Python for Data Analysis

Jules THUILLIER – Victor THUILLIÉ

DIA5

Context

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years.

The goal is to predict the number of shares in social networks (popularity).

There are 61 variables, 58 predictive, 2 non predictive and the target variable

Context

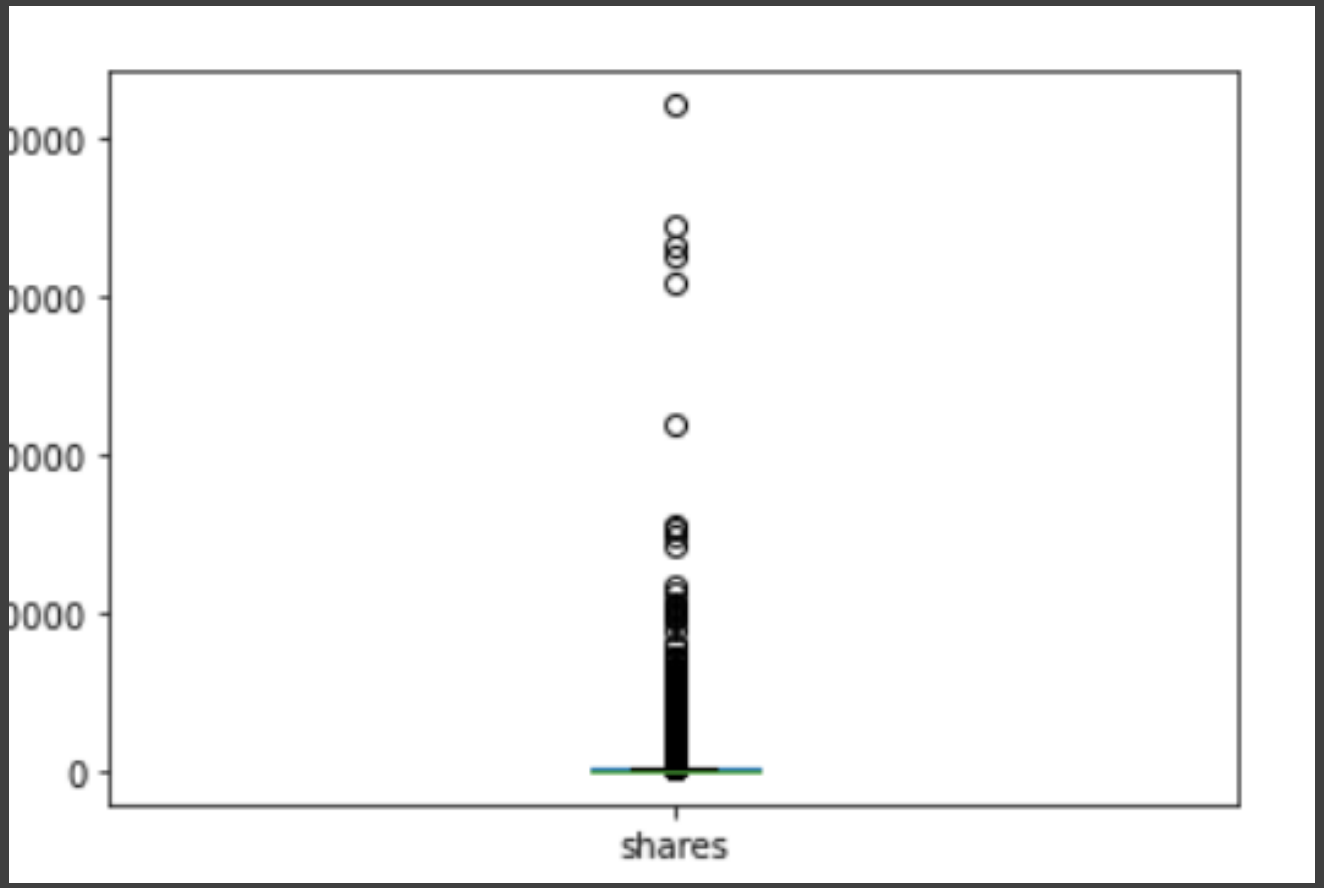
As a first observation, it would seem that there are only numerical values

At first, we will try to predict a continuous value (the number of shares), so we place ourselves in a case of regression

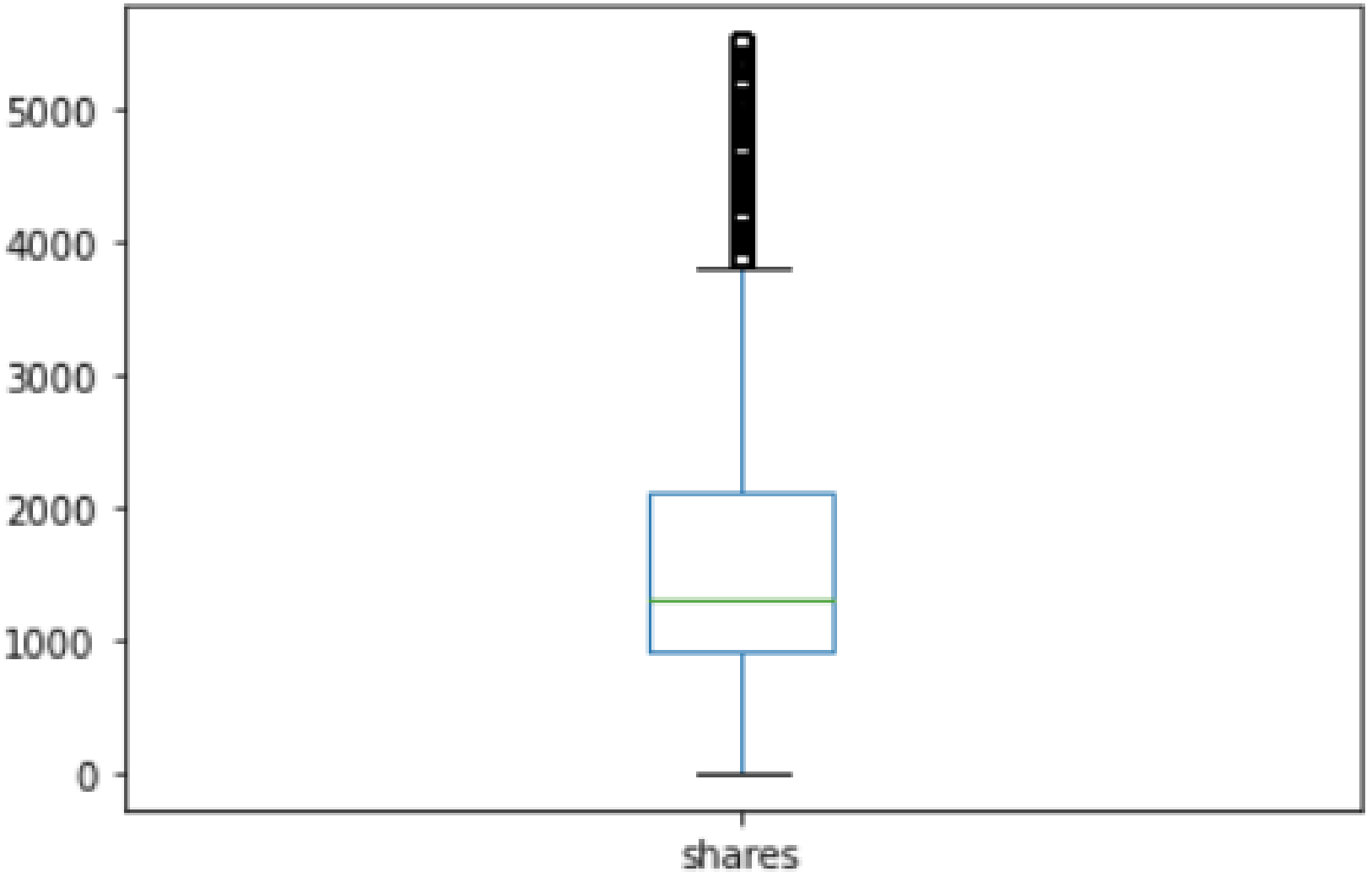
However, we can already add the 'popular' column for later. (We will explain it later)

Table of Content

Data Exploration – Outliers Values



- We can see that the boxplot is clearly flatten by the outliers values
- That's why we clean the data by removing it

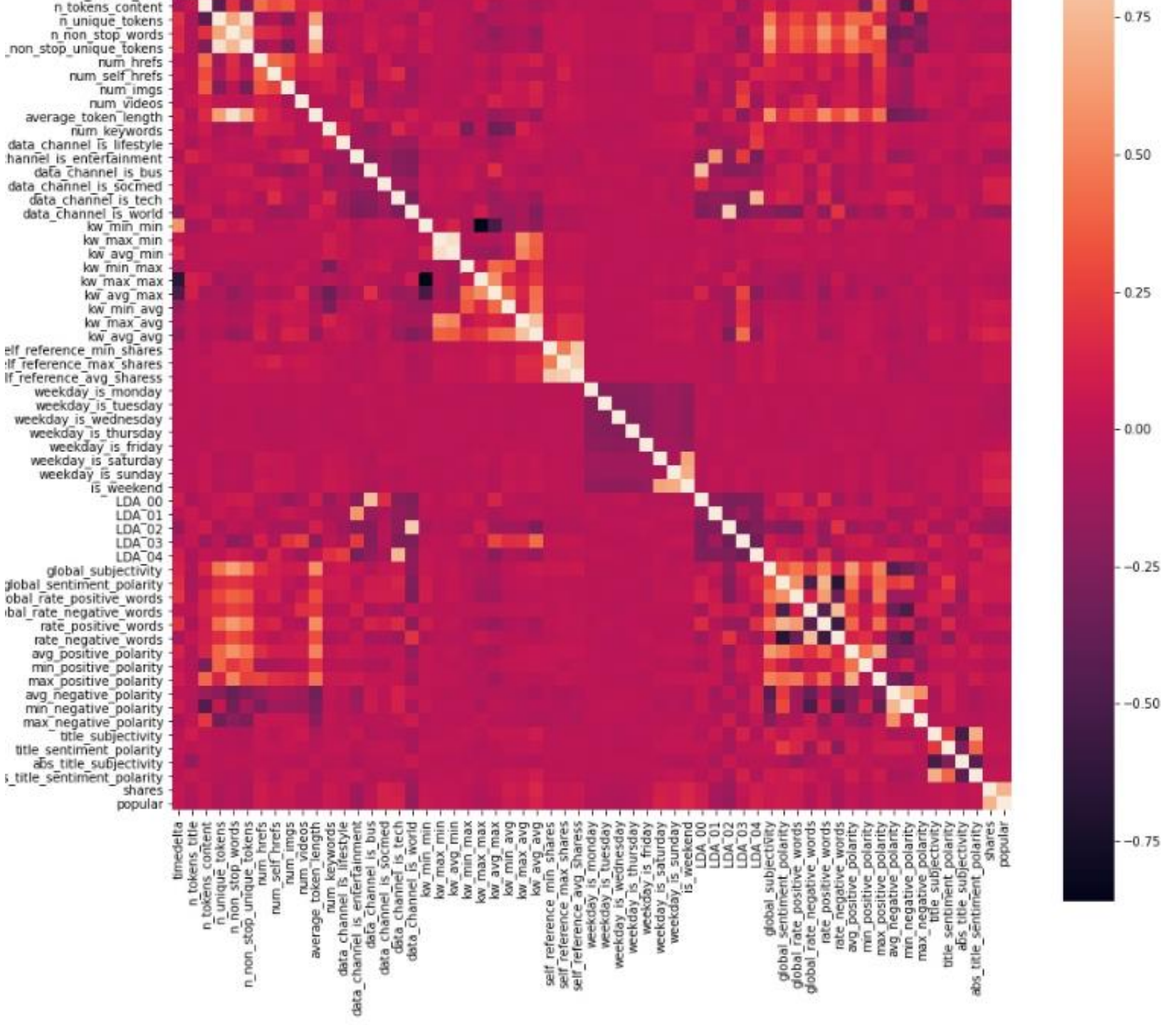


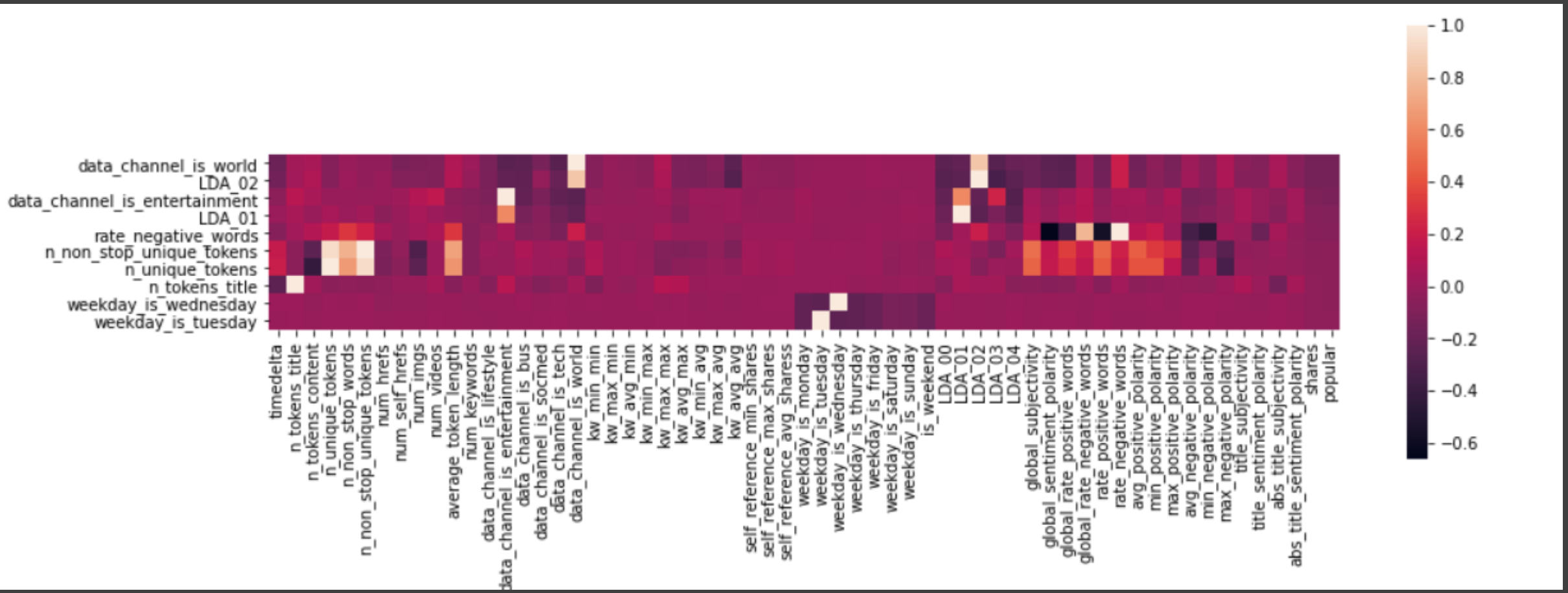
Data Exploration – Outliers Values

Data state after removing outliers

Data Exploration – Correlations Between Values

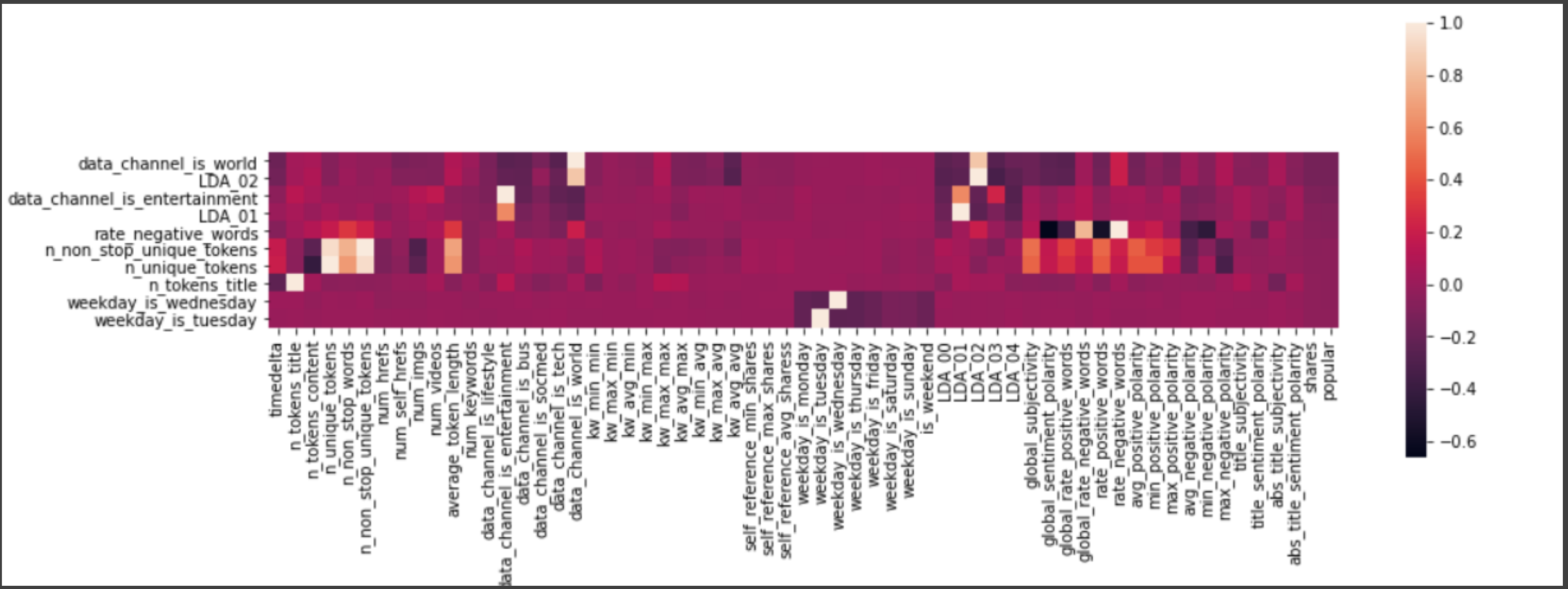
- There are a lot of variables but if you look closely, none of them seem to be correlated with the target variable.
- As this heatmap is not easy to read, we have two more, highlighting the 10 variables most correlated with the target variable and the 10 least correlated.





Data Exploration – Correlations Between Values

- On this heatmap we see that the 10 variables with the lowest correlation with the target variable do not have a significantly negative correlation.
- The lowest correlation value is around -0,14

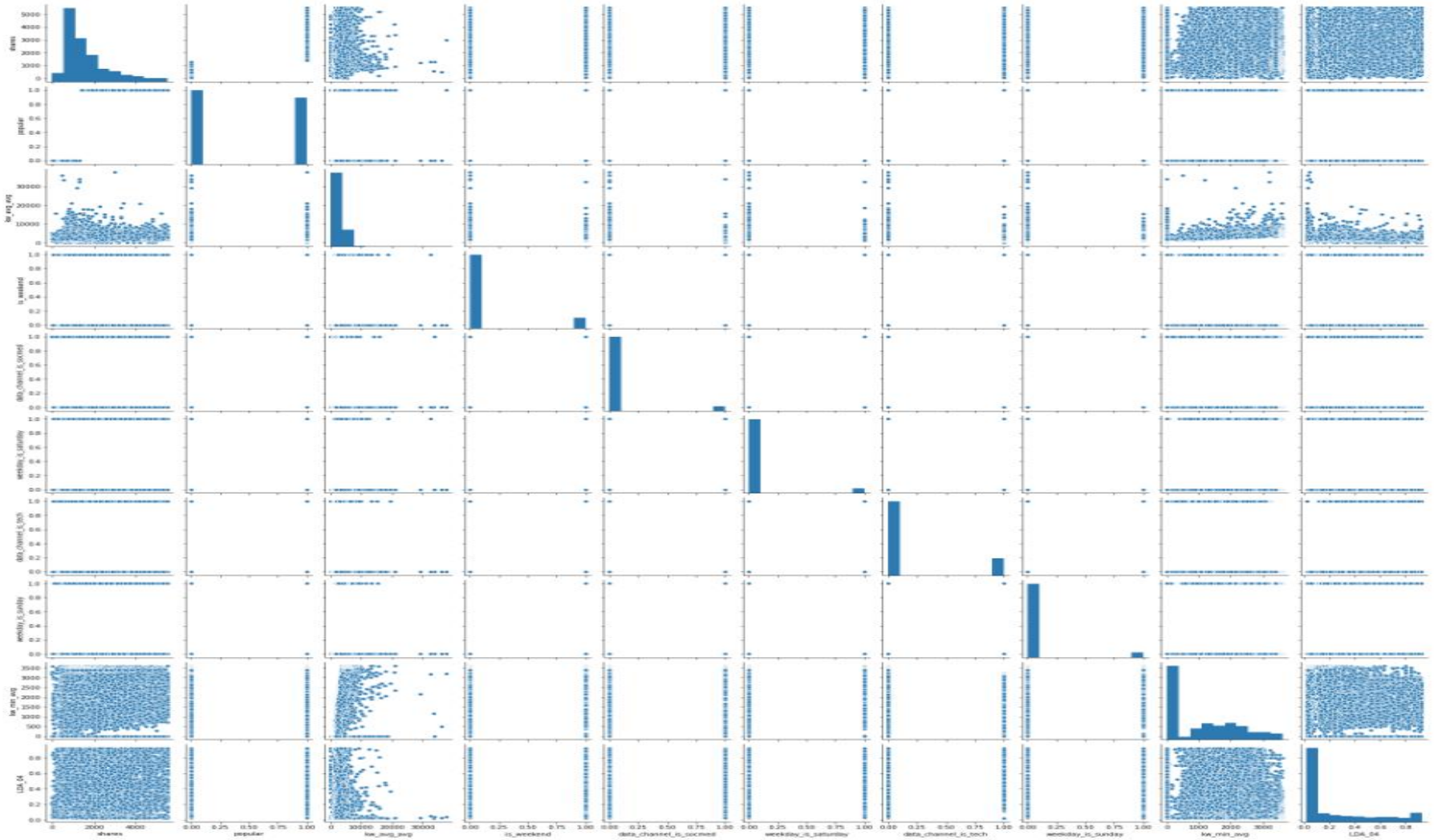


Data Exploration – Correlations Between Values

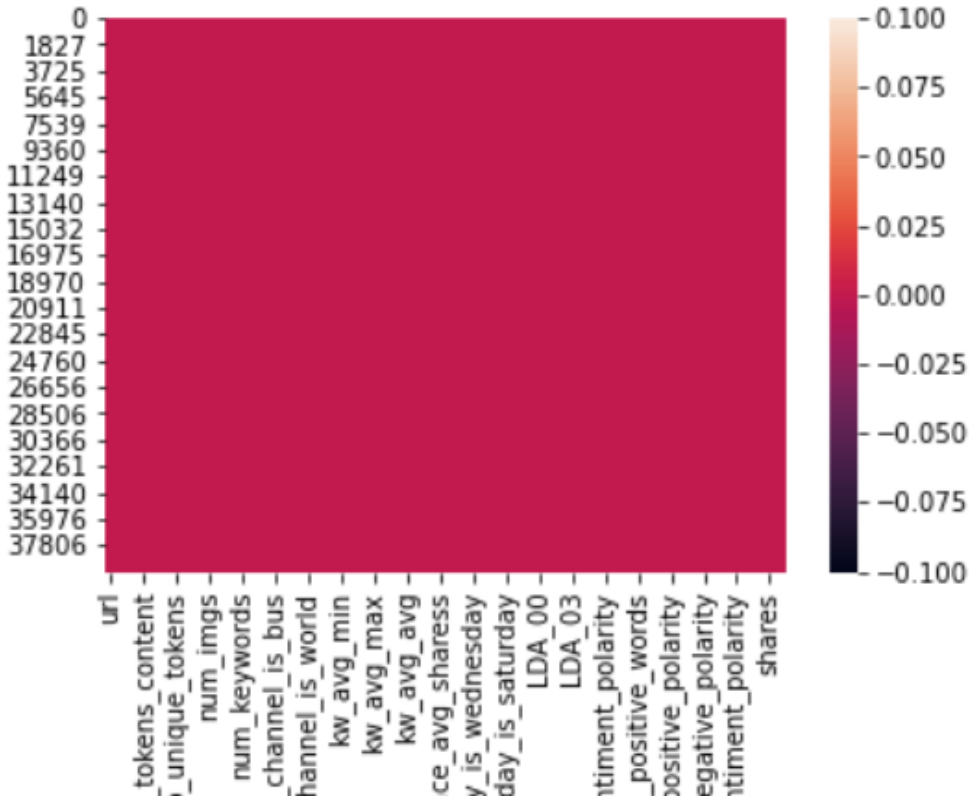
- Here we have the 10 variables most correlated with the target variable, as for the previous heatmap, we find that none of these variables are really correlated with the target variable.
- The highest correlation value is around 0,04

- Shares are on the first line, even if it’s quite difficult to read, we can clearly see that there is no really pattern or interesting pattern between variables that have the highest correlation with the target variable.

- We only have cloud of points or just a separated values.



Data Exploration – Missing Values



- By looking at the description of the dataset and this heatmap we can assert that there are no missing values

A brief first conclusion

- The predictor variables are all numeric
- There are no missing values
- The qualitative variables have already been treated (no need to apply some technique like One hot Encoding, Ordinal Encoder ...)
- There is almost no correlation between the predictive variables and the target variable (shares)
- It looks like the regression is going to be complicated


Regression

Regression

We’ll start by using 2 different types of algorithms

- Linear Regression
- Random Forest Regressor

We will quickly observe the score we obtain with these algorithms before continuing or seeking to optimize the hyperparameters.

Regression – Feature selection

- Even if we are not going to proceed to the hyperparameter tuning right away, we can already try to reduce the number of variables from which we will be working on.
- We use RFECV to do this. (Recursive Feature Elimination Cross Validation)
- It returns us a list of parameters that we will keep and can even give us a ranking of our parameters

Regression – Modeling

- As we don't obtain any good result in regression, we can try with a **classification objective.**
- We obtain a score of 0,11 in the best case with a Random Forest Regressor estimator.
- Now we will try to **determine if an article is popular** instead of the number of shares.

Classification

Classifiation

AS **WE CANNOT** PRECISELY **PREDICT THE NUMBER OF SHARES**

OUR NEW GOAL IS TO **PREDICT** WHETHER THE **ARTICLE IS POPULAR OR NOT.**

Classification

But what does popular mean ?

Classification

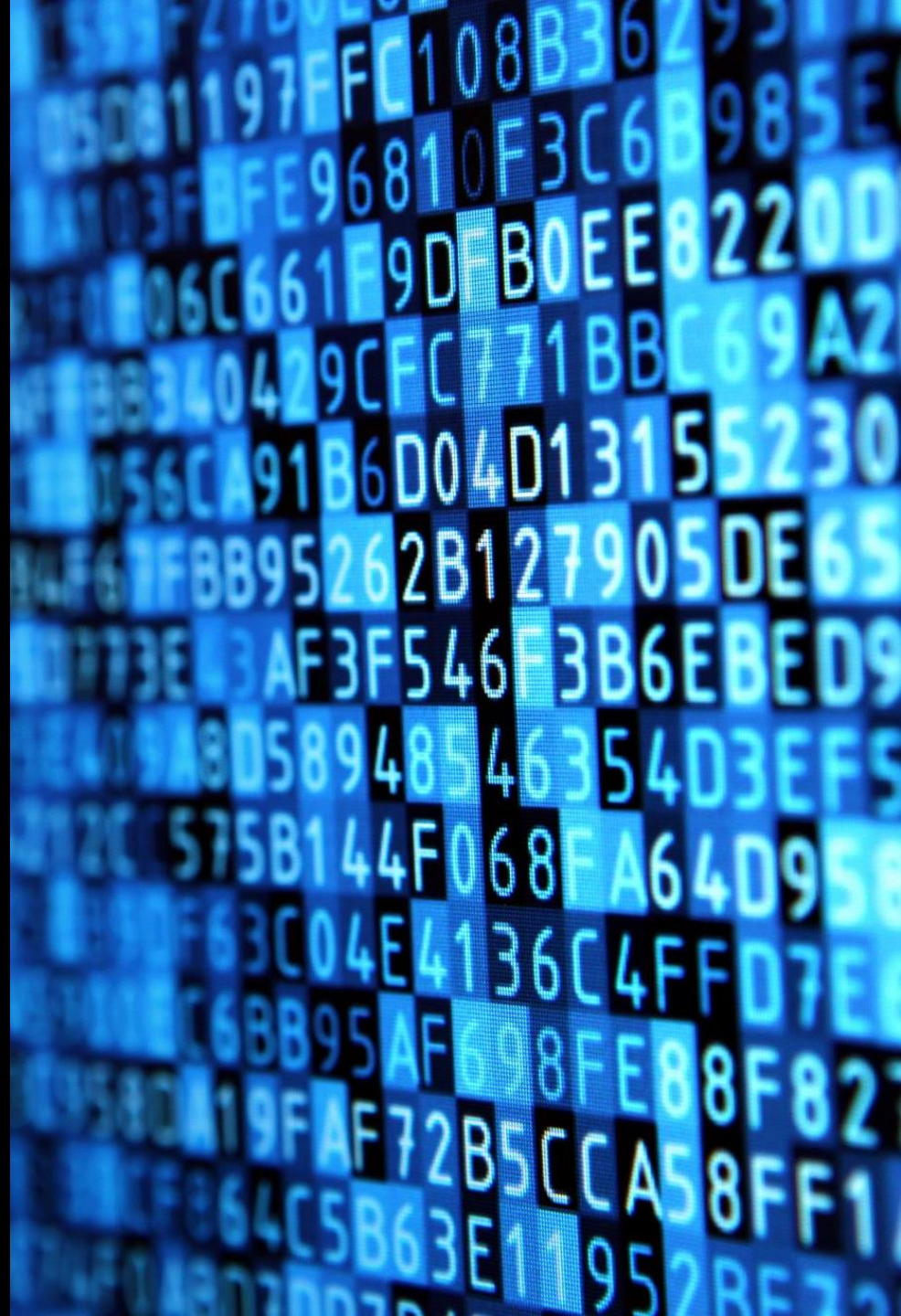
- That’s why we create the variable ‘popular’ as a new target variable for a classification problem.
- The popularity threshold is defined by the median of the target variable.
- if the number of shares is greater than this value, we consider the article to be popular, otherwise, it is not.

Classification

We'll start by using 2 different types of algorithms

- SGDClassifier
- Random Forest Classifier

We will also quickly observe the score we obtain with these algorithms before continuing or seeking to optimize the hyperparameters.



Classification – Feature selection

- As with regression, we use RFECV to reduce the amount of parameter to keep.
- However, we realize that the variables have almost no correlation with the target variable so we can hardly eliminate a lot of the variable.

Classification – Modeling

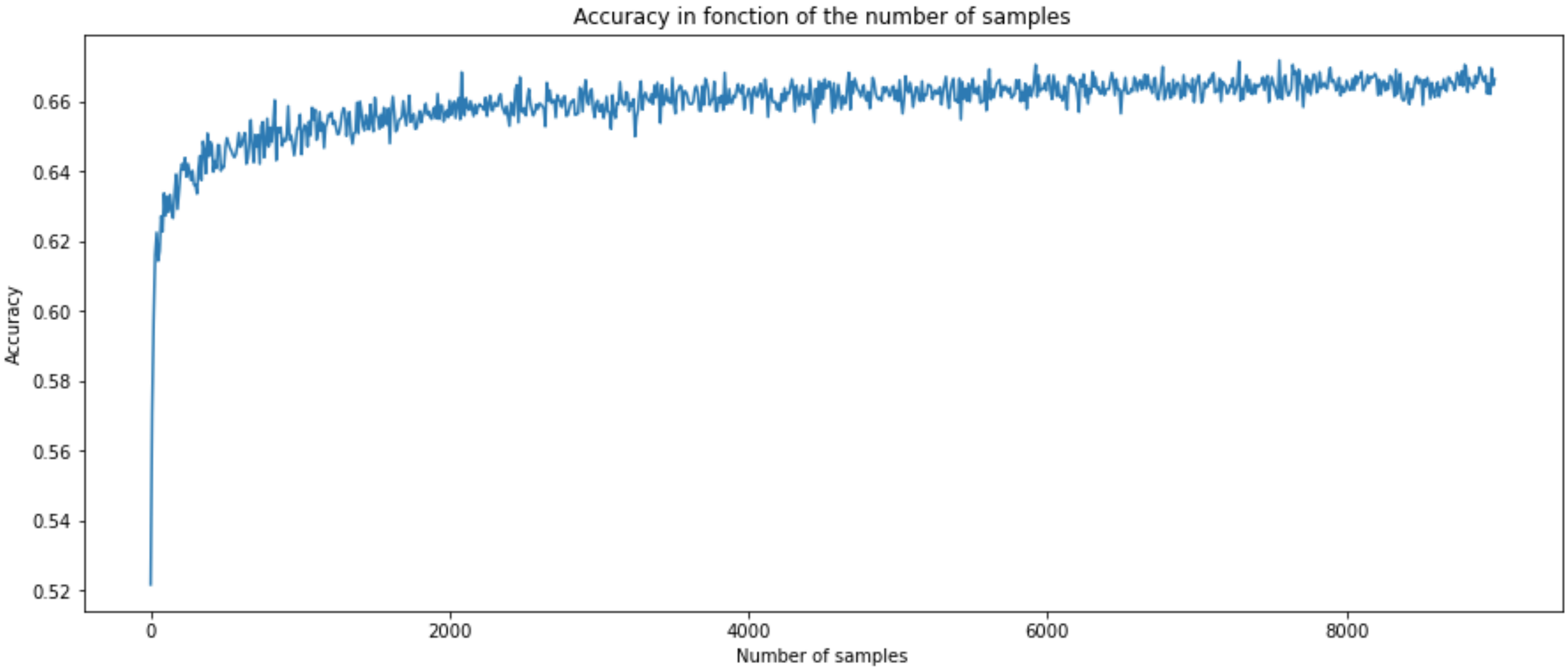
- We obtain far better result than before.
- We obtain at least a score of 0,6 with SGDClassifier and 0,65 with a Random Forest Classifier.

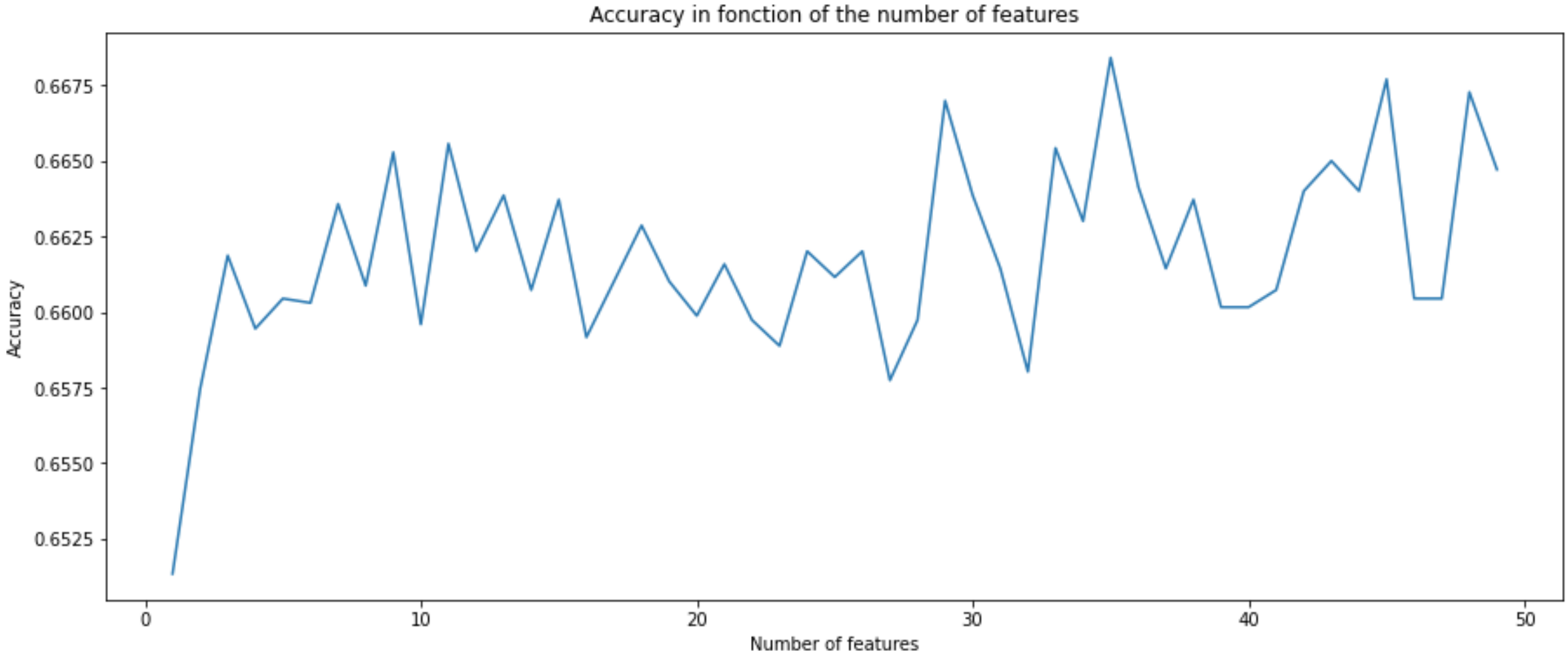
That’s why we decided to keep the Random Forest Classifier.

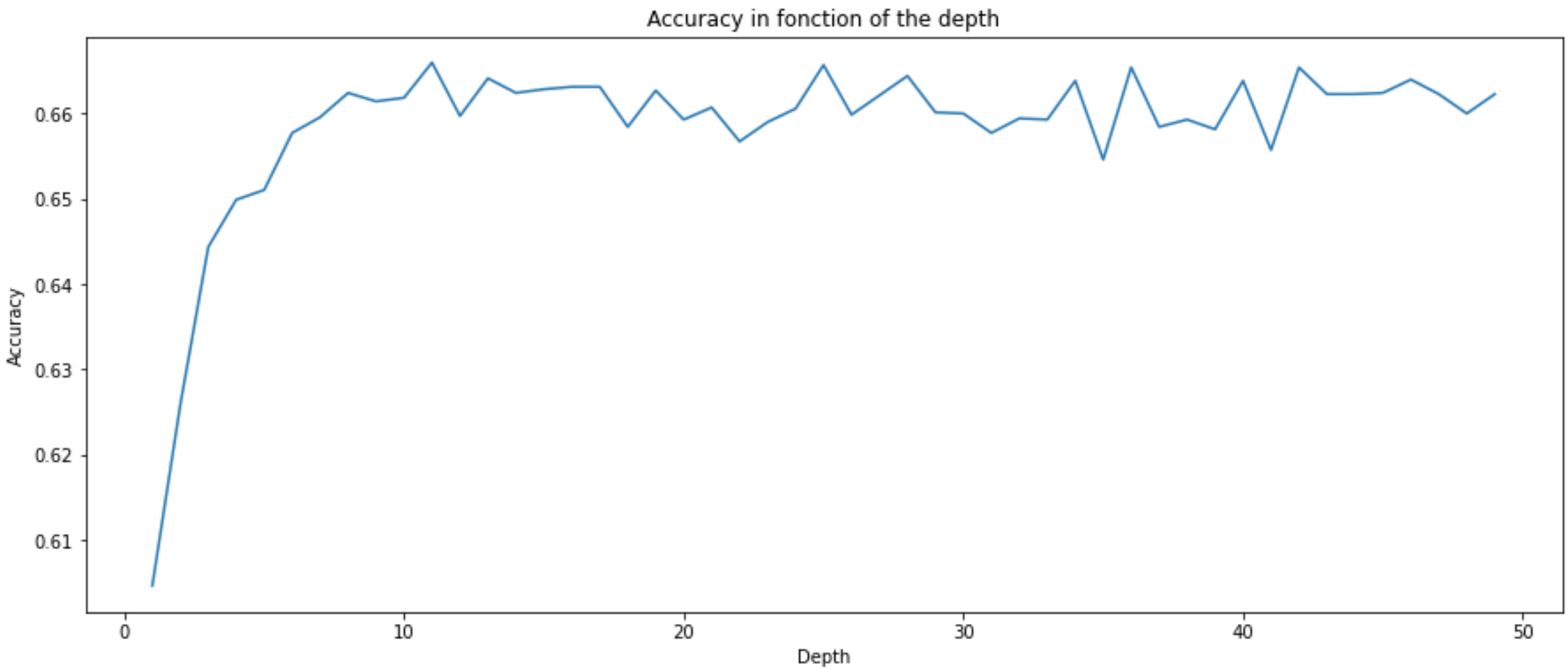
Classification – Hyperparameter Tuning

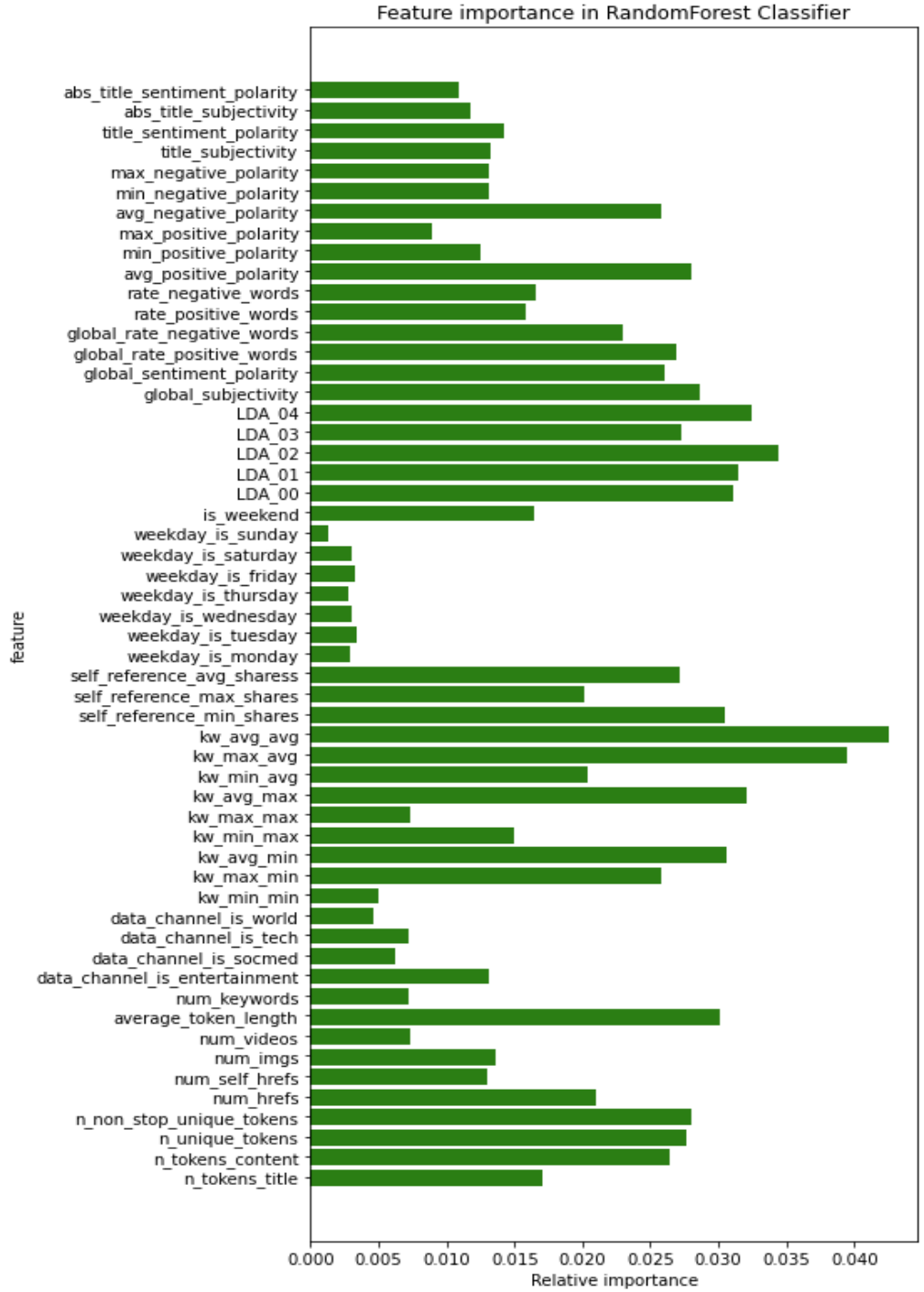
For the tuning of the hyperparameters we used :

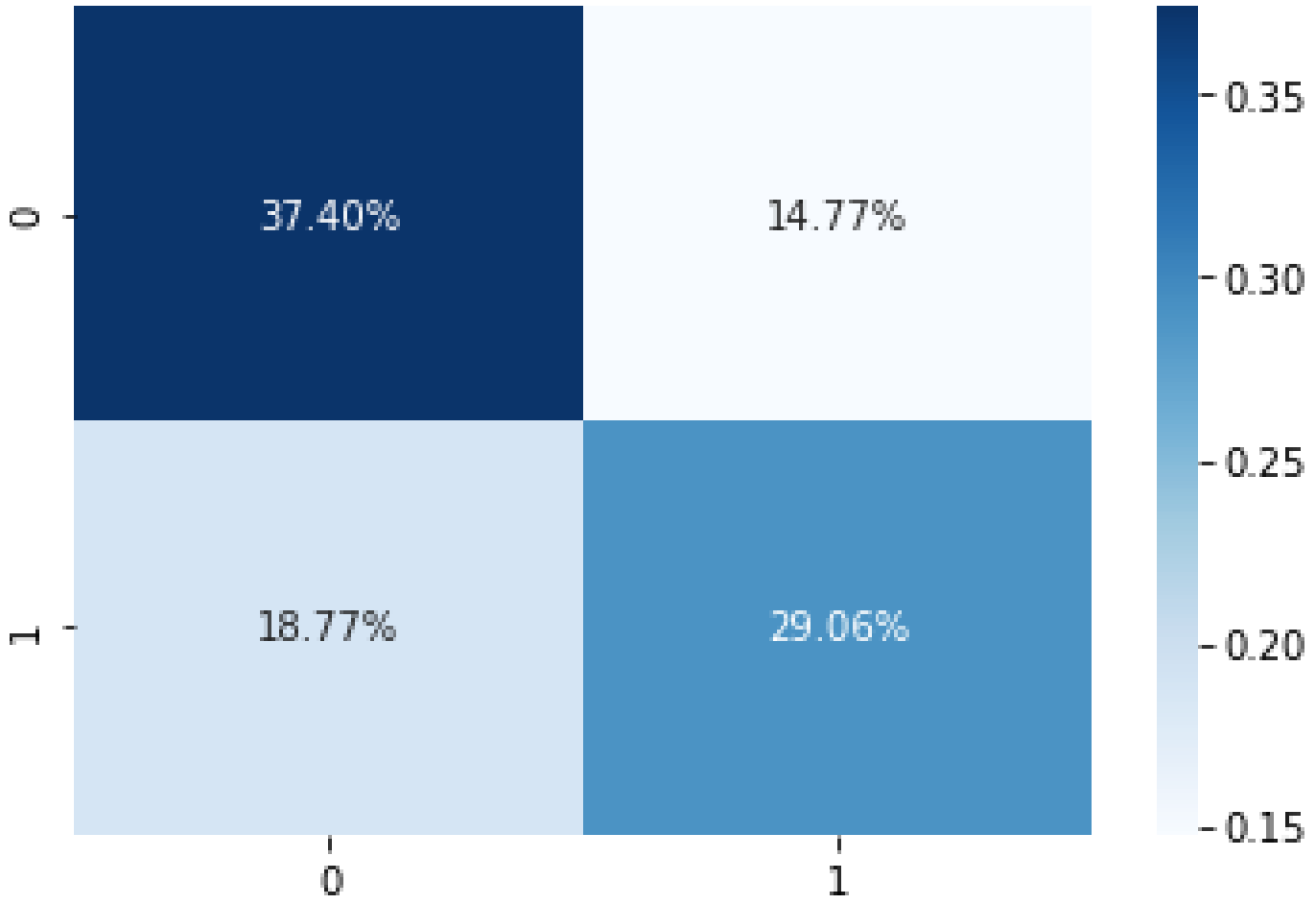
- Grid Search CV .
- We draw 3 learning curve.


Accuracy in fonction of the number of samples
Accuracy
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0
2000
4000
6000
8000
Number of samples

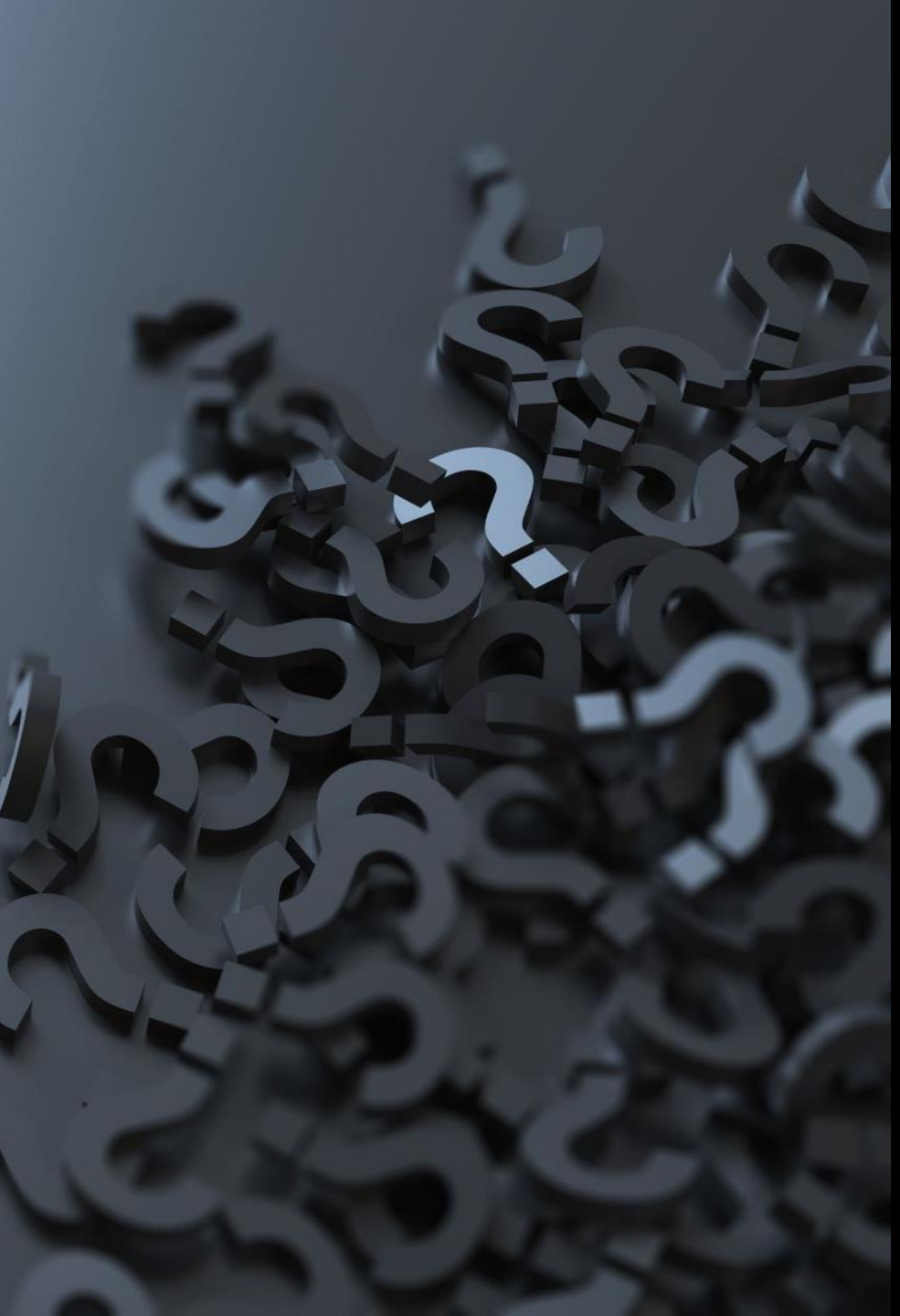

Accuracy in fonction of the number of features
Accuracy
0.6675
0.6650
0.6625
0.6600
0.6575
0.6550
0.6525
0
10
20
30
40
50
Number of features


Accuracy in fonction of the depth
Accuracy
0.66
0.65
0.64
0.63
0.62
0.61
0
10
20
30
40
50
Depth


Feature importance in RandomForest Classifier
abs_title_sentiment_polarity
abs_title_subjectivity
title_sentiment_polarity
title_subjectivity
max_negative_polarity
min_negative_polarity
avg_negative_polarity
max_positive_polarity
min_positive_polarity
avg_positive_polarity
rate_negative_words
rate_positive_words
global_rate_negative_words
global_rate_positive_words
global_sentiment_polarity
global_subjectivity
LDA_04
LDA_03
LDA_02
LDA_01
LDA_00
is_weekend
weekday_is_sunday
weekday_is_saturday
weekday_is_friday
weekday_is_thursday
weekday_is_wednesday
weekday_is_tuesday
weekday_is_monday
self_reference_avg_sharess
self_reference_max_shares
self_reference_min_shares
kw_avg_avg
kw_max_avg
kw_min_avg
kw_avg_max
kw_max_max
kw_min_max
kw_avg_min
kw_max_min
kw_min_min
data_channel_is_world
data_channel_is_tech
data_channel_is_socmed
data_channel_is_entertainment
num_keywords
average_token_length
num_videos
num_imgs
num_self_hrefs
num_hrefs
n_non_stop_unique_tokens
n_unique_tokens
n_tokens_content
n_tokens_title
feature
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Relative importance



Confusion Matrix



Conclusion

- At the beginning the objective was to predict the number of shares but it appears that despite the number of data, the predictive variables defined do not allow us to solve the problem of regression.
- By defining the popularity of an article, we can redefine our problem to a classification problem that we can answer with our data.