Precog Task Report

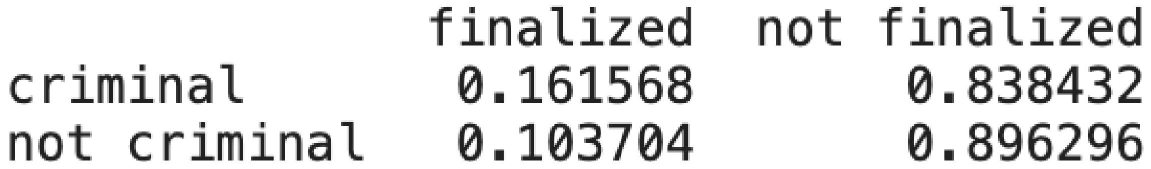
This report contains the data analysis and classification task for Precog. The task is to gain some insight from the dataset and then formulate a classification problem on the basis of it.

The dataset consists of indian court cases. The classification task decided is to classify the court cases based on their disposition. There were a number of labels denoting the disposition status. The task selected was to label the case as having reached a conclusion in the specific court (finalized) containing disposition status 'closed', 'decided', 'disposed', or as not finalized, containing the rest of the labels.

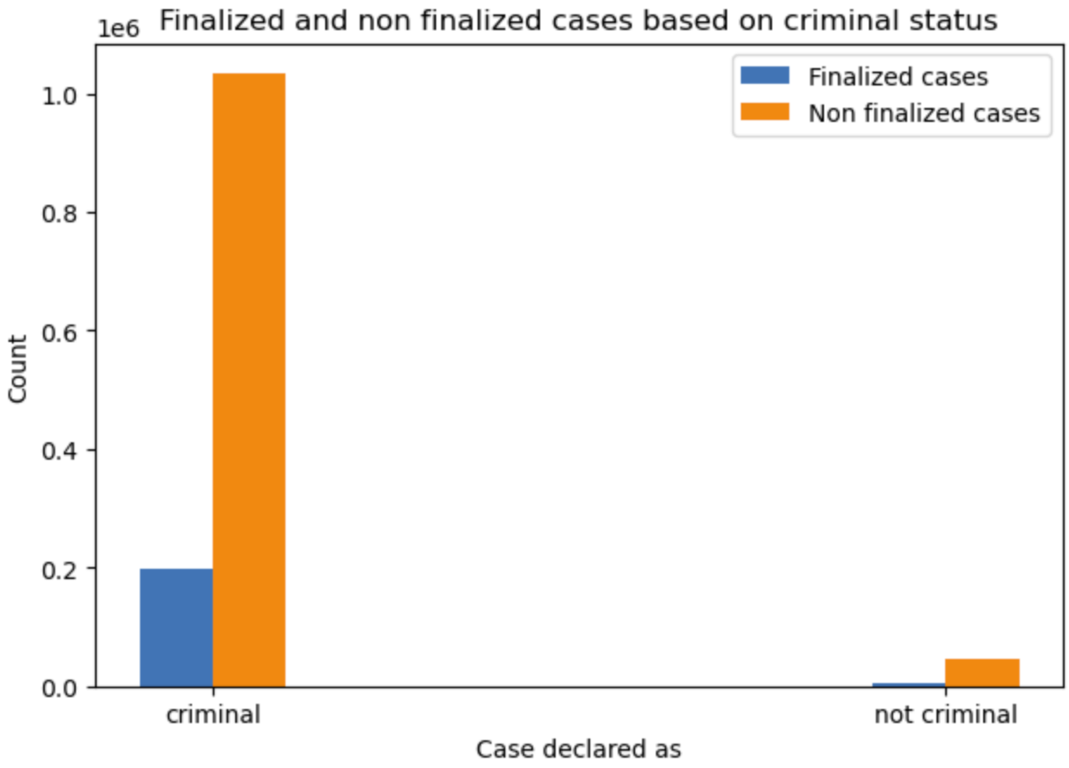
Due to the size of the dataset, the dataset used is for the 2015 cases. For the analysis, the data is analysed for the effect the gender of the judge, the state, and whether the case is deemed as a criminal case or not has on the case disposition. For classification, the methods used are logistic regression, KNN, Decision Trees and Random Forest

Analysis

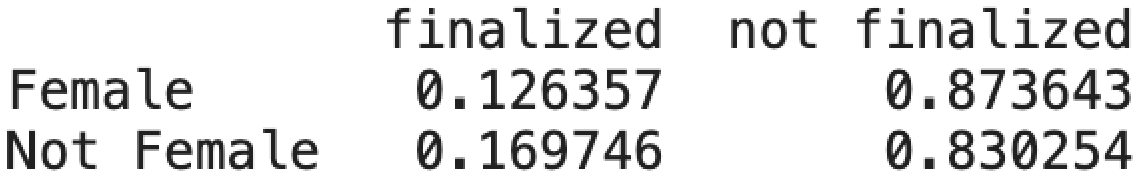
1.
The following graph and table show distribution of the cases into finalized and non finalized disposition according to the criminal value. There is a slight chance of more proportion of cases being finalized if the cases are criminal in nature, making it somewhat “easier” for cases to be driven to completion



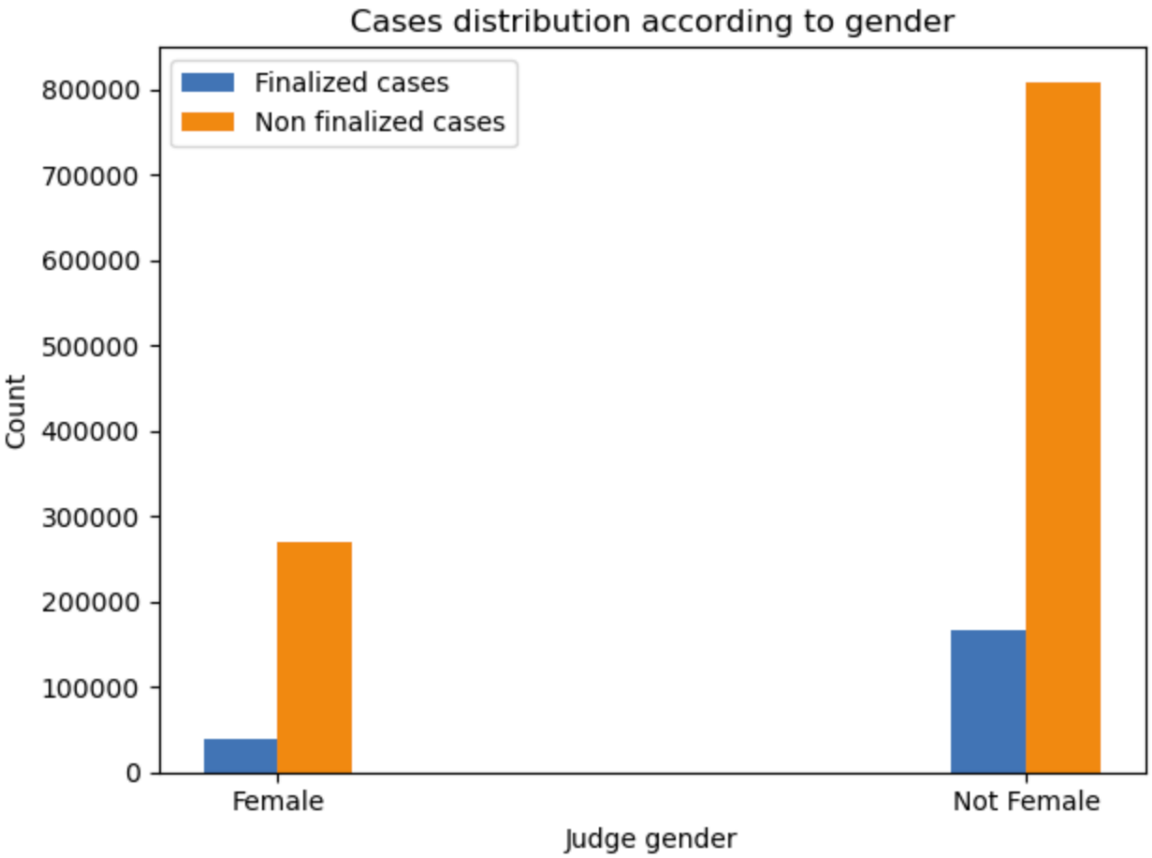
```
              finalized  not finalized
criminal       0.161568       0.838432
not criminal   0.103704       0.896296
```



2.
The following graph and table show the distribution of the cases into finalized and non finalized disposition according to the gender of the judge. There is a slight chance of more proportion of cases being finalized if the judge is not female, however the number of female judges is less, so this fact should be taken into account while making any observations

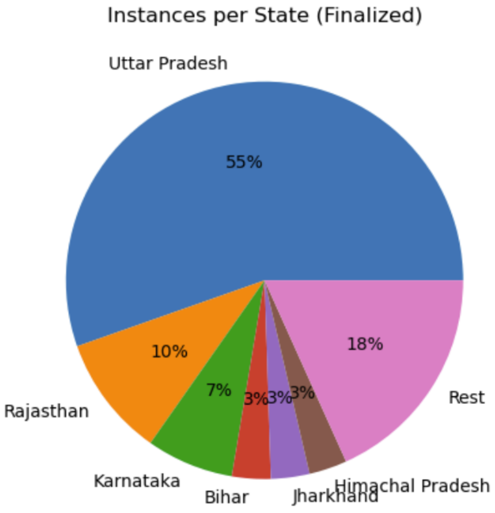


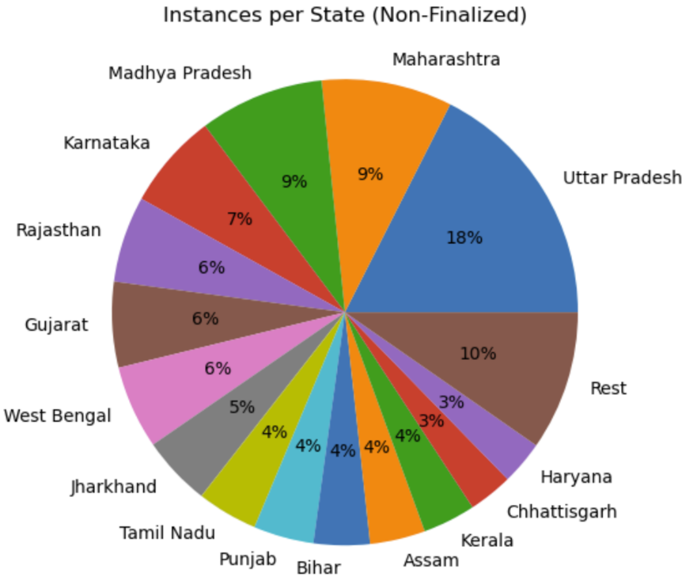
```
             finalized  not finalized
Female        0.126357       0.873643
Not Female    0.169746       0.830254
```



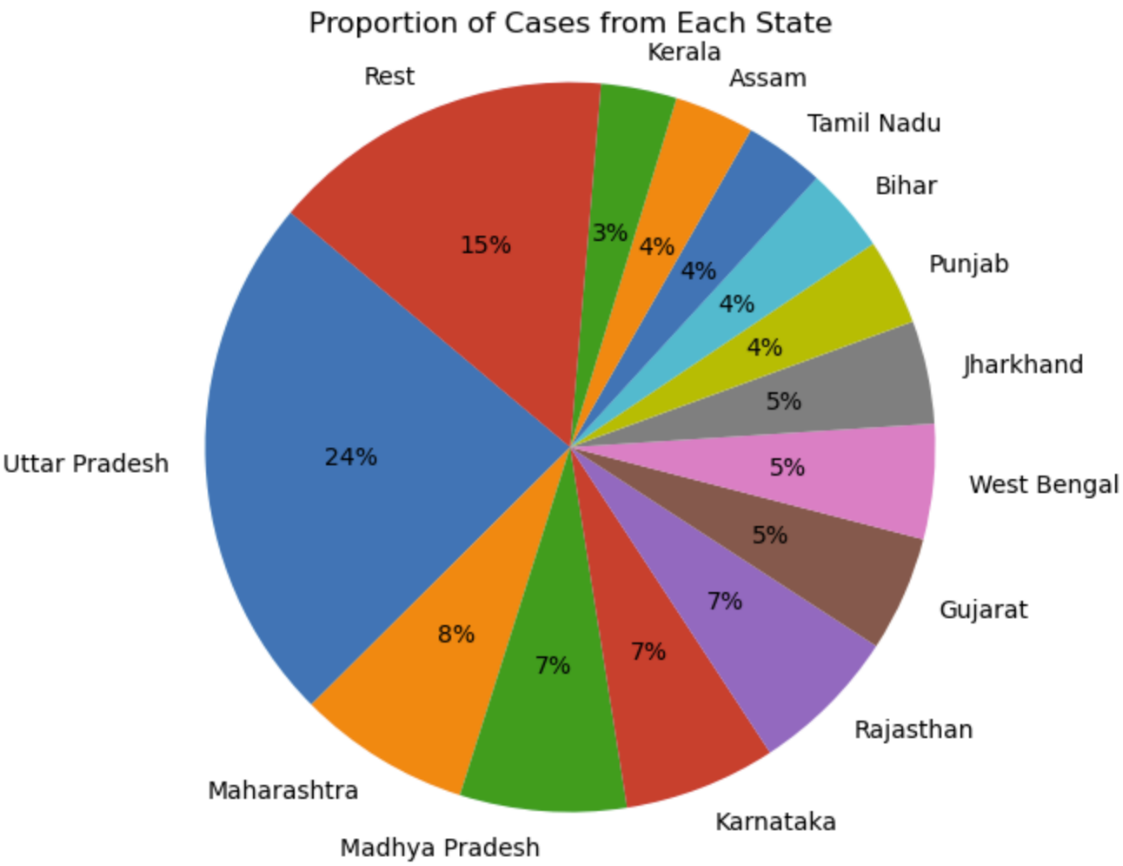
3.

Plotting the pie chart of the disposition status of cases according to state. The vast majority of cases in the finalized category come from Uttar Pradesh. In non-finalized cases the distribution is more equitable.





Plotting the pie chart of the cases according to state for more context. While Uttar Pradesh does account for 1/4 of the cases in the dataset, the number of finalized cases is still noteable.

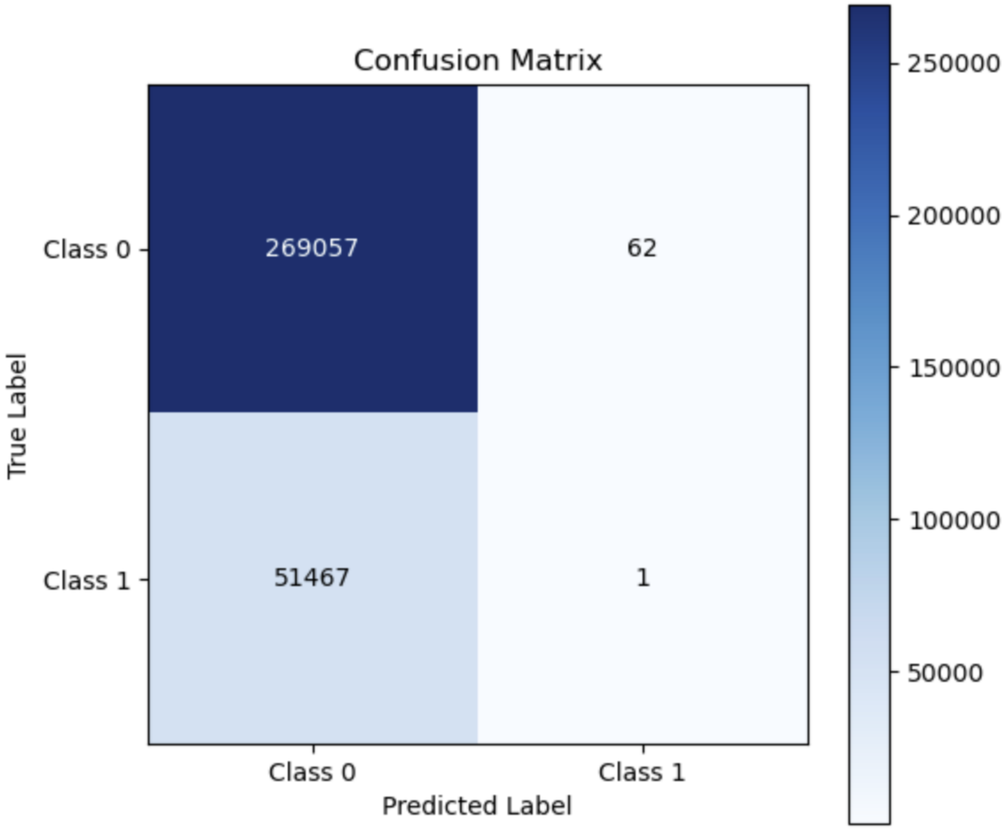


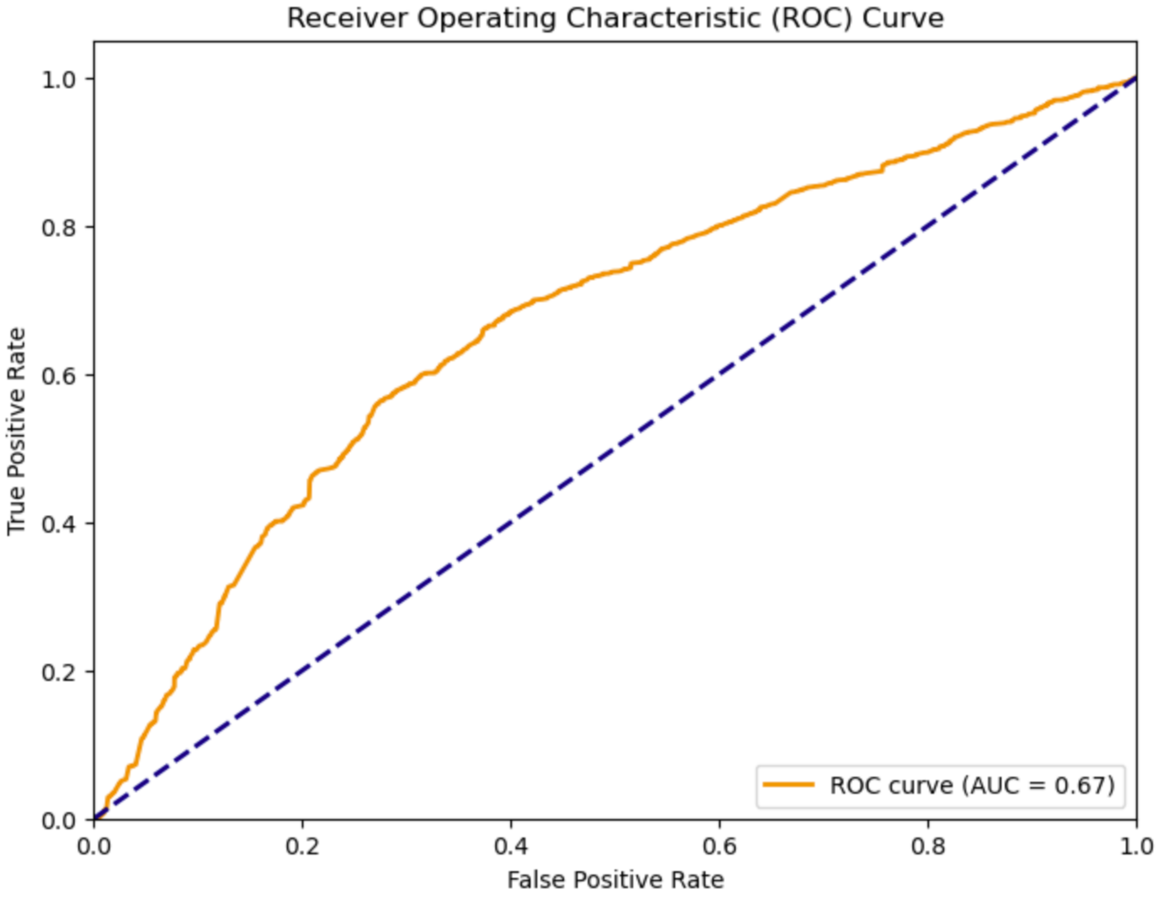
Classification Methods

The classification methods explored are mentioned below along with their metrics.

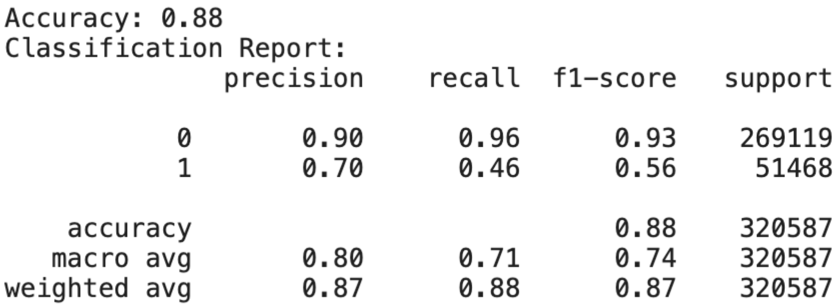
1.
Logistic regression

```
Accuracy: 0.84
Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91    269119
           1       0.02      0.00      0.00     51468

    accuracy                           0.84    320587
   macro avg       0.43      0.50      0.46    320587
weighted avg       0.71      0.84      0.77    320587
```

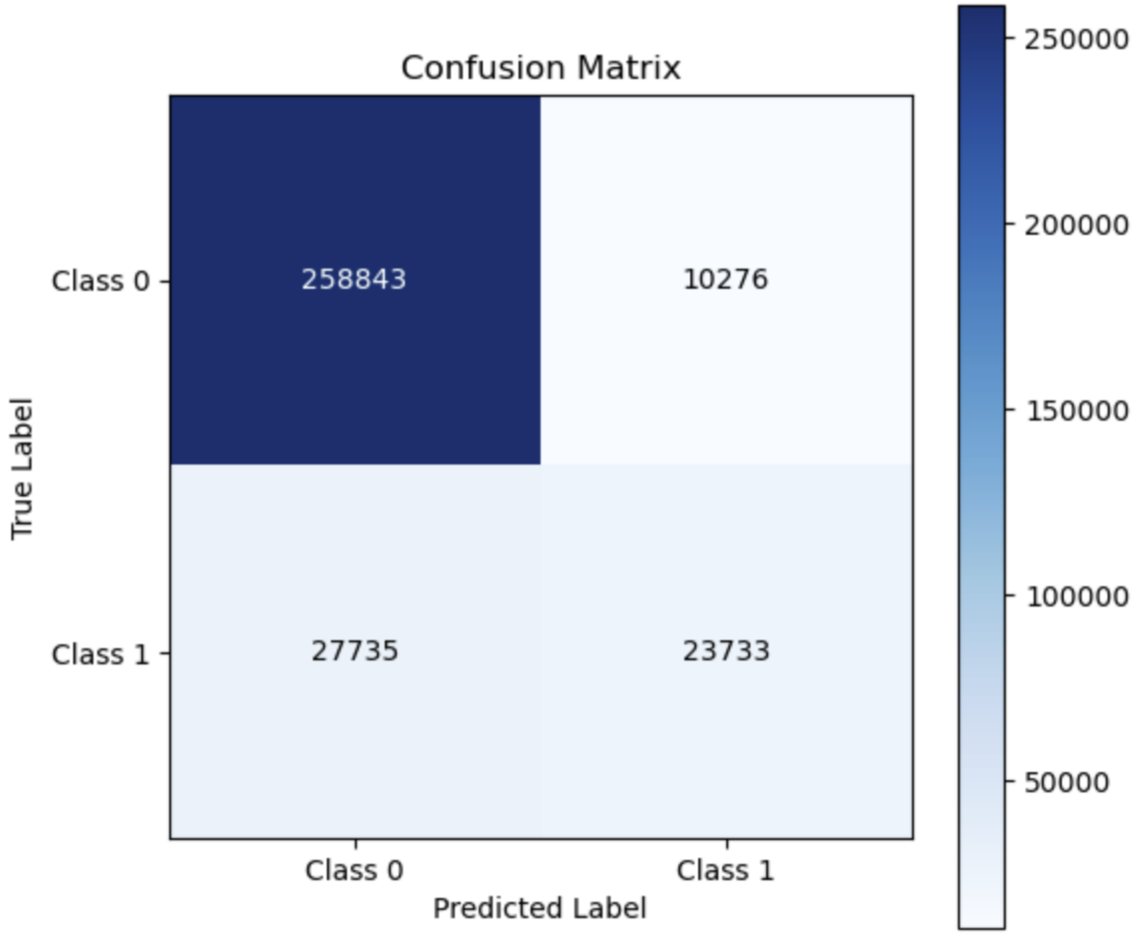


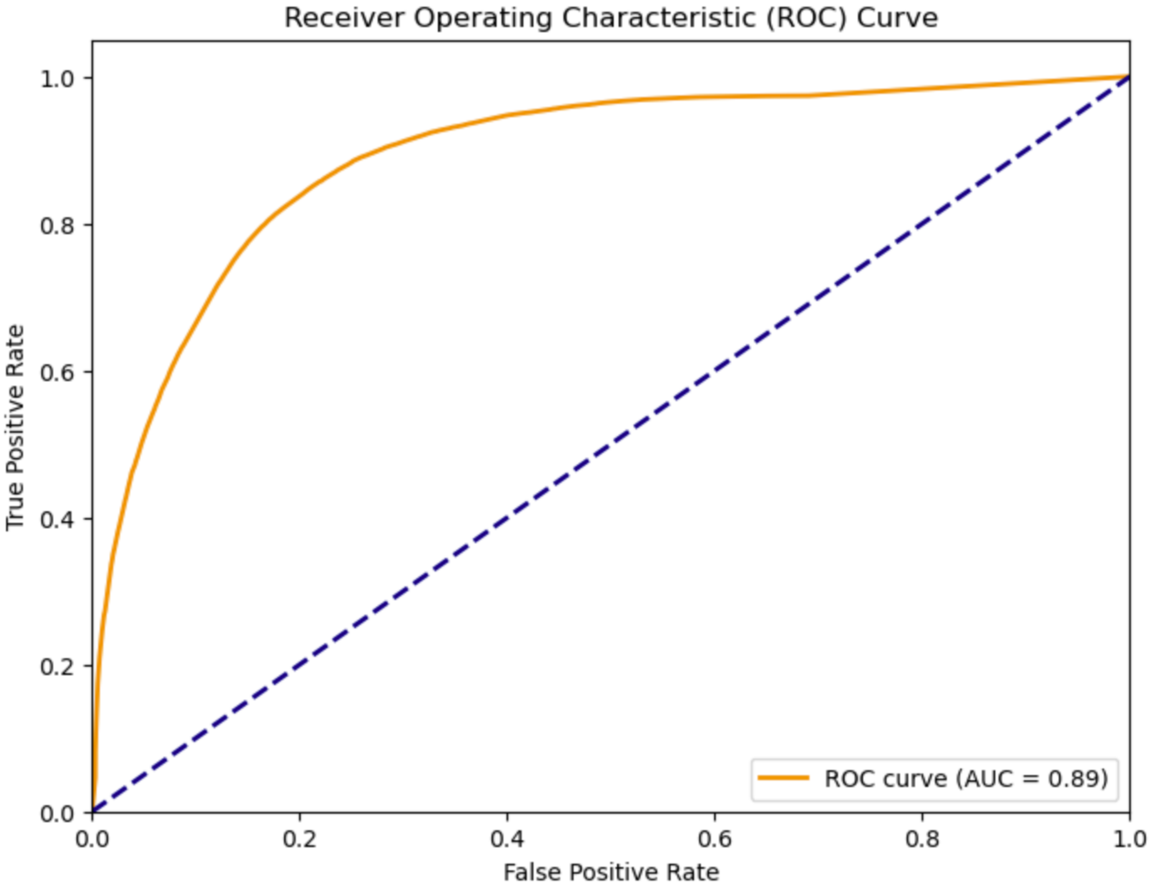


2.
Decision Tree



```
Accuracy: 0.88
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.96      0.93    269119
           1       0.70      0.46      0.56     51468

    accuracy                           0.88    320587
   macro avg       0.80      0.71      0.74    320587
weighted avg       0.87      0.88      0.87    320587
```

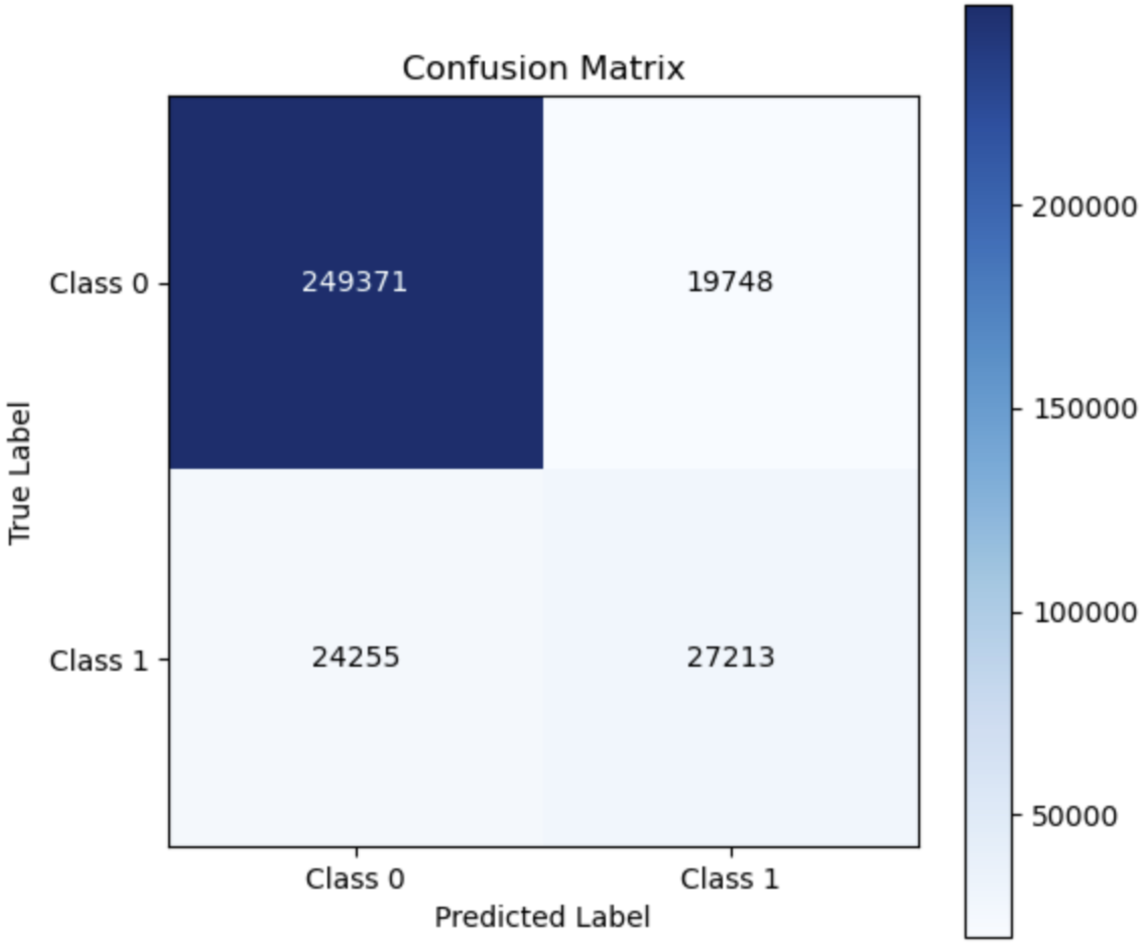


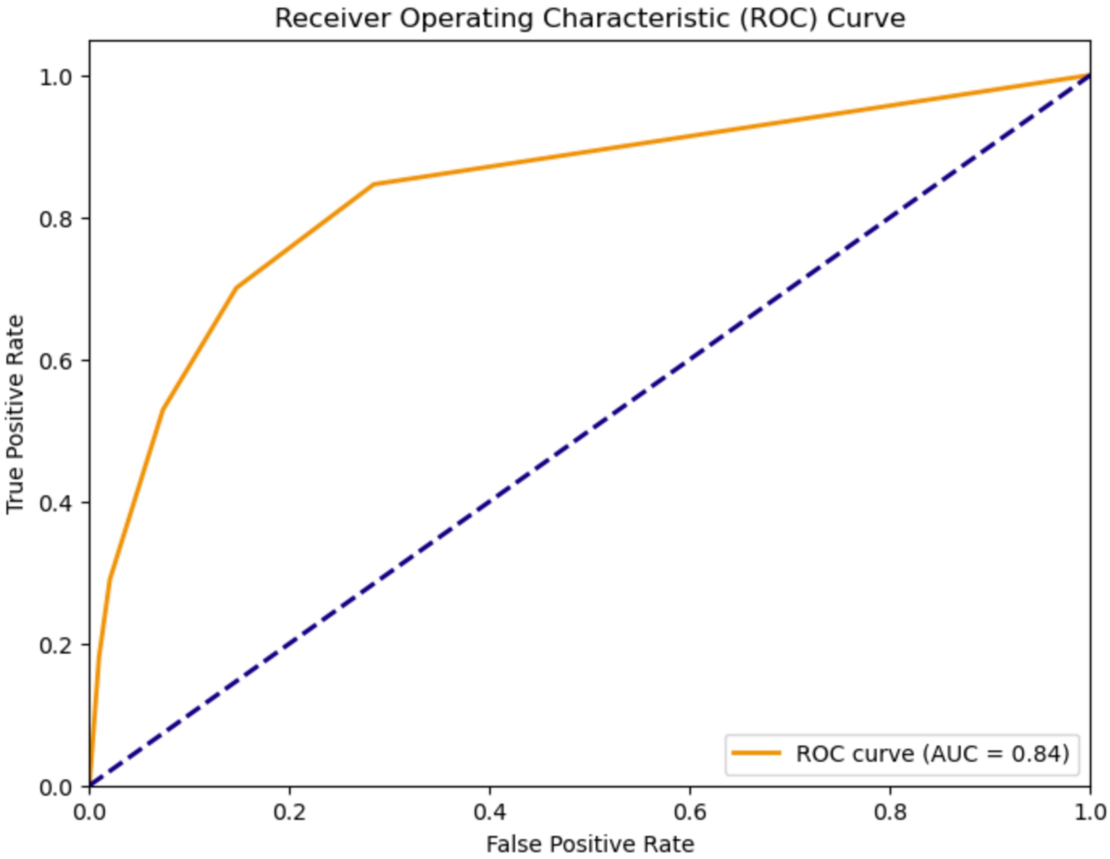


3.

K nearest neighbours, k = 5

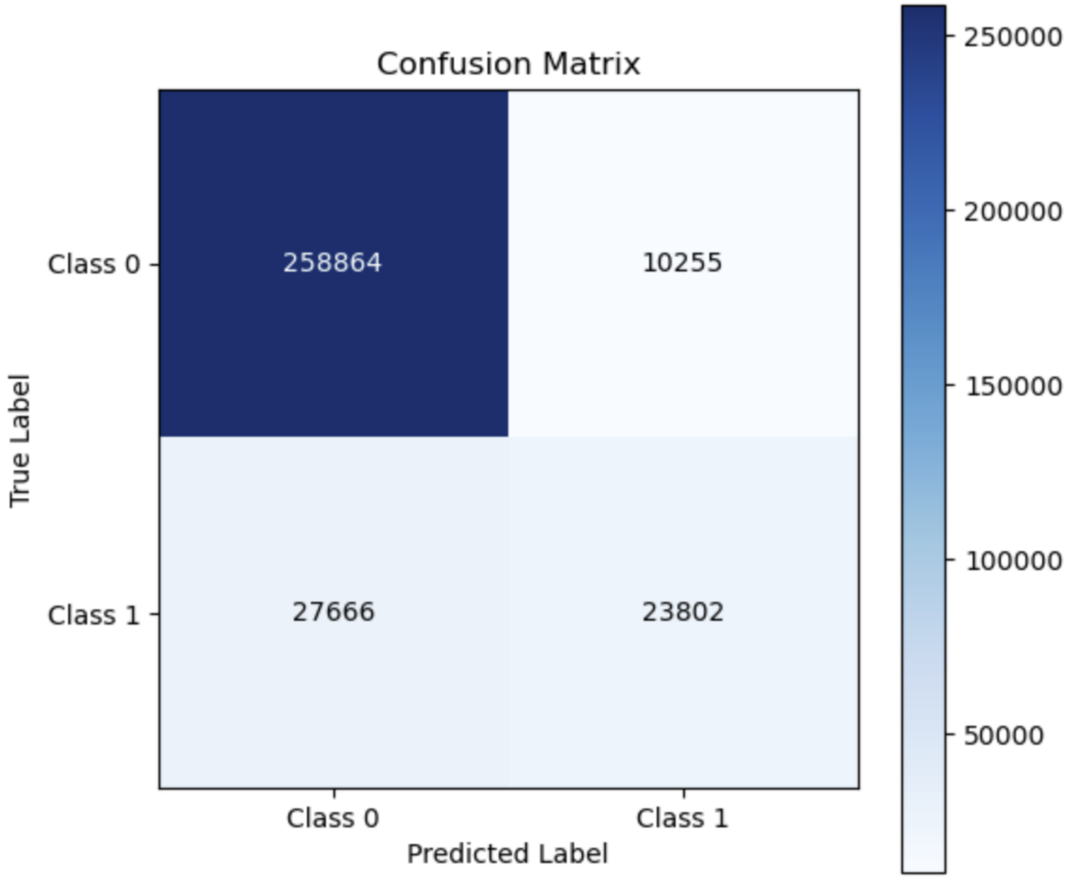
```
Accuracy: 0.86
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.93      0.92    269119
           1       0.58      0.53      0.55     51468

    accuracy                           0.86    320587
   macro avg       0.75      0.73      0.74    320587
weighted avg       0.86      0.86      0.86    320587
```

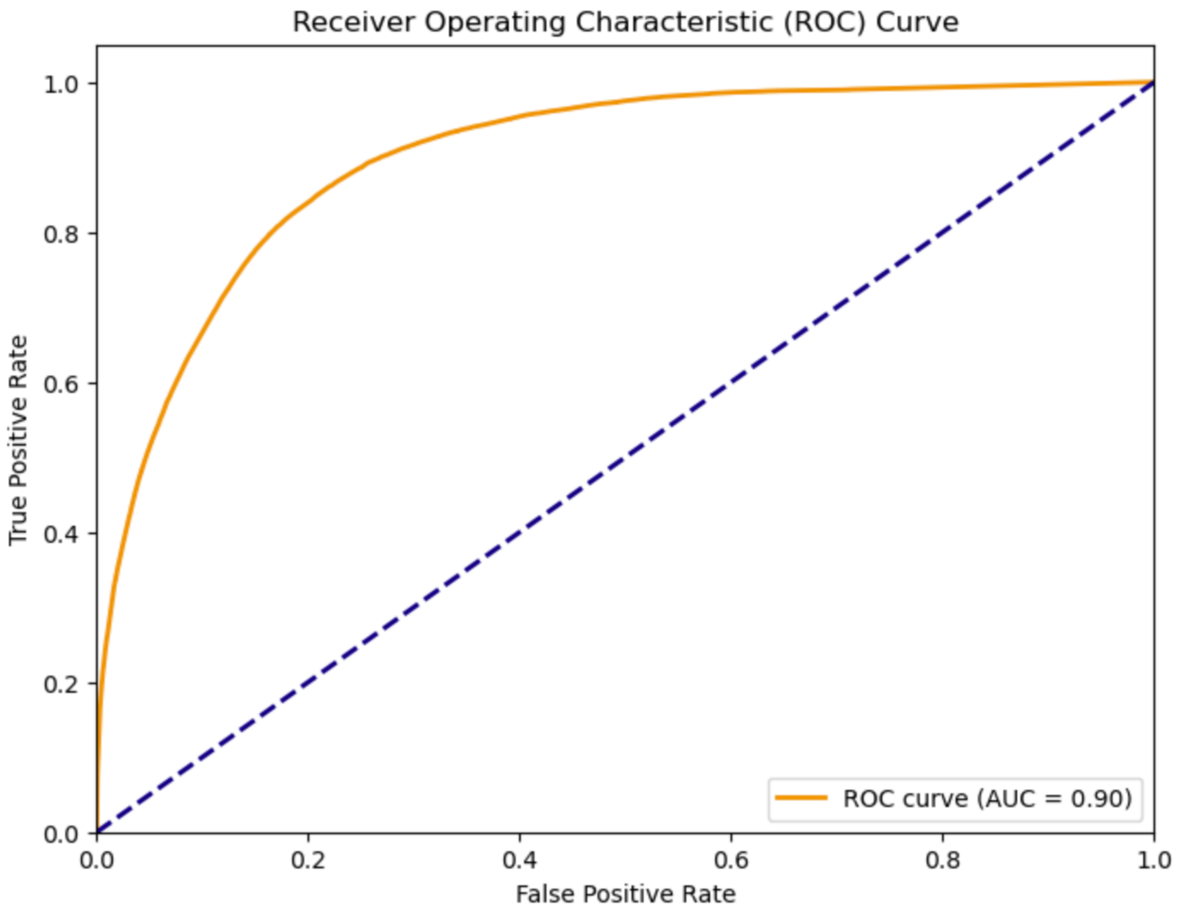




4.
Random forest

```
Accuracy: 0.88
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.96      0.93    269119
           1       0.70      0.46      0.56     51468

    accuracy                           0.88    320587
   macro avg       0.80      0.71      0.74    320587
weighted avg       0.87      0.88      0.87    320587
```





Observations:

1. Logistic Regression is the least effective of the classification techniques used. While the hyperparameters might be tuned better, the presence of a large number of features might deter it from performing well
2. Random forest performs the best classification and slightly edges out decision tree, which is to be expected as random trees by nature are an ensemble method. Both have similar accuracy scores with random forest having a slightly better value of AOC. Both these methods perform well in this situation as they are able to navigate through the large number of features
3. K nearest neighbours performs decently, although increasing the value of k beyond 5 makes the computation more complex and barely any change in the metrics value.
4. SVM and adaboost methods were thought of but were computationally extremely expensive while lending similar metrics, hence rejected from the final report.