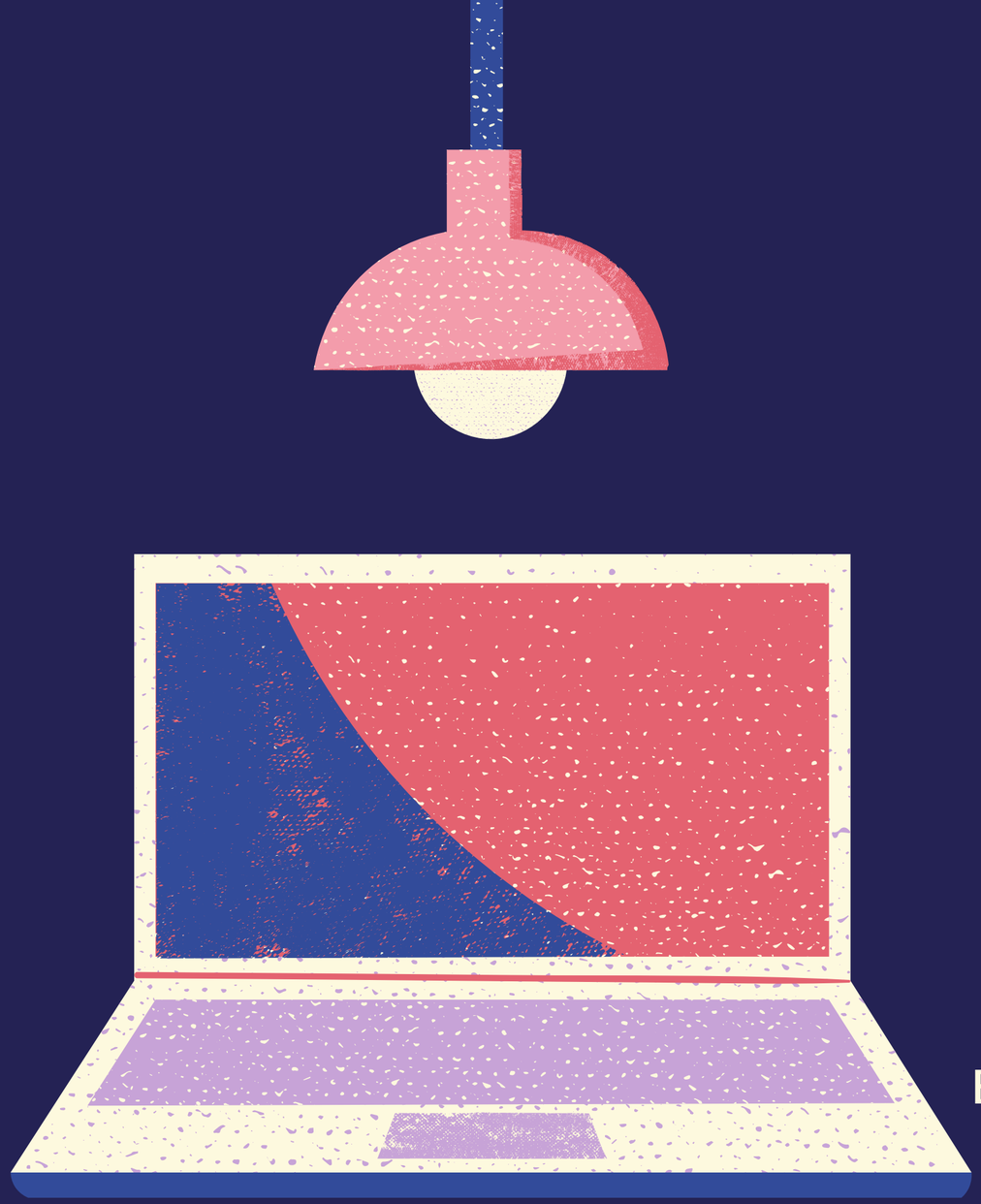


Text clarification 101

By team 10

Preprocessing and Data Cleaning

We worked on a 5 fictions books with different authors

- we have 1000 rows each row has 100 words
- we used label encoder "on Author"
- we cleaned the data using regex
- we removed the stop words too

Feature Engineering

"Feature selection" for the most 20% important features (SelectPercentile)

Before features reduction

BOW : (1000, 12202)
N-Gram : (1000, 80676)
TFiDF : (1000, 12202)

After features reduction

BOW : (1000, 2441)
N-Gram : (1000, 16135)
TFiDF : (1000, 2441)

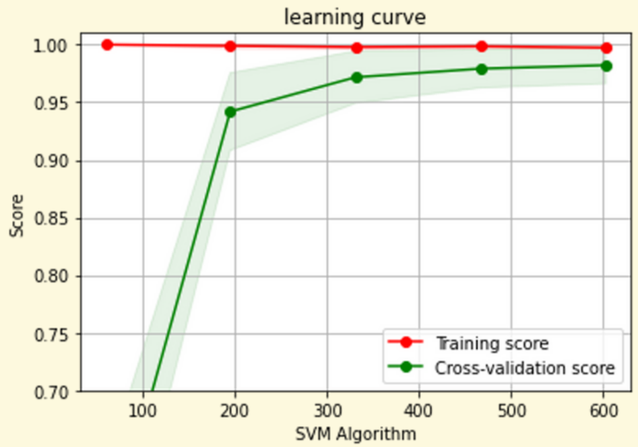
Modeling

- SVM
- D-Tree
- KNN

- Train 3 algorithms with 3 transformer
- calculate classification report for each model
- evaluate each model "cv ,mean accuracy and std "

print("%0.2f accuracy with a standard deviation of %0.2f" % (scores.mean(), scores.std()))

```
cross validation :[0.98507463 1.         1.         0.98507463 0.98507463 0.98507463
 0.95522388 0.97014925 0.98507463 0.98507463]
0.98 accuracy with a standard deviation of 0.01
```



Error Analysis and champion model

Display bais and variance for All models
to choose the champion model.

The champion model (SVM With TFiDF)

```
TFiDF with SVM
f1-score: 0.9818181818181818
```

We reduce the accuracy down for about 24% by reducing the Feature

```
TFiDF with SVM
f1-score: 0.7545454545454545
```