# MACHINE LEARNING

# LABORATORY SESSION 2: ML WITH PYTHON: SKLEARN

Pablo Rosales Rodríguez

Student Number: 213420

# INTRODUCTION

This assignment consists in solving a classification problem using Scikit-learn. The data set is already split into the training and testing subsets. The task is choosing a dataset, training a classifier and predicting the labels on the test set. Then, the algorithm must be tested using k-fold cross validation.

The problem is classifying the US presidential campaign tweets posted by the two candidates, Donald Trump and Hillary Clinton.

We first import the provided data and read them as pandas objects, so, later we can concatenate the training array with the testing one, obtaining a six columns structure as the following one.

| Tweet | Date | Retweets | Likes | Location | Author |
|---|---|---|---|---|---|

The classification performed in this assignment is based on a support vector machine. Support vector machines are linear classifiers that maximize the separation margin between classes.

In the figure [1], there is a plot with the retweets of each candidate. Donald Trump, represented in red, and Hillary Clinton, represented in blue.

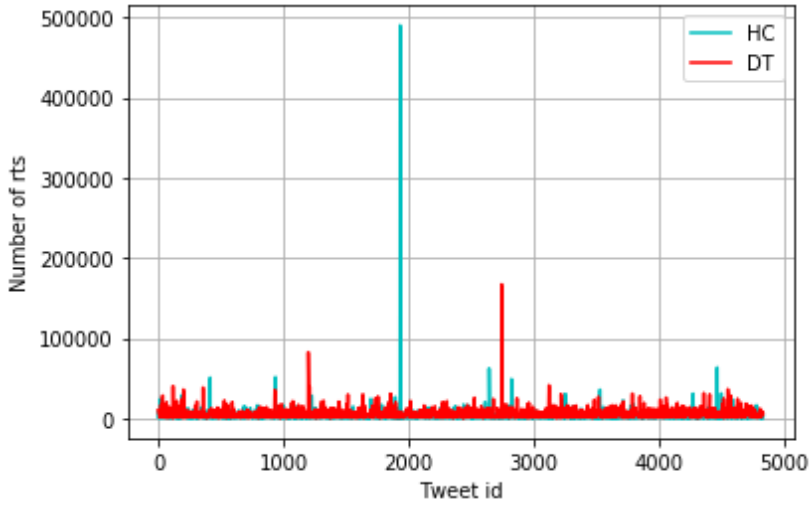


*Figure [1]. Retweets for each tweet posted.*

Donald Trump most retweeted tweet was the one with id 2743 (*"how long did it you're your staff of 823 people…"*) reaching 167274 retweets. Clinton's most popular tweet was retweeted a total number of 490180 times. Its correspondent index is 1932 (*"delete your account…"*).

# CLASSIFICATION

As it was mentioned before, the classifier used for this assignment is a support vector machine linear function.

First, the features used as inputs must be extracted. For it, a matrix was created with the following characteristics of the tweets:

| $\log(likes)$ | $\log(retweets)$ | Length of the words | Count of 'a' character | Count of tabs (spaces) | Count of dots | Month | Hour | Retweets |
|---|---|---|---|---|---|---|---|---|

This initial classification stands for the reference that must be optimized in the future one. After applying the classifier with a gamma parameter of 0.001 and a C=10, the accuracy obtained is 0.923. In the following tries, this accuracy should be higher.

## EVALUATING THE ALGORITHM USING K-FOLD CROSS VALIDATION

The main steps to be followed in order to perform the evaluation of the chosen algorithm with K-fold cross validation are:

- Split the dataset into *k* equal-sized subsets
- For each subset
  - Train the Predictor
  - Compute the score of the parameter (in the given case, accuracy, which is the one to be optimized)
- Return the average score computed over all the subsets

With the previously explained methodology, every subset is used for training, as well as for computing the score (testing).

In this assignment, the evaluation method applied is 3-Fold cross validation and the target variable is gamma. The optimized gamma obtained from the algorithm is 0.01. This value provides an accuracy of 0.95, much higher than the previous one.

The learning curves represented in the figure [2] demonstrate that the training score and the one from the cross validation are extremely close.

Another classifier was also used in order to determine the best C. A value of 10 was obtained, which is the one fixed by default.

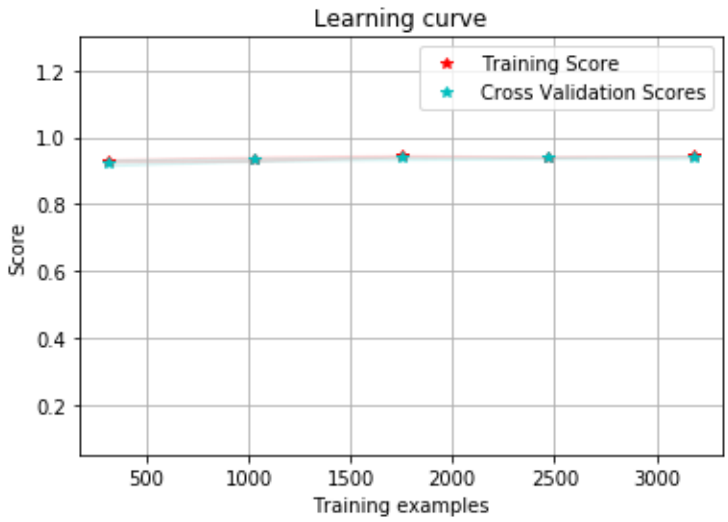


*Figure [2]. Learning Curves.*

## RESULTS

The final results are shown in the following table.

| | **Precision** | **Recall** | ***F1*** |
|---|---|---|---|
| **Donald Trum** | 0.94 | 0.96 | 0.95 |
| **Hillary Clinton** | 0.96 | 0.94 | 0.95 |

*Table [1]. Results for each candidate.*

In the table [2], the average results are shown.

| **Average Precision** | **Average Recall** | **Average *F1*** |
|---|---|---|
| **0.968** | 0.917 | 0.942 |

*Table [2]. Average Results.*

Finally, the average accuracy, the parameter of interest is 0.943.

In the figure [3], the whole results matrix, including the weighted averages, is shown as a screenshot.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| DT | 0.94 | 0.96 | 0.95 | 797 |
| HC | 0.96 | 0.94 | 0.95 | 799 |
| accuracy | | | 0.95 | 1596 |
| macro avg | 0.95 | 0.95 | 0.95 | 1596 |
| weighted avg | 0.95 | 0.95 | 0.95 | 1596 |

*Figure [3]. Results matrix.*

These results posted in the previous tables prove that the performance has improved.

## INTERESTING GRAPHS

Some other interesting plots are shown below.

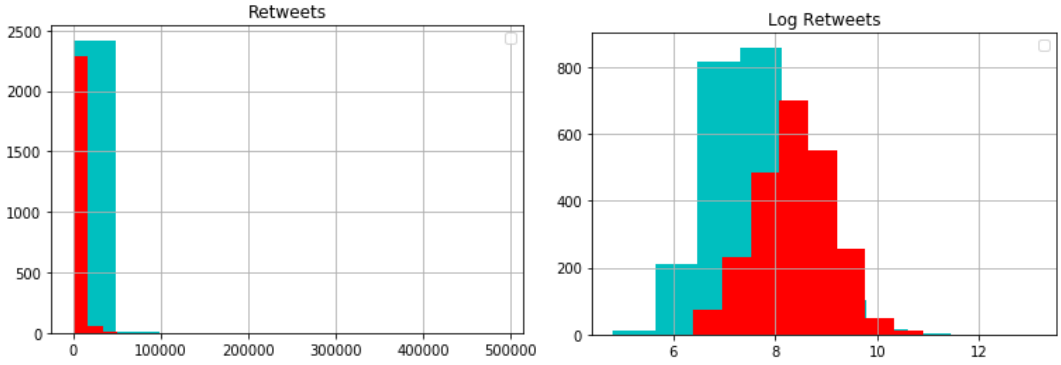


*Figure [4] and Figure [5]. Retweets for each candidate and their logarithmic function (Red for DT and Blue for HC).*

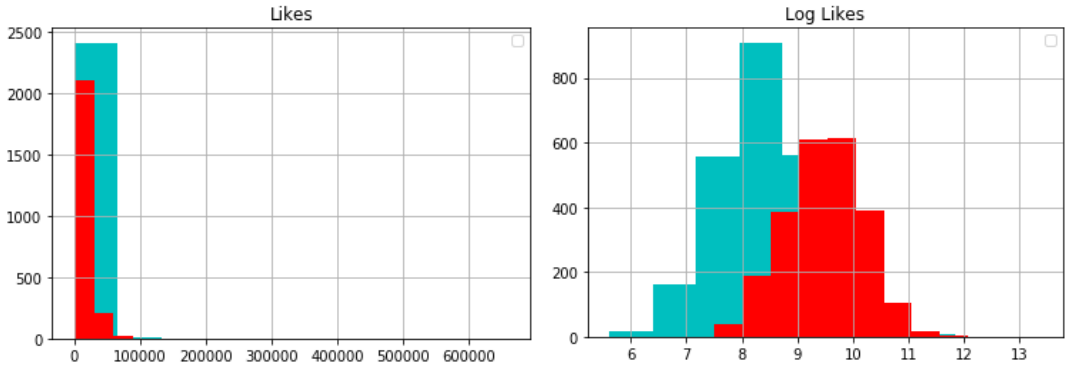


*Figure [6] and Figure [7]. Likes for each candidate and their logarithmic function (Red for DT and Blue for HC).*

## CONCLUSION

The chosen classifier, based on SVM theory provided a good performance, improving the parameters observed.