Regresión Lineal con Python

March 24, 2025

1 Introducción

La regresión lineal simple es un modelo estadístico utilizado para explicar la relación entre una variable independiente (predictora) y una variable dependiente (respuesta), mediante una línea recta. En este caso, se utilizó para predecir cuántas veces será compartido un artículo en línea (`# Shares`) en función de su cantidad de palabras (`Word count`).

2 Metodología

Se trabajó con un conjunto de datos reales de artículos. El archivo `articulos_ml.csv` contiene 161 registros con 8 columnas. Primero se filtraron los datos eliminando valores extremos para mejorar la visualización y la estabilidad del modelo.

El modelo fue entrenado usando la biblioteca `scikit-learn`. A continuación, se muestra un fragmento del código:

```
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score

X_train = np.array(filtered_data[["Word count"]])
y_train = filtered_data['# Shares'].values

regr = linear_model.LinearRegression()
regr.fit(X_train, y_train)
y_pred = regr.predict(X_train)

print("Coeficiente:", regr.coef_)
print("Intercepto:", regr.intercept_)
print("MSE:", mean_squared_error(y_train, y_pred))
print("R2:", r2_score(y_train, y_pred))
```

3 Resultados

Los resultados del modelo fueron los siguientes:

- **Coeficiente (pendiente):** 5.70
- **Intercepto:** 11200.30
- **Error Cuadrático Medio (MSE):** 372,888,728.34
- **Puntaje de varianza (R^2):** 0.06
- **Predicción para un artículo de 2000 palabras:** 22,595 # Shares

4 Conclusión

El modelo de regresión lineal simple aplicado a este conjunto de datos permitió observar una relación positiva entre la cantidad de palabras en un artículo y la cantidad de veces que se comparte. Sin embargo, el bajo valor de R^2 (0.06) indica que el modelo no explica bien la variabilidad de los datos, lo cual es común cuando solo se utiliza una variable predictora. Se recomienda considerar múltiples variables para mejorar la precisión del modelo en futuras implementaciones.