Pràctica 2

Christian Bevilacqua i Aregall

1. Descripció del dataset

El dataset amb el que treballem conté informació sobre les 50 publicacions més populars del subreddit r/popular, extretes mitjançant Web Scraping a través de la pràctica anterior (Pràctica 1). Aquest dataset és rellevant perquè reflecteix els interessos i les interaccions de la comunitat de Reddit en temps real. Hi ha diverses preguntes que es podrien respondre a partir de l'anàlisi d'aquestes dades, com ara són les següents:

1. Hi ha relació entre el subreddit i el nombre de vots?
2. Hi ha relació entre la longitud del títol i el nombre de vots?

Les variables que conté el data set, a mode de resum, són les següents:

- **Posició:** Posició del post en el rànquing (1-50)
- **Títol:** El títol del post
- **Subreddit:** Subreddit d'origen de la publicació
- **Autor:** Nom d'usuari de l'autor del post.
- **Vots:** Nombre de vots que ha rebut el post.
- **Enllaç\r:** URL directa a la publicació.

En quant al tamany, comentar que el dataset conté 29 observacions (1 per cada publicació popular i no hi arriben a 50 ja que en aquella data no hi havia tantes publicacions, encara que en el codi es pot veure que el límit està establert en 50) i un total de 6 variables cada observació.

	Posició	Títol	Subreddit	Autor	Vots	Enllaç\r
0	1	Trump to impose 50% additional tariffs on Chin...	StockMarket	RoyalChris	23903	https://i.redd.it/78sqwqccjfte1.png\r
1	2	Wingo Internet Deal! 🚀 Surfen mit bis zu 1 Gbi...	u_wingo-ch	wingo-ch	0	https://alb.reddit.com/cr?za=UksJga7zW0BnRrDD9...
2	3	Shot of a lifetime, captured from a car window	nextfuckinglevel	solateor	49826	https://v.redd.it/hm34sgm2zite1\r
3	4	Elon Musk gets trolled while attempting to liv...	Fauxmoi	bipartisanic	45033	https://v.redd.it/oauuwxojgite1\r
4	5	Trump Orders Four Mile Military Parade for his...	politics	Ok-Direction-4480	33447	https://www.thedailybeast.com/trump-orders-fou...

2. Integració i selecció

Com que el dataset ja ha estat extret d'una font agregada (r/popular) i presenta només les 29 publicacions més rellevants, no és necessari integrar altres conjunts de dades. En aquest cas, ens centrarem en una subselecció de variables útils per a una primera anàlisi descriptiva i visual, mantenint les variables amb informació rellevant per a l'estudi de popularitat, les quals seran:

- **Posició:** Ens ajudarà a ordenar els posts i estudiar la seva rellevància.
- **Subreddit:** Ens facilita agrupar continguts per temàtica.
- **Vots:** Entenem els vots com una mesura vital per a entendre la popularitat.
- **Títol:** Ens podria servir per analitzar paraules clau.

Amb la selecció d'aquestes variables, considerem que es poden analitzar diferents temes interessants com ara:

1. Quins subreddits tenen més posts dins del Top 50.
2. Hi ha relació entre la posició del post i el nombre de vots?

A continuació, tal i com es demana en l'enunciat, es mostrarà un resum de les dades que permeti veure a simple vista les diferents variables i els seus rangs de valors. A més a més, també es mostraran un parell d'exemples per a que quedi exemplificat d'una millor forma.

Estadístiques bàsiques de la variable numèrica (vots):
count 28.000000
mean 20957.392857
std 16076.105071
min 0.000000
25% 9560.250000
50% 18804.000000
75% 29127.250000
max 52115.000000
Name: Vots, dtype: float64

Valors únics en columnes categòriques:
Subreddits únics: 27
Autors únics: 26

Mostra de subreddits i autors més comuns:
Top 5 subreddits:
Subreddit
Anticonsumption 2
u_wingo-ch 1
StockMarket 1
Fauxmoi 1
politics 1
Name: count, dtype: int64

Top 5 autors:

Autor	
RoyalChris	3
wingo-ch	1
solateor	1
bipartisanic	1
Ok-Direction-4480	1

Name: count, dtype: int64

Exemple d'enllaços i títols:

	Títol \	Enllaç\r
0	Trump to impose 50% additional tariffs on Chin...	https://i.redd.it/78sqwqccjfte1.png\r
1	Wingo Internet Deal! 🚀 Surfen mit bis zu 1 Gbi...	https://alb.reddit.com/cr?za=UksJga7zW0BnRrDD9...
2	Shot of a lifetime, captured from a car window	https://v.redd.it/hm34sgm2zite1\r

3. Neteja de dades

Aquest apartat té com a objectiu garantir la qualitat i consistència de les dades abans de procedir amb l'anàlisi. Les dades s'han netejat mitjançant les següents accions:

Exercici 3.1 Tractament de valors nuls i buits

Identificació de valors nuls amb `df.isnull().sum()` per detectar columnes amb valors absents.

Detecció de cadenes buides amb `(df == "").sum()` per localitzar camps aparentment plens però que contenen valors buits.

Eliminació de files incompletes:

- S'han eliminat les observacions amb valors nuls o buits en columnes considerades crítiques (`Títol`, `Subreddit`, `Autor`, `Enllaç`).
- També s'han descartat les publicacions amb **0 vots**, ja que es considera que no aporten informació rellevant per a l'estudi de la popularitat.

Exercici 3.2 Conversió de tipus de dades

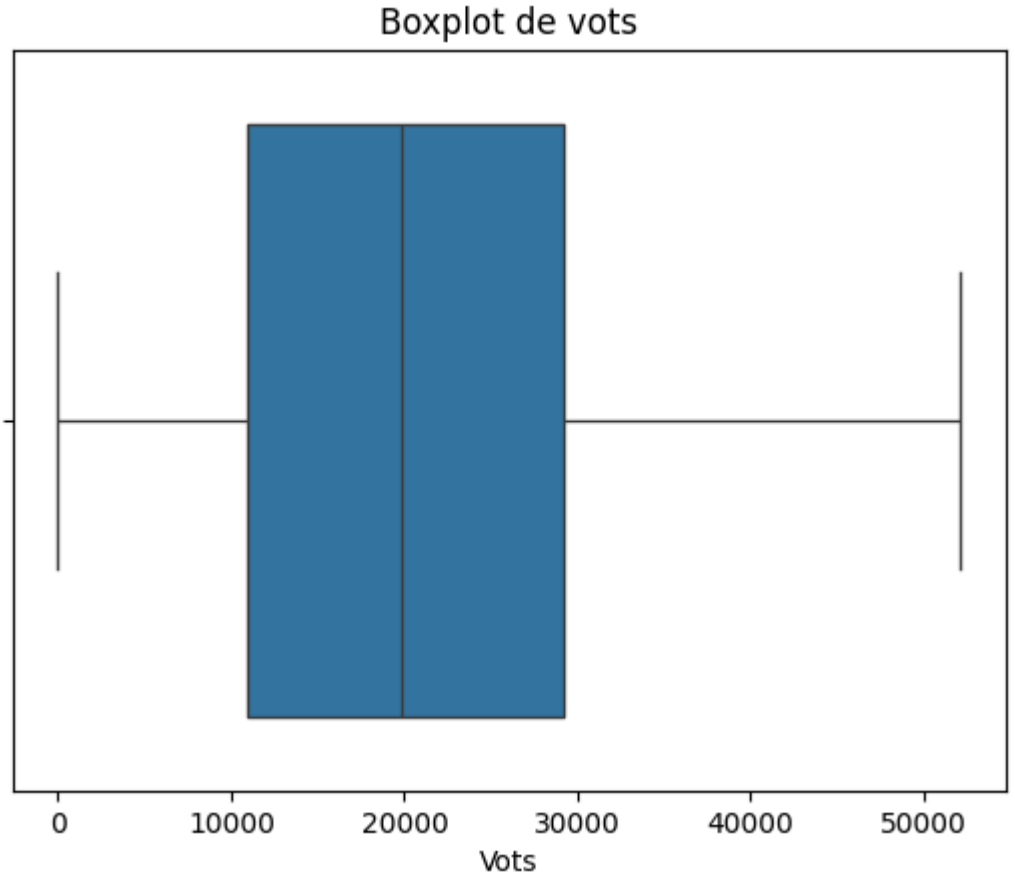
Per facilitar l'anàlisi posterior i assegurar el tractament correcte de les variables, s'han realitzat les conversions següents:

- La columna **Posició** s'ha convertit a tipus `int` utilitzant `pd.to_numeric()`, permetent tractar-la com una variable numérica ordinal.
- La columna **Vots**, clau per a l'anàlisi de popularitat, també s'ha convertit a `int`.
- La columna **Subreddit** s'ha transformat a tipus **categòric** (`category`), ja que representa una variable qualitativa amb valors repetits.

Aquestes transformacions asseguren que les operacions estadístiques i gràfiques es puguin aplicar de manera correcta i eficient.

Exercici 3.3 Identificació de valors extrems (outliers)

Per detectar possibles **valors extrems** en la variable `Vots`, s'ha utilitzat un **boxplot** mitjançant la llibreria `seaborn`. Aquest gràfic permet visualitzar la distribució dels vots i identificar observacions significativament allunyades de la resta.



Tot i que es poden detectar valors alts com a outliers, en aquest cas **no s'han eliminat**, ja que poden representar publicacions veritablement molt populars. S'ha considerat que aquestes dades són **rellevants** per a l'objectiu del projecte.

Exercici 3.4 Correccions addicionals

Per garantir la qualitat i fiabilitat de les dades, s'han implementat dues accions complementàries de neteja:

- **Eliminació de duplicats** mitjançant `drop_duplicates()`, per evitar que publicacions repetides distorsionin els resultats.
- **Validació dels enllaços**, conservant només aquells que comencen per l'URL oficial de Reddit (`https://www.reddit.com/`) per assegurar que la font és fiable.

Finalment, s'ha exportat el dataset net a un fitxer `.csv` amb separador `;` per a la seva posterior anàlisi.

4. Anàlisi de les dades

Exercici 4.1

Model supervisat: Classificació per popularitat

S'ha creat un model de **regressió logística** per predir si una publicació tindrà **alta popularitat** (definida com tenir més vots que la mitjana):

1. Es crea una nova variable binària `AltaPopularitat`.
2. Es codifiquen variables categòriques (`Subreddit`, `Autor`) amb one-hot encoding.
3. Es divideix el conjunt de dades en **entrenament** i **test**.
4. S'entrena el model i s'avalua amb `classification_report`.

A continuació implementaré els dos models per finalment poder comentar els resultats obtinguts:

```
Informe de classificació:
              precision    recall  f1-score   support

           0       0.44      1.00      0.62         4
           1       0.00      0.00      0.00         5

    accuracy                           0.44         9
   macro avg       0.22      0.50      0.31         9
weighted avg       0.20      0.44      0.27         9
```

El model supervisat de classificació logística no ha mostrat un rendiment satisfactori. Amb només 29 observacions i un nombre elevat de variables categòriques, el model no aconsegueix identificar correctament les publicacions amb alta popularitat. El fet que no hagi fet cap predicció per a la classe positiva evidencia problemes greus de desequilibri i sobreajustament. Per millorar el rendiment caldria probablement tenir una mostra més gran.

Tot i els esforços, el model ha mostrat **resultats molt pobres** degut a:

- Mida de mostra molt reduïda.
- Alt desequilibri entre classes.
- Massa variables categòriques en relació amb el nombre de mostres.

Conclusió: Caldria una base de dades més gran i tècniques per tractar desequilibris per millorar aquest model.

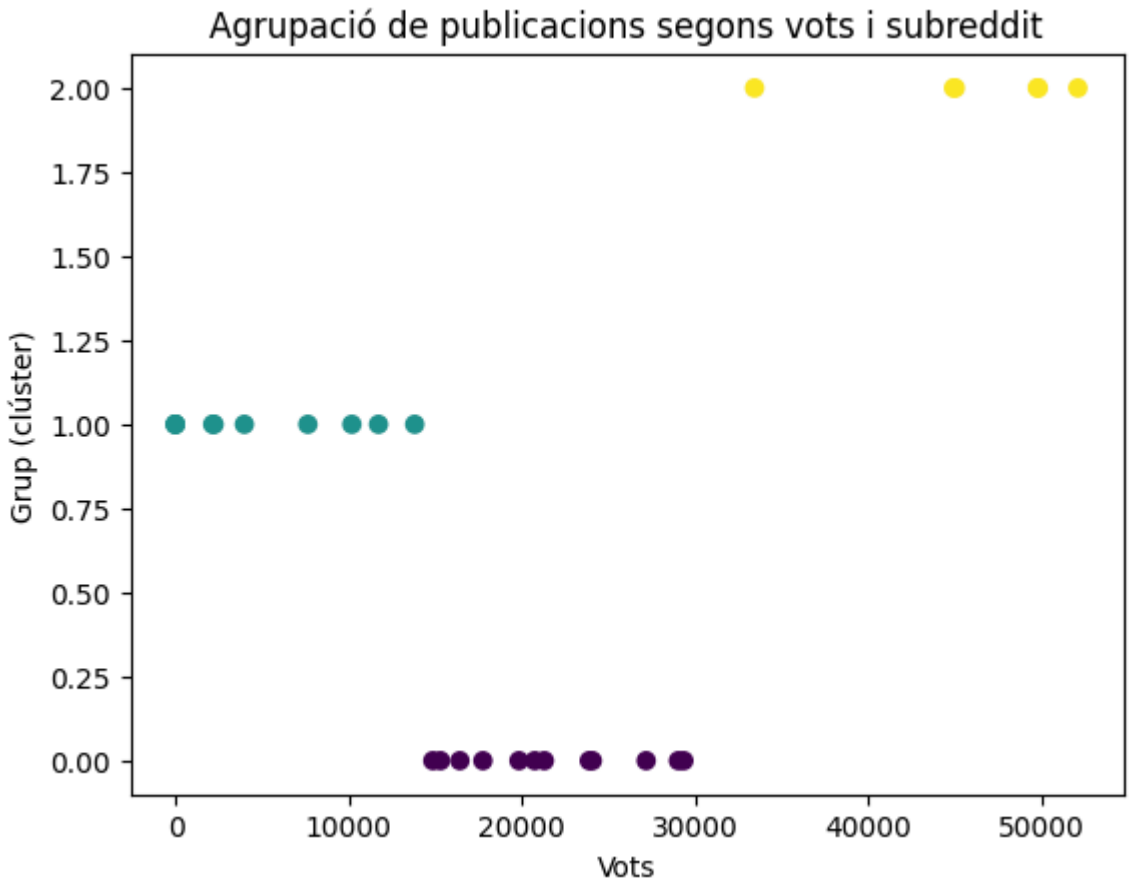
Model no supervisat: Agrupació amb K-Means

Per descobrir patrons ocults, s'ha aplicat un **model de clustering K-Means** amb 3 clústers:

- Variables emprades: `Vots` (numèrica) i `Subreddit` (codificada).
- Dades escalades amb `StandardScaler`.
- Els resultats mostren agrupacions clares segons **nivell de vots**.

> El **subreddit** no ha tingut gaire impacte en la formació dels grups, el que reforça la idea que el nombre de vots és la variable clau per explicar la popularitat.

S'ha representat visualment la distribució dels grups en funció dels vots.



En el model no supervisat s'ha aplicat K-Means amb 3 clústers, emprant com a variables els vots i el subreddit. El resultat mostra una agrupació clara basada en el nombre de vots, amb tres segments diferenciats: publicacions amb pocs vots, amb vots mitjans i amb vots molt elevats. Això indica que la variable "vots" té un pes decisiu en l'estructura de les dades. Tot i haver inclòs la informació del subreddit, sembla que no ha tingut un impacte rellevant en la formació dels clústers. Aquest resultat pot servir per caracteritzar millor els tipus de publicacions segons la seva popularitat.

Exercici 4.2 Contrast d'hipòtesis

Hipòtesi 1: El subreddit més popular rep més vots?

- Es compara el nombre de vots entre publicacions del subreddit més actiu i la resta.
- Donada la **petita mida de la mostra** i els resultats de les proves de normalitat (Shapiro) i variància (Levene), s'aplica la prova **Mann-Whitney U**.
- Resultat: **no s'observa una diferència significativa** (p = 0.582).

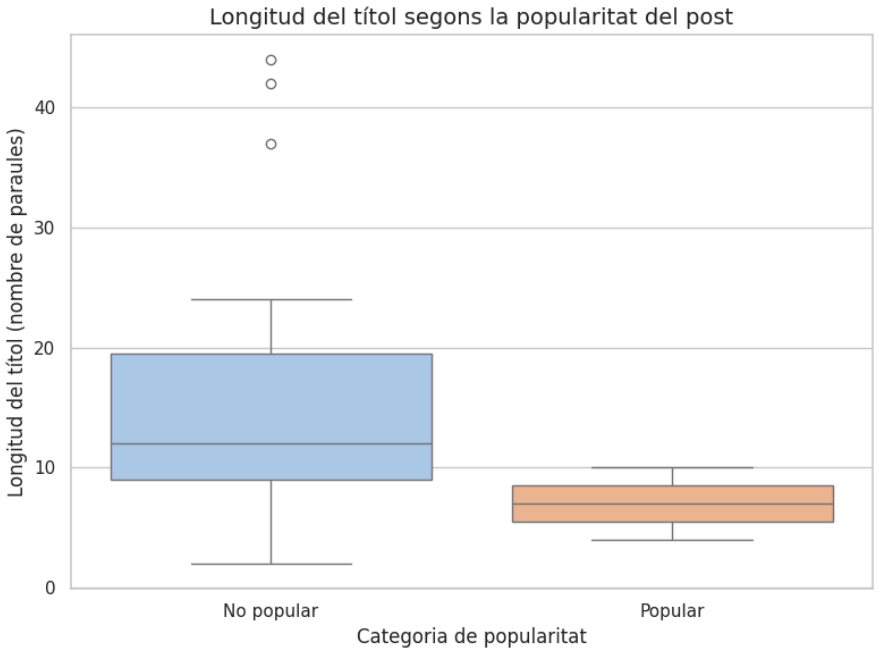
Conclusió: En aquesta mostra, el subreddit d'origen no sembla influir en el nombre de vots que rep una publicació.

Hipòtesi 2: La longitud del títol influeix en la popularitat?

- Es calcula la longitud del títol (en paraules).
- Es compara aquesta longitud entre publicacions populars i no populars amb un altre **test Mann-Whitney U**.
- Resultat: **tampoc es troben diferències significatives** (p = 0.2106).

A més, es representa un **boxplot** per il·lustrar gràficament la distribució de la longitud dels títols segons la popularitat, el qual **confirma visualment** la conclusió estadística.

Aquest gràfic mostra la distribució de la longitud dels títols per als dos grups. Si les caixes (i mediana) són molt semblants, això confirma visualment el resultat del test estadístic: no hi ha diferències rellevants entre els posts populars i no populars pel que fa a la longitud del títol.



Conclusió: En aquest conjunt de dades, la longitud del títol no sembla tenir cap impacte rellevant sobre la popularitat de la publicació.

5. Representació dels resultats:

Tal i com s'especifica en l'enunciat, aquest apartat l'he anat desenvolupant al llarg de la pràctica, ja que després de cada prova o anàlisi realitzat, s'han mostrat els resultats obtinguts.

6. Resolució del problema:

A partir de l'anàlisi de les 50 publicacions més populars de Reddit (r/popular), s'han aplicat diferents tècniques de tractament i anàlisi de dades amb l'objectiu de descobrir patrons i relacions entre variables com el nombre de vots, el subreddit, o l'autor de les publicacions.

A partir de l'elaboració d'aquesta pràctica, les conclusions principals que puc extreure són les següents:

1. **Exploració inicial i visualització:**

Les variables recollides (títol, subreddit, autor, vots i enllaç) mostren una gran variabilitat, especialment en el nombre de vots, amb valors que oscil·len entre menys de 1.000 i més de 50.000.

No tots els subreddits tenen el mateix pes o impacte pel que fa a interaccions.

1. **Model supervisat (classificació):**

El model aplicat (classificador binari) presenta un rendiment molt baix (precisió 0.44, f1-score 0.27), indicant que les dades disponibles no són suficients o prou representatives per predir correctament si una publicació tindrà molts o pocs vots.

A més, el desbalanceig en la mida de la mostra afecta clarament la qualitat del model predictiu.

1. **Model no supervisat (clústers):**

Amb k-means s'han identificat tres grups diferenciats de publicacions, segons el nombre de vots. Això indica que hi ha patrons que permeten agrupar les publicacions en funció de la seva popularitat, tot i que no necessàriament estan relacionats amb el subreddit.

1. **Contrast d'hipòtesis**

Contrast d'hipòtesis sobre el subreddit:

Després de comprovar la normalitat i homocedasticitat, s'ha aplicat la prova de Mann-Whitney U per contrastar si hi havia diferències significatives entre els vots de publicacions del subreddit més popular i la resta.

El resultat (p = 0.582) mostra que no hi ha una diferència significativa, per tant, el subreddit per si sol no sembla condicionar el nombre de vots.

Contrast d'hipòtesis sobre la relació entre post i nombre de vots: S'ha plantejat la hipòtesi que la longitud del títol del post podria estar relacionada amb el nombre de vots que rep. Per comprovar-ho, es va dividir la mostra en dos grups segons la popularitat (posts amb molts vots i posts amb pocs vots) i es va aplicar de nou la prova de Mann-Whitney U per comparar la longitud dels títols. Els resultats estadístics (Estadístic U: 11.50, p = 0.2106) indiquen que no hi ha diferències significatives entre els dos grups pel que fa a la longitud del títol. Per tant, aquesta variable tampoc sembla ser un factor rellevant per explicar el nombre de vots que rep un post.

Responen els resultats al problema?

En certa forma, considero que sí, els resultats aporten una resposta parcial però informativa, extraient les següents conclusions:

No s'ha pogut construir un model predictiu robust per anticipar la popularitat d'una publicació basant-se en les dades disponibles.

L'anàlisi ha permès identificar agrupacions i descartar que el subreddit sigui un factor determinant en el nombre de vots.

Així mateix, la longitud del títol no sembla tenir relació significativa amb la popularitat mesurada en vots.

Aquestes conclusions ens indiquen que la popularitat d'un post a Reddit depèn de factors més complexos i probablement qualitatius, com el contingut específic, el moment de publicació, o les dinàmiques pròpies de cada comunitat, que no s'han pogut captar amb les variables estudiades en aquesta pràctica.

8. Video

A continuacio deixo un link directe al video

https://drive.google.com/file/d/1Oue58QHZxonE5Cfadjy2ojgj_cBV0jJT/view?usp=drive_link