Regresión Lineal Múltiple IA ARLG

Abraham López

Marzo 2025

1 Introducción

1.1 ¿Qué es la regresión lineal múltiple?

La regresión múltiple es una técnica estadística utilizada para analizar la relación entre una variable dependiente y dos o más variables independientes. Esta metodología permite entender cómo estas variables explicativas influyen de manera conjunta en el resultado, lo cual es fundamental para el análisis de datos en situaciones complejas. En esencia, proporciona un modelo matemático que permite predecir el comportamiento de una variable dependiente a partir de las variables independientes. La ecuación general de la regresión lineal múltiple es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Donde:

- Y: Variable dependiente.
- β_0: Término independiente.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes de las variables independientes.
- $X_1, X_2, \ldots, X_n$: Variables independientes (predictoras).
- ϵ: Término de error (residual).

1.2 Diferencia entre regresión múltiple y regresión lineal

La regresión lineal es un caso particular de la regresión múltiple, en el cual se analiza la relación entre una variable dependiente y una sola variable independiente. En la regresión lineal, se busca identificar si existe una relación lineal entre estas dos variables, mientras que en este tipo de regresión se consideran múltiples variables independientes para determinar su influencia conjunta sobre la variable dependiente. Esta diferencia hace que la regresión múltiple sea más adecuada cuando se requiere evaluar el impacto de varios factores simultáneamente, mientras que la regresión lineal se limita a un análisis más simple entre dos variables.

2 Metodología

En este análisis, nos enfocamos en predecir el número de veces que un artículo es compartido en redes sociales (# Shares), utilizando como variables predictoras el número de palabras (Word count) y una combinación de otras características, como la cantidad de enlaces, comentarios e imágenes/videos.

Primero, se carga el archivo CSV en un DataFrame. Luego, se filtran los datos para eliminar valores atípicos, conservando solo las filas donde ‘Word count‘ es menor o igual a 35,000 y ‘ Shares‘ es menor o igual a 80,000.

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score

data = pd.read_csv("./articulos_ml.csv")
data.shape
filtered_data = data[(data['Word count'] <= 35000) & (data['# Shares'] <= 80000)]
```

Se crea una nueva variable llamada ‘suma‘, que es la suma de las columnas ‘# of Links‘, ‘# of comments‘ (rellenando los valores faltantes con 0) y ‘ Images video‘. Luego, se crea un nuevo DataFrame ‘dataX2‘ que contiene dos variables predictoras: ‘Word count‘ y ‘suma‘. Este DataFrame se convierte en un array de NumPy, que contiene las variables predictoras. Finalmente, se extrae la variable dependiente (‘# Shares‘) y se convierte en un array de NumPy .

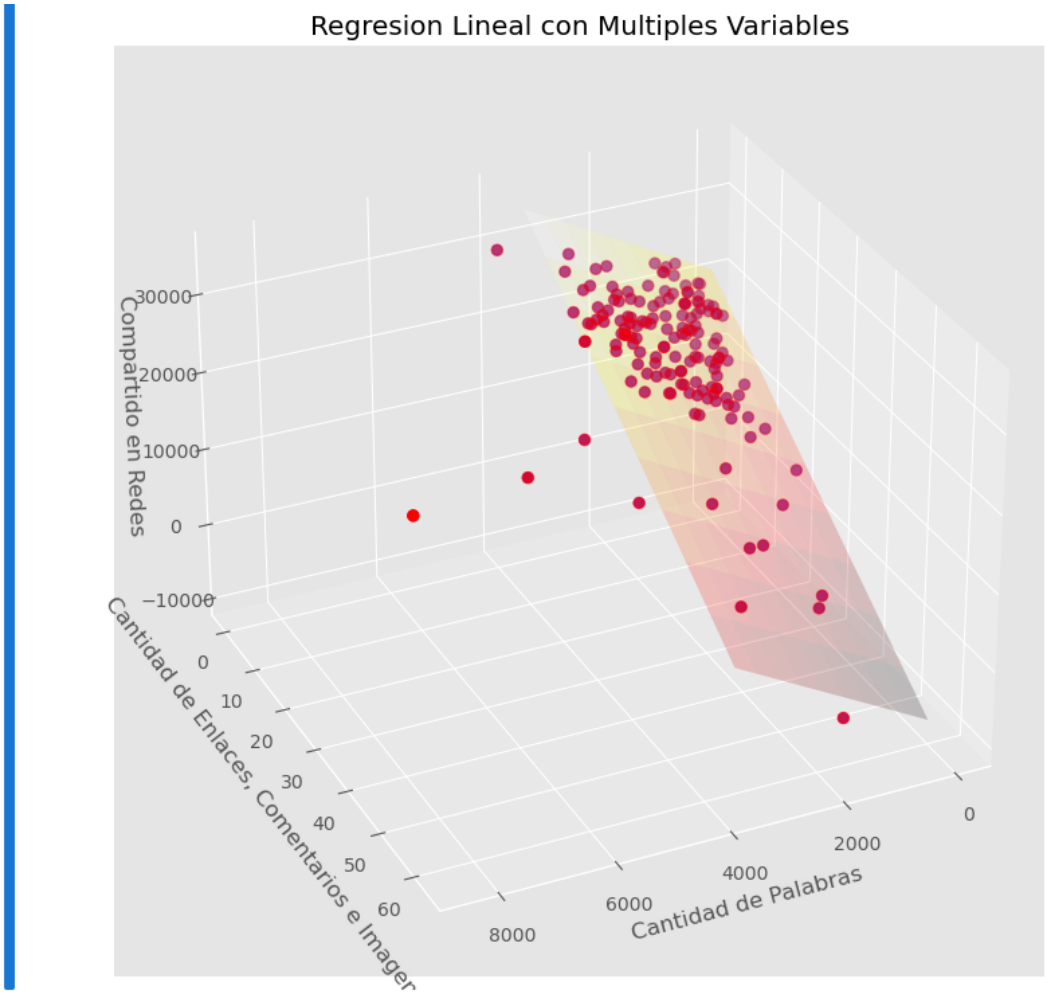
[3]:
```
suma = (filtered_data['# of Links'] + filtered_data['# of comments'].fillna(0) + filtered_data['# Images video'])
dataX2 = pd.DataFrame()
dataX2["Word count"] = filtered_data["Word count"]
dataX2["suma"] = suma
XY_train = np.array(dataX2)
z_train = filtered_data['# Shares'].values
regr2 = linear_model.LinearRegression()
regr2.fit(XY_train, z_train)
z_pred = regr2.predict(XY_train)
print('Coefficients: \n', regr2.coef_)
print("Mean squared error: %.2f" % mean_squared_error(z_train, z_pred))
print('Variance score:%2f' % r2_score(z_train, z_pred))
```

```
Coefficients:
 [   3.78192735 -508.3979127 ]
Mean squared error: 358158876.48
Variance score:0.075180
```

Se crea un modelo de regresión lineal múltiple y se entrena utilizando las variables predictoras y la variable dependiente. Una vez entrenado, se realizan predicciones utilizando el modelo.

Se crea un gráfico 3D para visualizar la relación entre las variables predictoras y la variable dependiente. En este gráfico, se grafican los puntos de datos reales y las predicciones.

[13]:
```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
xx, yy = np.meshgrid(np.linspace(0, 3500, num=10), np.linspace(0, 60, num=10))
nuevoX = (regr2.coef_[0] * xx)
nuevoY = (regr2.coef_[1] * yy)
z = (nuevoX + nuevoY + regr2.intercept_)
ax.plot_surface(xx, yy, z, alpha=0.2, cmap='hot')
ax.scatter(XY_train[:, 0], XY_train[:, 1] , z_pred, c='blue', s=30)
ax.scatter(XY_train[:, 0], XY_train[:, 1] , z_pred, c='red', s=40)
ax.view_init(elev=30., azim=65)
ax.set_xlabel('Cantidad de Palabras')
ax.set_ylabel('Cantidad de Enlaces, Comentarios e Imagenes')
ax.set_zlabel('Compartido en Redes')
ax.set_title('Regresion Lineal con Multiples Variables')
```



Finalmente, se realiza una predicción para un artículo con 2,000 palabras y una suma de 10 enlaces, 4 comentarios y 6 imágenes/videos.

```
[17]: z_Dosmil = regr2.predict([[2000, 10+4+6]])
      print(int(z_Dosmil[0]))

      18859
```

3 Resultados

Tras realizar el análisis de regresión lineal múltiple, se obtuvieron los siguientes resultados:

- **Coeficientes**. Los coeficientes del modelo son: [3.7819, -508.3979] El primer coeficiente (3.7819) corresponde a la variable 'Word count'. Esto indica que, por cada palabra adicional en un artículo, se espera que el número de shares aumente en aproximadamente **3.78 shares**, manteniendo constantes las demás variables. El segundo coeficiente (−508.3979) corresponde a la variable 'suma' (suma de enlaces, comentarios e imágenes/videos). Este valor negativo sugiere que, por cada unidad adicional en la suma de estas características, el número de shares disminuye en aproximadamente **508.40 shares**, manteniendo constantes las demás variables. Este resultado es contraintuitivo y podría indicar que la relación no es lineal o que hay otros factores no considerados en el modelo.
- **Mean Squared Error**. El error cuadrático medio (MSE) del modelo es: 358,158,876.48

Este valor representa el promedio de los errores al cuadrado entre los valores reales y los predichos. Un MSE tan alto indica que el modelo tiene un **error significativo** en sus predicciones, lo que sugiere que no se ajusta bien a los datos.

- **Variance Score**.El coeficiente de determinación (R^2) del modelo es: 0.07518 Este valor indica que solo el **7.52%** de la variabilidad en el número de shares es explicada por las variables predictoras ('Word count' y 'suma'). Esto sugiere que el modelo actual no es adecuado para predecir el número de shares de manera precisa, ya que la mayor parte de la variabilidad en los datos no es capturada por las variables incluidas.

4 Conclusión

El análisis de regresión lineal múltiple que realizamos nos permitió investigar la relación entre el número de shares y dos variables predictoras: el conteo de palabras y una combinación de otras características, como el número de enlaces, comentarios e imágenes/videos. Los resultados que obtuvimos muestran que el modelo no es el más adecuado para predecir con precisión el número de shares, lo que se evidencia en el bajo coeficiente de determinación ($R^2 = 0.07518$), que solo explica el 7.52% de la variabilidad en los datos. Además, el alto error cuadrático medio (MSE = 358,158,876.48) confirma que el modelo tiene un margen de error considerable en sus predicciones.

Aunque el número de palabras tuvo un efecto positivo en el número de shares (con un coeficiente de 3.7819), la suma de enlaces, comentarios e imágenes/videos mostró un impacto negativo (con un coeficiente de -508.3979), lo cual es un poco sorprendente y sugiere que la relación entre estas variables y el número de shares no es lineal, o que hay otros factores que no hemos considerado en el modelo.

En resumen, el modelo actual no logra captar la complejidad de los datos. Este análisis es solo un primer paso para entender qué factores influyen en la popularidad de los artículos, pero definitivamente necesitamos enfoques más sólidos para obtener resultados más precisos y útiles.