Regresión Lineal Múltiple

March 24, 2025

1 Introducción

La regresión lineal múltiple es una técnica de aprendizaje supervisado que busca modelar la relación entre una variable dependiente (respuesta) y dos o más variables independientes (predictoras). A diferencia de la regresión lineal simple, esta permite construir modelos más robustos al considerar múltiples factores que influyen en el resultado. En esta actividad, se utilizó un conjunto de datos de artículos web para predecir el número de veces que un artículo es compartido, considerando la cantidad de palabras y la suma de enlaces, comentarios e imágenes.

2 Metodología

Se empleó Python con la biblioteca `scikit-learn` para entrenar un modelo de regresión lineal múltiple. Primero, se filtraron los datos para trabajar con registros que tuvieran menos de 3500 palabras y menos de 80000 compartidos. Luego se creó una nueva variable que suma los enlaces, comentarios e imágenes, combinándola con el conteo de palabras como variables predictoras. El modelo se entrenó con estas dos variables y se evaluó utilizando métricas estándar.

Fragmento de Código

```
suma = (filtered_data["# of Links"] +
        filtered_data["# of comments"].fillna(0) +
        filtered_data["# Images video"])

dataX2["Word count"] = filtered_data["Word count"]
dataX2["suma"] = suma

regr2 = linear_model.LinearRegression()
regr2.fit(XY_train, z_train)
```

3 Resultados

El modelo generó los siguientes coeficientes para las variables:

- Coeficientes: [6.632, -483.407]
- Intercepto: 16921.89
- Error Cuadrático Medio (MSE): 352,122,816.48
- Puntaje R^2 (Varianza explicada): 0.11

Al probar el modelo con un artículo ficticio de 2000 palabras, 10 enlaces, 4 comentarios y 6 imágenes, se obtuvo una predicción de:

20518 compartidos

4 Conclusión

El modelo logró establecer una relación entre las variables predictoras y el número de compartidos, aunque el bajo valor de R^2 sugiere que no captura completamente la varianza de los datos. A pesar de ello, fue útil para aplicar la técnica de regresión múltiple y visualizar los resultados en un espacio tridimensional. Para mejorar el rendimiento, se recomienda agregar más variables relevantes o utilizar técnicas de regularización.