Court Case Dataset

Analysis and Insights

(The notebooks have no visible plots, all the plots have been given as separate files and also have been desribed in this Report extensively)

1. Gender Dispersion in Judge Positions

Question

Is there any discrimination in appointing females to high positions in the court? And if yes, then in which all states its more?

(image size is large, on next page)


female_judge
male
female
state_name
Goa
Chandigarh
Delhi
Manipur
Sikkim
Madhya Pradesh
Tripura
Chhattisgarh
Tamil Nadu
Jharkhand
Himachal Pradesh
Karnataka
Mizoram
Telangana
Meghalaya
Gujarat
West Bengal
Uttarakhand
Uttar Pradesh
Jammu and Kashmir
Bihar
Assam
Andhra Pradesh
Punjab
Haryana
Orissa
Rajasthan
Kerala
Maharashtra

[redacted]

Gender Analysis has been done for major Judge Positions across the dataset.

The objective was to analyze whether there is Gender Discrimination in appointing Judges to high positions in the court.

This has been analyzed for all the states given in the dataset.

Upon a closer look at the strip plot, we see that most of the states reflect a decent 50-50 male-to-female division with the premier reason of Reservation in positions given by the Government.

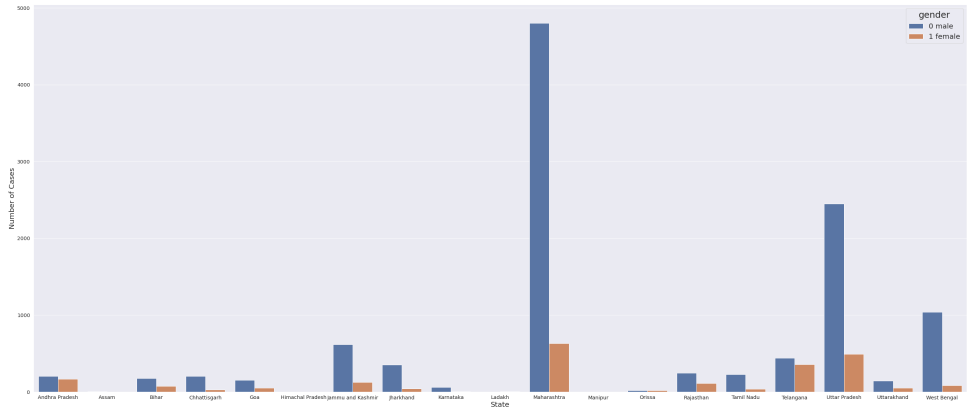
This has an exception for states like Rajasthan, Bihar, and Jammu Kashmir majorly due to their different reasons viz. Gender Discrimination history is associated with Rajasthan, slow development rate, the high corruption rate in Bihar, and political instability in Jammu Kashmir.

2. Caste-related Crime Analysis

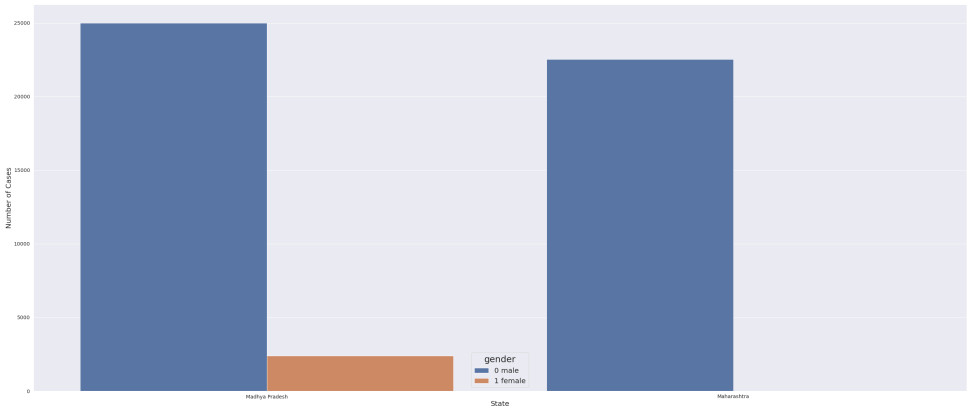
Question

Which states have the highest number of cases filed on Caste Atrocities and Crimes? And what has been their trend over the years?

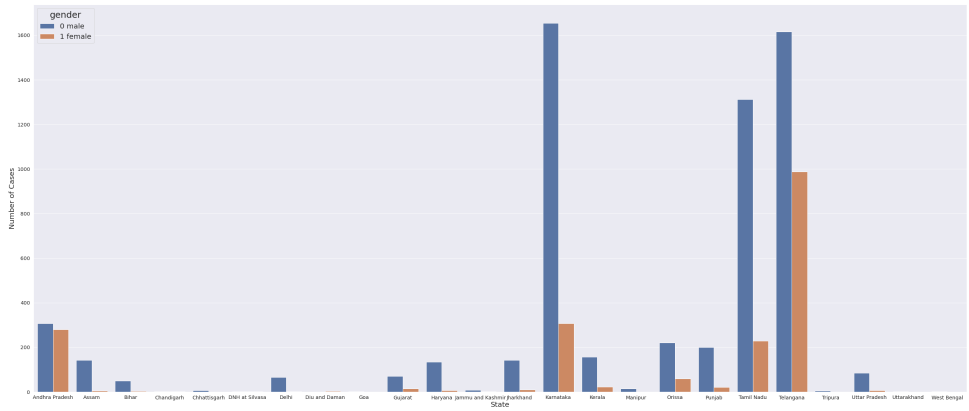
2016



2017

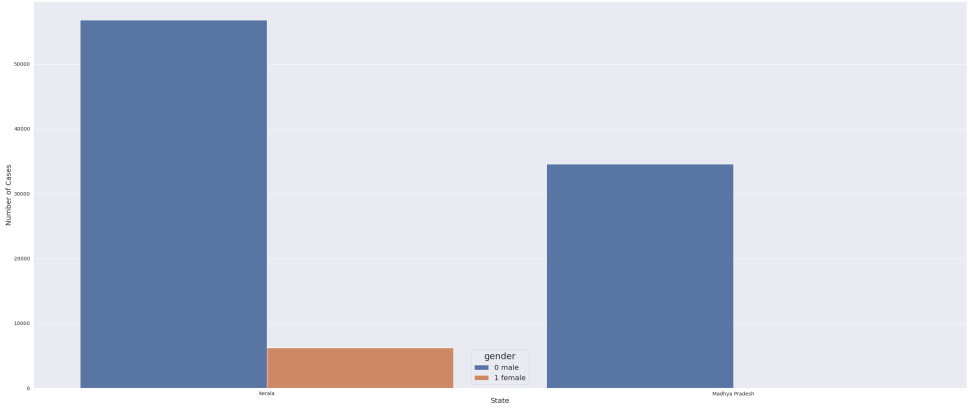


Outliers

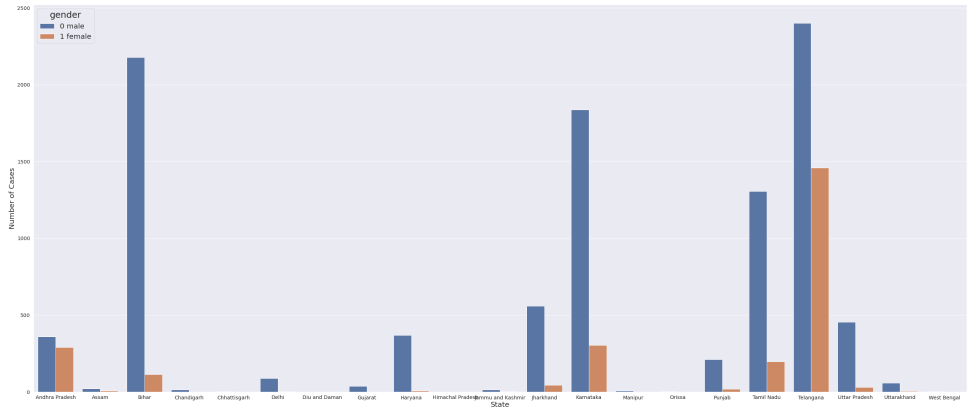


Without Outliers

2018



Outliers



Without Outliers

Analysis has been done for the Caste Related Crimes for the years 2016, 2017, and 2018, varying across all the states given in the dataset.

2018 - Kerala and Madhya Pradesh turn out to be the outliers for the category with an abnormally high number of cases filed.

2017 - Madhya Pradesh and Maharashtra turn out to be the outliers for the category with an abnormally high number of cases filed.

2016 - No such abnormalities are found.

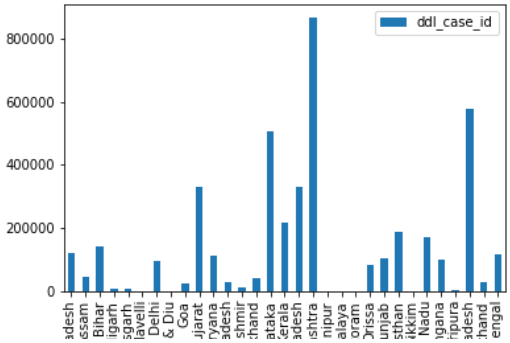
In general, it is observed that states with relatively more Tribal and working-class populations like Bihar, Jharkhand, Andhra Pradesh, Karnataka, and Telangana.

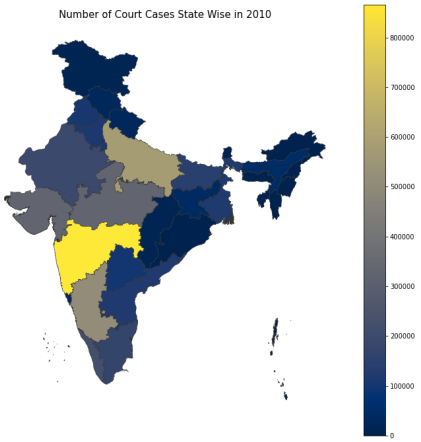
3. Cases Filed State-Wise Analysis

Question

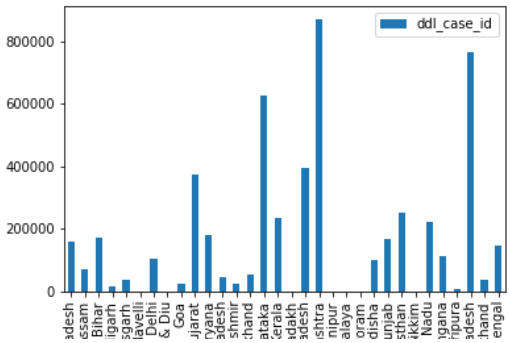
Which states have higher number of cases filed than the others over the years? It also analyses the same after normalizing the dataset with population of each state.

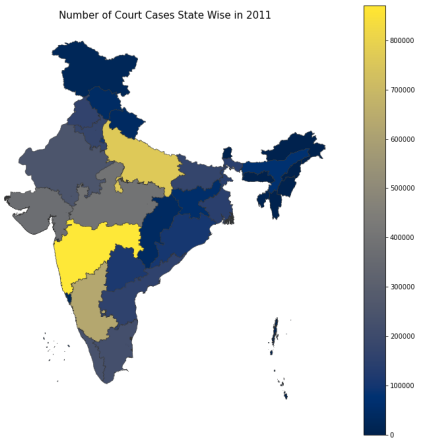
2010


ddl_case_id
800000
600000
400000
200000
0

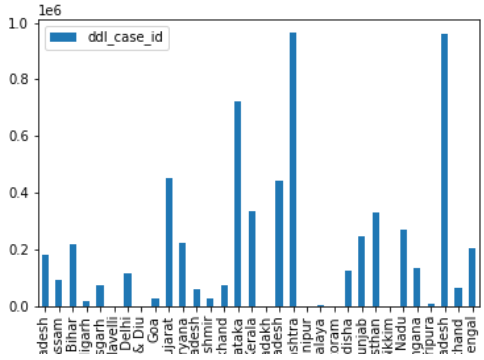

Number of Court Cases State Wise in 2010

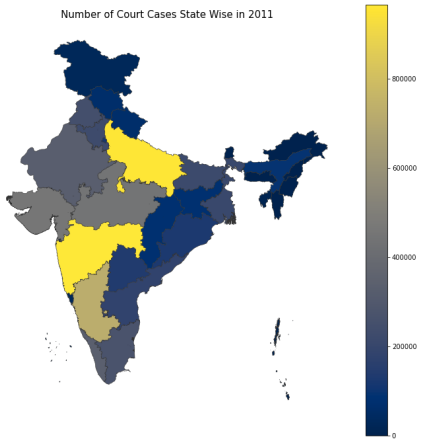
2011


ddl_case_id
800000
600000
400000
200000
0

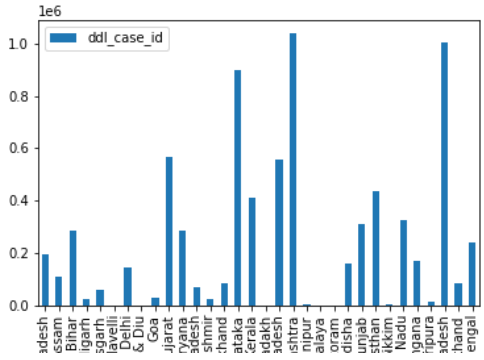

Number of Court Cases State Wise in 2011

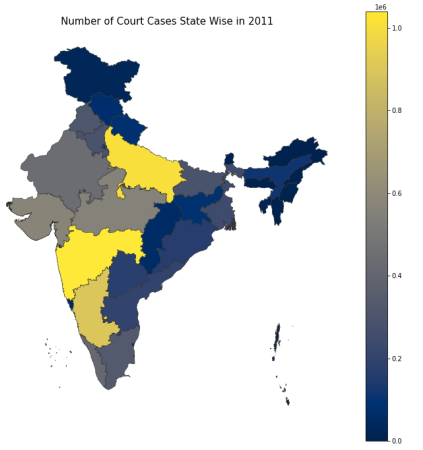
2012



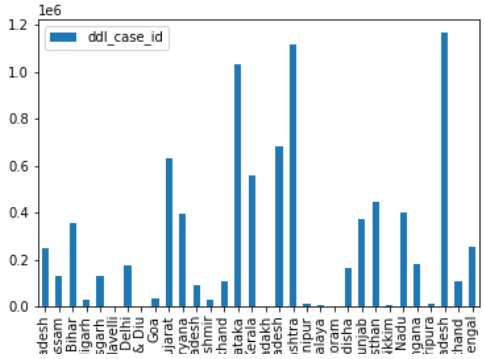


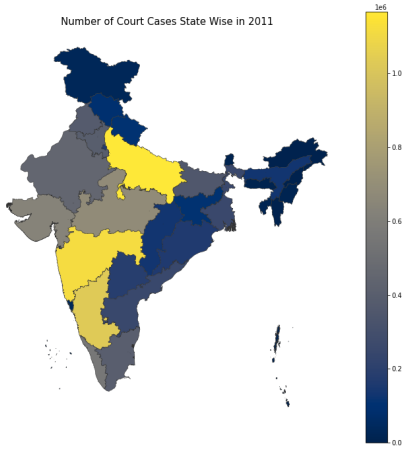
2013



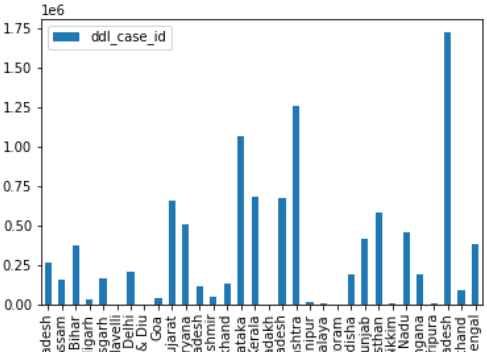


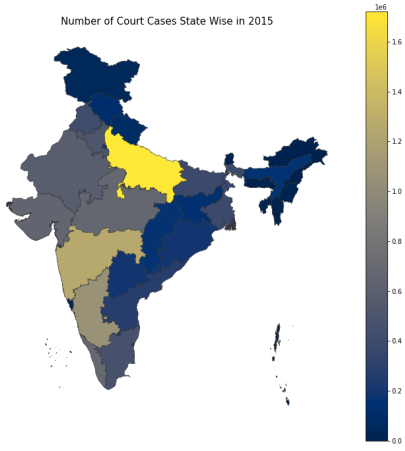
2014


1e6
ddl_case_id
1.2
1.0
0.8
0.6
0.4
0.2
0.0

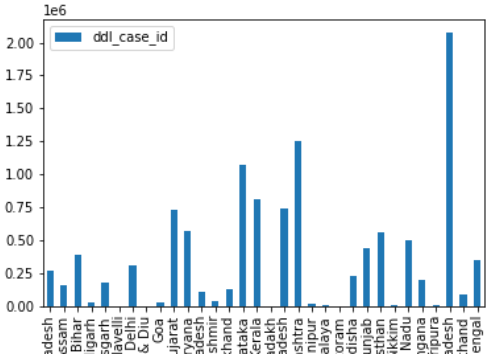

Number of Court Cases State Wise in 2011
1e6

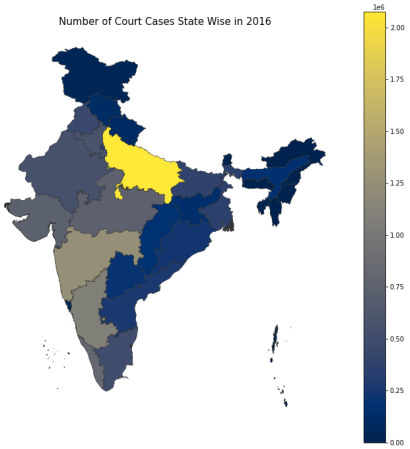
2015


1e6
ddl_case_id
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

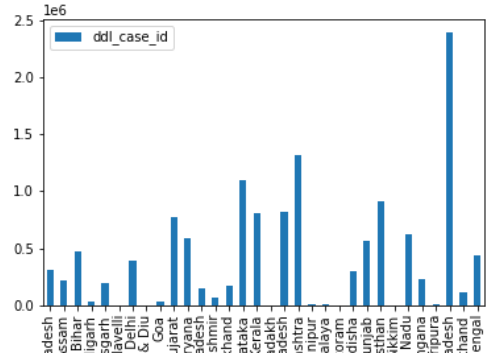

Number of Court Cases State Wise in 2015
1e6

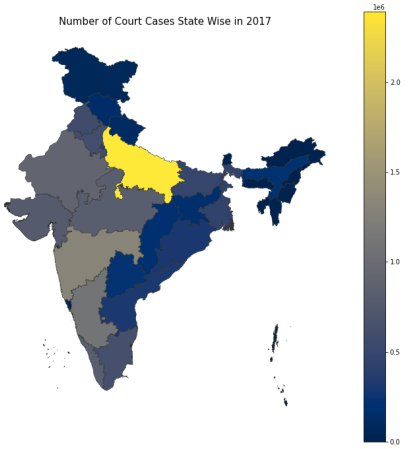
2016


1e6
ddl_case_id
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

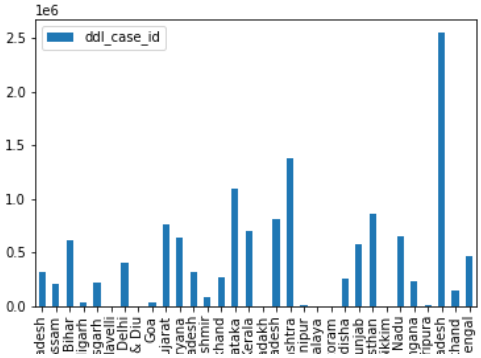

Number of Court Cases State Wise in 2016
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

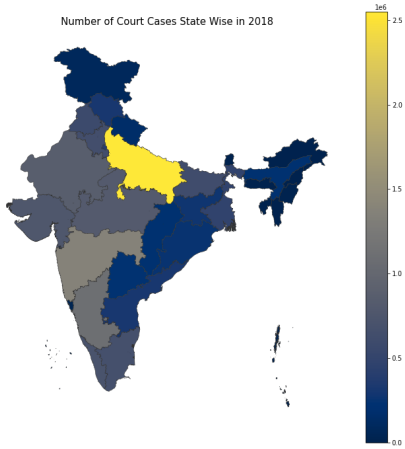
2017


1e6
ddl_case_id
2.5
2.0
1.5
1.0
0.5
0.0


Number of Court Cases State Wise in 2017
1e6
2.0
1.5
1.0
0.5
0.0

2018

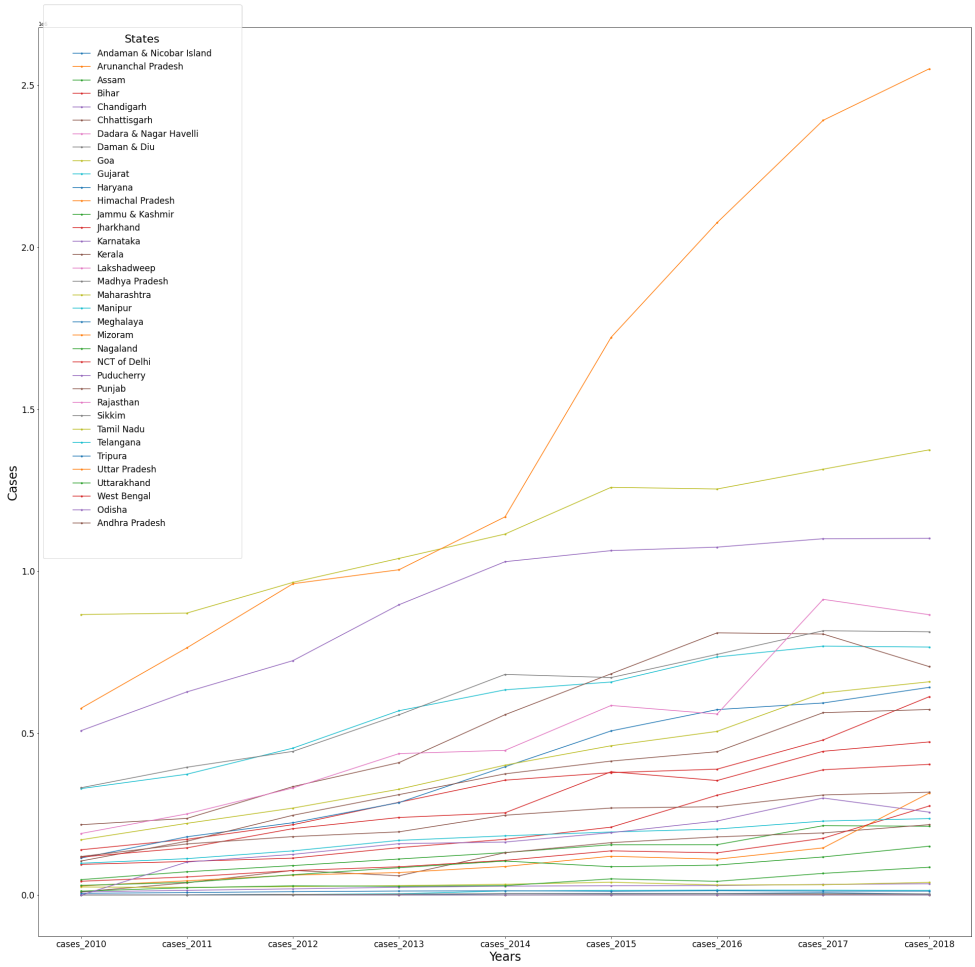




Analysis has been done for the number of cases filed across the states over the years and the same has been shown as a bar plot and on the Indian map.

It is evident that states with high populations like Uttar Pradesh, Maharashtra, and Karnataka have the highest number of cases filed.

Cumulative Trend



Cases filed over the years in various states

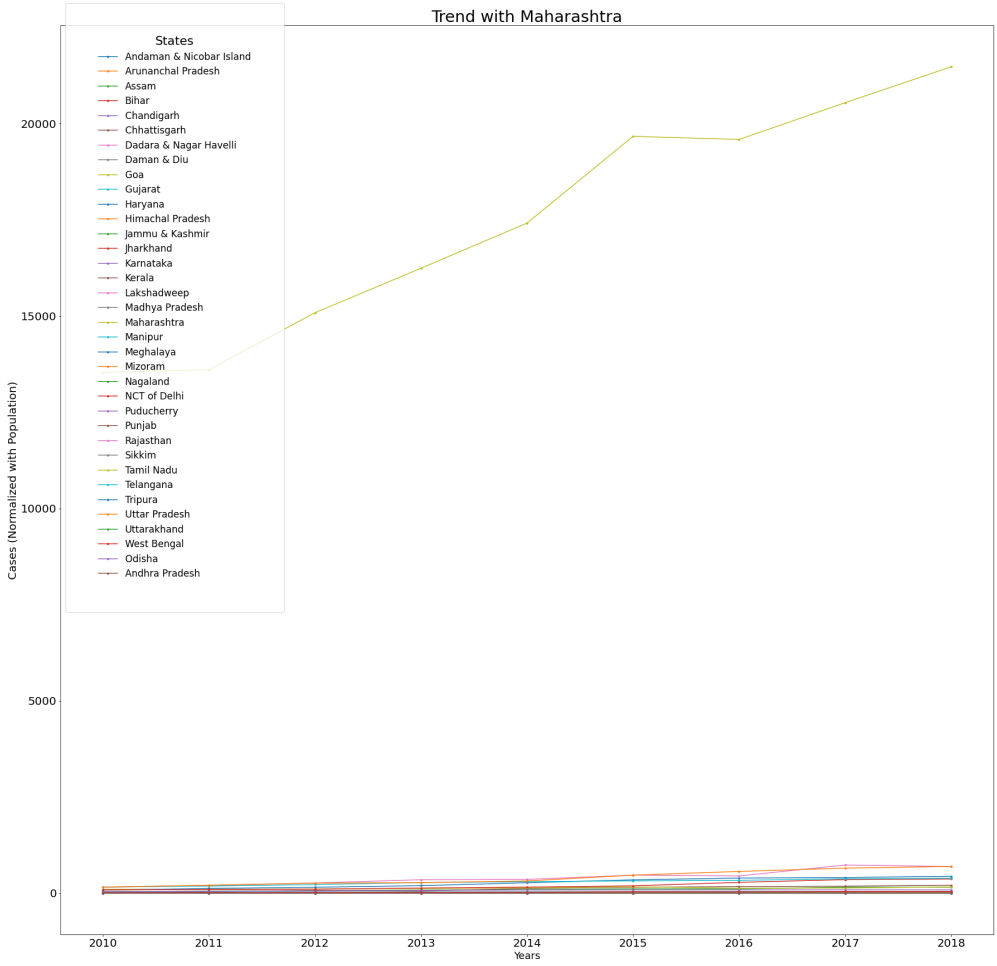
The same Case Distribution has been plotted as a line graph for all states with sufficient data over the years 2010-2018.

We can see the top 3 competitors as Uttar Pradesh, Maharashtra, and Karnataka.

Maharashtra and Karnataka have a steady increase in the number of cases filed over the years whereas Uttar Pradesh has had a massive spurt in the number of cases filed from the year 2014.

But this analysis is not completely free from biases as a major factor in the number of cases filed is the population of the state. Hence the dataset needs to be normalized with the population of each state to get a better prospect of the crime rate in each state.

Cumulative Trend (Normalized with Population)

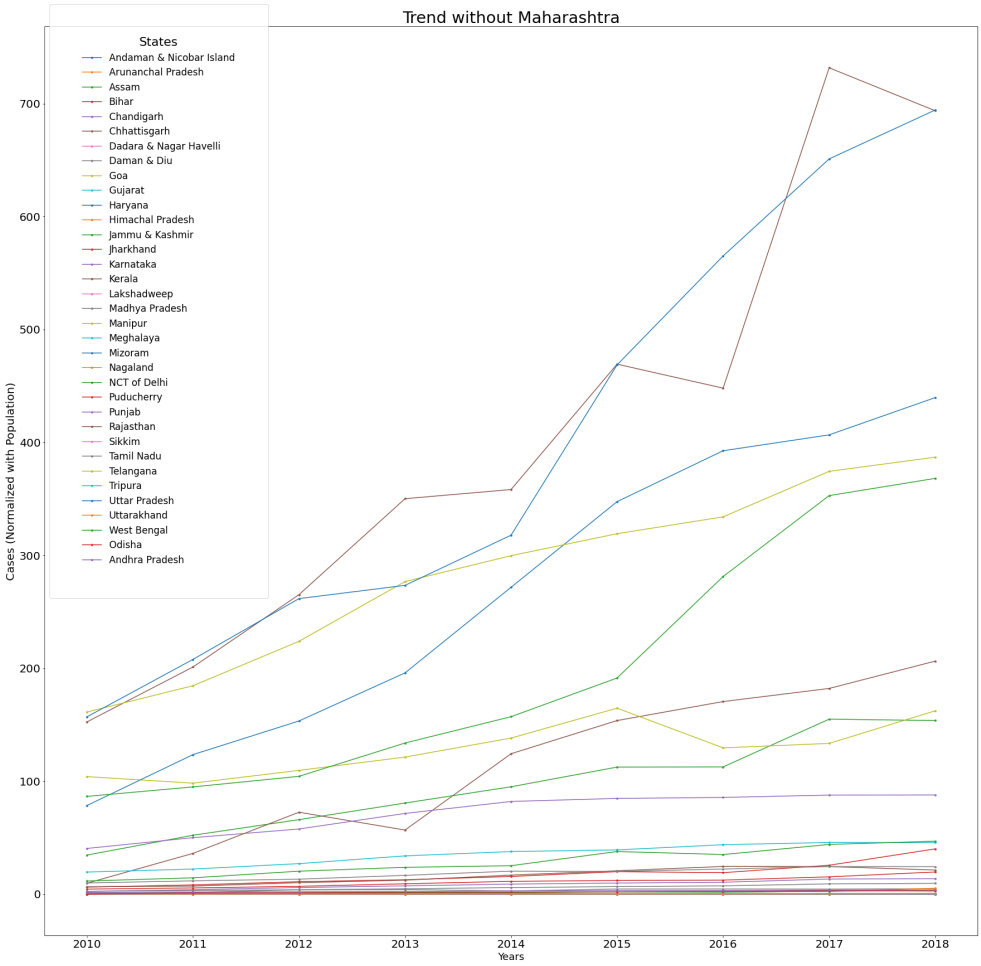


Cases filed over the years for states normalized with Population (2011)

From this trend set we see Maharashtra being a complete outlier.

Although this may also be due to the dataset imbalance towards Maharashtra as well.

Hence we remove Maharashtra from the DataFrame and plot the states again for a better understanding of other states.



Cases filed over the years for states normalized with Population (2011) (without Maharashtra)

Now we get a better picture of the trend with Uttar Pradesh and Rajasthan having the maximum number of cases filed per member of the state.

It is also found that the North-East States (like Tripura, Manipur, Mizoram) have a very low number of cases filed per member of their state.

4. Analysis on Gender Crimes

Extensive analysis has been done for gender crimes like (dowry, domestic violence, matrimonial issues, etc…) for all the states and over the years 2013-2018.

Questions

1. What is the State Wise Distribution of the cases involving gender crimes over the years?

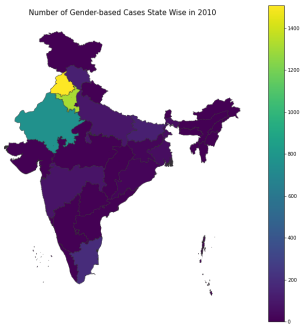
2. Which gender is more likely to be the alleged in such cases and what is its trend over the years?

3. Which gender is more likely to file the case in such cases and what is its trend over the years?

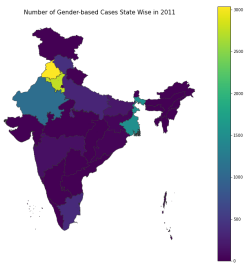
4. What is the usual outcome of such cases or what is the verdict distribution of such cases?

State Wise Trend

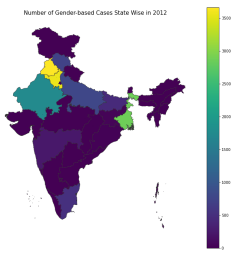
Over the Years 2010-2018



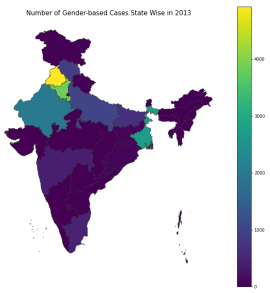
2010



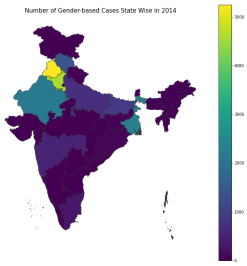
2011



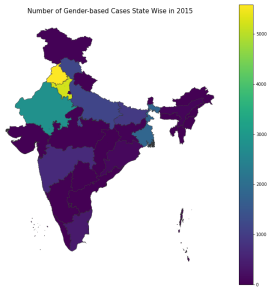
2012



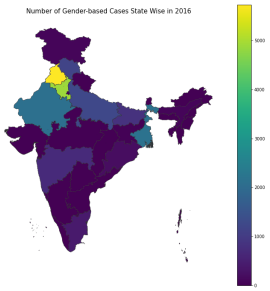
2013



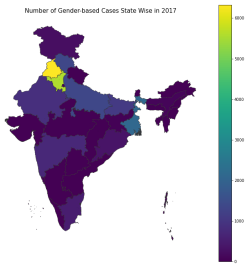
2014



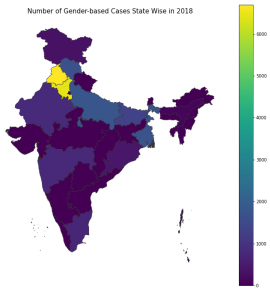
2015



2016



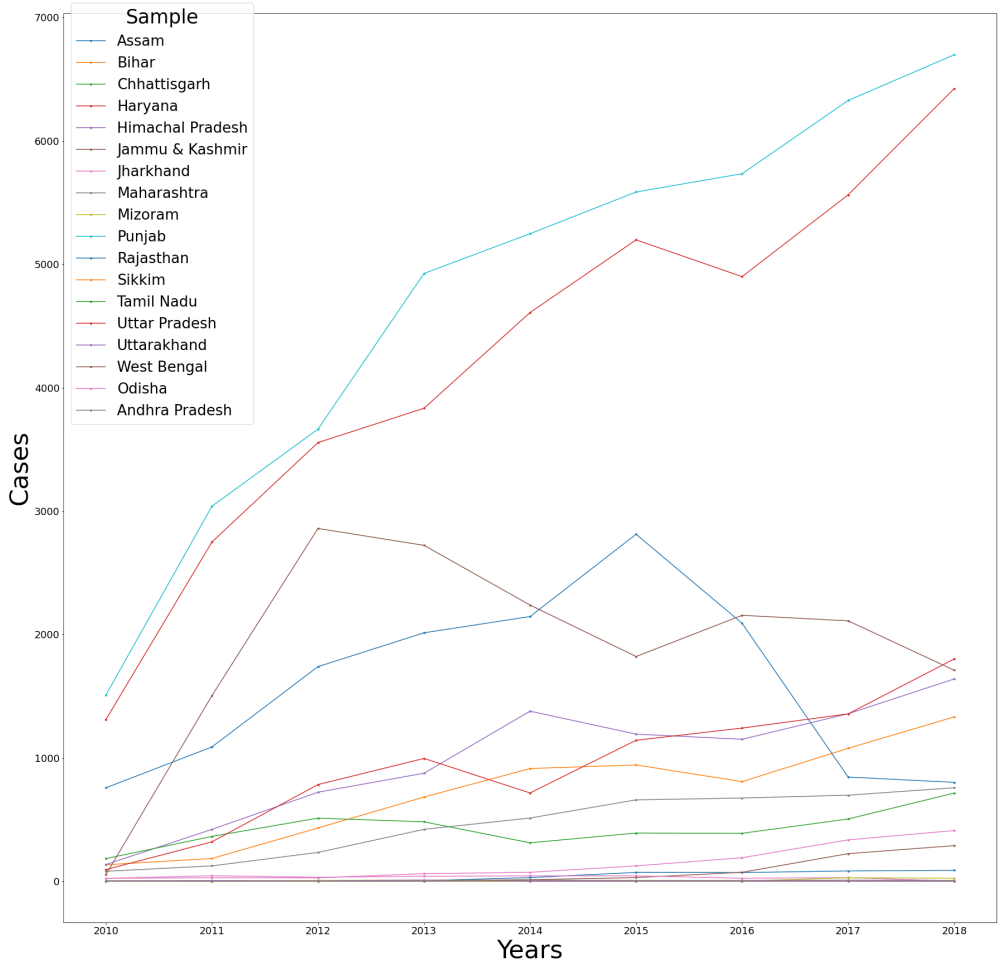
2017



2018

We see a general trend that Punjab, Haryana, Rajasthan, West Bengal and Uttar Pradesh have the highest number of cases filed based on Gender Crime over the Years.

Cumulative Trend



It is observed that Punjab and Haryana have a rapid and steep increase in the number of such crimes over the years. This may be due to the cultural history of these states involving the customs of dowry and other gender-oppressive customs.

Hence we see a huge number of cases filed in these states and at a high number, increasing continuously despite various gender-reforms and acts gender reforms.

Also, we see some states like Rajasthan and West Bengal showed a rapid increase in such cases in the middle years (2012-2015) but are reducing as well in recent years.

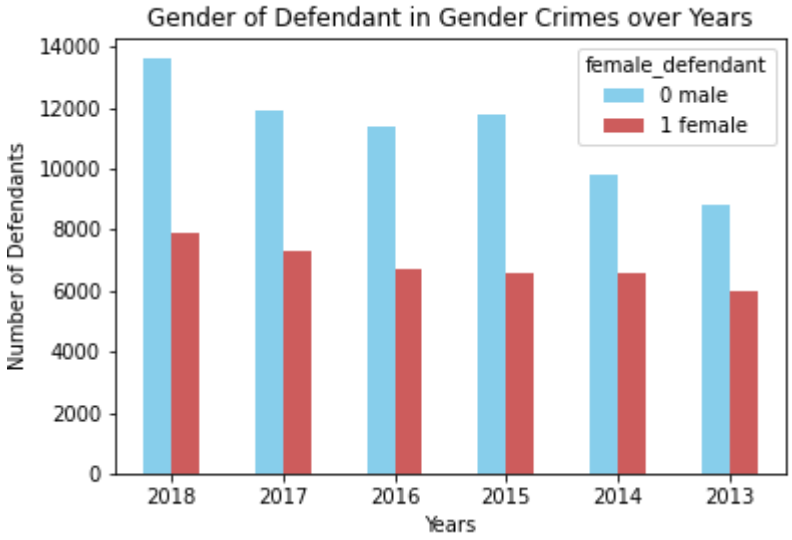
In general, we can say that these cases are rising over the years for most of the states but this may also be due to the fact that more and more people are becoming aware and responsible

for their rights and have been gradually given access to relatively fast judgments (especially after the advent of Fast-track Courts).

Hence they are becoming more vigilant and are becoming empowered to reach out to courts for crimes/discrimination against them.

Petitioner/Defendant Gender Distribution

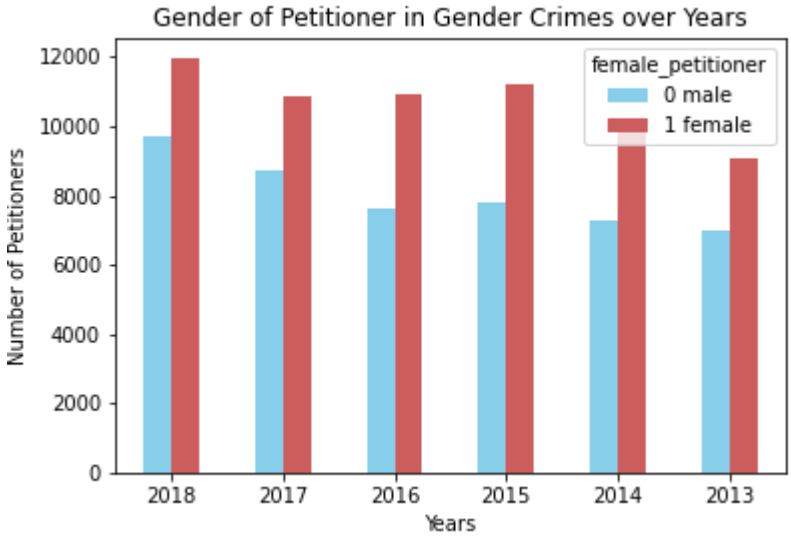
Defendant Distribution



Note: The years are decreasing on the x-axis

It is observed that there are a seemingly high number of male defendants in such cases. That is much more cases are filed on males than on females and about twice as many males are accused than females, which is increasing over the years.

Petitioner Distribution



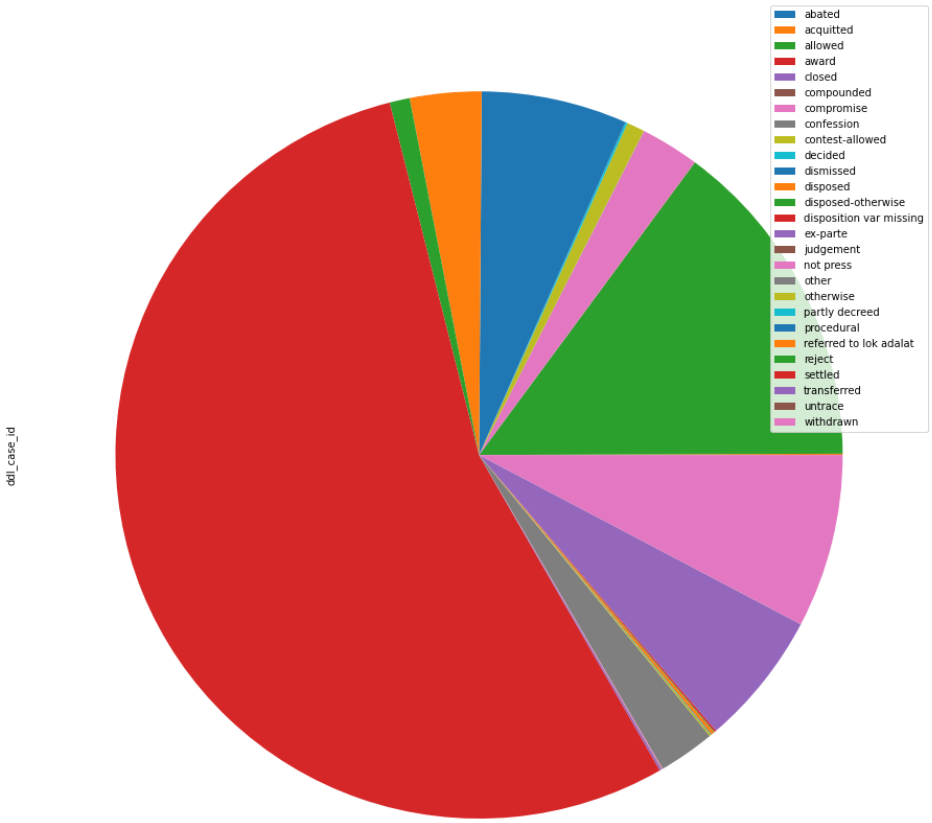
Note: The years are decreasing on the x-axis

It is observed that there are a seemingly high number of female petitioners in such cases. That is much more cases are filed by females than by males, which is also increasing over the years.

The difference between the genders filing the case is not as much as we see in defendants because, in many regions of India, families are still headed by male heads of the family.

Hence, many cases of crimes against females might also be filed by male heads of the family. But still in general we see females becoming more vigilant and empowered over the years to stand up and raise their voices for crimes committed against them.

Outcomes of Gender related Cases



It is observed that most of the cases have missing dispositions as an outcome in such cases which might be due to the slow and inefficient working of the system and also the attitude of the police and state towards such cases. Many such cases are diluted or dismissed due to prevalent corruption in the system. Many such cases are also diluted due to the condition of family honor and the risk of diminishing the dignity of the accused and the victim.

It is also observed that many cases result in the ‘offering of allowance’ to the victim by the accused under the category ‘contested and allowed’ or ‘allowed’.

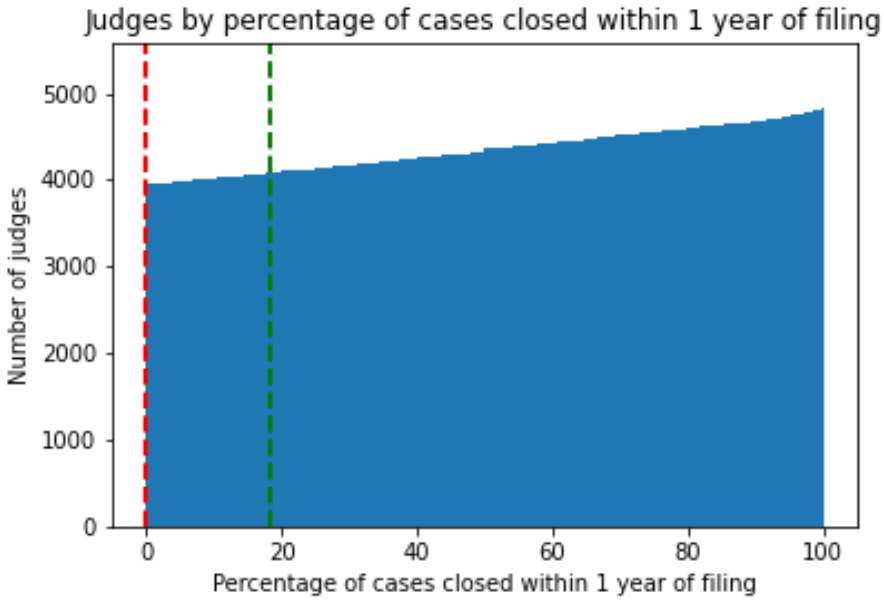
Some cases are also withdrawn due to the same dignity reasons cited above.

5. Duration Analysis

Question

How many judges are efficient enough to close the case within 1 year and how many take 5 years to close one?

Cases closed Within 1 Year

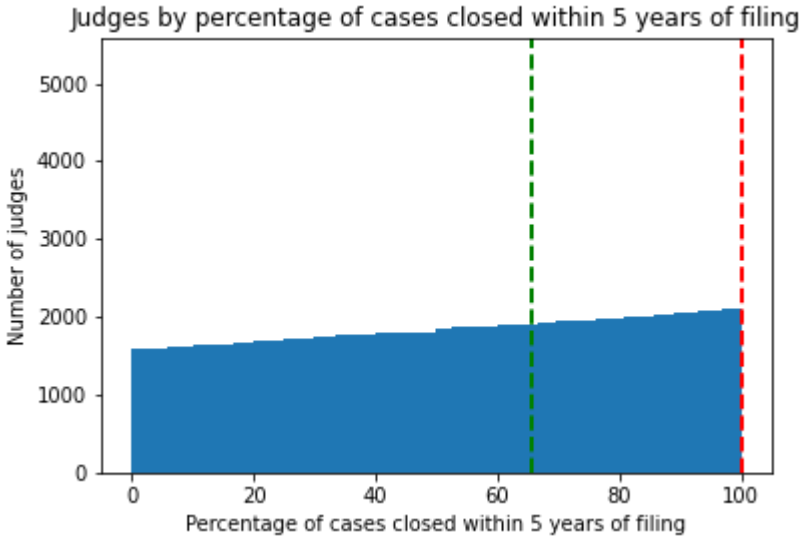


Number of Judges v/s Percentage of Cases closed within 1 year

Green Line - Mean

Red Line - Median

Cases closed within 5 Years



Number of Judges v/s Percentage of Cases closed within 5 years

Green Line - Mean

Red Line - Median

The cumulative histograms show the Number of Judges v/s Percentage of Cases closing within 1 or 5 Years of filing.

It is observed that most cases are assigned a verdict within 1 year of filing according to the dataset. This huge imbalance in the dataset of many cases falling within 1 Year category affects the training model of the Duration Classifier built later on, which is fixed by adding class weights to the Random Forest Classifier.

Classifiers

Much analysis was done to render and select only specific columns for training the model.

Features like ‘female_defendant’, ‘female_petitioner’ had a lot of NaN values which reduced the training rows by roughly 80%. Hence these features were needed to be discarded. They also had very less correlation to the target variable as well.

Features like ‘female_defendant_adv’ and ‘female_petitioner_adv’ had an extremely less correlation to the target variable and presented with a similar problem mentioned above.

Features like ‘criminal’ and ‘bailable_ipc’ presented with the issue of much NaN values as well as homogeneity. Non Null values in these cells were almost the same for a variety of data and over the range of categories of the target variable. Hence they were discarded too.

The dataset then presented with the problem of huge imbalance and skewness towards one of the categories in both of the classifiers.

Initially for the Duration Classifier a score of 0.82 was obtained but the Confusion Matrix showed that most of the classification was done in the category ‘Less than 1 year’ and other categories even had 0 True Positives. This although with a decently high score presented with a bad classification.

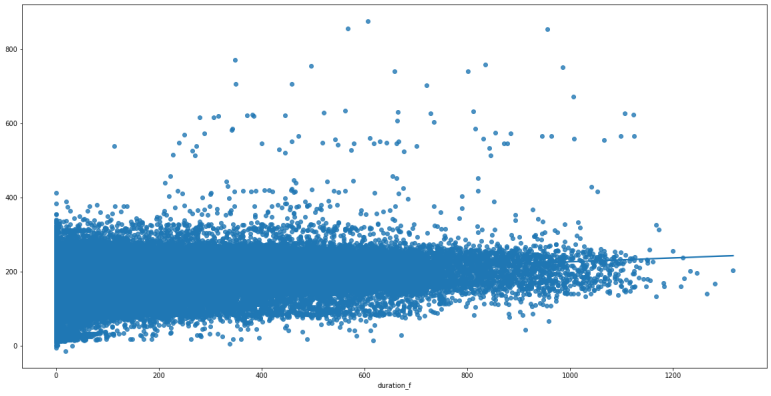
Hence, Class Weights were added to the classifier which were inversely proportional to the number of Testing rows in each Target Category for each subtree formed, using class_weights = ‘subsample’.

This although reduced the score but presented with a much more balanced classification.

Many Classification models were tried like, ‘Logistic Regression’, ‘K-Nearest Neighbours’, ‘Decision Tree Classifier’ for the Classification Problem each of which had a score roughly averaging to 0.42. But the best classifier turned out to be RandomForestClassifier.

After much analysis for the n_estimators an ideal value was chosen based on Binary Search.

For the duration classifier the initial model was to have integer outputs and build a Linear Regression model for the same.



This was roughly the output for the model. This had a Maximum Absolute Error (MAE) for the number of days for the case to close of **153.82.** But due to such a scattered dataset this model turned out to be inefficient for the purpose and hence, categories of duration were made and the approach changed from a regression model to a classification model.

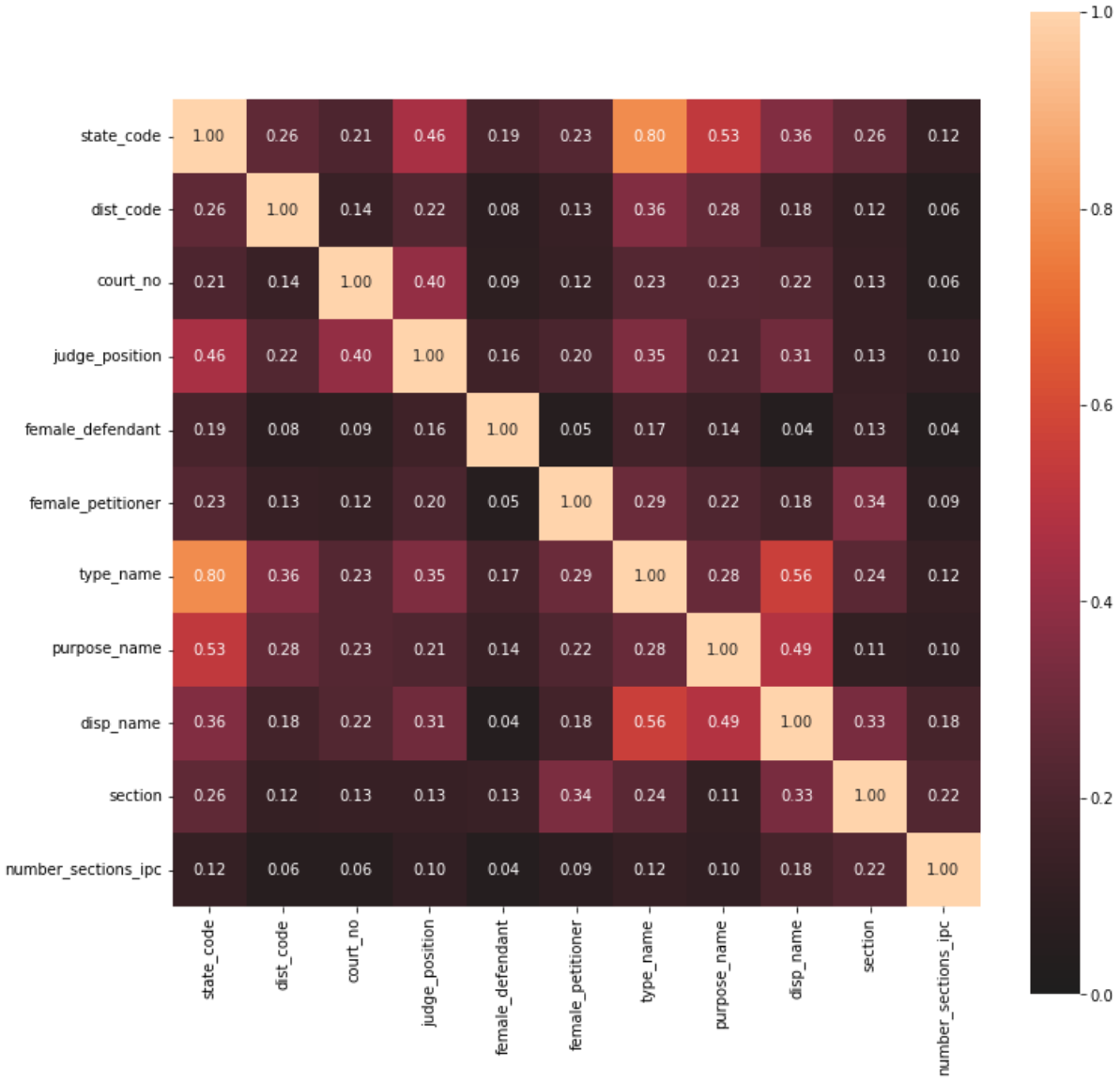
1. Outcome Classifier

A Classifier is built to predict the Outcome of the Cases under selected major categories namely - 'acquitted', 'bail granted', 'bail refused', 'bail rejected', 'convicted', 'dismissed', and 'settled'.

The dataset is pre-processed by removing the Null values and filtering non-relevant columns based on the correlation matrix and heatmap visualizations.

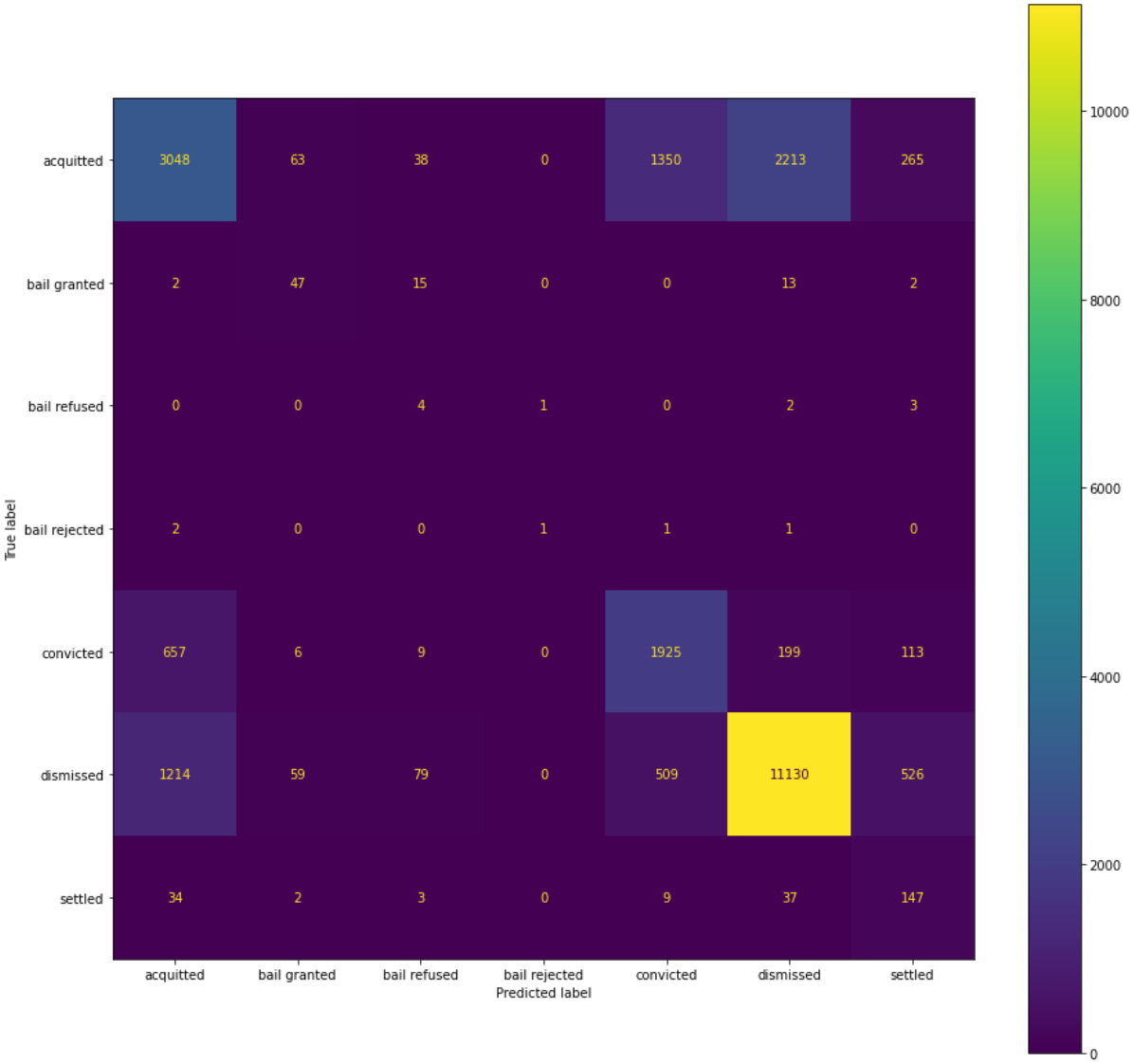
Dython Library is used for visualizing the heatmap and correlation matrix for 'nominal' variables.

Heatmap



- It is observed that the highest correlated variables to the target variable ‘disp_name’ turn out to be ‘type_name’ and ‘purpose_name’.
- The categories with strong and definitive outcome like ‘Dismissed’, ‘Acquitted’, ‘Convicted’ have a high True Positive. This indicates that the features affecting the outcome have a specific class or category which results in a definitive result.
- Like a Murder Case is very less likely to fall into a positive category associated with Bail, due to the ‘puspose_name’ , ‘type_name’ and ‘section’ associated with it. It’s highly likely to result in ‘Acquitted’ or ‘Convicted’.
- Hence the type and purpose of the case affect the outcome of the case majorly.

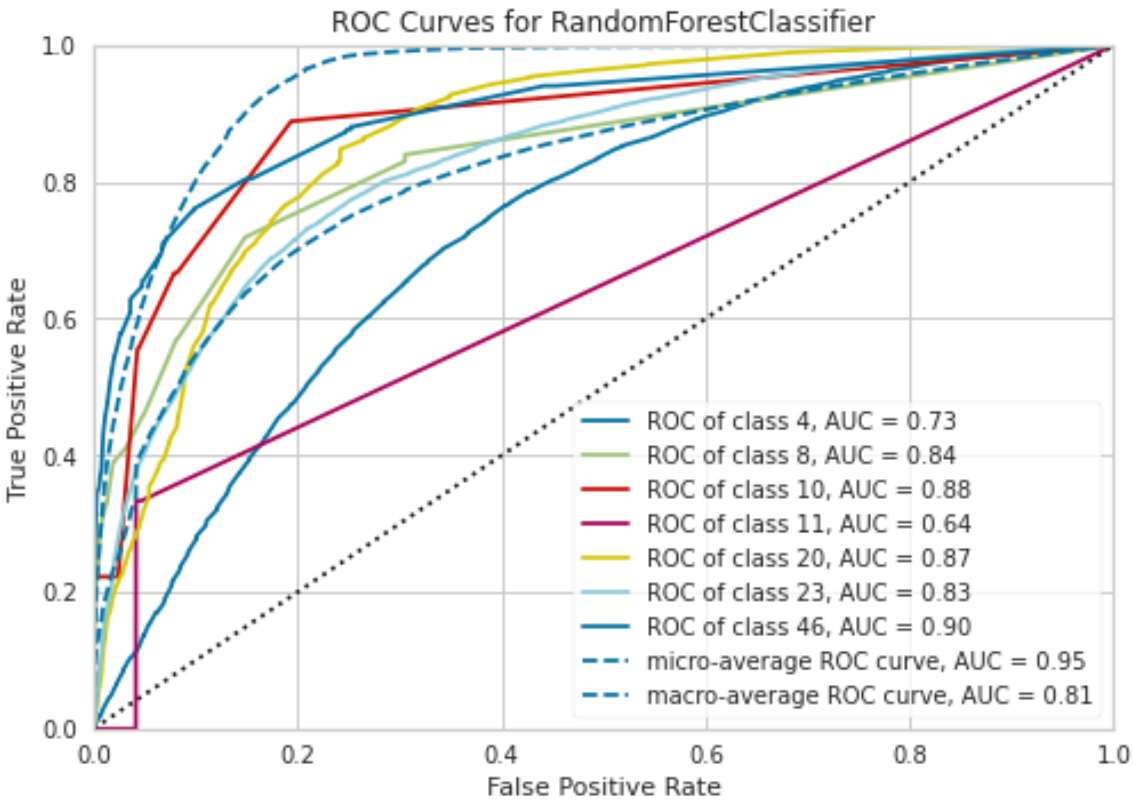
Confusion Matrix



The Confusion Matrix for the model is visualized with True Labels on the vertical axis and Predicted Labels on the horizontal axis.

Class Weights have been added to the Classifier due to dataset imbalance.

ROC Curves



Outcome of Case	Code
Acquitted	4
Bail Granted	8
Bail Refused	10
Bail Rejected	11
Convicted	20
Dismissed	23
Settled	46

The Score for the Classifier comes to be 0.6897045808925787.

2. Duration Classifier

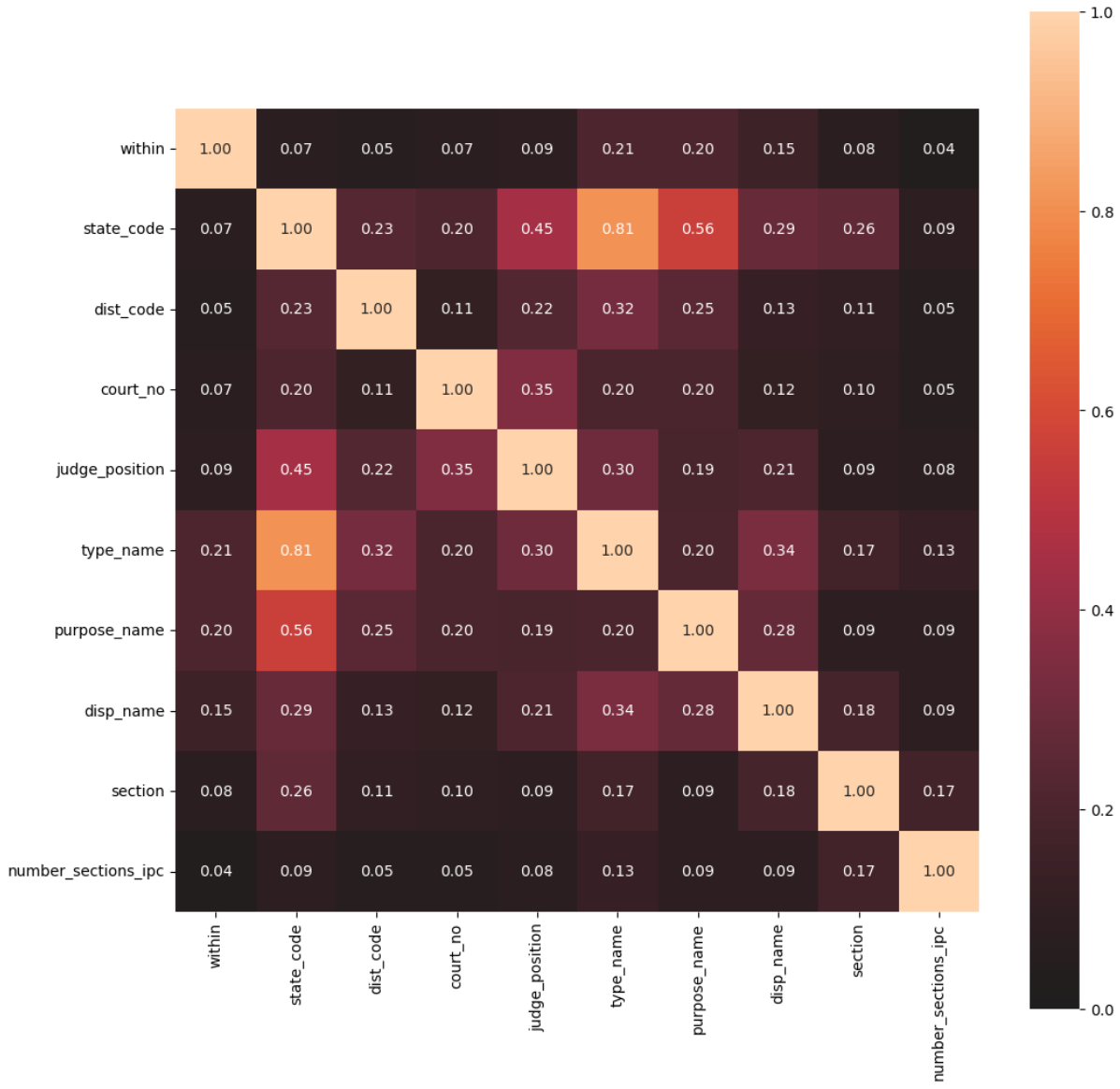
A Classifier is built to predict the Duration of the Cases which is segmented under selected categories namely - 'Within 1 Year’, ‘Between 1-2 Years’, ‘Between 2-3 Years’, ‘Between 3-4 Years’, ‘More than 4 Years’.

The dataset is pre-processed by removing the Null values and filtering non-relevant columns based on the correlation matrix and heatmap visualizations.

The gender of the petitioner and defendant has been removed as a feature as it contained many unknown and unclear values reducing the rows in the dataset heavily.

Dython Library is used for visualizing the heatmap and correlation matrix for ‘nominal’ variables.

Heatmap



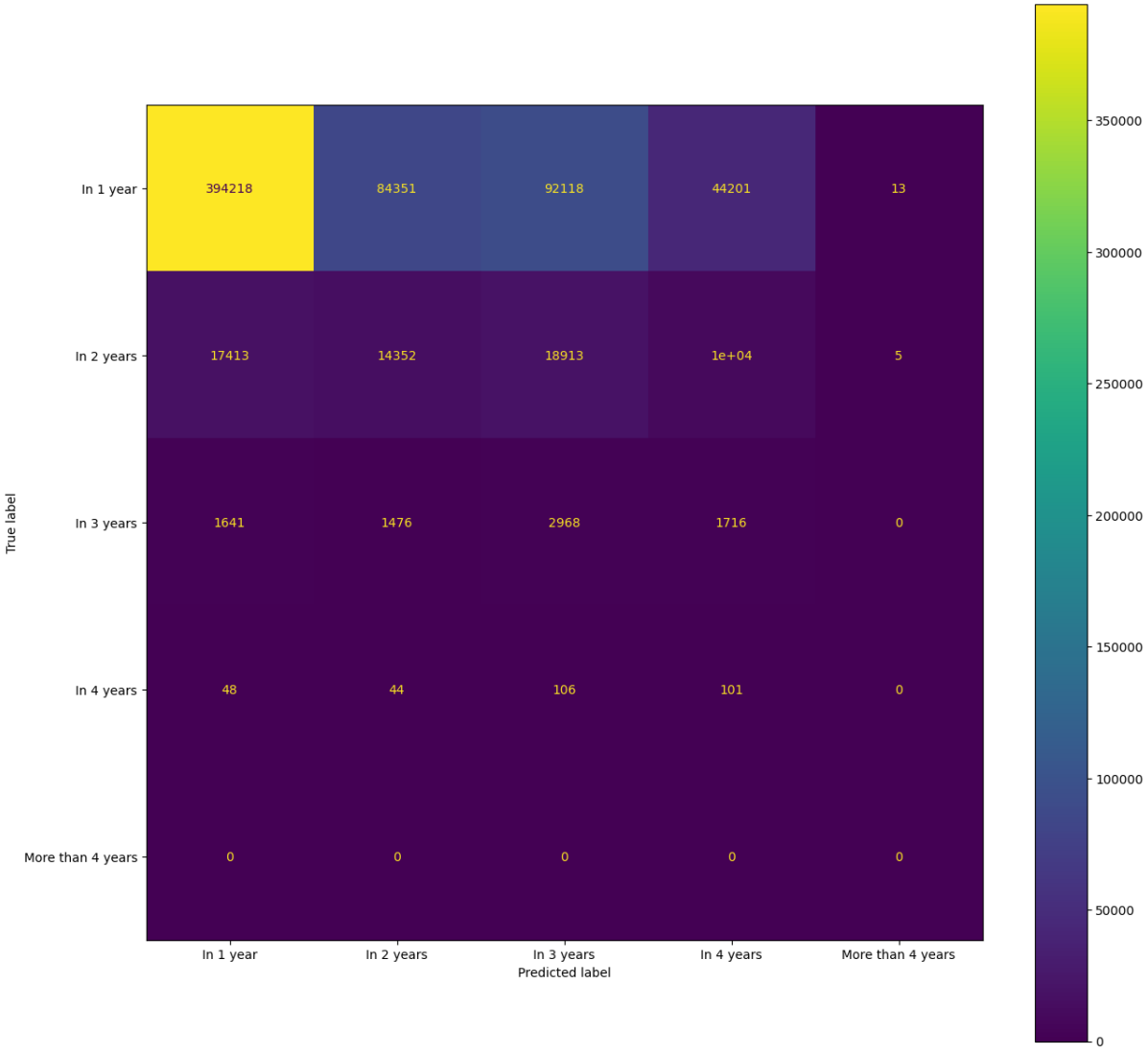
- It is observed that the highest correlated variables to the target variable ‘within’ turn out to be ‘type_name’ and ‘purpose_name’.
- Based on the Confusion Matrix, most cases are likely to be solved within 1 or 2 years.

  This does not very much depend on the state which the court belongs to.
- This highly depends on what type of case it is and what it resulted in.
- Like the cases of bail applications usually finish earlier than those of Murder Charges.

  Hence the type and purpose of the case affect the duration of the case majorly.

- To the contrary, it is also found that the number of sections of IPC imposed on the case does not heavily impact its closing duration.

Confusion Matrix



The Confusion Matrix for the model is visualized with True Labels on the vertical axis and Predicted Labels on the horizontal axis.

Class Weights have been added to the Classifier due to dataset imbalance.

The score for the Classifier on the Test Set is found out to be 0.6020534571647957.