Vi4: Crime Against Women – An Analysis

Tanvi Rajesh
Department of Computer Science and Engineering
PES University Bengaluru, India
tanvi.rajesh09@gmail.com

Vibha Murthy
Department of Computer Science and Engineering
PES University Bengaluru, India
vibha.harsha@gmail.com

Vidisha Chandra
Department of Computer Science and Engineering
PES University Bengaluru, India
vidishasateesh@gmail.com

Vishakha Hegde
Department of Computer Science and Engineering
PES University Bengaluru, India
vishakhahegde23@gmail.com

Abstract*— Statistics show beyond doubt that there has been a dramatic increase in violence against women in India in the past decades. Violence against women is a deep-rooted problem dating back to the 1900s. This includes a wide spectrum of crimes from mild teasing to rape and murder. These incidents take place in the streets, at workplaces, as well as their homes, in short everywhere. Women are neither safe in private nor in the public sphere. It is alarming that many crimes against women go unreported, fewer prosecuted, and only a small percentage of those accused get punished. With this paper, we have attempted to gain a better understanding of various crimes committed against women, and their demographics to bring to light the unseen correlations between different factors like the States of India, type of crime, reports filed, and cases convicted. Methodologically, this paper is based on data extracted from NCRB reports and supplemented by other published resources such as books, articles, and reports of governmental/ non-governmental organizations

***Keywords*—Crime, Abuse, Crime against women, Crime analysis, Cases, Conviction**

I. INTRODUCTION

Any kind of violence against women is a clear violation of the Fundamental Rights of women under Articles 14 and 15 of the Constitution of India. Despite innumerable efforts being made to maintain the rights, dignity, and well-being of women, the women of the patriarchal society of India are continually oppressed and are victims of various kinds of abuse, both within and outside their homes.

In recent years, there has been an alarming increase in the number of crimes committed, especially against women. Despite the seemingly increasing awareness concerning violence against women, cases of rape, kidnapping and abduction, dowry deaths, molestation, sexual harassment, cruelty by husbands and relatives, importation of girls, and trafficking continue to see an increase. This has far-reaching consequences on both the physical as well as mental well-being of women. Most cases of violence go unreported, but even the far and few ones that are reported do not get the required timely justice with the existing judicial system.

Of the six million crimes that police in India recorded between January and December of 2021, 428,278 cases involved crimes against women. Such alarming figures suggest that it is imperative to take stringent action immediately and bring justice to these affected women while increasing awareness about the same.

Thus, there is a need to better understand the correlation between factors such as type of crime and number of persons convicted to pinpoint where India's legal regime is lacking.

Gender discrimination has always been a major cultural and socio-economic issue in the Indian backdrop. Being one of the fastest-growing nations in the world India should aim to be ranked higher in the global gender gap index. Such studies that focus on analyzing crimes not only assist in a better perception of the safety and welfare of women of the nation but also help in elucidating the overall socio-demographics of the country.

With the data shown by the National Crime Records Bureau, the records of crimes reported against women have turned out to be higher than in the past few years. Cases of rape, murder, abduction, and trafficking are occurring much more frequently now. The government is trying to implement stricter laws and serious measures to prevent such crimes and ensure the safety of women. A huge amount of data is generated every year related to such different crimes in different regions of India. Analyzing this data is a task which can be handled with emerging methodologies, Data mining plays a huge role in the analysis of a large number of records, computing accurate results and revealing underlying patterns. Utilization of data at hand to predict future outcomes and probabilities of occurrences of crimes is an important aspect to address and may prove extremely beneficial in the future.

The absence of detailed studies on this issue has hampered the spread of awareness and proper understanding of the issue, this makes the issue more complex making it difficult to provide effective solutions to reduce if not abolish such violence. These instances of violence need to be thoroughly investigated so that ways and means can be devised to reduce their incidence.

II. RELATED WORK

Many papers and works have delved into topics related to crime in India. These include specific crime as well as crime deterrence and the need for reforms. We considered the studies that mainly focus on Crime Against Women in India. We observed a lot of the studies focused on the theoretical aspects related to women and crime whereas deeper state/district-level studies are lacking in terms of practical approaches and reports. Similar studies have been conducted that focus on predicting the positions of various states and regions in India considering the types of crime and factors that affect the crime rate. These factors include economic growth and status, education, and family background. Few papers also provide the factors which may improve this dire situation. A study suggests that in cases of reported violence against women, charges are dropped in many cases and only more severe assaults are prosecuted more vigorously.

R.N. Mangoli et.al [1] mainly focuses on analyzing rapidly increasing crime trend and where the working of the Government fails to handle these crimes. With the help of statistical reviewing and analysis, it aims to understand the current laws in India pertaining to combating such crimes.

“Crime against Women in India: An analysis”, a paper published by Gilani et al. [2] in the Journal of Society in Kashmir attempted to scrutinize various types of crimes committed against women and their consequences. It also aims to throw light on the correlation between law, crime, and women. This paper used a mixed approach to analyze the variety of crimes committed against women in India.

Another paper, ‘Crime against Women in India: A State Level Analysis’ was referred by us. Employing panel regression technique, this paper aims to identify the factors which can control crime against women. Chakraborty et al. [3] suggested that the states are harbouring different crimes against women, with a significant number of offences being related to dowry. The objective of this paper is split into 3 parts: (a) Studying trends in Sexual crimes, Offence related to Dowry and Other Crimes against women, (b) Understanding the relative position of the different states with respect to these crimes (c) Speculating methods to reduce the Crime rate.
The paper concluded that parental guidance and education can reduce crime against women. An interesting finding is that economic growth can initially encourage crime against women, but with further growth, crime may fall.

Our study uses a miscellaneous approach towards the same problem. By analyzing various crimes committed against women in India through the years 2001-2010 the trained model will predict the likelihood of arrests post-filing cases related to crimes committed against women for a given state. Also, considering discrepancies that may have occurred in the collection, documentation, and storage of the data due to factors like negligence, intimidation, cases being withdrawn etc., the model will shed light on the variation in cases initially filed vs cases pursued.
The study is based on quantitative data obtained from Kaggle and Crime in India reports prepared by National Crime Records Bureau, India.

III. ANALYSIS OF DATASET

A. *Datasets*

The datasets used in this project are taken from Kaggle
and are available at the following link: https://www.kaggle.com/datasets/rajanand/crime-in-india?resource=download&select=43_Arrests_under_crime_against_women.csv
https://www.kaggle.com/datasets/rajanand/crime-in-india?resource=download&select=42_Cases_under_crime_against_women.csv
This consists of two .csv files:
Cases_under_crime_against_women.csv and
Arrests_under_crime_against_women.csv.
The respective datasets are preprocessed and merged to form the final dataset.
Details of datasets:
The file Cases_under_crime_against_women.csv contains the details of cases under crimes against women in India from 2001 to 2010. This dataset has a total of 22 columns and 4165 rows.
The file Arrests_under_crime_against_women.csv contains details about arrests under crimes committed against women in India from 2001 to 2010. This dataset has a total of 16 columns and 4165 rows.
After merging the final dataset has a total of 34 columns and 4165 rows.

B. *Preprocessing*

Data obtained usually needs to undergo certain processes so that it can be a better fit for further analysis. These processes together constitute the preprocessing part. Data attributes may be incomplete, missing, or erroneous. Noise refers to errors and outliers which corrupt the data and if not handled properly may affect the predictions made. Throughout the process of preprocessing the raw data is transformed into a more comprehensible format by proper handling and treatment of all the above cases and more if present. The main steps under preprocessing are data cleaning, transformation, and reduction.

On analyzing both datasets there were no null values found. In dataset 1 (Cases_under_crime_against_women), we observed that Group_Name gives 11 values whereas Sub_Group_Name gives 12. This was due to an error in the dataset where ‘Importation of Girls’ was added as the Group_Name for the Sub_Group_Name ‘Dowry Prohibition Act’. On analyzing dataset 2 (Arrests_under_crime_against_women), no such error was found. To prevent any discrepancy in the merged data, we decided to drop the Group_Name column in both datasets.

It was also observed that dataset 1 had the Sub_Group_Name as ‘09. Dowry Prohibition Act’ and dataset 2 had Sub_Group_Name as ‘09. Dowry Prohibition’. To avoid any loss of data, all such values in dataset 2 are replaced with ‘09. Dowry Prohibition Act’.
We merge both datasets to get a combined ‘Crimes_against_women.csv’.

On conducting further analysis an anomaly was observed between the data entries 8 to 11 under Sub_Group_Name, rows with these entries mainly involved crimes like Sati. The other column entries for these values were 0 for the most part as the following crimes have been abolished back in the 1900s. Hence, the rows were dropped.

A similar process was applied for various columns as well. Columns with more than 50% null values were dropped.

C. *EDA and Visualization*

It was essential to understand the data and its features; provided in the dataset as it will play an important role in our model. For this, we have performed Exploratory Data Analysis and Visualization to spot the important features, elicit patterns and distinguish underlying correlations. We have used Python and its libraries such as Matplotlib, seaborn and pandas.

Figures 1a and 1b. Depict the number of cases of violence reported over the years 2001-2010. We can see that there is a steep increase in kidnapping reports and family cruelty cases. Rape and molestation also see a steady surge. Dowry deaths seem to be a relatively smaller number. Although sexual harassment cases appear to see a decrease, it could be that many women are not coming forward to report such cases. It looks as though people are more comfortable reporting kidnappings as opposed to crimes related to sexual violence. This can be linked to the general stigma in society revolving around sexual activities.

Figure 1a

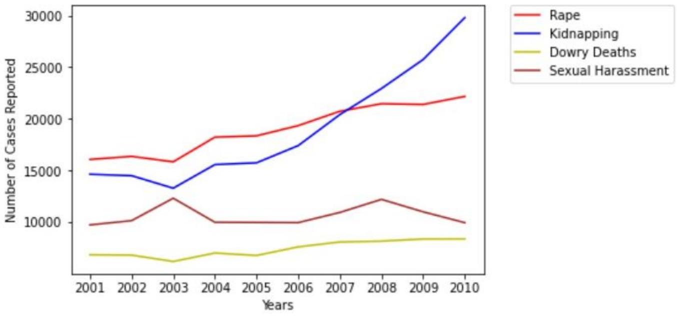


Figure 1b

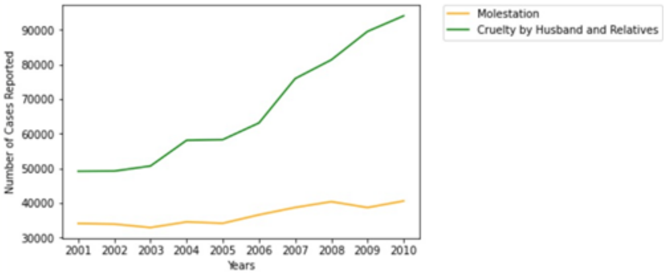
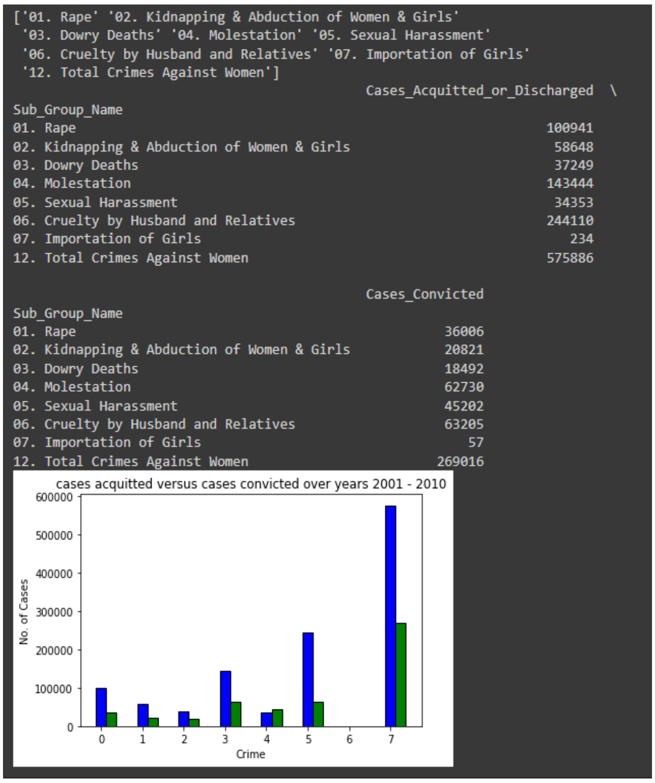


Figure 2. Shows the number of cases acquitted versus cases convicted over years 2001 - 2010. The x-axis represents the type of crime, and the y-axis gives us the number of cases. From this, it can be observed that the importation of girls (bar labelled 6 in the chart) has a relatively smaller number of cases when compared to the other crimes.

Figure 2



```
['01. Rape' '02. Kidnapping & Abduction of Women & Girls'
 '03. Dowry Deaths' '04. Molestation' '05. Sexual Harassment'
 '06. Cruelty by Husband and Relatives' '07. Importation of Girls'
 '12. Total Crimes Against Women']
                                              Cases_Acquitted_or_Discharged  \
Sub_Group_Name
01. Rape                                                             100941
02. Kidnapping & Abduction of Women & Girls                           58648
03. Dowry Deaths                                                      37249
04. Molestation                                                      143444
05. Sexual Harassment                                                 34353
06. Cruelty by Husband and Relatives                                 244110
07. Importation of Girls                                                234
12. Total Crimes Against Women                                       575886

                                              Cases_Convicted
Sub_Group_Name
01. Rape                                                36006
02. Kidnapping & Abduction of Women & Girls             20821
03. Dowry Deaths                                        18492
04. Molestation                                         62730
05. Sexual Harassment                                   45202
06. Cruelty by Husband and Relatives                    63205
07. Importation of Girls                                   57
12. Total Crimes Against Women                         269016
```

Figure 3. Shows us that the number of Cases_Reported has increased over the years however, so has the crime.

Figure 3

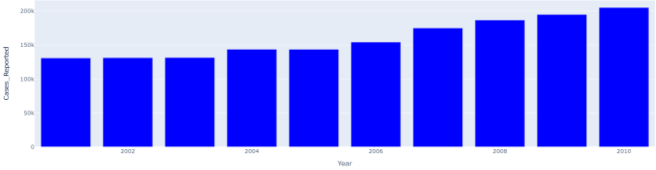
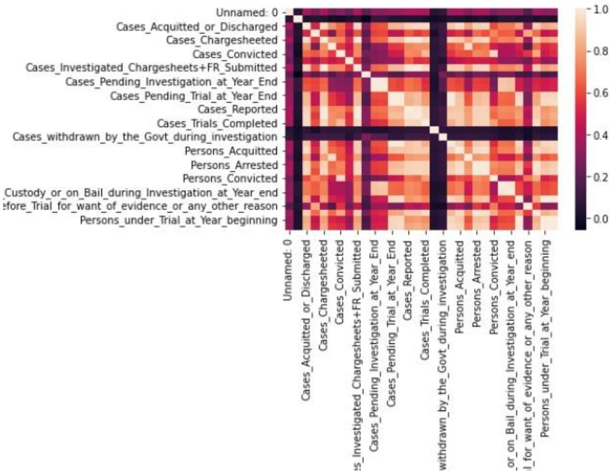


Figure 4. Depicts a heatmap showing the correlation matrix for the features in the dataset. A dataframe containing only positive correlations above 0.9 was constructed to analyze the heatmap better. This depicts few highly correlated features as well as interactions between several factors such as the Cases_Investigated_Chargesheets+FR_Submitted and Cases_Sent_for_Trial (0.997099) Cases_Pending_Trial_from_the_previous_year and Persons_in_Custody_or_on_Bail_during_Trial_at_Year_End (0.955157) Cases_Sent_for_Trial and Cases_Reported (0.979532)

Figure 4



IV. PROBLEM STATEMENT

To study the trends in crime against women state-wise and bring to light gaps in the judicial workings that handle such cases by analyzing patterns in data regarding cases across 10 years and predicting important aspects such as future state-wise crimes reported, probability of seeing progress in cases, the efficiency of various states in handling cases.

Our project will comprise looking into several aspects of crime against women and trying to find patterns and then use these patterns to predict the cases reported for a particular crime given the type of crime, state name and the year. Furthermore, predicting the Cases Convicted considering all the parameters mentioned above and using the predicted cases reported as a hidden variable. With this, we hope to point out how the Indian judicial system has evolved as well as how much awareness has been created with respect to reporting these atrocities and seeking justice. This sort of data can prove to be useful in deploying police forces, investing in surveillance equipment etc.

V. PROPOSED METHODOLOGY

The dataset contains both numerical as well as categorical values. The categorical values have been encoded using various encoding techniques as per the requirement of the models implemented. The dataset is divided into testing and training sections individually for each model to avoid any discrepancies. The problem at hand has 2 subparts:

A. Predicting the Cases Reported given the Area_Name (name of state), Year and Sub_Group_Name (type of crime)

B. Predicting the Cases Convicted given the same parameters and Cases Reported as a hidden variable.

To solve the above-mentioned problem statements, we intend to use several different models, the explanation of the models and the experimental results of the same have been recorded.

We decided to use the predicted values for Cases Reported in A. as input to further predict the Cases Convicted. The idea behind using this approach was to acknowledge the fact that we may have access to only basic data such as State, Type of Crime, and Year, in a real-life scenario and hence, are trying to make valid and accurate predictions using only this data.

A. *Models Implemented for Part A*

1. *MLR*

Due to many correlations, we decided to start with a

In multiple linear regression (MLR) a response variable y is measured, which can be explained as a linear combination of the x-variables which are linearly independent.

On plotting the correlation matrix Figure 4, we noticed a high correlation between the Cases_Reported and several other features. We predicted the Cases_Reported using the Area_Name, Year, Sub_Group_Name and the Cases_Sent_for_Trial.

Since we are dealing with categorical data(Area_Name and Year, Sub_Group_Name) whose features are crucial to make predictions. Those non-numeric values are treated as dummy variables using encoding methods such as label encoding.

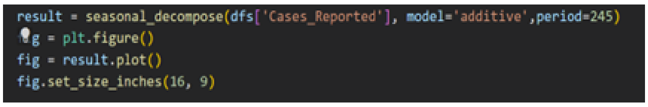
We considered 90% as the training data and tested on the remaining 10% post-encoding the columns.

2. *SARIMAX*

Figure 1a & 1b. On decomposing the data to check the trend and seasonality it was observed that both were present.

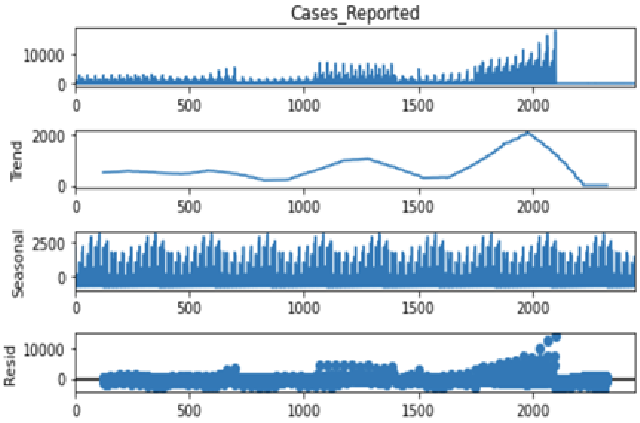
An additive model was used to decompose the data due to the presence of 0 entries in the data. The period was set to 245 as it is the number of entries per year.

Figure 1a.



```
result = seasonal_decompose(dfs['Cases_Reported'], model='additive',period=245)
fig = plt.figure()
fig = result.plot()
fig.set_size_inches(16, 9)
```

Figure 1b.



Label encoding was used to encode the categorical values of Area_Name (0-34) and Sub_Group_Name (0-6). SARIMAX was first implemented with the three main features as the exogenous variables. However, due to poor results, Cases_Sent_for_Trial was also included in the set of exogenous variables as it is strongly correlated with Cases Reported, the variable to be predicted.

The data was split into training and testing parts in a 70-30 ratio. With the help of the auto Arima the values for p,q,d (0,0,1) were estimated and the same was used in the final model prediction.

3. *RANDOM FOREST*

Random forests are an ensemble learning method for classification, regression and various other tasks.

A common problem with decision trees alone is overfitting, random forests reduce overfitting since the final classification occurs through majority voting and regression problems return the mean of the individual trees.

a) *One-Hot Encoding*

One hot encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

The 2 categorical features, Area_Name and Sub_Group_Name and are then split into 35 and 7 columns respectively with 0s and 1s.

b) *Label Encoding*

There are several methods to deal with categorical data. Here we go for the simplest method where every category is given a unique integer based on alphabetical ordering, known as label encoding.

We started by label encoding the 2 features of Sub_Group_Name and Area_Name.

B. *Model Implemented for Part B*

1. *ARIMAX*

Arima is a statistical analysis model that is used to model time series data and predict future trends. Arima can handle trends in data and provide an accurate forecast of the future based on these trends.

Our data consists of the year and the type of crimes occurring in each state in that year. Therefore, we decided to implement this model. The data was grouped by state followed by a group by type of crime in order to predict the number of cases convicted in that state. Some exogenous variables used were Cases_Trials_Completed and Persons_Convicted.

2. *KNN*

K Nearest Neighbors is a supervised learning classifier which makes classifications and predictions based on the distance between points in the dataset. For regression problems, KNN uses feature similarity to predict values of new data points.

KNN works well when the dataset size is smaller and progressively becomes slower as the dataset grows. But since our dataset was relatively smaller and non-linear with no assumptions about underlying data, we decided to use KNN. We used Euclidean distance as the measure the calculate the distance between each data point and K-nearest neighbours. The KNeighboursRegressor function from sklearn was used with K=5 found using hit-and-trial method.

3. *MLR and LASSO Regression*

On plotting the correlation matrix (Figure 4), we found some variables which were highly correlated with ‘Cases Convicted’ and hence decided to use regression analysis to predict new values. Multiple Linear Regression estimates the relationship between one dependent variable and multiple independent variables. Lasso regression is a modification of this which uses regularization and shrinkage thus increasing the accuracy.

The columns ‘Area_Name’ and ‘Sub_Group_Name’ are encoded using Label encoding for this purpose. The Lasso regression uses an alpha of 1.0. Alpha denotes the penalty term which indicates the amount of shrinkage in the equation. We also used repeated K Fold cross-validation to improve the performance of the model.

4. *Random Forest*

We implemented a Random Forest model after encoding the State as well as the Type of Crime in order to include these categorical data as input to the model. Our research suggested that Random Forest as a model has been popularly used for regression problems too and has given good results.

VI. EXPERIMENTAL RESULTS

The main criteria for the evaluation of these models were the normalized RMSE values which gave us an overall understanding of how well each model performed with respect to one another.

$$\text{Normalized RMSE} = \frac{RMSE}{\max val - \min val}$$

A. *Results for Part A*

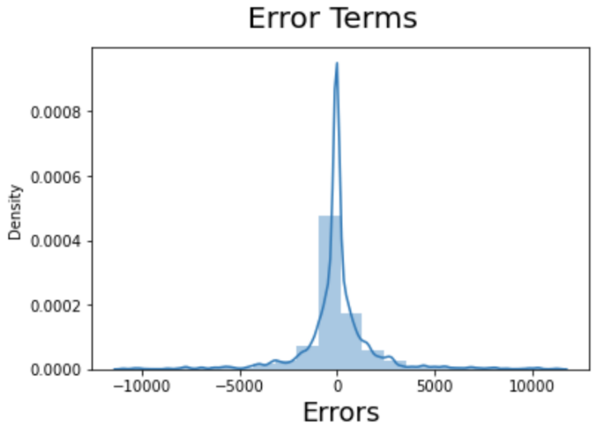
1. *MLR*

Considering the practicality of the problem, it did not seem feasible to predict the number of cases reported from the number of cases sent for trial. Hence, we forgo the MLR model.

The RMSE of the MLR model is 243.81.

The normalised RMSE is 0.0158. The closer the value to 0, the more accurate the model is.

Figure 1.



2. *SARIMAX*

We learnt that as the number of exogenous variables were increased, we saw an improvement in the RMSE values for this model. However, due to a shortage of data over the years, we decided to test other models.

3. *RANDOM FOREST*

a) *One-Hot Encoding*

We would normally expect one-hot encoding to improve the model but the model with one-hot encoding performed significantly worse than the one with label encoding.

In Random Forest with one hot encoding, we are inducing sparsity into the dataset due to the presence of 35 area names and 6 types of crime. Hence the accuracy of the model was compromised drastically.

The RMSE of the model produced undesirable results with a value of 1865. The normalised RMSE is 0.1048.

b) *Label Encoding*

In Random Forest with label encoding, the features are aptly converted into numeric values without increasing the number of columns.

The RMSE of the model above is 199.3.

The normalised RMSE is 0.011. The closer the value to 0, the more accurate the model is. Thus the random forest with label encoding gives us the best model.

The Mean Absolute Error calculated from the predictions made: 86.8.

MAPE (Mean Absolute Percentage Error) cannot be calculated as values in the dataset are 0 and hence, MAPE will give a "division by zero" error.

B. *Results for Part B*

1. *ARIMAX*

A problem we faced with this model was the lack of year data for each state and each type of crime. As we had data from the years 2001 – 2020 only, the model could not be trained efficiently. We concluded that this approach was not appropriate or adequate to deal with the data we had at hand, and might prove useful when extra data is collected.

2. *KNN*

K Nearest Neighbours gave an RMSE of 214.5014 and a normalised RMSE of 0.0709. This seemed to be a good result but we decided to try out some other models before drawing conclusions.

3. *MLR and Lasso Regression*

Lasso Regression gave MAE of 91.474, RMSE of 201.61 and a normalised RMSE of 0.0648.

Multiple Linear Regression gave an RMSE of 201.67 and a normalised RMSE of 0.0666.

Lasso performed slightly better than MLR because of the regularization and cross-validation. Regression seemed to work better than KNN.

4. *Random Forest*

The Random Forest algorithm gave a good result for Part A, hence we decided to try it out for Part B too. We got an RMSE of 167.984 on the test set and a normalized RMSE value of 0.05555. This model proved to be the one that performed the best.

Models for Cases Reported	
Model	**Normalized RMSE**
MLR	0.0158
SARIMAX	0.0172
Random Forest with One Hot Encoding	0.1048
Random Forest with Label Encoding	0.0111
Models for Cases Convicted	
ARIMAX	Nil
MLR	0.0666
Lasso Regression	0.0648
KNN	0.0709
Random Forest	0.0555

VII. CONCLUSION

This analysis was done by using datasets from 2001-2010. We worked on this problem keeping in mind the limitations of the data provided to us in the dataset and the conditions during its collection. The data received was not efficient hence extensive data cleaning and preprocessing was performed before its use. Therefore, the predicted result may differ slightly. The objective of the analysis was to provide an efficient and clear understanding of the issue which would encourage modification of the existing solutions and laws to curb such activities. As future work, our model can be used to predict the crime rate for future years based on past data and compare this prediction with the crime rate obtained from the newer datasets. This will aid in the formulation of effective policies. Detection technologies take a step forward in incident detection to provide public safety resources at the earliest. This analytics helps identify trends and patterns to improve operational effectiveness and efficiency. Proactive policing can help stop crime before it happens.

VIII. CONTRIBUTIONS OF TEAM MEMBERS

Each of us was involved in the data preprocessing and exploratory data analysis. For the modelling, we divided the 2 problem statements into 2 sections and each contributed respectively.
Cases Reported models:
Tanvi – SARIMAX, MLR
Vidisha – Random Forest 1 & 2
Cases Convicted models:
Vibha – ARIMAX, Random Forest
Vishakha - KNN and MLR/ Lasso Regression
All of the team members were involved in the writing of the
report and creation of the video. Decisions with respect to choosing datasets, models and other aspects of this project were taken together over several calls. The code has been uploaded onto Github with appropriate documentation and comments.

IX. PEER REVIEW

The suggestions we received during the insightful and informative peer review were regarding the usage of different forecasting models to predict the cases reported. To address this suggestion, we implemented the SARIMAX as well as the ARIMAX models to forecast cases reported as well as cases convicted. However, we ran across issues such as lack of sufficient data to perform time series forecasting as several group-by statements had to be executed. The performance of these models was sub-par. Another suggestion was regarding the usage of MLR to predict cases convicted from cases reported due to the glaringly obvious correlation between these two data items.

X. ACKNOWLEDGEMENT

We would like to express our profound gratitude to Dr Gowri Srinivasa for providing us with this opportunity to implement a practical project and gain perspective on the numerous methodologies of Data Science that can prove to be beneficial in solving society’s issues. We would also like to thank the Teaching Assistants for their invaluable support throughout this process as well as this semester.

REFERENCES

[1] https://ijcst.journals.yorku.ca/index.php/ijcst/article/view/23401/21601

[2] https://www.researchgate.net/publication/359279537_Crime_against_women_in_India_An_analysis_Journal_of_Society_in_Kashmir

[3] https://vc.bridgew.edu/cgi/viewcontent.cgi?article=2436&context=jiws

[4] https://psycnet.apa.org/record/2000-14050-025

[5] Mapping Crime against Women in India: Spatio-Temporal Analysis, 2001-2012 (waset.org)

[6] Crime Against Women: Analysis and Prediction Using Data Mining Techniques | IEEE Conference Publication | IEEE Xplore

[7] (PDF) Crimes against Women: An Overview of Indian scenario (researchgate.net)

[8] "Crime against Women in India: A State Level Analysis" by Chandrima Chakraborty, Anam Afreen et al. (bridgew.edu)

[9] F12970486S419.pdf (ijitee.org)