Email Spam Classification

Fundamentals of Data Science

A.Y. 2022/2023

Prof. Galasso

Group Members:

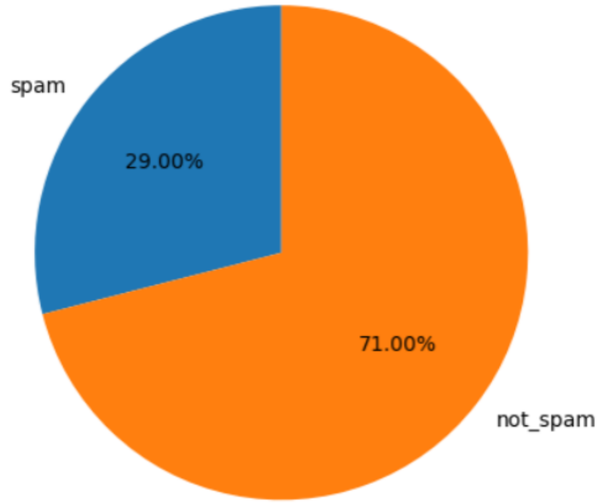
- Marina Iuliana Aur, 1809715
- Sophia Balestrucci, 1713638
- Viktoriia Vlasenko, 2088928
- Michele Musacchio, 2070948
- Ilaria Gagliardi, 1796812

Analyze our dataset

4]:

```
emails.info()
```

```
RangeIndex: 5172 entries, 0 to 5171
Columns: 3001 entries, the to Prediction
dtypes: int64(3001)
memory usage: 118.4 MB
```




```
# Check data completeness
print(emails.isna().sum())
print(f'total NULL sum: {sum(emails.isna().sum())}')
```

```
the           0
to            0
ect           0
and           0
for           0
             ..
military      0
allowing      0
ff            0
dry           0
Prediction    0
Length: 3001, dtype: int64
total NULL sum: 0
```

Prepocessing the fields

1. Harmonise letter case - parse words into lowercase format
2. Remove numbers, symbols and non alphabetic characters
3. Are words unique?
4. Remove the stopwords (NLTK and spiCy libraries)
5. Remove all words of lenght 1
6. Lemmatization (spiCy library)
7. Standardization (scikit learn library)

Final dataset: from 3000 to 2261

```
In [32]: n_features_before_lemmatization = len(nlp(" ".join(emails_raw.columns[:-1])))
         print(f"Number of features before lemmatization = {n_features_before_lemmatization}")

Number of features before lemmatization = 2774
```

```
emails_raw
```

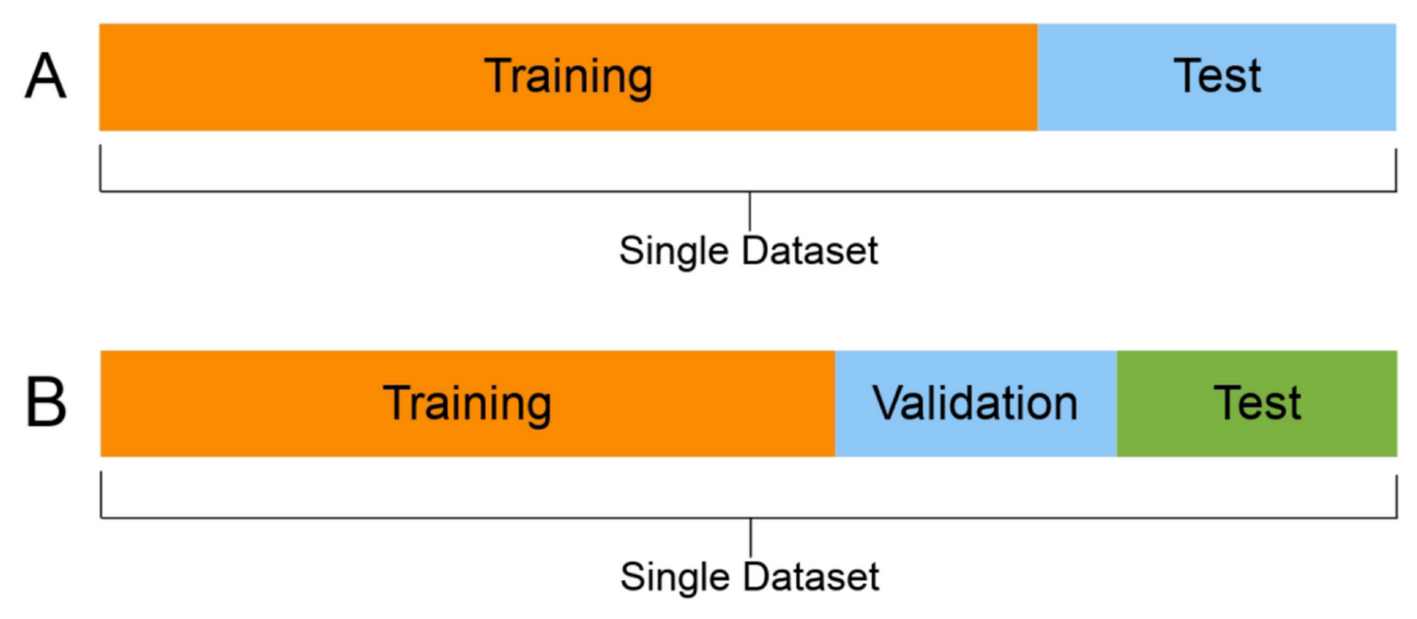
	abdv	ability	able	accept	acceptance	access	accord	account	accountant	accounting	...	yesterday	yet	york	young	yvette	zajac	zero	zivley	z
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	1	0	1	...	1	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5167	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
5168	0	1	4	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	
5169	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
5170	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
5171	0	1	3	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	

5172 rows × 2261 columns

```
emails_raw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5172 entries, 0 to 5171
Columns: 2261 entries, abdv to zonedubai
dtypes: int64(2261)
memory usage: 89.2 MB
```

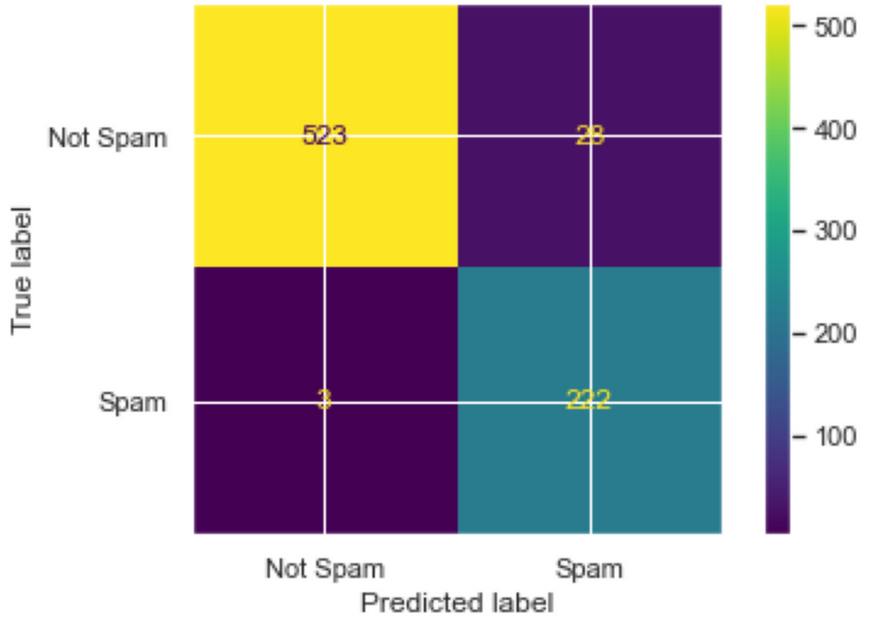
Dataset Splitting

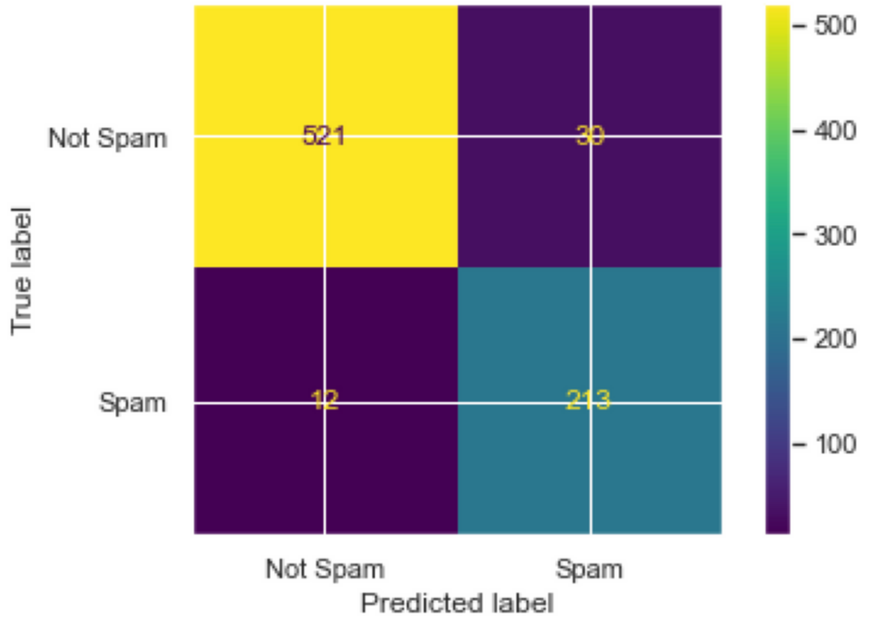


- Train: 72%
- Validation: 12%
- Test: 15%

Model Building: Implemented Classifiers

- Logistic Regression
  - First order boarders
- Naive Bayes
  - Laplace smoothing
  - Logarithm for numerical stability
- Predict and predict probability functions





Model Building:
Existing Classifiers

- Classifiers provied by Scikit-learn: AdaBoost, ExtraTrees, KNeighbors
- Improvement of the performance (accuracy) of each classifier through

  GridSearchCV

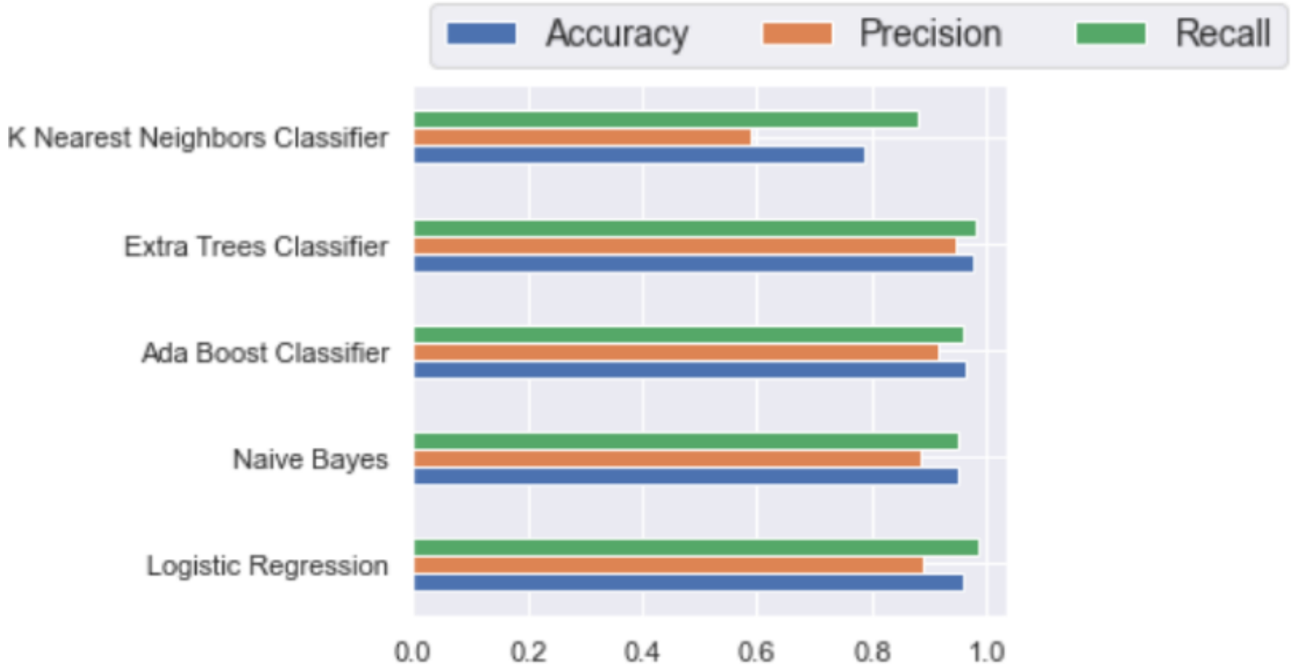
```
Best accuracy score for AdaBoostClassifier: 0.955835 using {'n_estimators': 75}
Best accuracy score for ExtraTreesClassifier: 0.973236 using {'n_estimators': 175}
Best accuracy score for KNeighborsClassifier: 0.884891 using {'n_neighbors': 1}
```

- Training/Test

Performance Evaluation

- Accuracy, Precision, Recall

	Accuracy	Precision	Recall
Logistic Regression	0.960052	0.888000	0.986667
Naive Bayes	0.948454	0.882231	0.948889
Ada Boost Classifier	0.962629	0.915254	0.960000
Extra Trees Classifier	0.978093	0.944444	0.982222
K Nearest Neighbors Classifier	0.788660	0.591045	0.880000



- Computation Time

	time (seconds)
Logistic Regression	189.275045
Naive Bayes	79.577362
Ada Boost Classifier	17.836216
Extra Trees Classifier	14.323022
K Nearest Neighbors Classifier	0.441461

- FAR (“Type II Error”)

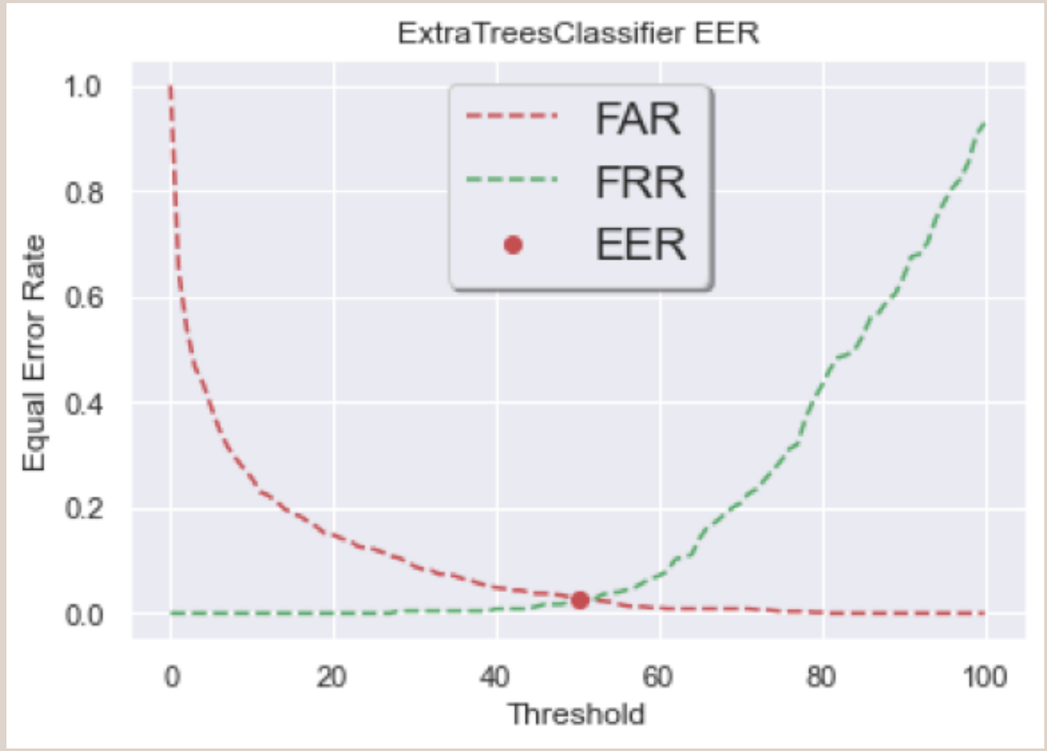
$$FAR = \frac{FP}{FP + TN}$$

- FRR (“Type I Error”)

$$FRR = \frac{FN}{FN + TP}$$

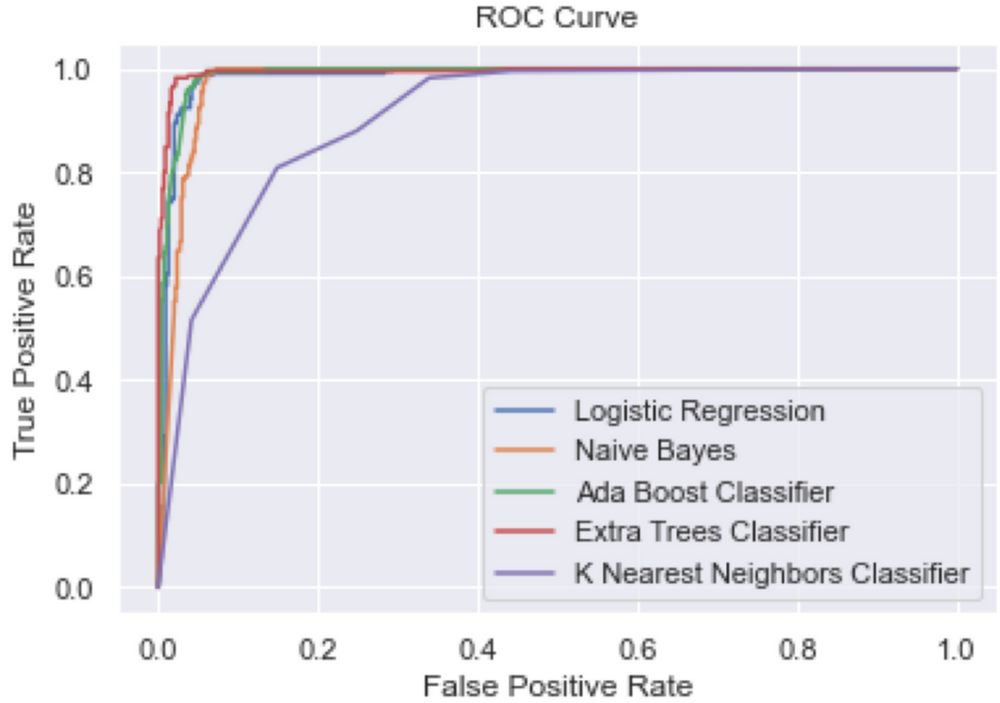
- EER (Crossover Error Rate):

where the FAR and FRR are equal.



	EER%
Logistic Regression	3.86
Naive Bayes	5.39
Ada Boost Classifier	3.81
Extra Trees Classifier	2.20
K Nearest Neighbors Classifier	17.00

- ROC curve
  - True Positive Rate vs. False Positive Rate on different thresholds
- Cross Validation
  - Choose the best model on validation dataset
  - Score evaluate as accuracy



	accuracy
Logistic Regression	0.528788
Naive Bayes	0.578788
Ada Boost Classifier	0.962121
Extra Trees Classifier	0.983333
K Nearest Neighbors Classifier	0.837879

ExtraTrees is the winner

Thank You