Regresion Lineal

Adrián Esteban Morales Rodriguez

April 1, 2025

1 Introducción

La regresión lineal simple es una regresión lineal con una variable independiente, también llamada variable explicativa, y una variable dependiente, también llamada variable de respuesta. En la regresión lineal simple, la variable dependiente es continua.

La regresión lineal simple ayuda a hacer predicciones y a comprender las relaciones entre una variable independiente y una variable dependiente. Por ejemplo, podrías querer saber cómo afecta la altura de un árbol (variable independiente) al número de hojas que tiene (variable dependiente). Recopilando datos y ajustando un sencillo modelo de regresión lineal, podrías predecir el número de hojas en función de la altura del árbol. Esta es la parte de "hacer predicciones". Pero este enfoque también revela cuánto cambia, por término medio, el número de hojas a medida que el árbol crece en altura, que es como también se utiliza la regresión lineal simple para comprender las relaciones.

2 Metodología

Para la realización de esta actividad tome los pasos a seguir por el libro de Aprende Maching Learning. Este ejercicio consta de predecir, a partir de los datos sobre articulos, cuantas veces será compartido el articulo en redes sociales. Utilice lo requerido para la práctica como Python, librerias como Scikit-Learn, compilado con Anaconda.

- Paso 1

  Importar las librerias:

```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
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
```

- Paso 2

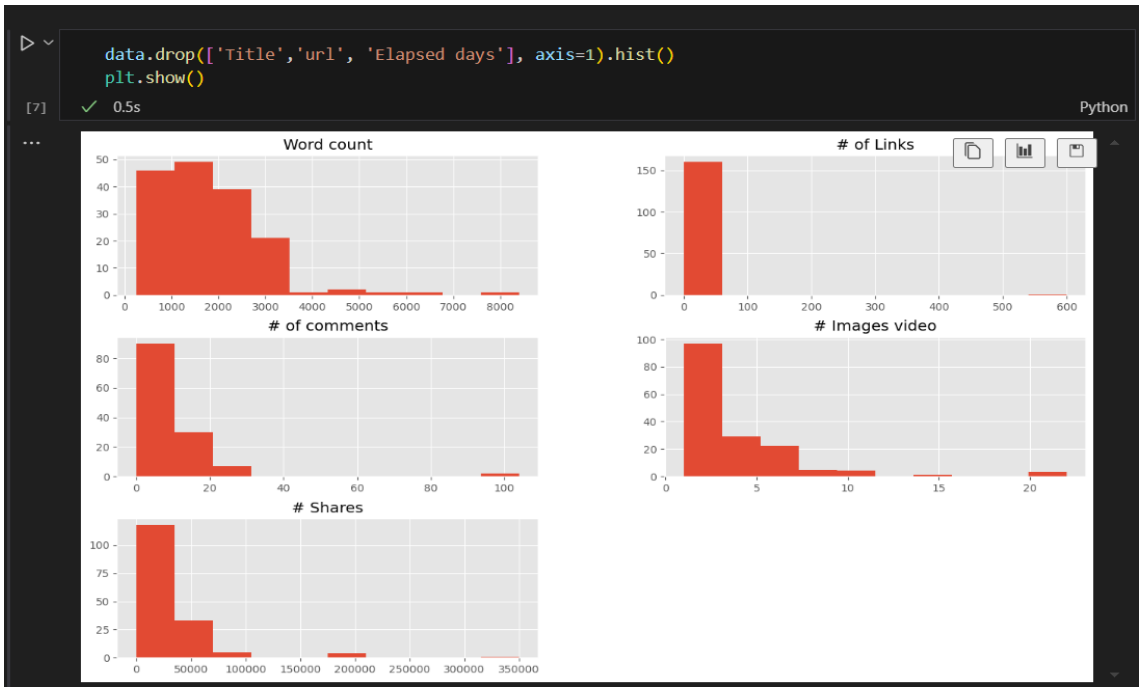
  Leer el archivo .csv y ver el tamaño

```
data= pd.read_csv("articulos_ml.csv")
data.shape
```
✓ 0.4s

```
(161, 8)
```

- Paso 3

  Visualizar las características de entrada



- Paso 4

  Filtrar los datos de cantidad de palabras para quedarnos con los registros con menos de 3,500 palabras y también con los que tengan Cantidad de compartidos menos a 80,000. Pintando en azul los puntos con menos de 1808 palabras (la media) y en naranja los que tengan más.

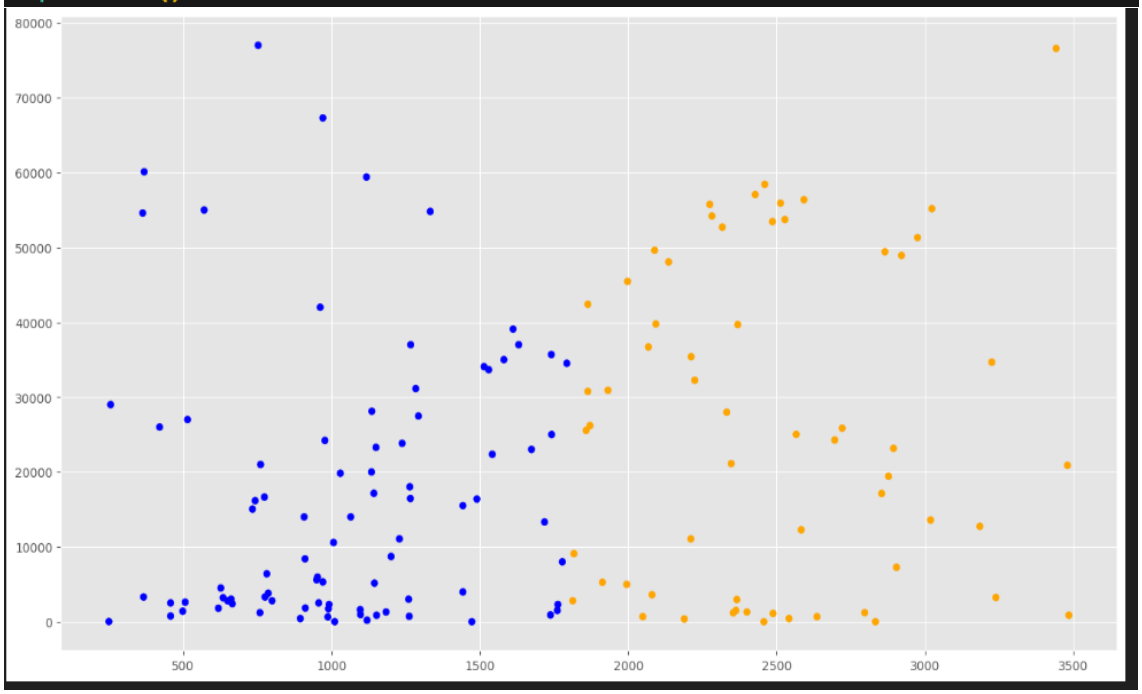
```
filtered_data = data[(data['Word count']<=3500)&(data['# Shares']<=80000)]

colores=['orange','blue']
tamanios=[30,60]

f1=filtered_data['Word count'].values
f2=filtered_data['# Shares'].values

asignar=[]
for index,row in filtered_data.iterrows():
    if(row['Word count']>1808):
        asignar.append(colores[0])
    else:
        asignar.append(colores[1])

plt.scatter(f1,f2,c=asignar,s=tamanios[0])
plt.show()
```



- Paso 5

  Crear los datos de entrada solo Word Count y como etiquetas los # Shares,

crear el objeto LinearRegression y hacerlo encajar con el método fit().

```
#Asignamosnuestravariabledeentrada Xpara entrenamientoylasetiquetasY.
dataX= filtered_data[["Word count"]]
X_train=np.array(dataX)
y_train=filtered_data['# Shares'].values

#Creamoselobjetode RegresiónLinear
regr=linear_model.LinearRegression()

#Entrenamosnuestromodelo
regr.fit(X_train,y_train)

#Hacemoslaspredicciones queen definitivauna línea(enestecaso,alser2D)
y_pred=regr.predict(X_train)

#Veamosloscoeficientes obtenidos,Ennuestro caso,seránlaTangente
print('Coefficients:\n',regr.coef_)

#Esteeselvalordonde cortael ejeY(enX=0)
print('Independentterm:\n',regr.intercept_)

#ErrorCuadradoMedio
print("Meansquarederror: %.2f" % mean_squared_error(y_train,y_pred))

#PuntajedeVarianza. Elmejorpuntajeesun1.0
print('Variancescore: %.2f' % r2_score(y_train,y_pred))
```

✓ 0.0s

```
Coefficients:
 [5.69765366]
Independentterm:
 11200.30322074163
Meansquarederror: 372888728.34
Variancescore: 0.06
```

3 Resultados

Finalmente obtuve la siguiente predicción de la posible cantidad de compartidos en redes sociales con un articulo de 2000 palabras.

```
y_Dosmil = regr.predict([[2000]])
print(int(y_Dosmil))
```

✓ 0.0s

```
22595
```

4 Conclusión

La regresión lineal simple tiene limitaciones, como la necesidad de una relación lineal entre las variables y la sensibilidad a valores atípicos. Sin embargo, sigue siendo una técnica esencial para el análisis de datos y la toma de decisiones. Tras la realización de esta actividad solo tuve algunos incovennientes con el archivo el cual no lo lograba leer, pero pude solucionarlo y continue sin problema.