Actividad 9: Programando Regresión Lineal en Python

García Herrera Carlos Eduardo

March 2025

1 Introduction

La regresión lineal es un modelo estadístico que busca establecer una relación lineal entre una variable dependiente (o de salida) y una o más variables independientes (o explicativas). En términos simples, intenta encontrar la "mejor línea" que prediga el valor de la variable dependiente a partir de las variables independientes.

2 Metodología

2.1 Parte 1: Creacion del Ambiente Virtual

Las siguiente lineas de codigo, se encargan de crear un ambiente virtual llamado VirtualEnv con las librerias necesarias para ejecutar el script

```
#Automatic creation of a virtual environment to run the script and install the libraries
import subprocess
import os
import venv
import sys
script_dir = os.path.dirname(os.path.realpath(__file__))
env_name = os.path.join(script_dir, "VirtualEnv")
if os.path.exists(os.path.join(script_dir, "VirtualEnv")):
    #Checks if the VirtualEnv is activated
    if sys.prefix == sys.base_prefix:
        print("Activating the Virtual Environment...")
        python_exe = os.path.join(env_name, "Scripts", "python")
        subprocess.run([python_exe, __file__])
else:
    print("Installing the Required Libraries on a New Virtual Environment")
    venv.create(env_name, with_pip=True)

    libraries = ["scikit-learn", "matplotlib","seaborn","pandas","numpy"]
```

```
    for lib in libraries:
        subprocess.run([os.path.join(env_name, "Scripts", "pip"), "install", lib],
        check=True)

    python_exe = os.path.join(env_name, "Scripts", "python")
    subprocess.run([python_exe, __file__])
```

2.2 Parte 2: Analisis de los datos

Se cargan los datos contenidos en articulos_ml.csv y se analizan sus distintas propiedades

```
#cargamos los datos de entrada
data = pd.read_csv("./articulos_ml.csv")
#veamos cuantas dimensiones y registros contiene
print("Dimensiones de los datos:")
print(data.shape)
#Veamos los primeros registros
print("Primeros 5 registros:")
print(data.head())
# Ahora veamos algunas estadísticas de nuestros datos
print("Estadististicas de los Datos:")
print(data.describe())
# Visualizamos rápidamente las caraterísticas de entrada
data.drop(['Title','url', 'Elapsed days'],axis=1).hist()
plt.show()

# Vamos a RECORTAR los datos en la zona donde se concentran más los puntos
# esto es en el eje X: entre 0 y 3.500
# y en el eje Y: entre 0 y 80.000
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

# Vamos a pintar en colores los puntos por debajo y por encima de la media de Cantidad
de Palabras
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()
```

2.3 Parte 3: Regresion Lineal

A continuacion se hace el calculo de la regresion Lineal:

```
#Regresion Lineal
# Asignamos nuestra variable de entrada X para entrenamiento y las etiquetas Y.
dataX = filtered_data[["Word count"]]
X_train = np.array(dataX)
y_train = filtered_data['# Shares'].values

# Creamos el objeto de Regresión Lineal
regr = linear_model.LinearRegression()

# Entrenamos nuestro modelo
regr.fit(X_train, y_train)
# Hacemos las predicciones que en definitiva una línea (en este caso, al ser 2D)
y_pred = regr.predict(X_train)

# Veamos los coeficientes obtenidos, En nuestro caso, serán la Tangente
print('Coefficients: \n', regr.coef_)
# Este es el valor donde corta el eje Y (en X=0)
print('Independent term: \n', regr.intercept_)
# Error Cuadrado Medio
print("Mean squared error: %.2f" % mean_squared_error(y_train, y_pred))
# Puntaje de Varianza. El mejor puntaje es un 1.0
print('Variance score: %.2f' % r2_score(y_train, y_pred))

x = np.linspace(min(f1), max(f1), 100)  # Genera 100 en el mismo rango de los datos
m = regr.coef_  # Pendiente
b = regr.intercept_  # Intersección con el eje Y
y = m * x + b

plt.scatter(f1, f2, c=asignar, s=tamanios[0])
plt.plot(x, y, label=f'$y = {m}x + {b}$', color='red', linewidth=2)
plt.show()

#Vamos a comprobar:
# Quiero predecir cuántos "Shares" voy a obtener por un artículo con 2.000 palabras,
# según nuestro modelo, hacemos:
y_Dosmil = regr.predict([[2000]])

print('Predicción para 2000 palabras:', int(y_Dosmil[0]))

input("Press any key to Exit")
```

3 Resultados

Al ejecutar el script de pyhton la informacion obtenida es la siguiente:

3.1 Analisis de los Datos

La informacion de los datos es la siguiente:

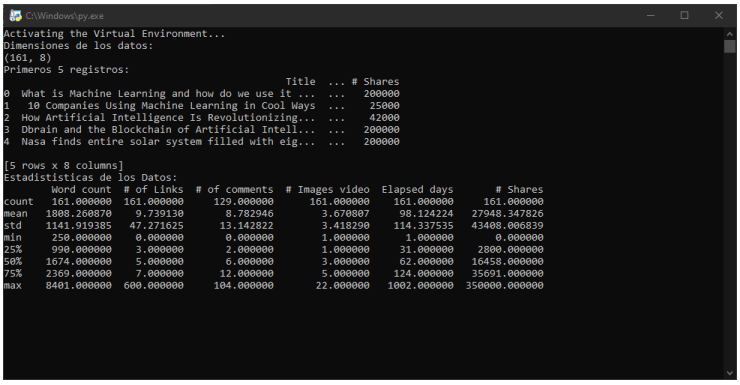

Figure 1: Informacion de los Datos

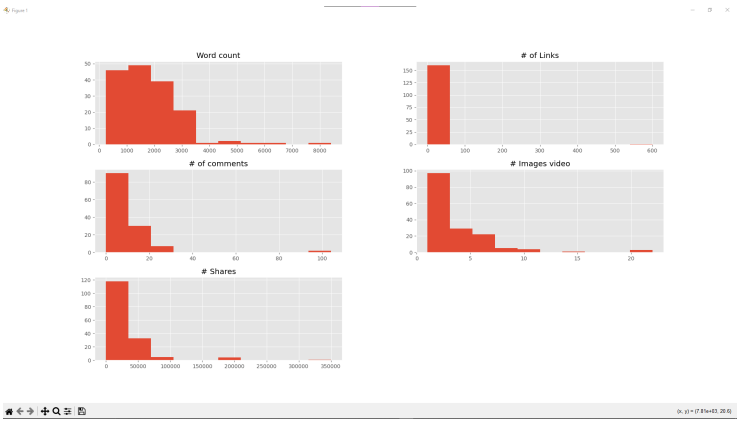

Figure 2: Numero de elementos por Post

3.2 Regresion Lineal

Los resultados de la regresion lineal fueron los siguientes:

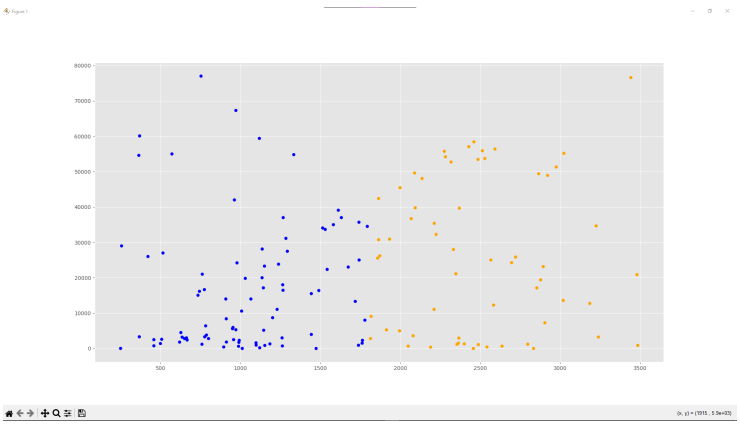

Figure 3: Numero de Palabras vs Numero de Compartidos

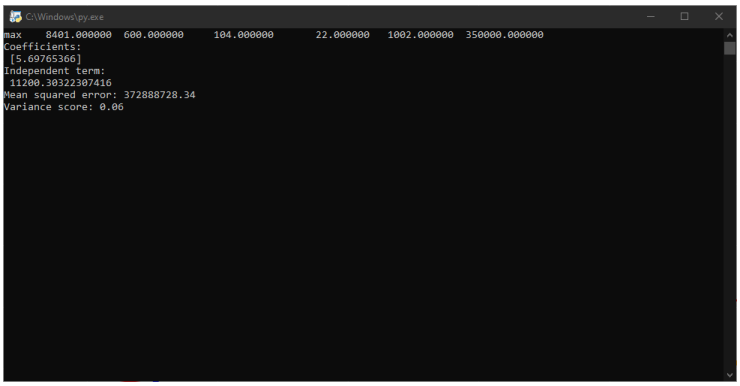

Figure 4: Datos calculados de la regresion Lineal

4 Interpretacion de Resultados

La ecuación general de una recta es de la forma:

$$y = mx + b$$

Donde:

- m es la pendiente de la recta.
- b es el término independiente o intercepto.

Sustituyendo los valores obtenidos de nuestro modelo:

$$y = 5.69765366x + 11200.30322307416$$

- **Coeficiente (pendiente)**: $m = 5.69765366$

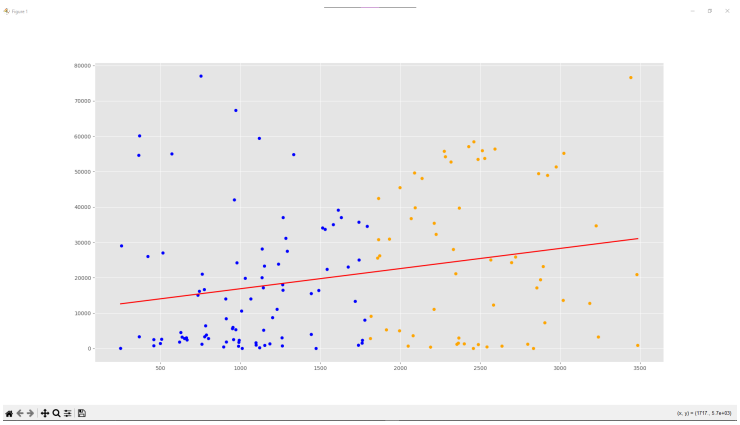

Figure 5: Grafica de la Regresion Lineal sobre los datos

- **Término independiente (intercepto)**: $b = 11200.30322307416$
- **Error cuadrático medio**: 372888728.34
- **Puntaje de varianza (R^2)**: 0.06

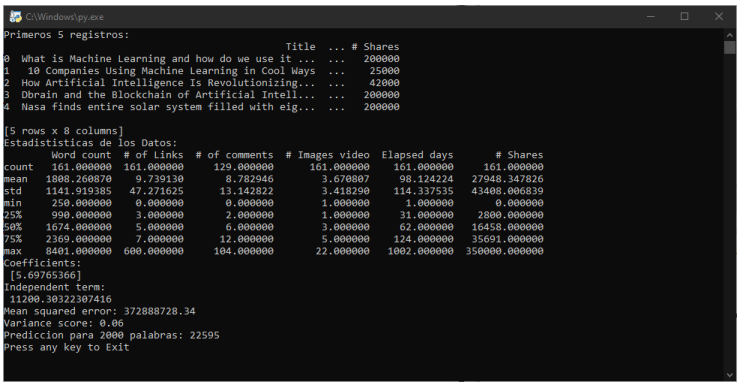

Figure 6: Prediccion Calculada para 2000 palabras

5 Conclusion

La regresión lineal es una técnica estadística utilizada para modelar la relación entre una variable dependiente y una o más variables independientes. A través de esta técnica, podemos ajustar una línea recta a los datos para hacer predicciones o entender cómo varía la variable dependiente en función de las independientes. Es especialmente útil cuando la relación entre las variables es lineal.

Sin embargo, su efectividad depende de la calidad de los datos y de la suposición de que la relación entre las variables es realmente lineal.