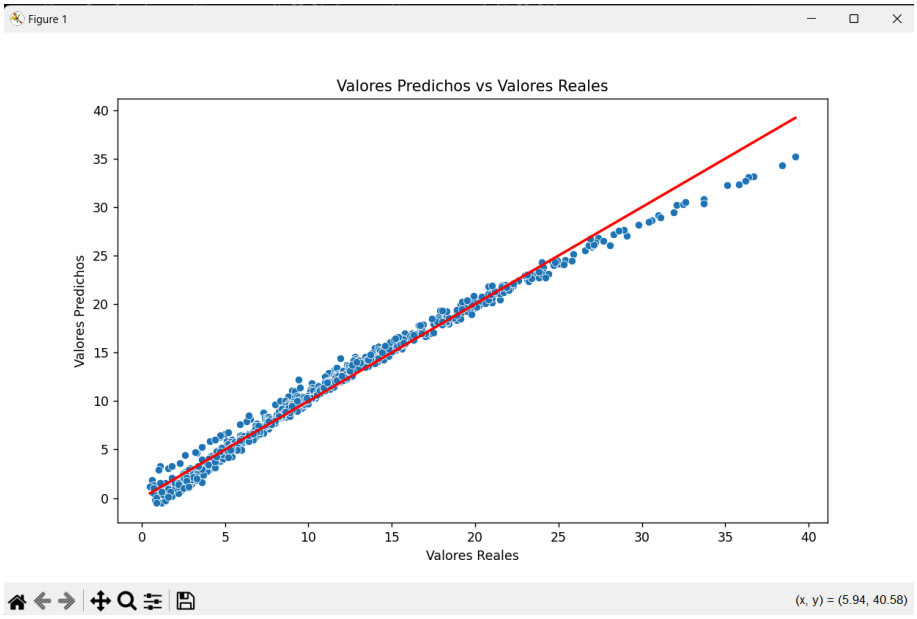
El modelo fue evaluado mediante dos métricas principales:

- Error Cuadrático Medio (MSE): Mide el promedio de los errores al cuadrado entre los valores reales y los predichos.

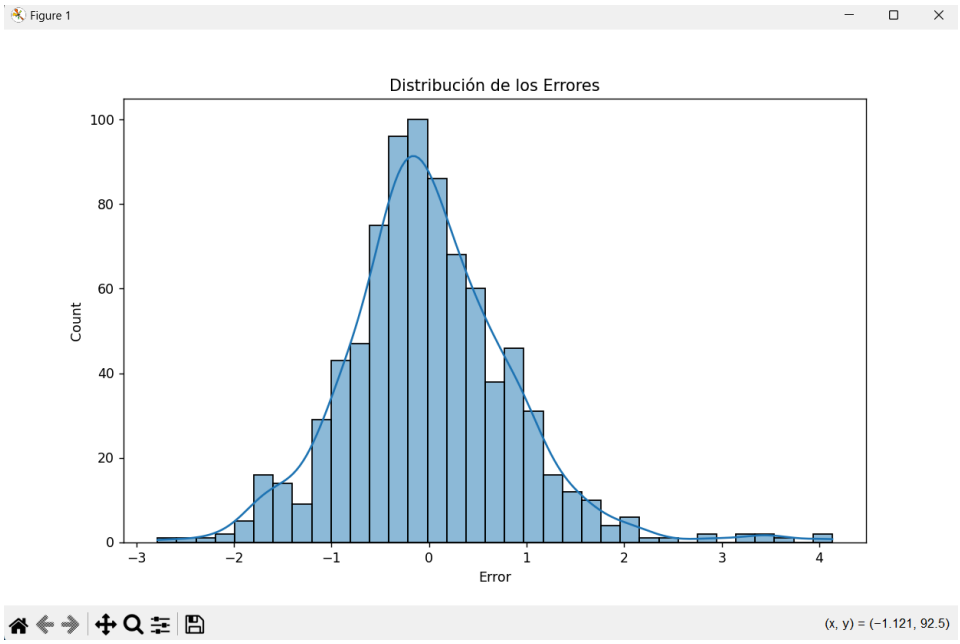
- Coeficiente de Determinación (R^2): Indica la proporción de la varianza en y explicada por X, proporcionando una medida de la calidad del ajuste del modelo

```
Coeficientes: [-3.21419675e+01 -1.89105704e-01 -5.40922924e-02  5.66491118e+00
 -2.15463284e-03  3.80477673e-02  1.00158398e-02]
MSE: 0.7304816319701989
R²: 0.9867093756251311
```

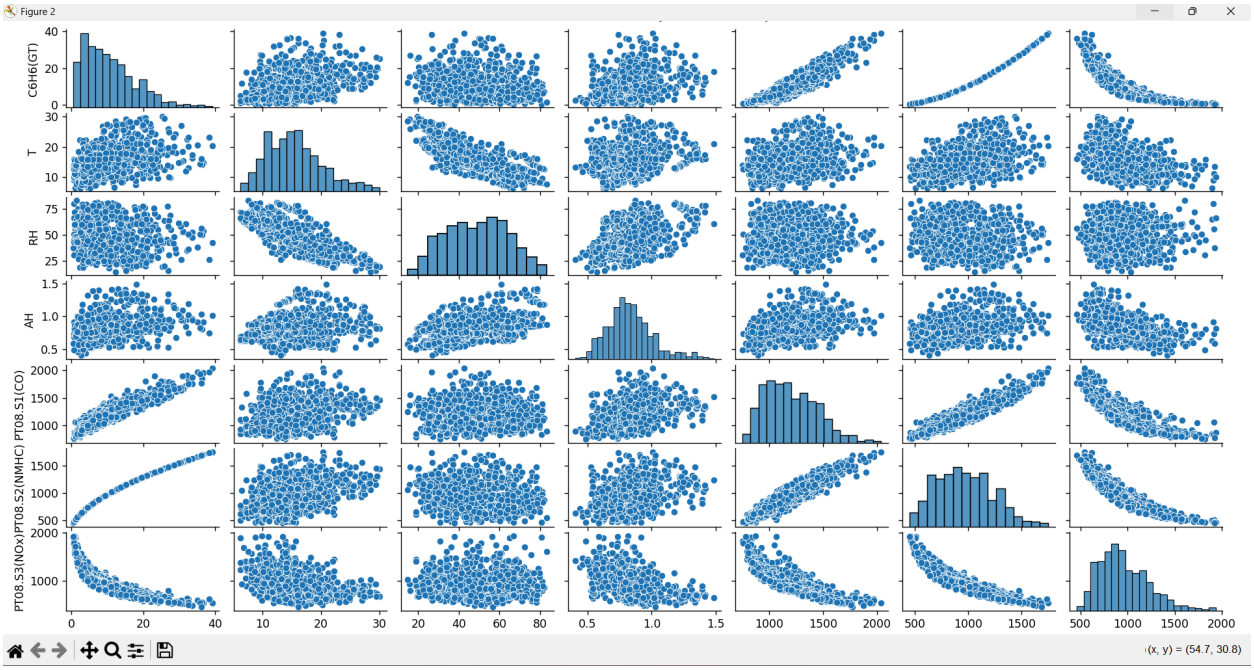
Visualización de los Resultados



El gráfico de dispersión muestra una comparación entre los valores reales de C6H6(GT) y las predicciones realizadas por el modelo. La mayoría de los puntos se alinean cerca de la línea de identidad (roja), lo que indica que las predicciones del modelo son bastante precisas.



Esta gráfica de distribución de errores del modelo muestra los mismos errores distribuidos de manera normal, lo cual indica una buena calidad del modelo.



Este pairplot permite visualizar las relaciones entre todas las variables predictoras y la variable objetivo (Benzeno). Ayuda a identificar posibles correlaciones entre las variables y patrones que podrían afectar el modelo.

Discusión

El análisis ha demostrado que la regresión lineal es un enfoque eficaz para modelar la concentración de Benceno en función de otras variables atmosféricas y de sensores. Sin embargo, se observaron algunos puntos que no se alinean perfectamente con la línea de identidad en el gráfico de dispersión, lo que sugiere la existencia de outliers o la posibilidad de mejorar el modelo con enfoques adicionales, como la inclusión de términos polinómicos o transformaciones de variables.

Referencias

- UCI Machine Learning Repository. Air Quality Data Set. Disponible en: https://archive.ics.uci.edu/ml/datasets/air+quality