APC-CAS KAGGLE "ONLINE NEWS POPULARITY"

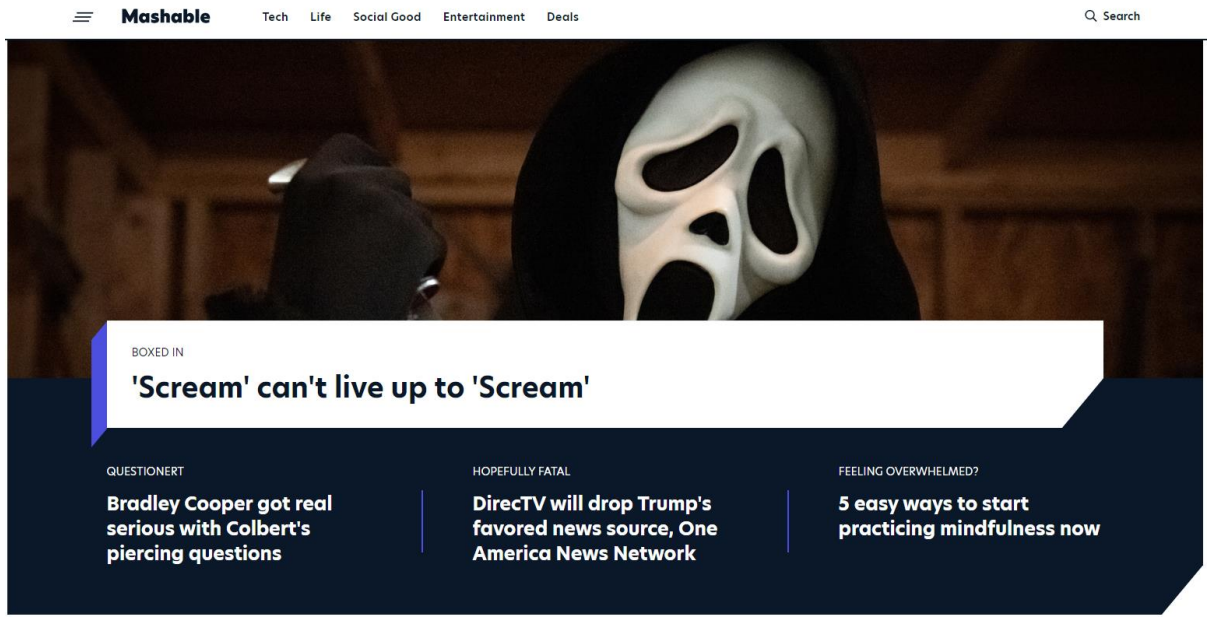
Nom: Javier Alejandro Camacho Machaca

NIU: 1566088

Link Github

INTRODUCCIÓ

- En aquest cas Kaggle ens donen un ‘target’ que es el nombre de vegades que es comparteix un article.
- I la resta son característiques dels articles de la pàgina web **Mashable.**

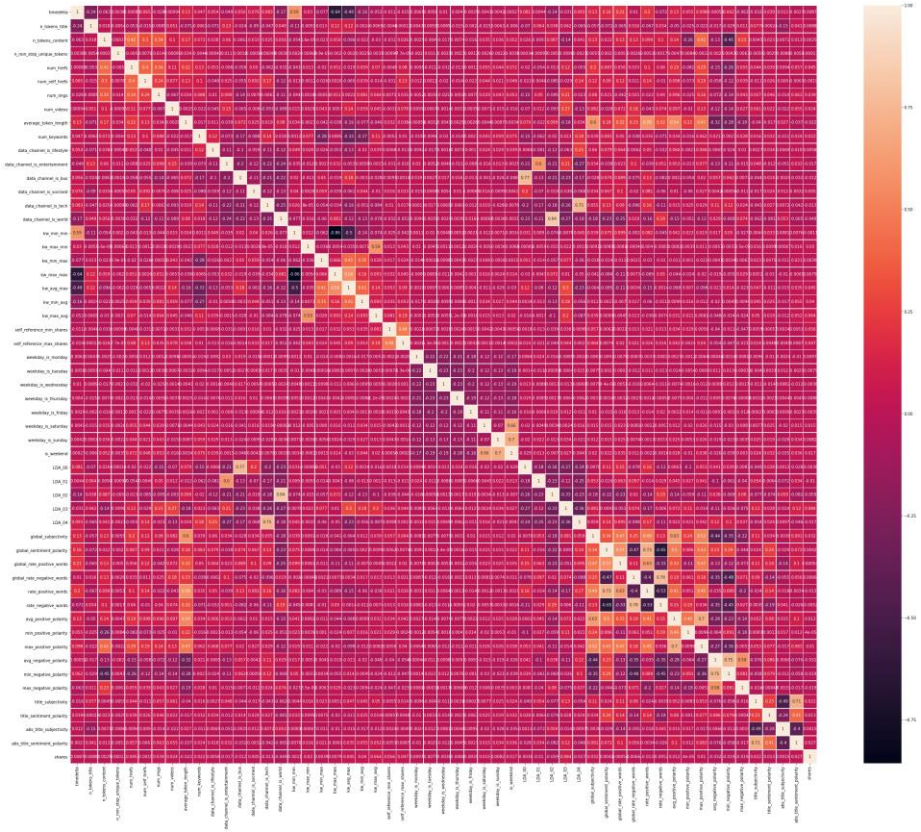


ANÀLISI DE DADES

- Hi han 60 columnes (exceptuant el 'target') que descriuen un article, i aquests estan definits en un '.names' en el github.
- Exemple: **average_token_length:** Longitud mitjana de les paraules en el contingut

	url	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words	n_non_stop_unique_tokens	shares
0	http://mashable.com/2013/01/07/amazon-instant-...	731.000	12.000	219.000	0.664	1.000	0.815	593
1	http://mashable.com/2013/01/07/ap-samsung-spon...	731.000	9.000	255.000	0.605	1.000	0.792	711
2	http://mashable.com/2013/01/07/apple-40-billio...	731.000	9.000	211.000	0.575	1.000	0.664	1500
3	http://mashable.com/2013/01/07/astronaut-notre...	731.000	9.000	531.000	0.504	1.000	0.666	1200
4	http://mashable.com/2013/01/07/att-u-verse-apps/	731.000	13.000	1072.000	0.416	1.000	0.541	505

PREPROCESSAT



- Eliminar les variables que estaven directament correlacionades.
- Valor Atípic > Q3 * 1.5(Q3-Q1)
- Passar les dades a 1 i 0, 1 si es supera el valor atípic i 0 si no.
- Utilitzar això com una mida de que un article sigui viral/popular.

MÈTODE D'APRENENTATGE

	Cross-Validation Score
Random Forest (min_samples_split=9, random_state=42)	88.5%
Decision Tree(random_state=42)	79.7%
Ada Boost(n_estimators=100, random_state=42)	88.4%

RESULTATS I CONCLUSIONS

- Com a resultat i conclusió final, podem arribar a dir que el Random Forest Classifier dona els millors resultats, encara que el Ada Boost es casi igual.
- El Random Forest Classifier es una millor elecció tenint en compte les característiques de la Base de Dades.