Spam Detector

Data Mining & Machine Learning

Artificial Intelligence and Data Engineering - Project

Francesco Campilongo

2022

Goal

Given two different datasets, E-Mail and SMS. Train five different classifiers to find the undesired SMSs and E-Mails.

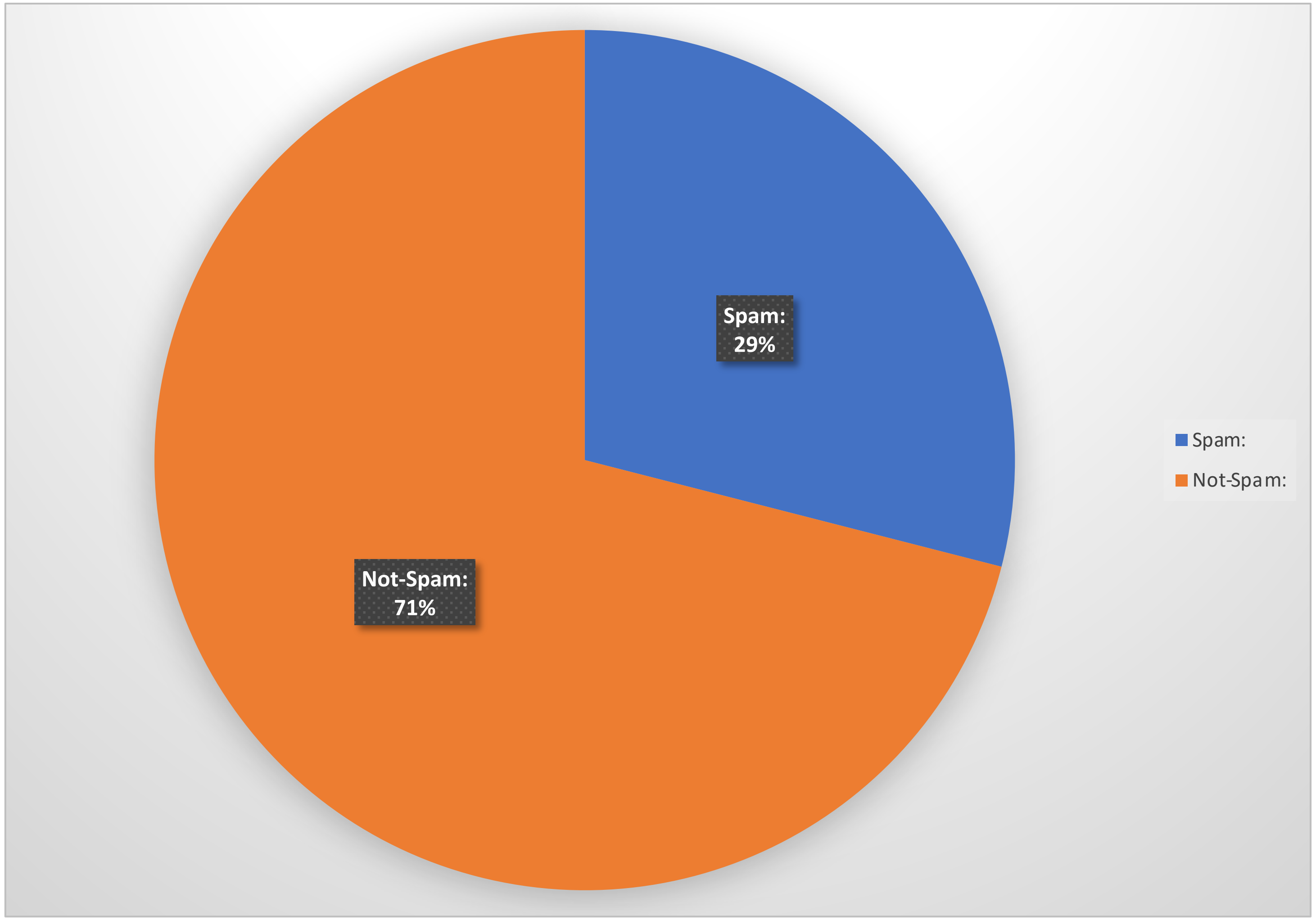
Classifiers chosen:

- Support Vector Machine
- Naïve Bayes
- K Nearest Neighbour
- Random Forest
- AdaBoost

And find the better one.

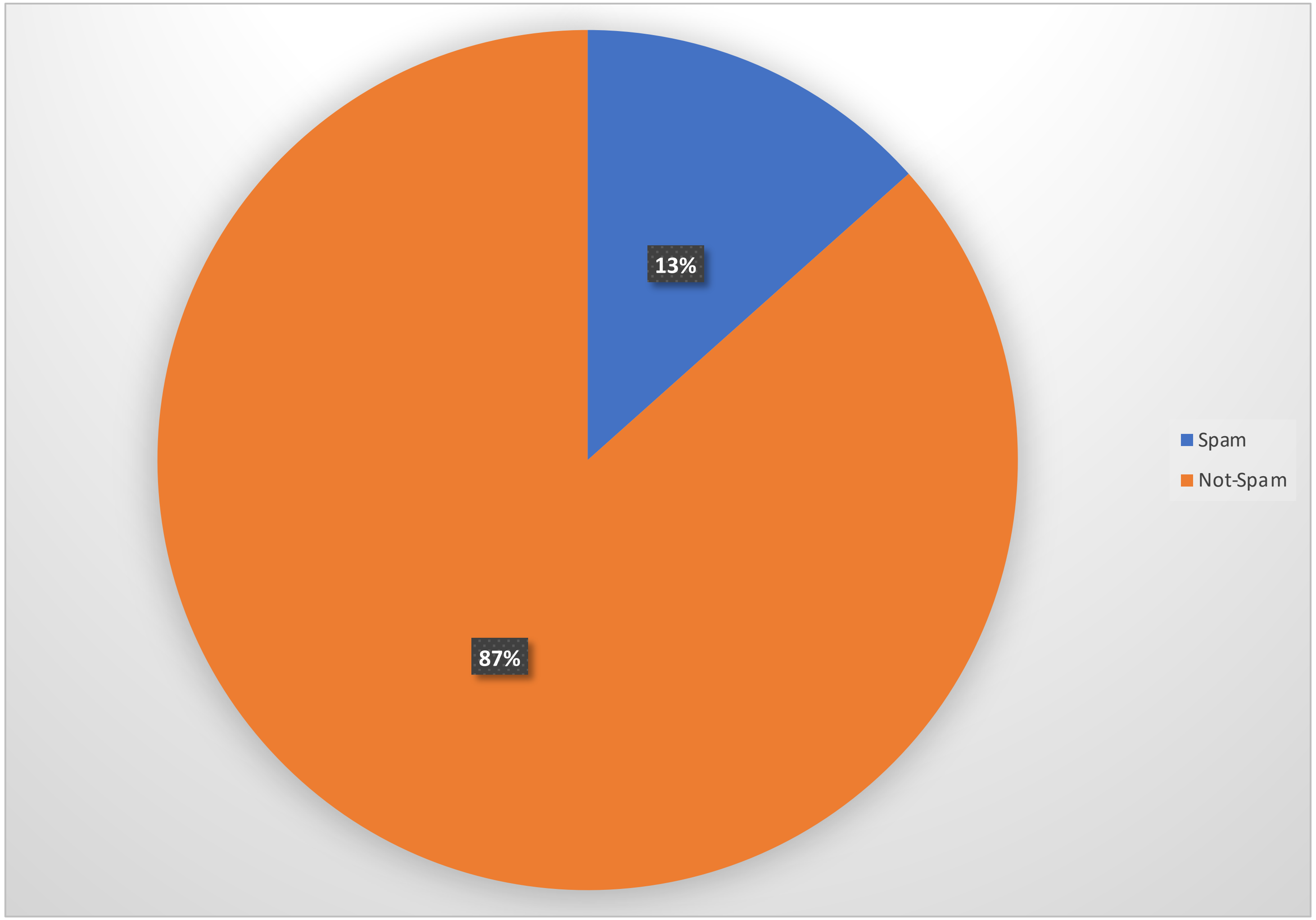
E-Mail Dataset

- The E-Mail dataset found was already pre-labelled and distributed on the platform Kaggle.com.
- The tuples found were 5171, divided as shown into the following diagram.



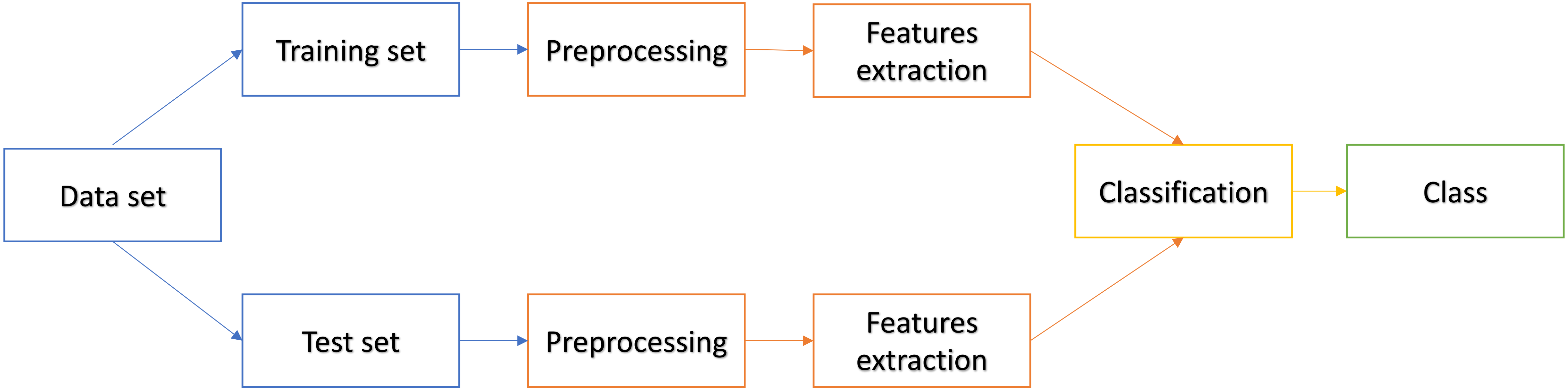
SMS Dataset

- The SMS dataset found was already pre-labelled and also in this case, distributed on the Kaggle.com platform.
- The tuples found were 5572, divided as shown into the following diagram.



Classification Steps

- The dataset split was carried out directly from the KFold function from SK learn.
- The Test set was used to understand the goodness of the model used.
- The classification step was made with five different classifier as previously mentioned.



Preprocessing

- Removing some noisy column (both on E-Mail and SMS datasets)
- Renaming the remaining column
- Transform all the text in lower case
- Remove stopwords
- Stemming

Features Extraction

- *CountVectorizer*: tokenization, stopword filtering and relevant tokens identification
- *TfidfTransformer*: TF and idf calculation
- *fit_transform*: to get the actual features, from a count matrix to a tf-idf representation

TF-IDF, calculated as: $\mathbf{w_{q,j} = Tf_{q,l} \bullet w_q}$ with $\mathbf{Tf_{q,l}}$ as the frequency of the word s_q in the text j and $\mathbf{w_q = \ln (N_{tr}/N_q)}$ where $\mathbf{N_{tr}}$ is the number of labelled texts and $\mathbf{N_q}$ is the number of texts containing stem s_q.

Classifiers Evaluation - E-mail

In order to find the best classifier, a KFold cross validation was carries out, with k=5.

All the metrics value in the following table are an average per iterations.

Classifier	Accuracy	Precision	Recall	F1 Score
SVM	0,98822	0,97104	0,98866	0,97978
MNB	0,87586	0,99886	0,57288	0,72794
KNN	0,96248	0,97524	0,89338	0,93252
RF	0,94522	0,96072	0,84576	0,89946
Adaboost	0,95978	0,93246	0,92854	0,93036

Classifiers Evaluation - SMS

In order to find the best classifier, a KFold cross validation was carries out, with k=5.

All the metrics value in the following table are an average per iterations.

Classifier	Accuracy	Precision	Recall	F1 Score
SVM	0,98634	0,98714	0,91284	0,94776
MNB	0,96106	1	0,71198	0,8305
KNN	0,95242	0,99792	0,6479	0,78418
RF	0,97346	0,98818	0,81452	0,8919
Adaboost	0,9745	0,95884	0,84962	0,89998

Conclusion

- From this analysis the best classifier for this kind of dataset is the **Support Vector Machine**.
- With all the metrics calculated the SVM classifier is as good to find the True positive as to the True Negative, compered to the other algorithm used.
- It is also consistent through the iterations.
- The SVM is also the classifier with significantly higher F1 Score.