

# 1 Que es la regresión lineal

La regresión lineal es un método estadístico que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. La regresión lineal se ajusta a una línea recta o superficie que minimiza las discrepancias entre los valores de salida previstos y reales.

## 1.1 Por que es importante la regresión lineal?

La regresión lineal se aplica a varias áreas de estudios empresariales y académicos, utilizándose principalmente en la predicción de tendencias futuras.

# 2 Metodología

El código del siguiente programa, como lo ejecuto desde la terminal, solo puede darme los datos en la terminal, por lo que no es necesario todas las importaciones originales en el código:

```
1 import numpy as np
2 import pandas as pd
3 from sklearn import linear_model
4 from sklearn.metrics import mean_squared_error, r2_score
```

El código restante y que es importante para obtener los datos que deseamos es el siguiente:

```
1 data = pd.read_csv("./articulos_ml.csv")
2 filtered_data=data[(data['Word count']<=3500)&(data['# Shares']<=80000)]
3
4 f1=filtered_data['Word count'].values
5 f2=filtered_data['# Shares'].values
6
7 dataX = filtered_data[["Word count"]]
8 X_train=np.array(dataX)
9 y_train=filtered_data['# Shares'].values
10
11 regr=linear_model.LinearRegression()
12
13 regr.fit(X_train,y_train)
```

En el código hacemos un filtrado de los datos que obtuvimos del csv, mas específicamente, eliminamos los artículos que tienen una cantidad de palabras mayores a 3500 y los que se hayan compartido mas de 80000 veces. después usando los datos de la cantidad de palabras y el numero de compartidos, se crea y entrena el modelo de regresión lineal.

### 3 Resultados

Al ejecutar el programa obtenemos los siguientes valores

Coefficients	5.69765366
Independent term	11200.30322307416
Mean squared error	372888728.34
Variance score	0.06

Table 1: Caption

### 4 Conclusión

En base a los datos obtenidos podemos concluir que el modelo de regresión lineal no es efectivo para predecir el numero de veces que sera compartido basándonos solo en la cantidad de palabras. Esto lo podemos ver explicando que es cada resultado, veremos los mas relevantes: El termino independiente es de 12200, esto es el valor predicho de compartidos cuando hay cero palabras, lo cual no tiene sentido, un articulo con cero palabras no puede tener tantos compartidos. El error cuadrático medio también es bastante alto, por lo que sera normal que el modelo cometa muchos errores en predicciones.