FINAL PROJECT

TRISTAN LE-GOFF

NIELS LORGNET

DIA4

THE DATASET

- We are working on the Online News Popularity dataset, dealing with popularity of articles
- 60 variables
- Around 40 000 rows
- Classifiaction problem

- Our goal is to understand the factors impacting on the number of views of an article, according to several variables:
- the day it is published
- the number of words or references
- the subject of the article
- the subjectivity, or the polarity

DATA VISUALISATION

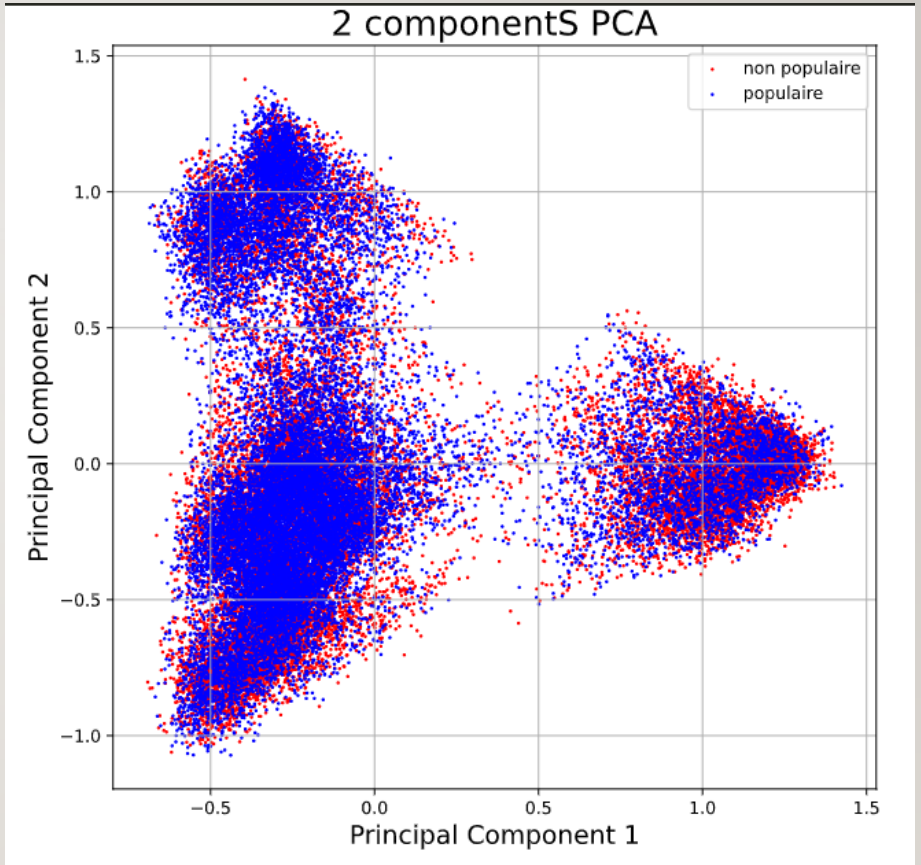
- First, we looked at the correlation between the different variables, in particular the number of shares with others.
- We did this as we are studying the number of shares and we are going to predict it later.

	shares
shares	1.00
kw_avg_avg	0.11
LDA_03	0.08
kw_max_avg	0.06
self_reference_avg_sharess	0.06
self_reference_min_shares	0.06
self_reference_max_shares	0.05
num_hrefs	0.05
kw_avg_max	0.04
kw_min_avg	0.04
num_imgs	0.04
global_subjectivity	0.03
kw_avg_min	0.03
kw_max_min	0.03
abs_title_sentiment_polarity	0.03
num_videos	0.02
title_subjectivity	0.02

Part of the table of correlation between shares and other variables

PCA

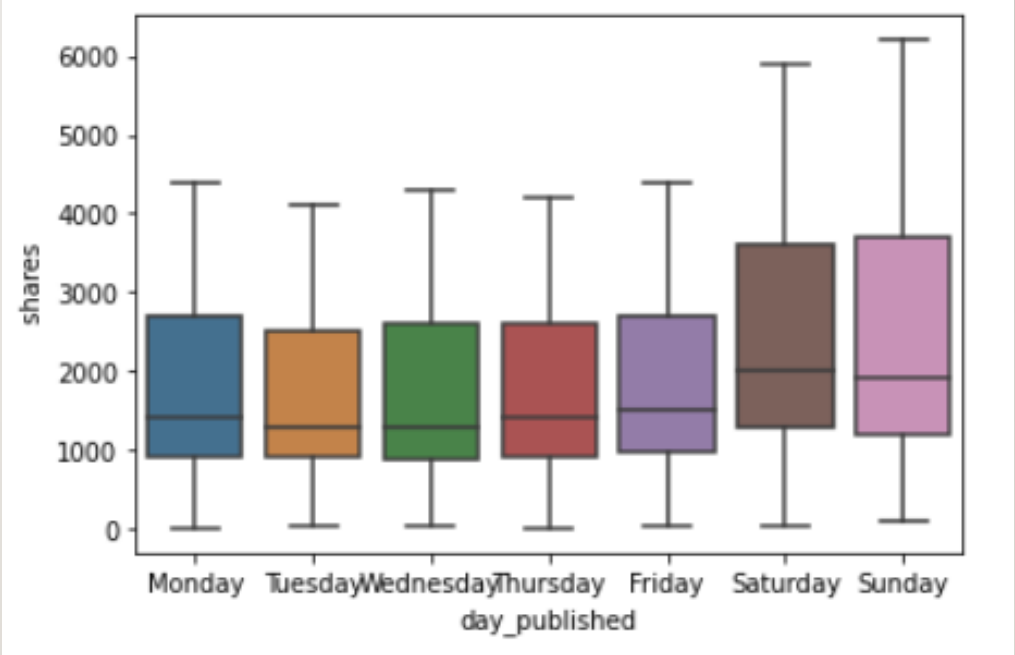
- We wanted to know if a PCA could be effective
- With two components we only explain 20% of the total variance. It was predictable in view of the correlation matrix. We will therefore not be able to use it subsequently



PCA with 2 components on our the dataset

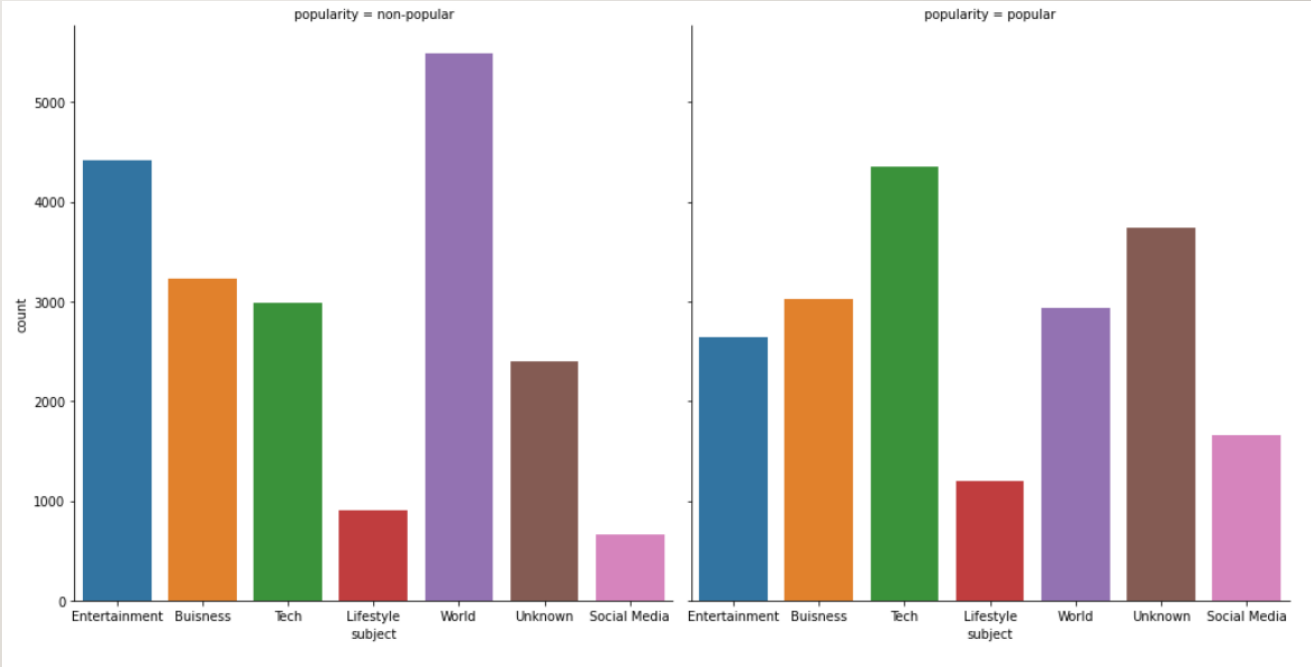
VARIABLES CHANGED

- We grouped the variables for the day it was published in a global variable, in order to have a better visualisation.
- We did the same with the subject of the article.
- Here is an example of what we can do easier by grouping these variables.



DATA INTERPRETATION

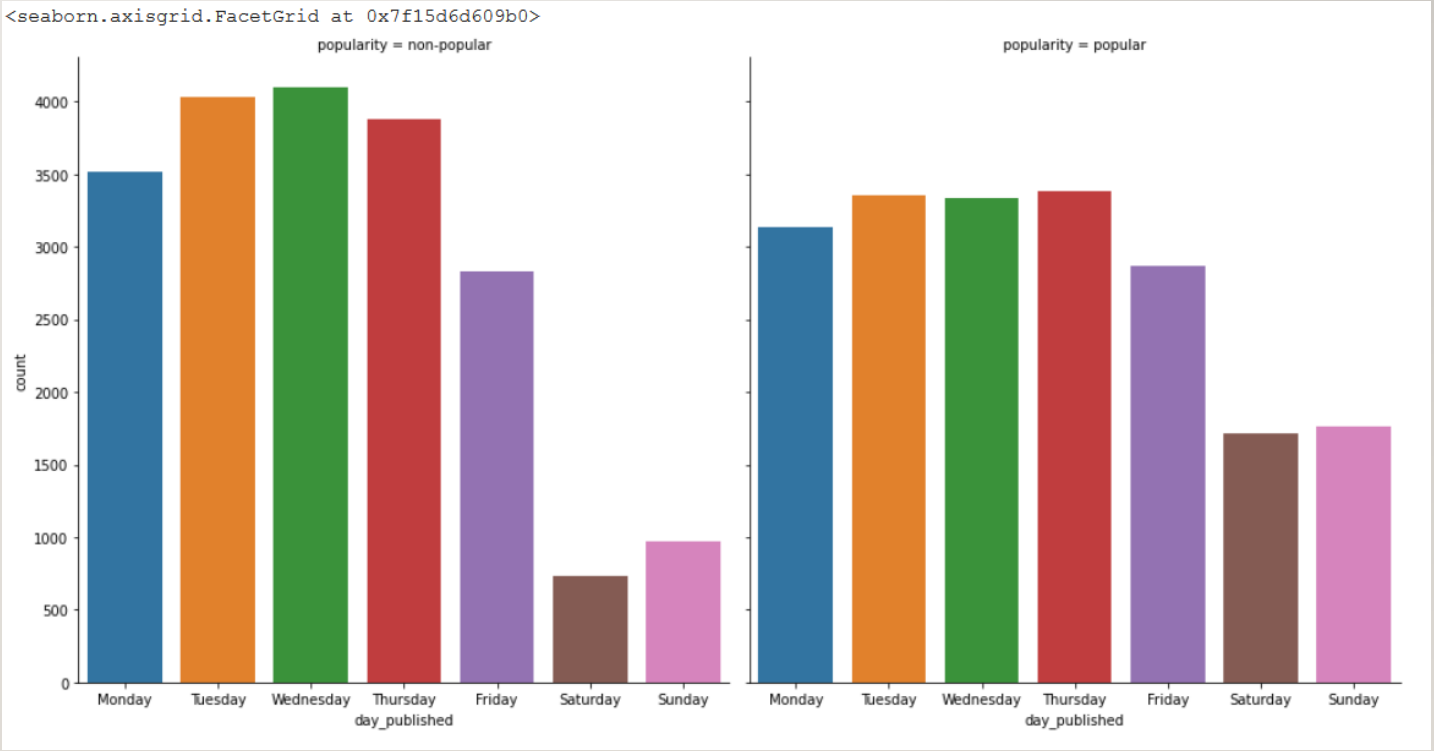
- Here was a key thing to understand about our data. The most popular subjects are not necessary the ones with the highest count of ‘popular article’. We need to compare the left and right to understand it correctly.
- We explained this in our code in details.



Comparison of the popularity between different subjects

DATA INTERPRETATION

- Same as for the subject, the best day to publish is not the one with the highest count of ‘popular articles’ but rather the one with the highest ratio between ‘popular’ and ‘non popular’ (week-end).
- Or we could simply look at the average number of shares… (graph on page 4)



DATA INTERPRETATION

As a conclusion we can say that there are clearly subjects with a higher popularity than others, same for the day of publication.
There is no quantitative variable impacting that much the popularity of the article as the correlation coefficients remain always very low between ‘shares’ and other variables.
Therefore, it might be hard to fit a good prediction model on this dataset as we will see from now on.

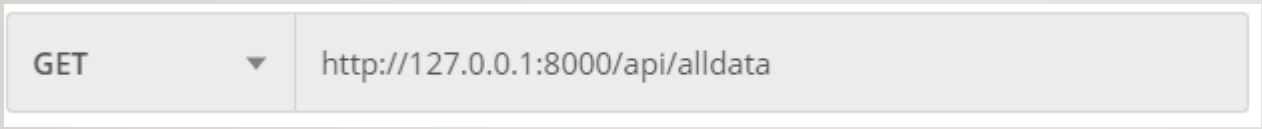
MACHINE LEARNING

SVM	Log Regression	KNN	Random Forest
32%	60%	58%	67%

Above are the results we got with our models. The random forest is by far the best model but it also takes a lot more time than the other models. We were expecting this kind of results according to the dataset and the variables we analysed previously.

DJANGO API REST

- We have created a django api project which contains 2 api routes.
- The first API routes is : http://127.0.0.1:8000/api/alldata

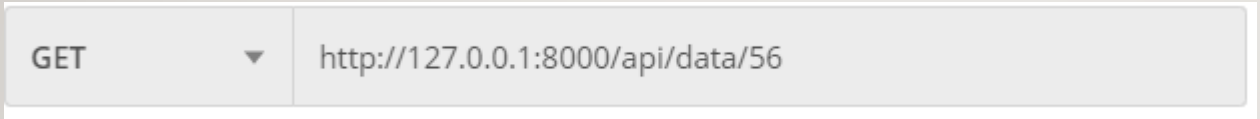


This routes have 3 request :

- Get => Show all articles data
- Post => Post the data that is in the body request and predict the variable « popularity » with our model of machine learning.The data is after send in the database
- Delete => Delete all the datas of the database

DJANGO API REST

- The second API route is : http://127.0.0.1:8000/api/data/[id]



This routes have 3 request :

- Get => Show the article with specific id
- Put => Updates the data by replacing the old one with the body of the request
- Delete => Delete the data with specific id

- To try all the requests we can use PostMan :

As we can see we put in the body of the request all the datas of a new article without the label « popularity ». After POST this request the api route send us a response with the data and the label popularity with our machine learning model. This also adds it to the database.

Th database is a simple SQL Lite database.