# Precog Task - 2 Report

Ishan Kavathekar
ishan.kavathekar@research.iiit.ac.in
2022121003

## 1. Problem Statement

Classify the case as criminal case or non-criminal case based on the following features:

- State code
- District code
- Court number
- Judge designation
- Defendant gender
- Advocate's gender
- Petitioner gender
- Type name
- Purpose name
- Display name
- Act

## 2. Technologies and libraries used

- Python
- NumPy
- Pandas
- Seaborn
- Matplotlib

## 3. Methodology

File: acts_section.csv and cases_2010.csv
Corresponding notebook: Precog_task_2.ipynb

1) The acts_section.csv file is a huge file hence it needs to be read in chunks. The dataset contains ~ 80 millions cases. Out of those, I chose 10 million of them at random. This file was merged with the cases_2010.csv corresponding to the similar case ID. Necessary preprocessing and data cleaning was done.
2) Data in the column was standardised where 0 stands for male, 1 for female, -9998 for unclear and -9999 for missing value. The cases with unclear or missing names were dropped because they did not provide any information for the model. All of the duplicate rows were dropped.
3) The ordinal data was converted to data type int. Categorical data was encoded using Ordinal encoder using sklearn.
4) The data was split into a training set and testing set in the ratio 3:1.
5) The training set was trained on various models such as:
    - Logistic Regression
    - K Nearest Neighbors Classifier
    - Decision Tree Classifier
    - Random Forest Classifier
    - Gaussian Naive Bayes

6) The accuracy and confusion matrix of these tests was recorded.

## 4. Results

It was observed that Decision Tree and Random Forest give the most accurate results while Naive Bayes model gave the least accuracy.

| | model_name | Accuracy | AUC |
|---|---|---|---|
| **0** | Logistic Regression | 0.836342 | 0.70 |
| **1** | SVM | 0.906104 | 0.83 |
| **2** | KNeighbor | 0.985601 | 0.98 |
| **3** | Decision Tree | 0.999797 | 1.00 |
| **4** | Random Forest | 0.999189 | 1.00 |
| **5** | Naive Bayes | 0.833908 | 0.73 |

Fig 1 Model vs Accuracy

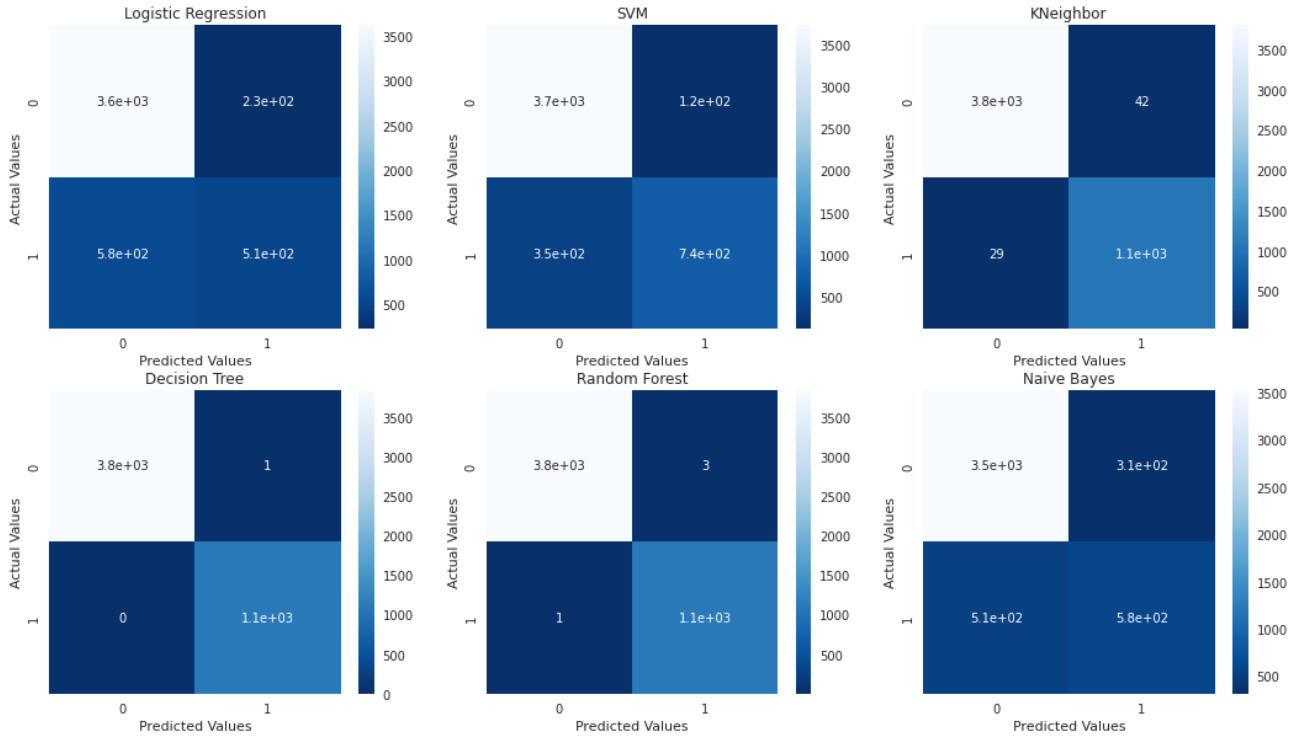


Fig 2 Confusion matrix for various models

## 5. Conclusion

The decision tree model yields the highest accuracy (99.97%) with an AUC of 1.00

## 4. References

The following articles and websites were referred to while conducting the analysis and preparing this report:


1) https://medium.com/analytics-vidhya/optimized-ways-to-read-large-csvs-in-python-ab2b36a7914e
2) https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html
3) https://www.geeksforgeeks.org/python-pandas-dataframe-astype/
4) https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/