**Report**

Hassan Mahmoud Abdelkhaleq Sarwat

1- Which techniques you have used while cleaning the data if you have cleaned it?

I tried removing numbers and, the pound sign, and comas but after testing I realized it decreased the accuracy instead of improving, so I decided to leave the data as is.

2- Why have you chosen this classifier? (E.g. I used Multinomial Naive Bayes because it is easy to interpret with text data and there are more than two outcomes).

I used linear SVM as we have little amount of data and it scored better than naive bayes, also there were many references to it scoring better than naïve bayes.

3- How do you deal with (Imbalance learning)?

Applied smote to the features of underrepresented classes

4- How can you extend the model to have better performance?

I realized that some job titles include locations and salaries, completely removing those from the dataset should help out our model. Another way is to add more data, we can gather by web scraping from job listing websites

5- How do you evaluate your model? (i.e. accuracy, F1 score, Recall)

Accuracy score: 93%

F1 score: 0.9

The model is lacking in some regards when it comes to accuracy, looking at the confusion matrix I can’t really say that it is biased towards a certain class.

6- Limitations of Methodology

I believe given new samples not seen in the training data in accounting class, the model might not perform very well.