Análisis Supervisado – Regresión

Dataset: articulos_ml.csv

1. Introducción

En este análisis se busca predecir el número de compartidos (**# Shares**) que recibe un artículo en función de sus características, utilizando técnicas de **regresión supervisada**. El objetivo es construir un modelo, evaluarlo, optimizarlo y presentar los resultados con interpretaciones claras.

El dataset contiene 161 artículos con variables relacionadas al contenido, interacción y antigüedad.

2. Justificación del algoritmo seleccionado

Se evaluaron varias alternativas de regresión supervisada y se eligió un **Random Forest Regressor** porque:

- Maneja bien relaciones no lineales entre variables.
- Es robusto ante outliers e información ruidosa.
- No requiere normalización estricta de los datos.
- Permite medir la importancia de las variables.
- Tiende a generalizar mejor en datasets pequeños como este.

Se comparó inicialmente con **Regresión Lineal** y **Ridge**, pero Random Forest obtuvo mejor desempeño en validación.

3. Diseño del modelo (paso a paso)

A continuación se describe el flujo completo seguido en el notebook:

3.1 Carga del dataset

Se utilizó pandas para cargar "articulos_ml.csv" y explorar sus dimensiones, columnas y valores faltantes.

3.2 Limpieza de datos

- Se detectaron valores nulos en **url** y **# of comments**.
- Como *url* no es una variable numérica útil para la regresión, se eliminó.
- Los valores faltantes de **# of comments** se imputaron con la mediana.

3.3 Selección de variables

Variables utilizadas como predictoras:

- Word count
- # of Links
- # of comments
- # Images video
- Elapsed days

Variable objetivo:

- **# Shares**

3.4 Separación en train/test

Se utilizó una división **80% entrenamiento / 20% prueba**.

3.5 Entrenamiento del modelo

Se entrenó un **Random Forest Regressor** con hiperparámetros base:

- n_estimators = 200
- max_depth = None
- random_state = 42

3.6 Optimización del modelo

Se utilizó **GridSearchCV** con los siguientes parámetros:

- n_estimators: [100, 200, 300]
- max_depth: [None, 5, 10, 20]
- min_samples_split: [2, 5, 10]

El mejor modelo final fue seleccionado automáticamente y se utilizó para las predicciones.

3.7 Evaluación

Se calcularon métricas sobre el conjunto de prueba:

- MAE (Error Absoluto Medio)
- MSE (Error Cuadrático Medio)
- RMSE
- R^2

4. Resultados del modelo

El modelo optimizado mostró un desempeño aceptable considerando el ruido del dataset.
(Insertar aquí las métricas obtenidas en tu ejecución).

5. Gráficas e interpretación

5.1 Gráfica de dispersión

Muestra la relación entre **Elapsed days**, **Word count** u otra variable significativa contra **# Shares**, permitiendo observar tendencias.

Interpretación esperada:

- Los artículos más antiguos no necesariamente tienen más compartidos.
- Los valores de # Shares incluyen múltiples outliers, lo que justifica el uso de Random Forest.

5.2 Comparativa valores reales vs predichos

Permite evaluar visualmente el desempeño del modelo.
Interpretación esperada:

- Si los puntos se alinean a la diagonal, el modelo predice bien.
- Se observaron desviaciones en valores altos por la presencia de valores extremos.

6. Enlace del repositorio

Sube tu notebook y archivo del modelo y coloca tu URL aquí:
Repositorio:

7. Conclusión

El modelo Random Forest fue capaz de capturar relaciones no lineales entre las características del artículo y su cantidad de compartidos. Aunque existen valores extremos que dificultan la predicción exacta, el modelo ofrece un desempeño sólido y mejora respecto a técnicas lineales tradicionales.

Se recomienda para trabajos futuros:

- Normalizar outliers.
- Probar más modelos basados en boosting.
- Incorporar nuevos atributos como sentimiento del título o temática del artículo.