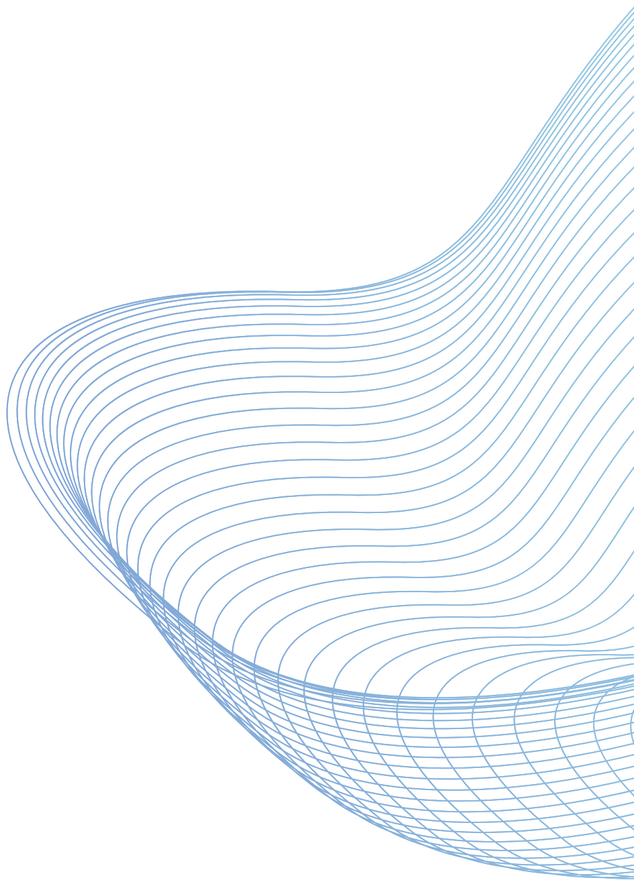
ML PROJECT PRESENTATION

FAKE NEWS DETECTION USING ML TECHNIQUES



MOTIVATION

- Fake News is a challenging problem in today's times.
- Social Media websites are flooded with much misinformation, which can prove fatal.
- Twitter particularly struggles with the fake news problem.
- However, there is a certain regular pattern in fake news. Some individuals are more likely to spread fake news.
- We can use Machine Learning to identify such patterns and try to predict fake news.

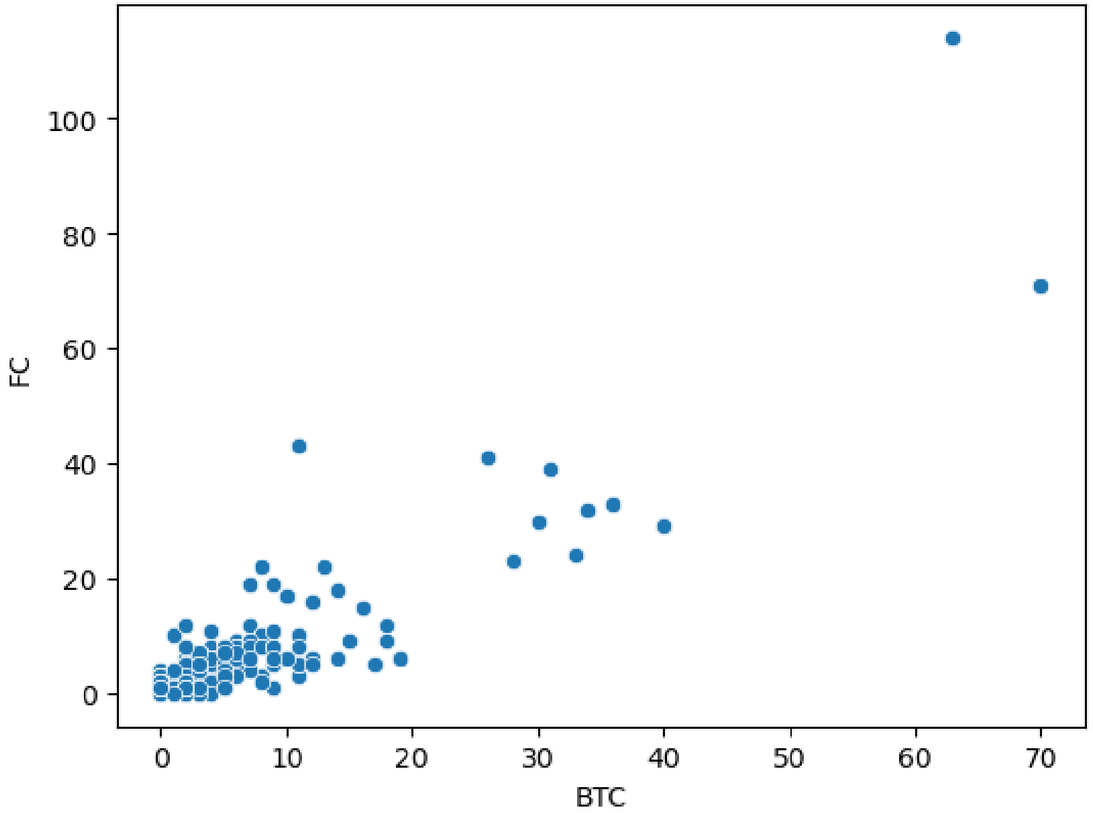
LITERATURE REVIEW

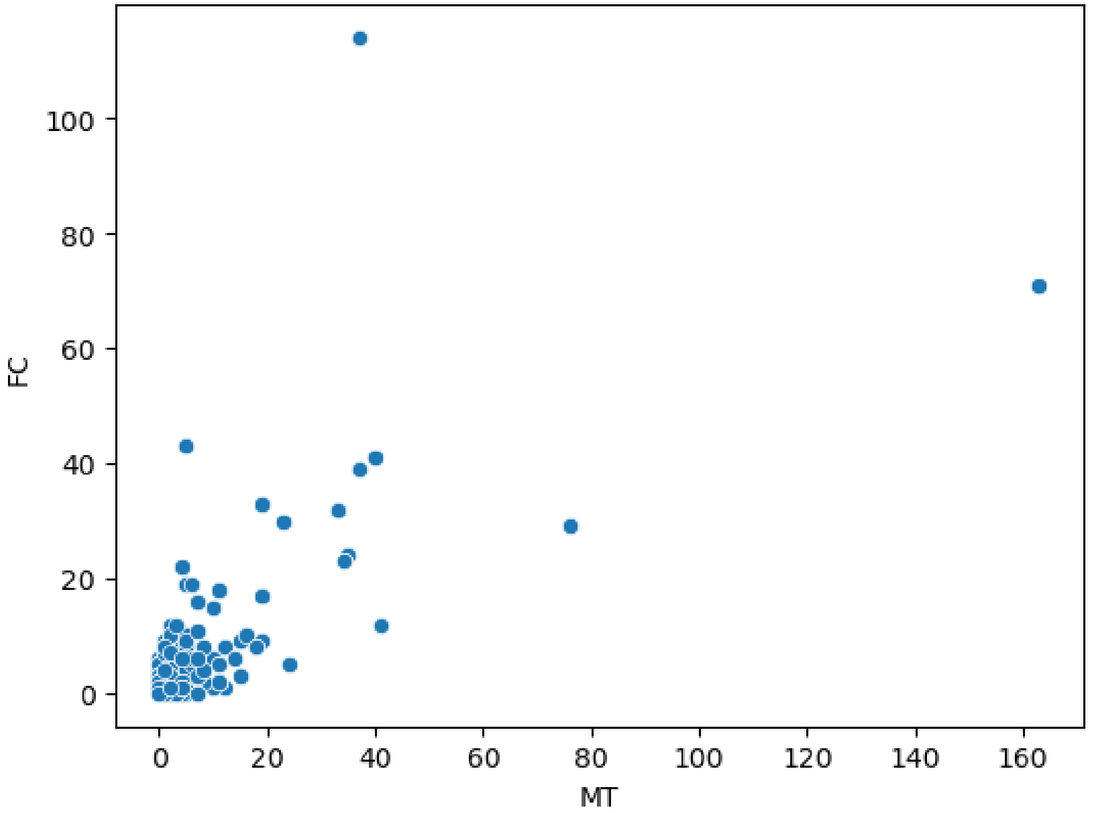
- Researchers used ML models like logistic regression and deep learning models.
- Researchers describe the use of detecting fake news.
- They discuss vectorization techniques like TF-IDF and BOW to convert text to numeric values.
- Researchers discussed the importance of addressing the bias using lexical and sentiment analysis.
- They experimented with several models like SVM, Random Forest etc.

DATASET USED

- Used Liar Dataset
- It contains sentences with their speakers and their affiliations with labels representing fake news or not.
- The dataset contains 16 columns and 12788 rows.
- Some columns are labels, statement speakers, etc.

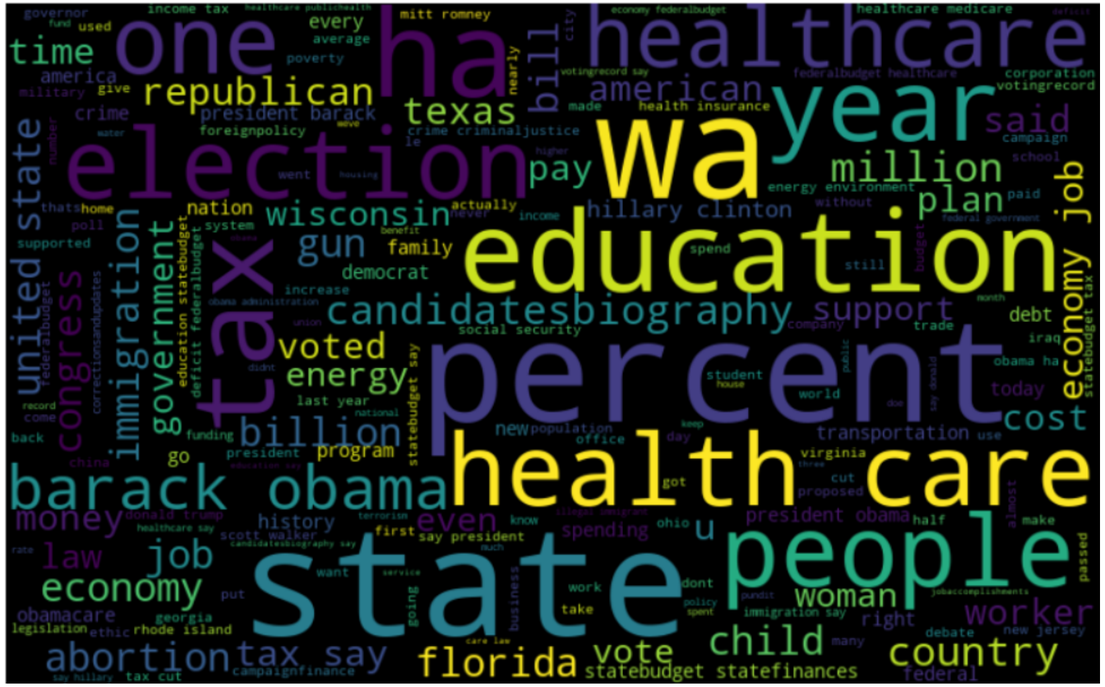
EDA SOME SCATTERPLOTS !!!





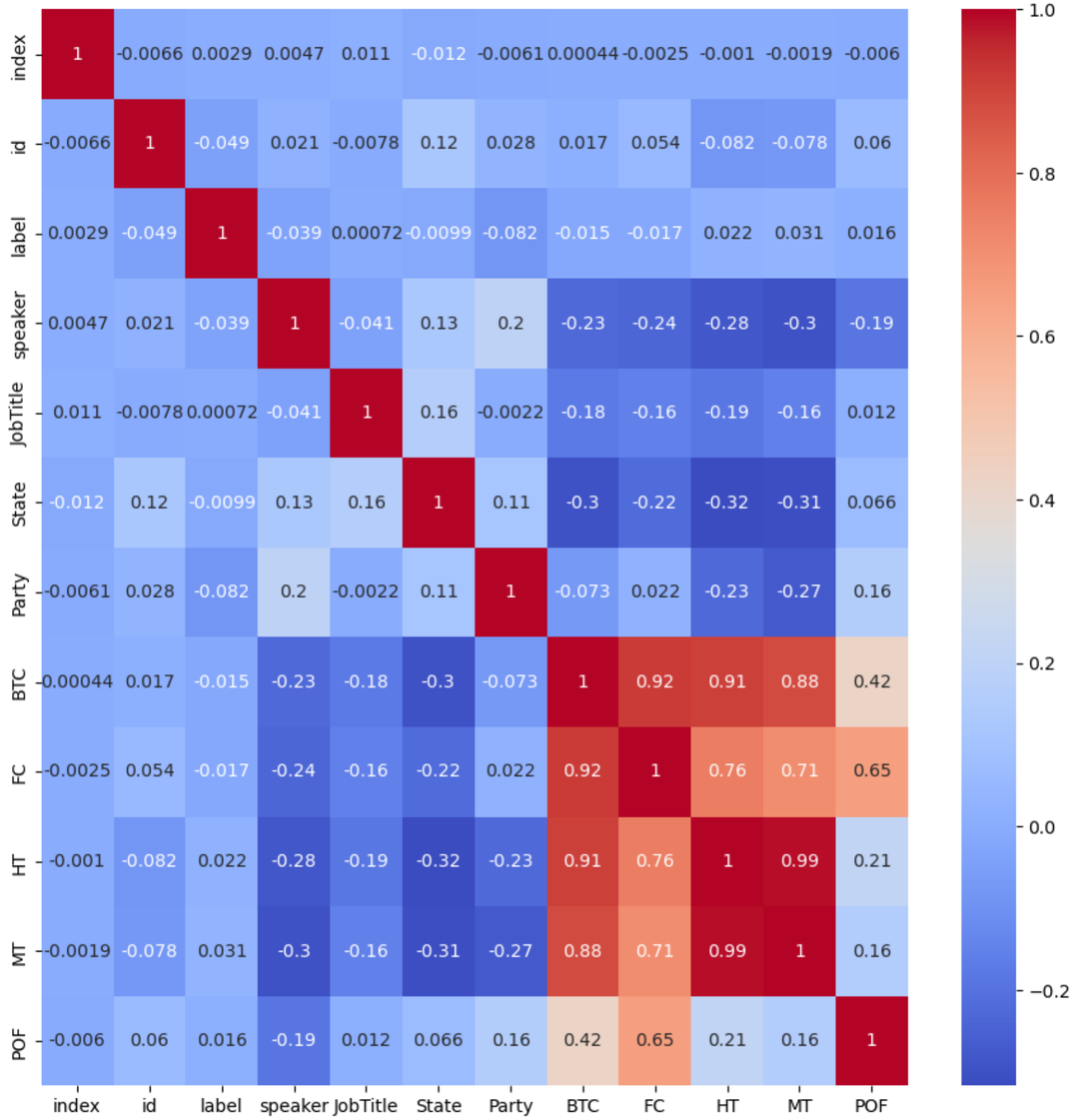
DATA PREPROCESSING

- We clean the data and remove punctuation marks, white spaces, etc
- We tokenize the data using NLP
- We use TF-IDF and BOW to vectorize the tokenized data into numeric form.
- We use wordcloud to visualize the word frequencies.
- We use a label encoder on Political Party and speaker column.



DATA PREPROCESSING - 2

- We dropped 12 columns out of 16 columns
- The decision to keep which columns were taken based on the heatmap.
- Columns with high correlation were dropped.
- We take 4 partitions of data with and without party and speaker using TF-IDF and BOW.



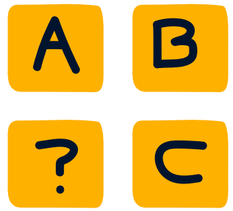
MORE ON NLP !!!

- **We use TF-IDF and BOW in NLP vectorisation.**

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

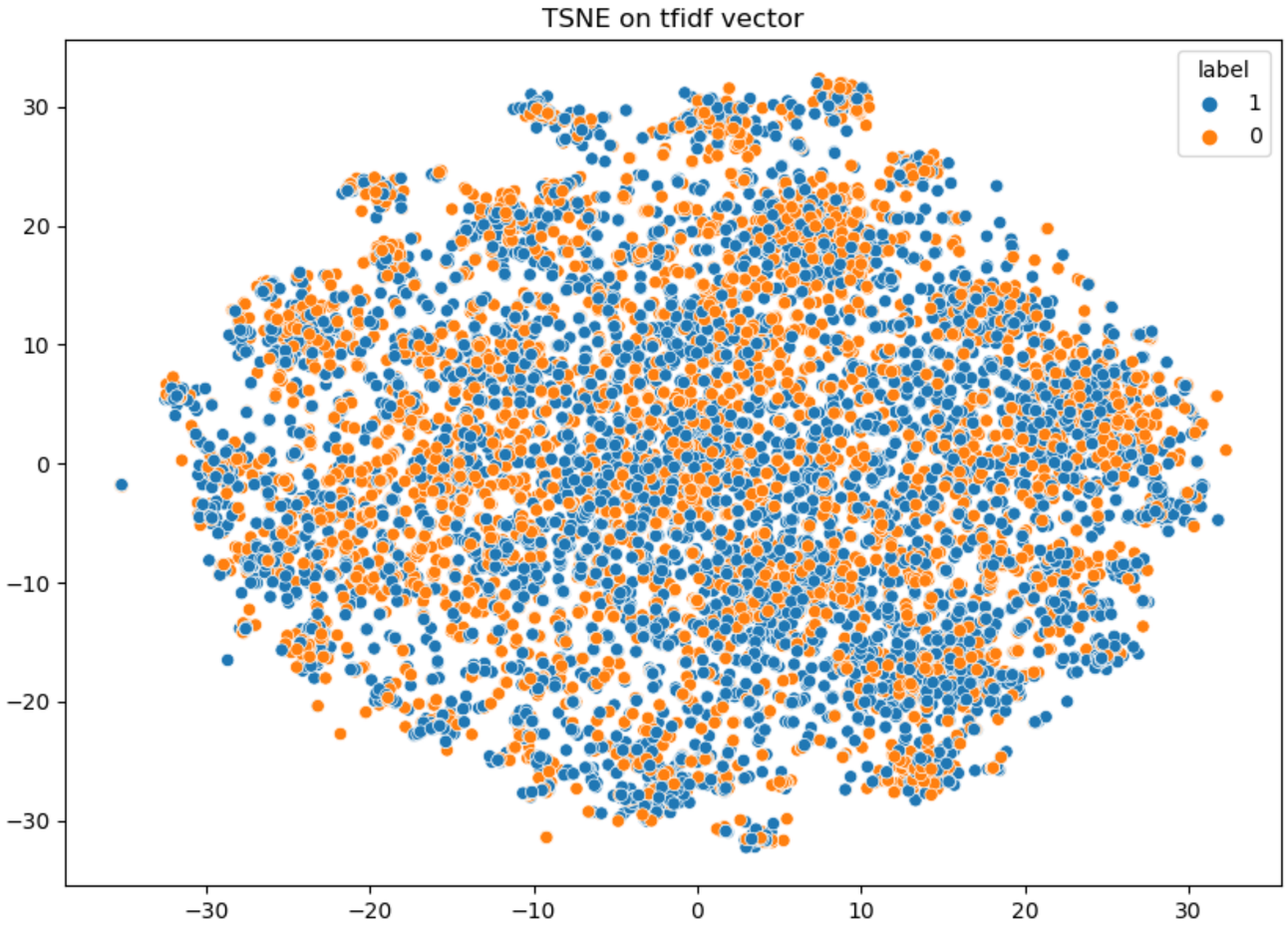
$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

$$TF\text{-}IDF = TF * IDF$$



PERFORMING TSNE !!

- **Following is the result of TSNE on TF-IDF Vector**



ML MODELS

1

We use Grid Search to find the best hyperparameters

2

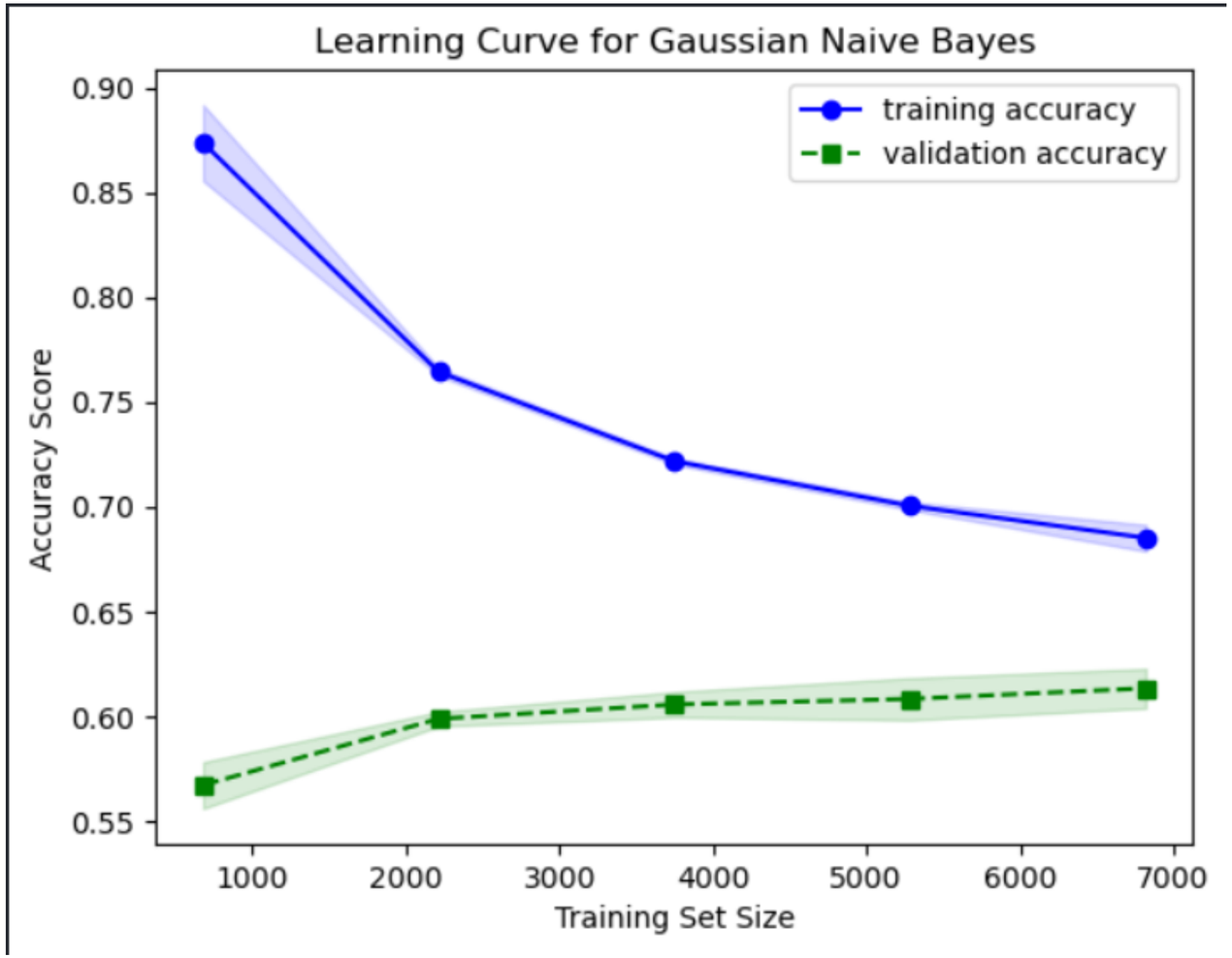
We report the accuracies and loss curves for various ML models like SVM, Neural networks etc.

3

The results are for TF-IDF in NLP without considering the party and speaker name.

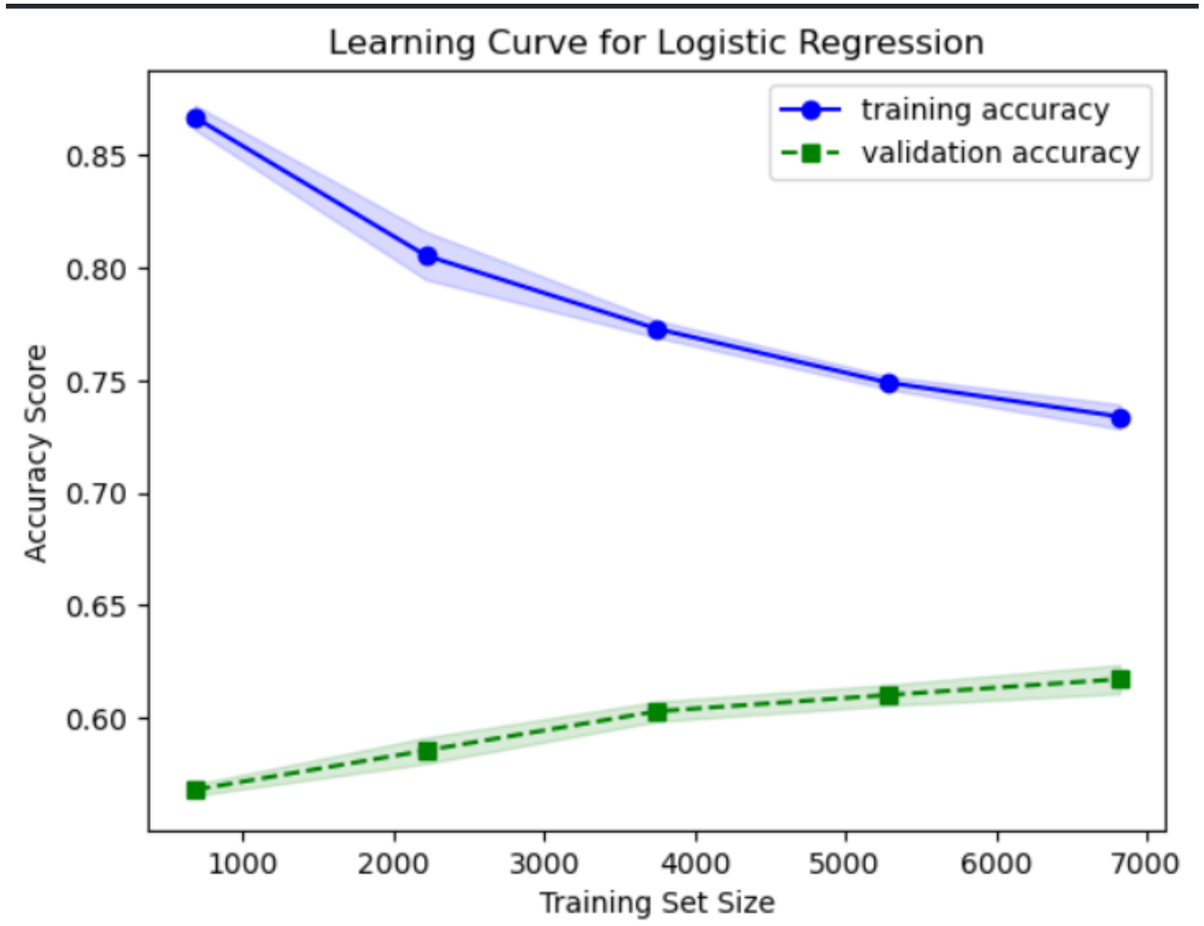
GAUSSIAN NAIVE BAYES

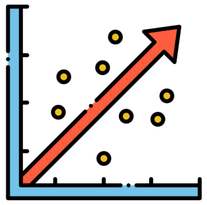
- **The accuracy using Gaussian Naive Bayes is 58.09%**



LOGISTIC REGRESSION

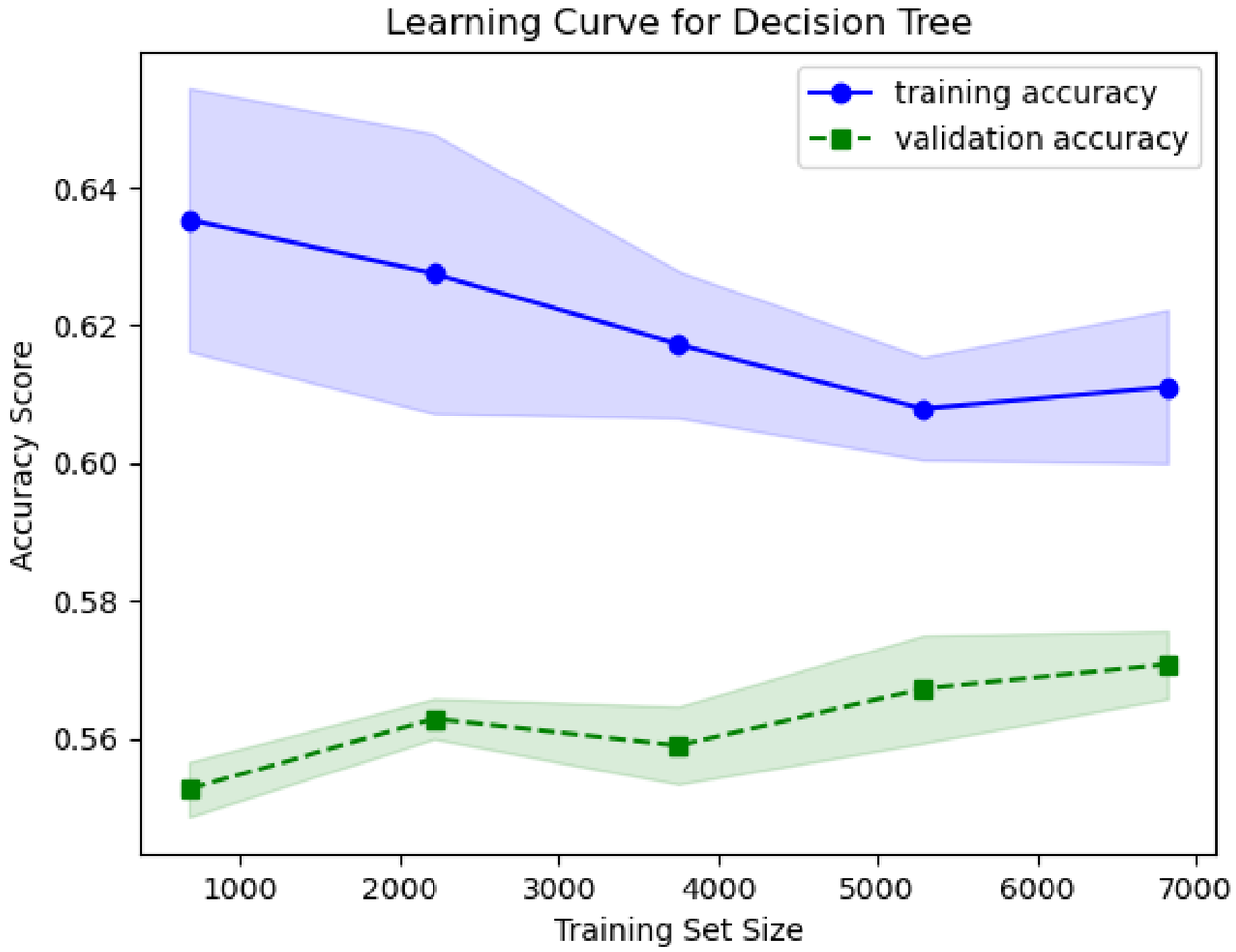
- **The accuracy using Logistic Regression is 61.74%**





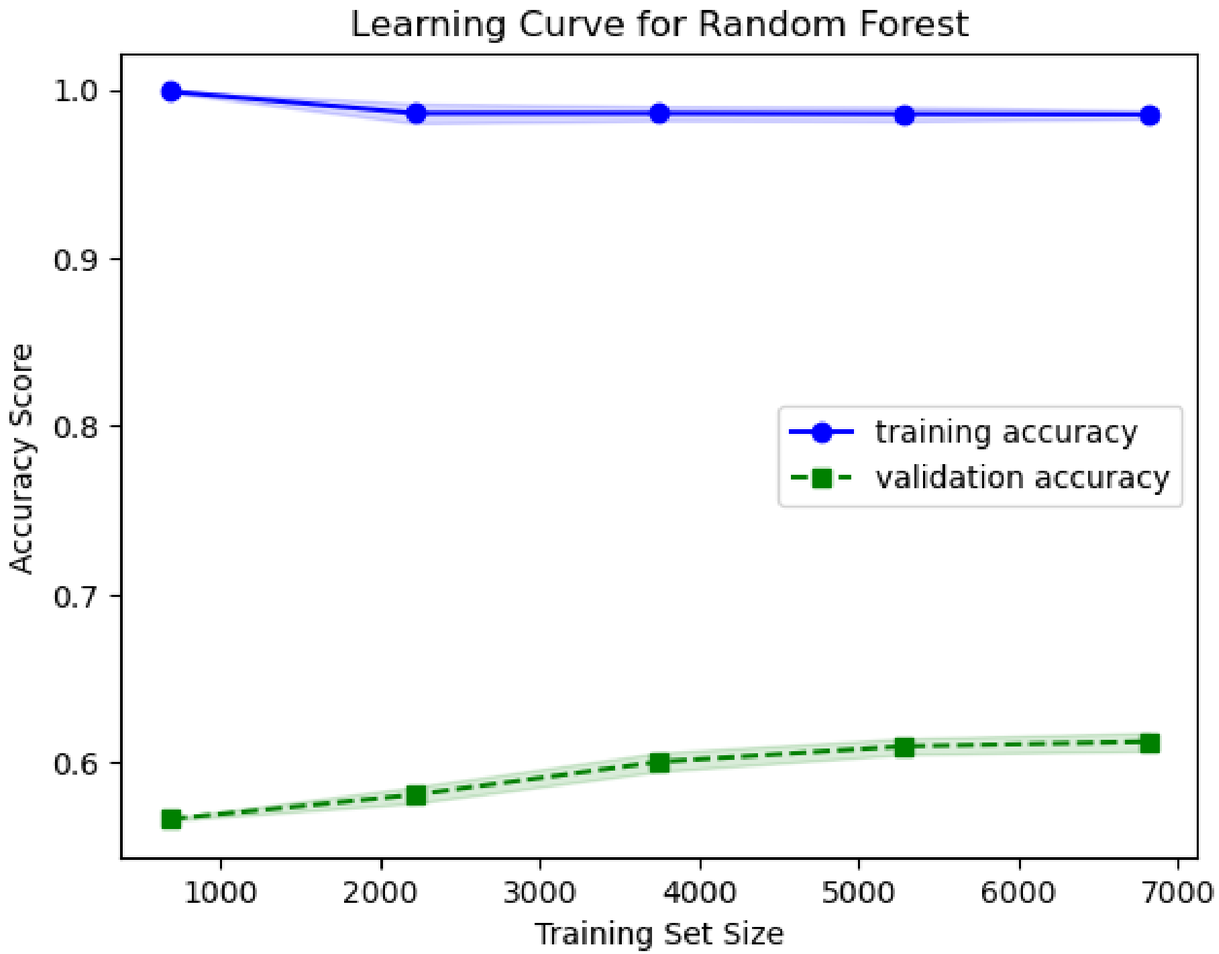
DECISION TREE

- **The accuracy using the Decision Tree is 56.91%**



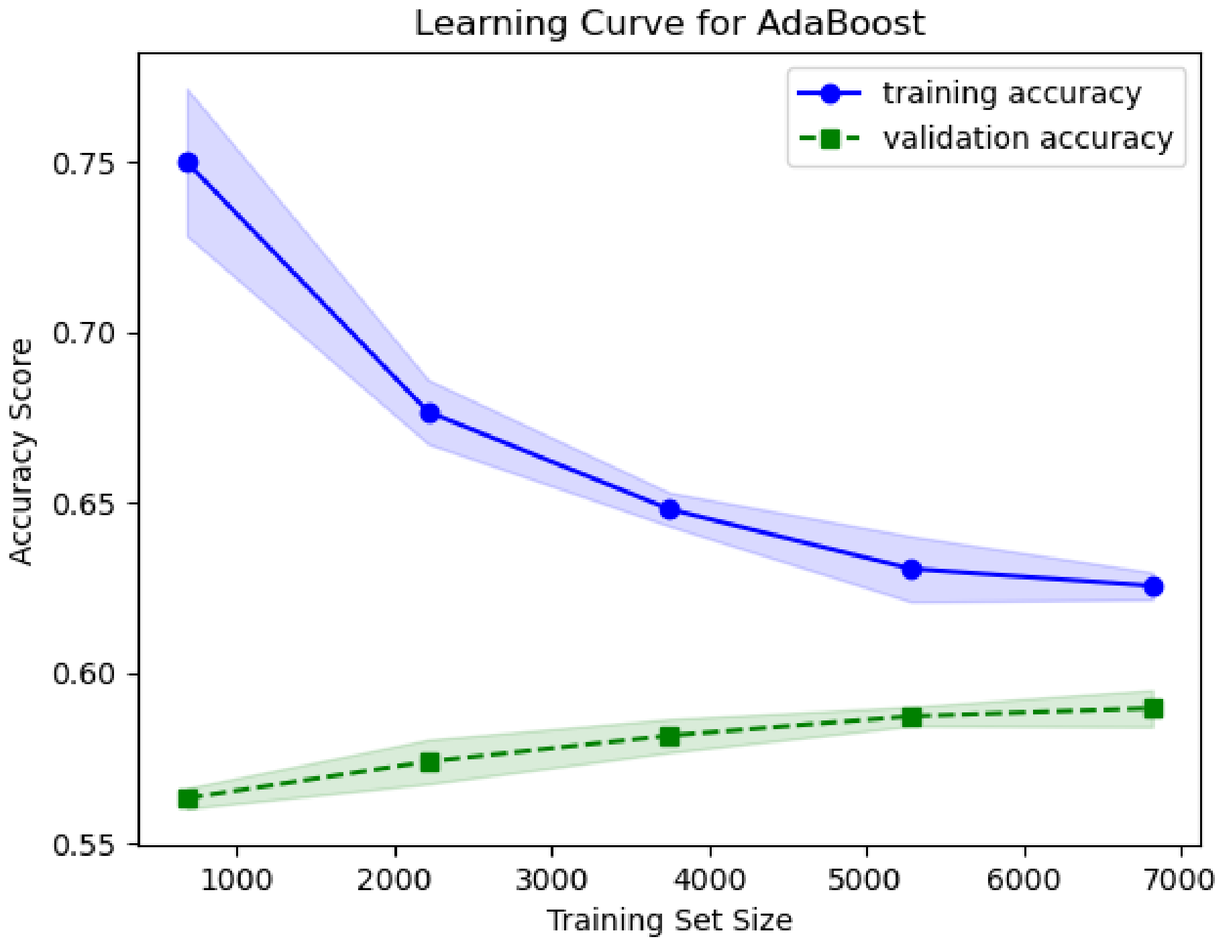
RANDOM FOREST

- **The accuracy using Random Forest is 59.57%**



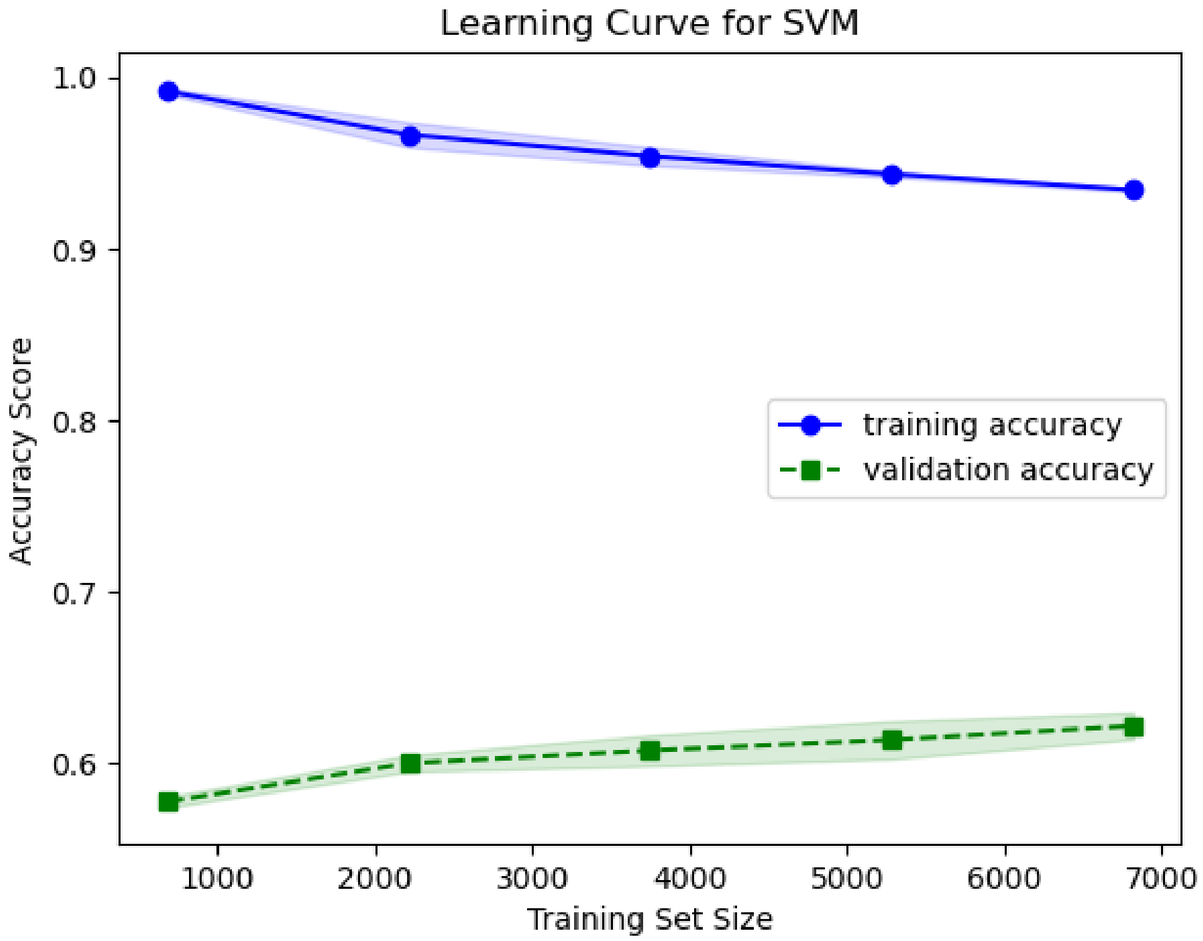
ADABOOST

- **The accuracy using Adaboost is 58.93%**



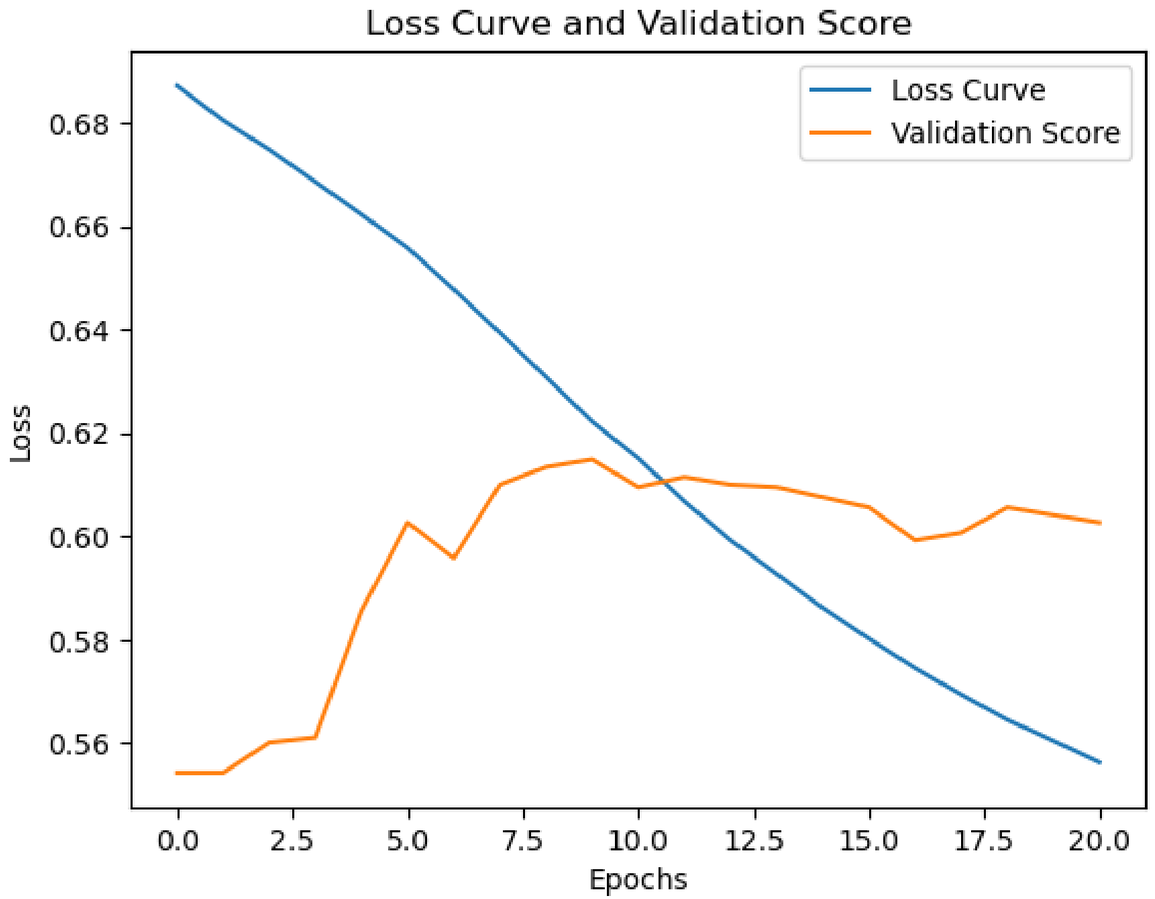
SVM

- **The accuracy using SVM is 59.34%**



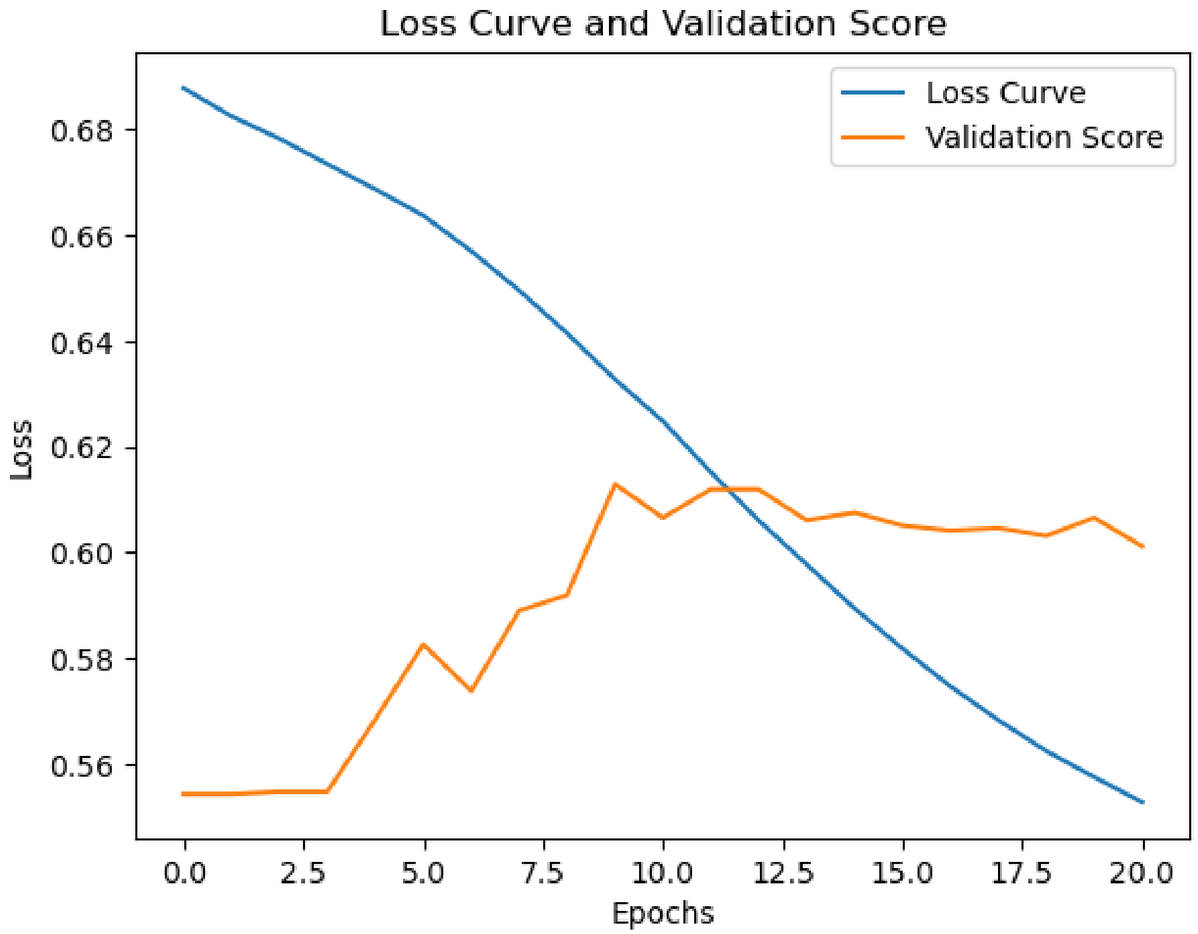
MLP

- **The accuracy using MLP is 58.95%**



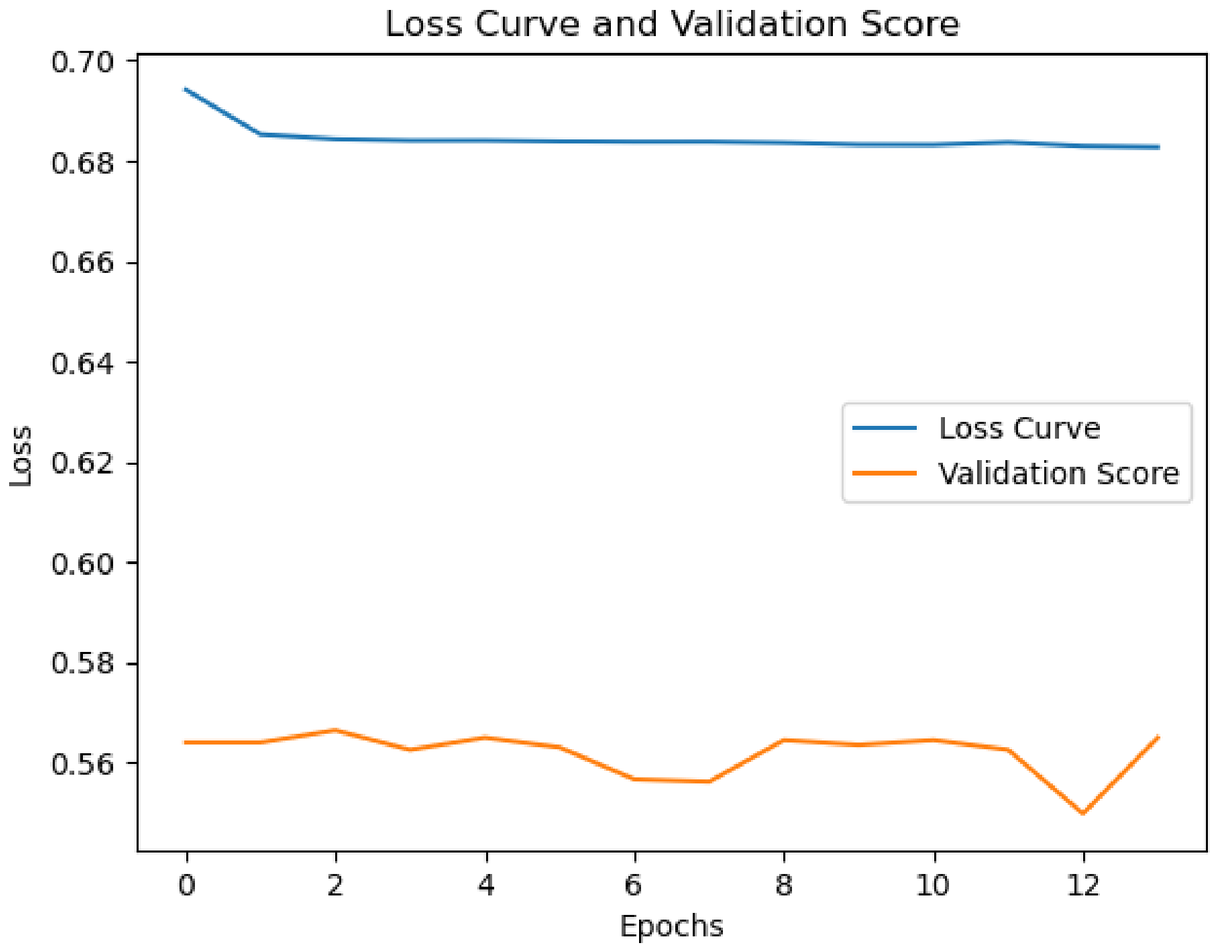
MLP WITH PCA

- **The accuracy using MLP along with PCA is 57.70%**



MLP WITH TSNE

- **The accuracy using MLP along with TSNE is 55.27%**



RESULTS SUMMED UP !!

Accuracy Scores for every model				
Model	Tfidf without speaker and party	Bow without speaker and party	Tfidf with speaker and party	Bow with speaker and party
Naïve Bayes	58.09	54.65	60.04	54.41
Logistic Regression	59.57	58.405	60.98	59.34
Decision Tree	56.91	57.23	58.95	59.89
Random Forest	59.57	60.59	60.906	62.93
Adaboost Classifier	57.38	58.17	60.43	60.75
SVM	59.34	58.405	58.32	58.09
MLP Classifier	58.95	59.65	57.38	59.42
MLP with PCA	57.70	59.42	57.93	58.56
MLP with TSNE	55.27	56.91	54.73	59.34

LIMITATIONS

- The accuracies are close to 60%, which is not much efficient.
- This is because it's impossible to solve this problem using standard ML and NLP Techniques.
- It is impossible to predict whether the news is fake without knowing the ground truth at that time.

CONCLUSION

- **We can predict whether the given news is fake or not with an accuracy better than a random guess i.e. 50%.**
- **The best accuracy was 62.93% using Random Forest, BOW vectorisation, with speaker and party and Gini gain as feature selection criteria.**

FUTURE WORK

- We can extend the scope by also incorporating visual and audio content in news articles.
- We will try to incorporate languages other than English.
- Develop an interactive system where users can give a news article as input and can receive a credibility score for that article, suggesting its credibility.

REFERENCES

- https://arxiv.org/pdf/1705.00648.pdf%E2%80%8B
- https://www.researchgate.net/publication/336436870_Fake_News_Detection_Using_Machine_Learning_approaches_A_systematic_Review
- https://paperswithcode.com/paper/liar-liar-pants-on-fire-a-new-benchmark
- https://github.com/manideep2510/siamese-BERT-fake-news-detection-LIAR