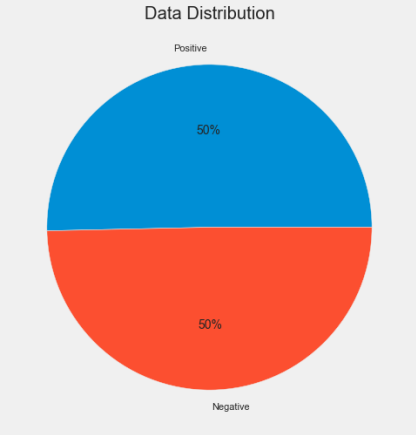
IMDB Reviews
NLP Classification project

Abstract:
Our goal for this NLP Classification Model is to predict the IMDB reviews if it was positive or negative as our dependent is the actual review and as our independent value is the sentiment which is positive or negative. This model will help us identify which review is good or bad even if the rating is high

Data:
Our IMDB Reviews data form Kaggle website, and we will have around 100,000 data points and four features the most important one were:
- sentiment: positive/negative
-the actual reviews
After the data cleaning we got 50,000 data rows pf reviews and sentiment



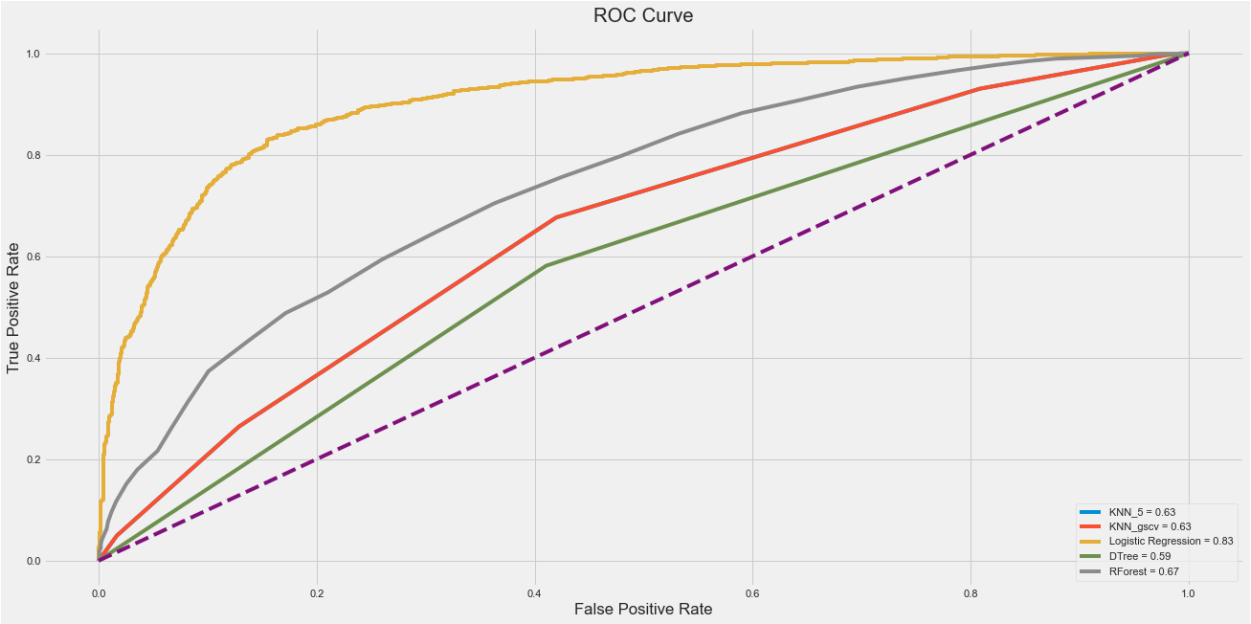
As we can see our data is balanced.

Algorithms and methodology:
For feature engineering we removed Unsupervised rows and then did feature selecting for two columns which were the actual reviews and the sentiment (Positive, negative), also we did feature reduction since we had more than 20,000 features we applied Principle component analysis PCA we reduced the features to 900, after that we initialized and fitted the data to four models KNN, Logistic regression, Decision tree, Random forest, after that we did GridSearchCV to determine the best number of neighbors and here is the result and the Logistic regression model looks the best between them:

Models Results

Model	Accuracy	Precision	Recall	F1
KNN_5	62.83	61.76	67.64	64.57
KNN_gscv	62.83	61.76	67.64	64.57
Logistic Regression	82.77	80.35	86.82	83.46
DTree	58.53	58.67	58.12	58.39
RForest	66.9	67.71	64.78	66.21

ROC Curve:



Tools:

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Sklearn
- SpaCy

Contributors:

- **Faisal Alasgah**
- **Ali Altamimi**
- **Saleh Aljomyl**