Annals of Data Science (2024) 11(1):379–410
https://doi.org/10.1007/s40745-022-00424-6

Machine Learning Algorithms for Crime Prediction under Indian Penal Code

Rabia Musheer Aziz[1] · **Prajwal Sharma**[1] · **Aftab Hussain**[1]

Received: 29 January 2022 / Revised: 27 May 2022 / Accepted: 2 June 2022 / Published online: 6 July 2022

Abstract

In this paper, the authors propose a data-driven approach to draw insightful knowledge from the Indian crime data. The proposed approach can be helpful for police and other law enforcement bodies in India for controlling and preventing crime region-wise. In the proposed approach different regression models are built based on different regression algorithms, viz., random forest regression (RFR), decision tree regression (DTR), multiple linear regression (MLR), simple linear regression (SLR), and support vector regression (SVR) after pre-processing the data using MySQL Workbench and R programming. These regression models can predict 28 different types of IPC cognizable crime counts and also a total number of Indian Penal Code (IPC) cognizable crime counts region-wise, state-wise, and year-wise (for all over the country) provided the desired inputs to the model. Data visualization techniques, namely, chord diagrams and map plots, are used to visualize pre-processed data (corresponding to the years 2014 to 2020) and predicted data by the relatively best regression model for the year 2022. For the chosen data, it is concluded that Random Forest Regression (RFR), which predicts total IPC cognizable crime, fits relatively the best, with a 0.96 adjusted r squared value and a MAPE value of 0.2, and among regression models predicting region-wise theft crime count, the random forest regression-based model relatively fits the best, with an adjusted R squared value of 0.96 and a MAPE value of 0.166. These regression models predict that Andhra Pradesh state will have the highest crime counts, with Adilabad district at the top, having 31,933 predicted crime counts.

Keywords Random forest regression (RFR) · Decision tree regression (DTR) · Indian Penal Code (IPC) · Support vector regression (SVR) · Mean absolute percentage error (MAPE) · Natural language processing (NLP)

✉ Rabia Musheer Aziz
rabia.aziz2010@gmail.com

[1] VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, Bhopal, M.P. 466116, India

1 Introduction

Crime is a very important aspect of concern for the economy of any country. As per the world bank data, in the year 2019, the crime rate per 100 k population in India was 3.07. With such an unignorable crime rate the number of cases registered increases day by day, eventually increasing the data associated with crime in India. On the one hand, where NCRB maintains the Indian crime data whereas, on the other hand, the roles of data analysts and scientists come into play, who can extract knowledge from the data to make useful insights [1–4]. Use of spatiotemporal crime data can be used by data analysts in order to perform time-series analysis on it and this data can also be used to build regressors that can predict different types of IPC cognizable crime count and total IPC cognizable crime counts region-wise, state-wise, and year wise for the country [5–8]. The predicted data from such regressors can be further used by analysts to draw useful insights which can help police and other law enforcement agencies to control and prevent crime. It can be said without any doubt that without a data-driven approach it is impossible to achieve such goals as the crime rate is affected directly and indirectly by numerous factors which also change from time to time [9–12]. For building regressors (crime count predictors) different regression algorithms have been used which are discussed in Sect. 2 of this paper. Different regression templates based on these regression algorithms are built in our proposed approach in order to efficiently build regressors with different transformed and derived data sets. These templates can further be used by other aspirants to build regression models by making necessary changes.

1.1 Types of Crime

We've developed models for overall IPC crime count projections as well as various individual crime count predictions. These crimes may be grouped into the following categories:

1.1.1 Theft

The act of taking away another person's property or services with any taken consent from that person. Auto theft is a form of theft in which the property taken away is/are automobile(s). In cases of burglary mostly an individual or group of individuals illegally enter a building in order to commit theft. Punishment for this country differs from country to country. Although this is not considered a major problem among all types of crimes number registered under this crime is relatively high [13].

1.1.2 Violent Crimes

An unlawful act or attempt by a person or a group of people to hurt or threaten another person or group of people via the use of violence. Robbery, dacoity, rioting, rape, domestic violence, attempted murder, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable homicide, culpable

homicide, culpable in most situations, the punishment for this crime is a lengthy jail sentence or the death penalty. [14].

1.1.3 Assault on Women

Assaults against women are violent crimes committed against women or girls. This includes acid assaults, rape, and domestic violence against women, among other things. The primary motivations for these crimes are physical pleasure, entitlement, superiority, and so on [15].

1.1.4 Kidnapping

The unlawful act of transporting, taking away, or holding/confining a person without his/her will/consent. Kidnapping and abduction of women and girls, kidnapping and abduction of children, etc. comes under this crime [16].

1.1.5 Death Caused by Negligence

This crime is said to be committed when someone dies due to negligence or careless act of some other person who did not intend to kill the individual who died. Sometimes buildings are demolished, major accidents, patient death, etc. types of events are happened due to negligence [17–19].

1.2 Related Works

Until now, the literature has attempted to shed light on the decision-making processes in the realm of crime by examining crime data using different techniques. Various researchers designed models for crime detection and identification in Indian cities with the help of Soft Comput. and machine learning techniques [20–23]. Tayal et al. used k-means clustering with Google Maps for crime detection and map visualization of clusters. KNN classification is used for crime identification, and then they used WEKA software to verify their results. They achieved around 94% accuracy in their results [24]. Abdul Awal et al. proposed a linear regression model fitted to the Bangladesh crime data set to forecast future crime trends in Bangladesh. They used this model to forecast crime for specific crimes such as robbery, dacoity, women's and children's repression, murder, etc. for different regions [25]. Sunil Yadav et al. proposed a linear regression model fitted to the 14 years of Indian crime data (2001–2014) to use the model to predict crime rates in different states for the years after 2014. They have also used the apriori algorithm for association rule mining, k-means for clustering, and Naive Bayes for classification of the data set to increase the accuracy of the predictive model [26]. Suhong Kim et al. proposed a machine learning-based approach for crime prediction. They used two different data sets obtained by two different approaches to data preprocessing. They used K-Nearest Neighbour and boosted decision tree classification algorithms to build classifiers for both the data sets. They achieved 39% to 44% accuracy by using both the classifiers. The accuracy of the built classifiers is too

low, but they can be used as a basic framework for further use [27]. Hitesh Kumar Reddy and Toppi Reddy et al. proposed numerous data visualization techniques to visualize raw data and different machine learning algorithms to predict crime distribution over an area. They built classifiers for crime prediction by fitting K-Nearest Neighbor and Naive Bayes classification algorithms to the U.K. crime data collected from the official U.K. police department website [28, 29]. Mamta Mittal et al. proposed an approach to monitor the effect of the economic crisis on the crime rate in India. They have used linear regression, decision tree regression, random forest regression algorithms, and a neural network algorithm on an Indian crime data set collected from NCRB to study the correlation between unemployment and different crimes, viz., theft, burglary, and robbery. Further, they used Granger causality to study the causal relationship between parameters affecting the Indian economy. They observed that the rate of unemployment is a significant factor that affects the crime rate in India [30]. Priyanka Das et al. proposed numerous classification techniques for crime prediction and analysis in India. They used KNN, decision, Random Forest, Naive Bayes, and adaptive boost classification algorithms to classify processed crime data collected from the official site of NCRB. They also prepared comparison tables for different classifiers, which were compared based on accuracy, recall, F-measure, and other parameters [31]. Sohrab Hossain et al. proposed an approach to predict criminal activities with the help of supervised learning algorithms. K-nearest neighbors (KNN) and decision trees are used on the San Francisco criminal activity data set of 12 years to predict crime. Classifiers built using KNN and decision trees have low accuracy, so they used a random forest algorithm with Adaboost to increase the accuracy. Log–loss is used to measure the performance of classifiers by penalizing false classification. The classifier built with a random under sampling method for the random forest algorithm gives the best accuracy. The final accuracy is 99.16% with a 0.17% log loss [32]. Marcus Pinto et al. proposed an approach to minimize the negative actions or crimes that are harmful to human society with the help of machine learning algorithms. The KNN, Decision Tree, and Multivariate Linear Regression classification algorithms are used with the New York crime data set for the year 2019. With the help of past crime data, these models show trends or patterns in a crime, which helps correlate the factors that help predict future crimes. The decision tree gives the best prediction (99.95%) for the borough's correct name. It is because the decision tree deals better with a large dataset that has many nodes with different layers and a small target, which in this case has only five boroughs. There is a limited number of possible decision points, which is why it gives the best accuracy. KNN comes in as the second-best prediction model with an accuracy of 99.65%, and multivariate linear regression comes last with an accuracy of 98.03%. The common point between all three models is that they give a good accuracy rate on a limited target [33]. Wheeler et al. using the machine learning algorithm give the most accurate long-term prediction of crime in microcities compared to the other popular techniques, and how their advanced model is improving their interpretability, helping to open the "black box" of random forests. Using this model, they estimate in terms of forecasting future crimes using different measures of predictive accuracy. Their model is accurate in predicting crime in micro places, but they are unable to understand why these places are predicted to be very risky [34]. Some of the researchers proposed various approaches to estimate the accurate

crime rate, types of different crimes, and hot spot locations from the past pattern with the help of machine learning and deep learning techniques, proposed model not only applicable for crime prediction domain its also suitable for other domain problems [35–38]. Wajiha Safat et al. used machine learning algorithms viz., SVM, Naive Bayes, KNN, decision tree, Multilayer Perceptron neural net, random forest, logistic regression, XGBoost for crime prediction, and deep learning algorithm LSTM for time series analysis and ARIMA for forecasting on Chicago and Los Angeles crime data. They also performed exploratory analysis. LSTM gave satisfactory results, with acceptable root mean square error and mean absolute error. They concluded various useful inferences based on their experimental results [39].

1.3 The Objectives of the Proposed Work Are

- To create effective predictive models in R that can be further used by others for performing regression analysis on similar crime data, which can be achieved by making necessary changes.
- To build predictive models with Indian district-wise crime data (2012–2020) that can predict/forecast total IPC and specific crime counts, region-wise and all over the country.
- To use different regression models and choose the best one for predicting different types of crime and total IPC crime counts.

In this paper, we have proposed the use of different regression algorithms, namely, simple linear regression, multiple linear regression, DTR, SVR, and RFR, to build predictive models which can predict the total number of IPC crime counts and crime counts of different types of crime (murder, rape, kidnapping and abduction, riots, etc.) region-wise and all over the country. We have used district-wise crime data from 2012 to 2020, which we collected from the official website of NCRB. Figure 1 shows the flowchart of the proposed work. The rest of the paper is organized as follows: Sect. 2: Methodology, Sect. 3: Experimental Setup, Sect. 4: Experimental Results and Discussion, Sect. 5: Conclusion and future scope.

2 Method Used

2.1 Regression Algorithms Used

The following is an explanation of the regression techniques that we utilized for predictive modelling in this paper.

2.1.1 Simple Linear Regression

Linear regression is the statistical model used to predict the relationship between the independent and dependent variables. It is used to estimate the value of the variable with the help of a continuous variable. The accuracy and goodness of fit are measured by loss, R-square value, and adjusted R-square value. The higher the R-square value,

the more the data is fitted to the model. On increasing the dataset to measure the model goodness, we see an adjusted R-square value. The representation of the linear regression is given as [22]:

$$Y = a^*X + C \tag{1}$$

where Y: Dependent Variable, X: Independent Variable, C: Intercept, and a: Coefficient of X.

2.1.2 Multiple Linear Regression

Multiple Linear Regression is one of the important regression algorithms in which there is a relationship between the single dependent variable and multiple independent variables. In this algorithm, the dependent variable must be a continuous value, and the independent variable may be continuous or categorical. The representation of the multiple linear regression is given as [40, 41]:

$$Y = a^*X_1 + b^*X + c^*X_3 + \cdots + C \tag{2}$$

where Y: Dependent Variable, X_1, X_2, X_3, …: Independent Variables, C: Intercept, and *a, b, c*, …: Coefficient of X_1, X_2, X_3,…

2.1.3 Decision Tree Regression

A regression tree is one of the two terms encompassed in the umbrella term CART, which stands for Classification and Regression Trees. Decision tree algorithms are generally used for classifying the labeled data, but they can be used for regression analysis as well. In the training process, initially, the whole data set is considered as the root node and then different split boundaries are created among the data points at successive levels of the tree. A decision tree split, in simple words, is a criterion for splitting a parent node and it leads the tree either to a new split (at an internal node) or to a region of data points (at a leaf node). The best split is decided based on the information gain value of the split. The split with the highest information gain is chosen as the best split. There are different decision tree algorithms. They all differ in the formula used to calculate the information gain of a split. In the case of decision tree classification, the leaf nodes are the labels of the labeled data, and in the case of decision tree regression, the leaf nodes are the average value of the data points belonging to the region, which is the result of splits in the path from the root to the corresponding leaf node [42, 43].

Information gain for a feature X is calculated as the difference between the entropy in the data segment before the split and the total entropy of partitions after partitioning the data segment, and is given by,

$$\text{InformationGain}(X) = \text{Entropy}(S1) - \text{Entropy}(S2), \tag{3}$$

where S1 is the data segment before partition and S2 is the set of partitions after partition.

The entropy of a data segment S1 (no partitions) is given by,

$$Entropy(S1) = \sum_{k=1}^{n} -p_i log(p_i) \tag{4}$$

where n is the total number of class levels, p_i is the percentage of data points in class level *i*.

The entropy of a set of partitions S2 is given by the weighted sum of the entropy of all the partitions,

$$Entropy(S2) = \sum_{k=1}^{n} w_i entropy(p_i) \tag{5}$$

where w_i is assigned as per the proportion of records falling in the partition and P_i refers to partition i.

2.1.4 Random Forest Regression

Random Forest Regression is an ensemble learning-based method. Ensemble learning means making use of multiple algorithms or running the same algorithm multiple times to yield a much more powerful algorithm. In the case of a random forest algorithm, we make use of the decision tree algorithm multiple times to create a forest of decision trees. During the random forest regression training process, k data points are chosen at random from the training set and a decision tree is built with these data k points; this process is repeated n times to create a forest of n decision trees built on n randomly chosen sets of data points. This forest of decision tree regressors is our random forest algorithm-based regressor. When a new data point is provided to this model, all n decision trees predict the value of the dependent variable, and the average of all these predictions is given as the final predicted value of the dependent variable [34, 44].

2.1.5 Support Vector Regression

A Support Vector Machine (SVM) is a supervised machine learning model that can be used to perform data classification and regression analysis. For the given sets of vectors in multi-dimensional space (values of a dependent variable depending on multiple independent variables), SVM finds a hyperplane (in the case of two-dimensional space, the hyperplane will be a line) that separates the labelled data in space belonging to different class levels. This hyperplane is more often referred to as the maximum margin hyperplane (MMH) or maximum margin classifier. With the use of a slack variable, SVM can deal with linearly as well as non-linearly separable data [45–47]. Though the kernel SVM model is generally preferred for non-linearly separable data, the maximum margin hyperplane is the best possible plane in space to separate the vectors. Support vectors are the sets of vectors from each class that are closest to the maximum margin

hyperplane. Each class has at least one support vector, but any class can have more than one support vector. MMH can be found by SVM as well. This way, SVM can deal with data sets with a high number of features. Identification of support vectors depends upon vector geometry and involves some tricky maths [48, 49].

3 Experimental Setup

This section describes about the work flow of this research.

3.1 Metrics Used for Evaluating Regression Models

3.1.1 R squared

R squared value for a regression model is a statistical measure used to evaluate the fitness of the regression model. The fitness of the regression model refers to how well the regression model curve is fitted to the training data. The formula for the calculation of the R squared value is given by,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{6}$$

where SSres is the squared sum of residuals, that is, the sum of squares of differences between the actual value and predicted value and SStot is the squared sum of differences between the actual value and the average value of the actual values. R squared can also take a negative value when the regression curve is even worse than the average curve. R squared generally ranges from 0 to 1. Closer to 1 better is the fitness of the regression model.

3.1.2 Adjusted R Squared

Adjusted R squared is a version of R squared measure which is also used to measure the goodness of fit for the regression model and is preferred over R squared for the same. R squared can be misleading when it comes to the evaluation of fitness when a new set of predictors are introduced or there is a large number of predictors. As the number of predictors increases, the R squared value also increases which might lead to a poor model. By keeping the number of predictors in the account, adjusted R squared solves this problem. The formula for adjusted R squared is given by,

$$AdjR^2 = 1 - \left(1 - R^2\right)\left[\frac{n-1}{n-p-1}\right], \tag{7}$$

where n is the data sample size and p is the number of independent variables. Closer to 1 better is the fitness of the regression model.

3.1.3 MAPE (Mean Absolute Percentage error)

Mean absolute percentage error is a measure used to evaluate the prediction accuracy of a predictive model. The formula for MAPE is given by,

$$M = (1/n) \sum_{k=1}^{n} |(A_k - F_k)/A_k|, \tag{8}$$

where A_k is the actual value and F_k is the predicted value and n is the data sample size.

3.2 Data Used

The raw data set is taken from the official site of NCRB. The data set contains 9017 records and 33 columns (variables), each record is distinct on the basis of STATE/UT (state or union territory name), DISTRICT (district name), and YEAR (year) variable values. Out of the other 30 variables, 28 Variables represents the number of cases registered under 28 different types of cognizable IPC crime (viz. murder, rape, attempt to murder, riots, kidnapping and abduction, etc.) in the corresponding region and year: And other two variables are:

Other IPC Crimes: Count of crimes other than the above mentioned 28 crimes.
Total IPC Crimes: Count of total cognizable IPC crimes.

Thus, this data set contains information about total number of cognizable IPC cases and number of cases registered under different types of cognizable IPC crime from 2012 to 2020 for 808 Indian districts and 35 states and union territories.

3.2.1 Data Pre-processing

Firstly, data pre-processing techniques were utilised to clean the raw data set for further analysis. Data pre-processing is about removing or replacing the missing values in the data set and performing necessary steps to transform the data to make it compatible with the machine learning algorithm to be used. Replaced NA/na values (missing values) in a column by the mean of the values in the corresponding column. Regression algorithms need numerical data to work with and label encoding will be a wrong approach. Therefore, in this work used dummy data frame () method from the dummies library in R to create dummy variables for each state and district, which leads to the creation of 35 state dummy columns and 808 district dummy columns. For a feature containing n class levels n number of dummy variables. A dummy store either a value 0 or 1, where 0 refers to the record that does not belong to the corresponding class level and 1 refers to a record that belongs to the corresponding class level. Data pre-processing is about removing or replacing the missing values in the data set and performing necessary steps to transform the data to make it compatible with the machine learning algorithm to be used.

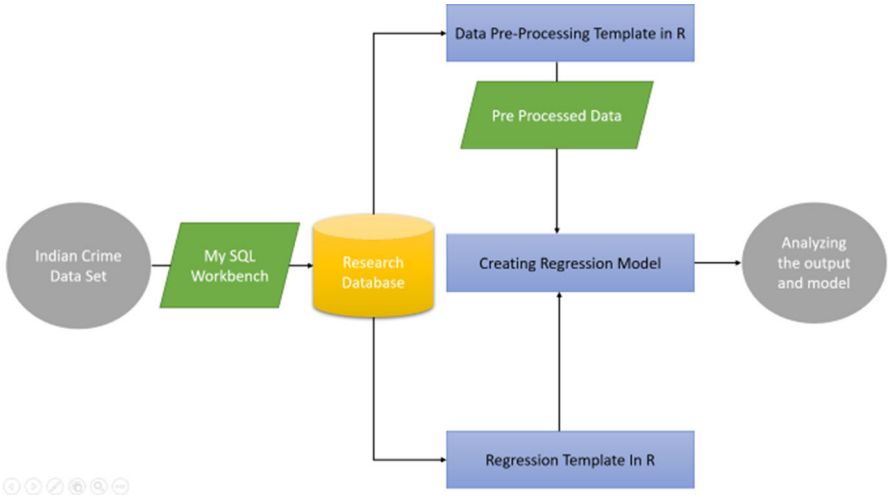

Fig. 1 The flow chart of proposed work

The first data set containing information about district-wise registered cognizable IPC crimes from 2012 to 2020 is available from the official site of NCRB. Then data cleaning and data transformation are performed as per requirement using R and MySQL Workbench. The missing values in the data set are replaced with the mean value of the corresponding column using R, and then the data set is imported into a database using MySQL Workbench. With the help of structured queries, we produced state-wise and year-wise crime data after this dummy coding was performed on the categorical variables in these data sets before using them to build regression models. After producing the desired derived data sets, these data sets are then used to build the desired regression models.

Derived state-wise, district-wise, and year-wise crime data are then used to build different regression models on them using different regression algorithms, namely, simple linear regression, multiple linear regression, decision tree regression, support vector regression, and random forest regression algorithms. For each derived data set, these algorithms are used to build regression models by choosing 1 of the 29 dependent variables (28 variables corresponding to different types of cognizable IPC crime and 1 variable corresponding to total cognizable IPC crimes) as the dependent variable. After this, by making necessary changes in R scripts as per the next chosen dependent variable, regression models predicting crime counts state-wise, district-wise, and year-wise are built. Each R script is written to yield a trained, tested and evaluated regression model. In the results section of this paper, the accuracy and goodness of fit of regression models predicting district-wise total cognizable IPC crimes and thefts are compared based on MAPE, r squared and adjusted r squared measures. The comparison is shown with the help of comparison Tables 1 and 2. Among these models, the best model is chosen based on the lowest MAPE value and the highest adjusted r squared value. The relatively best model among these is then used to predict district-wise total IPC crime counts and theft crime counts, which are visualized using leaflet map plots in R.

Table 1 Comparison among regression models predicting District Wise Total IPC Crimes

Regression model used	R-squared value	Adjusted-R-squared value	MAPE
Multiple linear regression	0.8935085	0.893493	1.99711
Support vector regression	−0.06258563	−0.06274019	4.794302
Decision tree regression	0.5735719	0.5735099	4.543087
Random forest regression	0.9631605	0.9631551	0.2027437

Table 2 Comparison among regression models predicting District Wise Theft crimes

Regression model used	R-squared value	Adjusted-R squared value	MAPE
Multiple linear regression	0.9185906	0.918579	0.8956748
Support vector regression	−0.07252166	−0.0726745	0.7894
Decision tree regression (min. Split = 2)	0.7131494	0.7131085	0.5951368
Random forest regression	0.9666091	0.9666044	0.16571

4 Experimental Results and Discussion

The results of eight regression models, namely multiple linear regression, decision tree regression, random forest regression, and support vector regression models, which can predict total cognizable IPC crime counts district-wide, as well as four other models based on the same four regression algorithms, which can predict theft crime counts district-wide, are described in this section.

4.1 Plots to Visualize Fitness of Regressions

Figures 1a, b and 2a, b depict the total number of IPC criminal cases versus year plots, whereas Figs. 3a, b and 4a, b depict the total number of theft crime cases versus year plots.

In all these scattered plots, red-coloured dots represent the data points from the actual training data, and blue dots represent the data points predicted by the corresponding regressor built on the training data. All these plots are created to visualize how well the regression curve has fitted the training data, but this cannot be achieved by these plots because in these plots, for a particular year, there is not a distinct data point, rather there are hundreds of data points representing data for different districts for that year. So, it is not possible to tell which data point belongs to which district, making it impossible to visualize fitness from the plot. So, to evaluate the fitness of these regression models, statistical measures like r squared and adjusted r squared are used (Fig. 5).

Figures 6, 7, 8 and 9 present the fitness of regression curves based on different regression models that can predict the crime rate per 100 k in India for a year. The

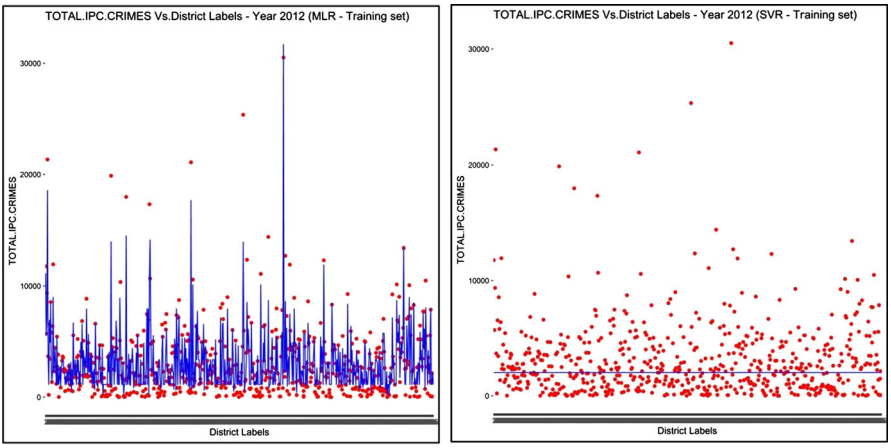

Fig. 2 Total IPC crimes Vs district labels with MLR model and SVR model for year 2012

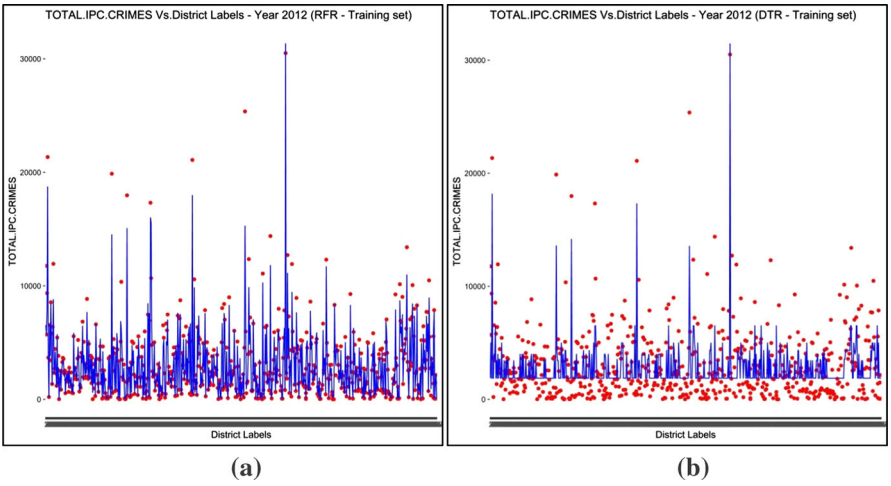

Fig. 3 Total IPC crimes Vs district labels with RF and decision tree for year 2012

plots in the figures have a blue line as the regression curve made by the training data predicted by the regressor, and red dots are the actual data points from the training set.

4.2 Comparing the Regression Models

The regression models are compared based on their accuracy and fitness to the data, for this MAPE and adjusted r squared values are used respectively. Table 1 shows the comparison of regression models built to forecast total IPC crime counts. Table 2 shows a comparison of regression models built to forecast theft crimes. Table 3 shows

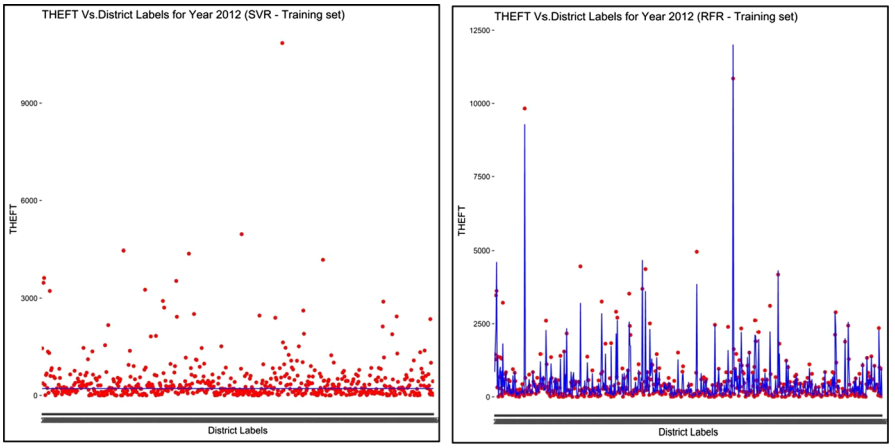

Fig. 4 Thefts Vs district labels with SVR and RF model for year 2012

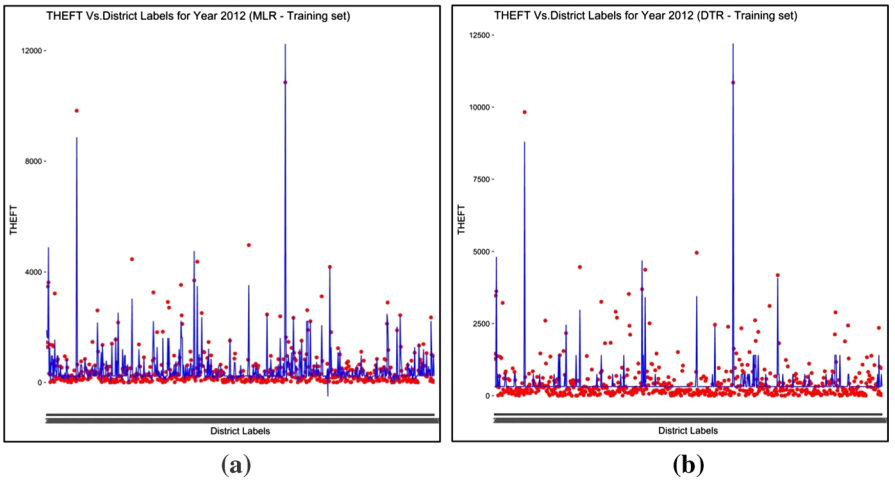

Fig. 5 Thefts Vs year with MLR and decision tree model

the comparison among regression models built to predict the crime rate for a given year.

Table 1 shows that the regression model built using the random forest for district-wise total IPC crime count forecasting is relatively the best model, with a 0.9631551 adjusted r squared value and a mean value of 0.2027437. From Table 2, it can be concluded that a regression model built using the random forest for district-wise theft crime count forecasting is relatively the best model with a 0.9666044 adjusted r squared value and a 0.16571 MAPE value. In Table 3, it can be concluded that a regression model built using support vector linear regression is relatively the best for predicting

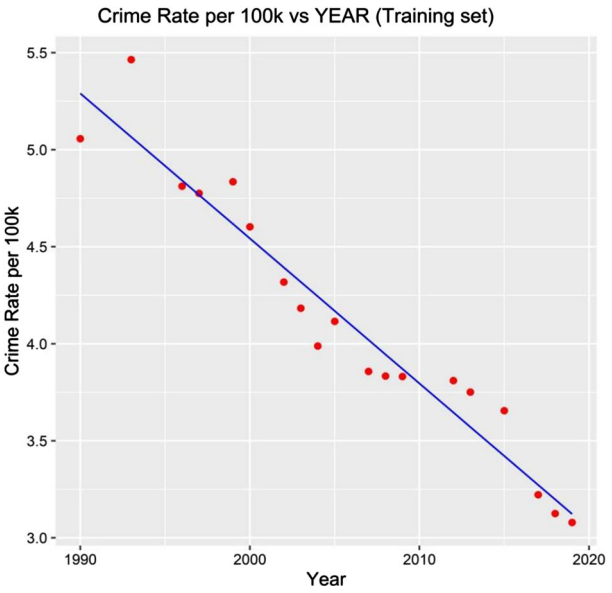

Fig. 6 Crime rate per 100 k Vs year (SLR model)

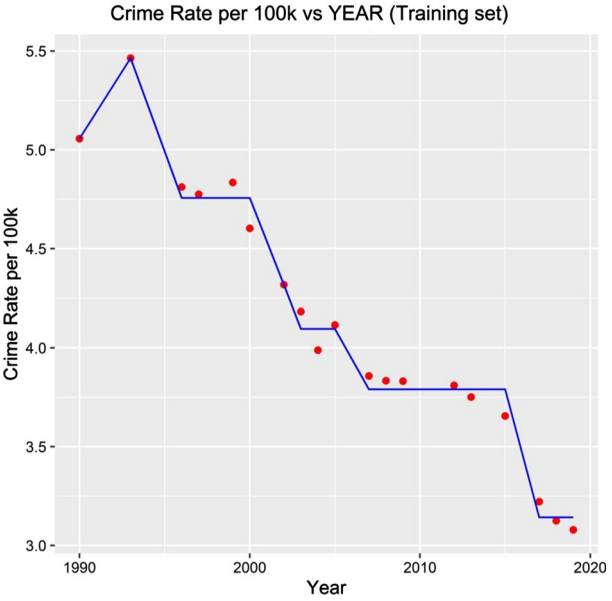

Fig. 7 Crime rate per 100 k Vs year (DTR model)

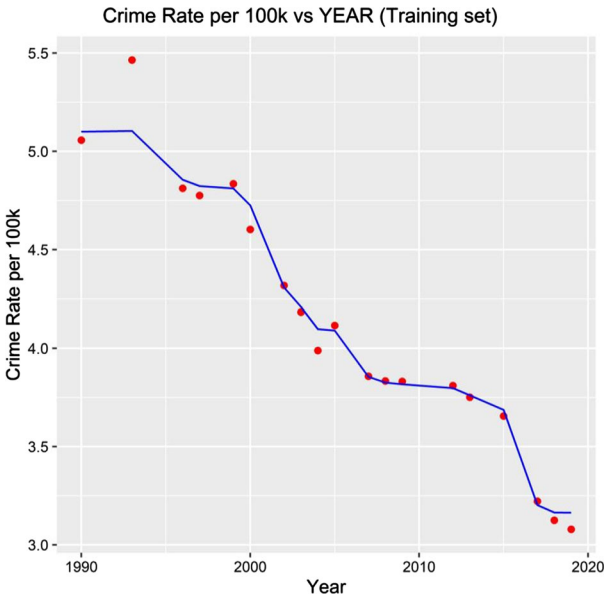

Fig. 8 Crime rate per 100 k Vs year (RFR model)

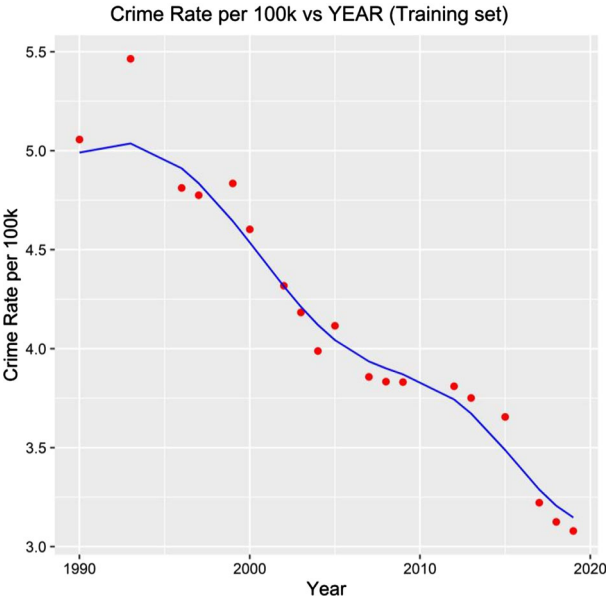

Fig. 9 Crime rate per 100 k Vs year (SVR model)

Table 3 Comparison among regression models predicting crime rate per 100 k population

Regression model used	R-squared value	Adjusted-R squared value	MAPE
Simple Linear Regression	0.9345979	0.9307507	0.0243525
Decision Tree Regression	0.9884693	0.987791	0.03548595
Random Forest Regression	0.9779519	0.976655	0.02685891
Support Vector Regression	0.9584286	0.9559832	0.01971657

the crime rate per 100 k for a given year, with an adjusted R squared value of 0.9559832 and a 0.01971657 MAPE value.

4.3 Map Plots

Figures 10 and 11 are the map plots plotted using the leaflet library in R based on the district-wise total IPC crime counts for the year 2022 and district-wise theft crime counts for the year 2022 predicted by regression models built using a random forest algorithm.

Observations:

i. Figure 10 shows 50 districts with the highest predicted total IPC crime count, such that the predicted crime count for these 50 districts/regions is greater than the mean of the predicted total IPC crime counts for 827 districts/regions for the year 2022. The larger the radius of the circular mark, the higher value of the predicted crime count.
ii. Figure 11 shows 50 districts with the highest predicted theft crime counts, such that the predicted theft crime count for these 50 districts/regions is greater than the mean of the predicted theft crime counts for 827 districts/regions for the year 2022. The larger the radius of the circular mark, the greater the value of the predicted crime count.
iii. These leaflet plots are created such that when a viewer hovers the mouse over the circular markers, you can see the predicted crime count, and when you click on the markers, you see the details about that region, viz., address, latitude, and longitude coordinates, and crime count, as shown in Fig. 11.
iv. As Fig. 11 shows, the random forest model predicts that the Adilabad district/region of Andhra Pradesh state will have the highest crime count in India for the year 2022, with a predicted crime count of 31,933.
v. As Fig. 10 shows, the random forest model predicts that the Anantapur district/region of Andhra Pradesh state will have the highest theft-specific crime count in India for the year 2022, with a predicted theft crime count of 12,000.
vi. Figure 11 shows that in the year 2022, Andhra Pradesh, Bihar, and Assam will be the states from which the top 50 districts will belong. Also, Andhra Pradesh state will have regions with the highest crime counts as per the model's predictions.
vii. As per the model's predictions, in the year 2022, Andhra Pradesh state will have the regions with the highest theft-specific crime counts (Fig. 10).

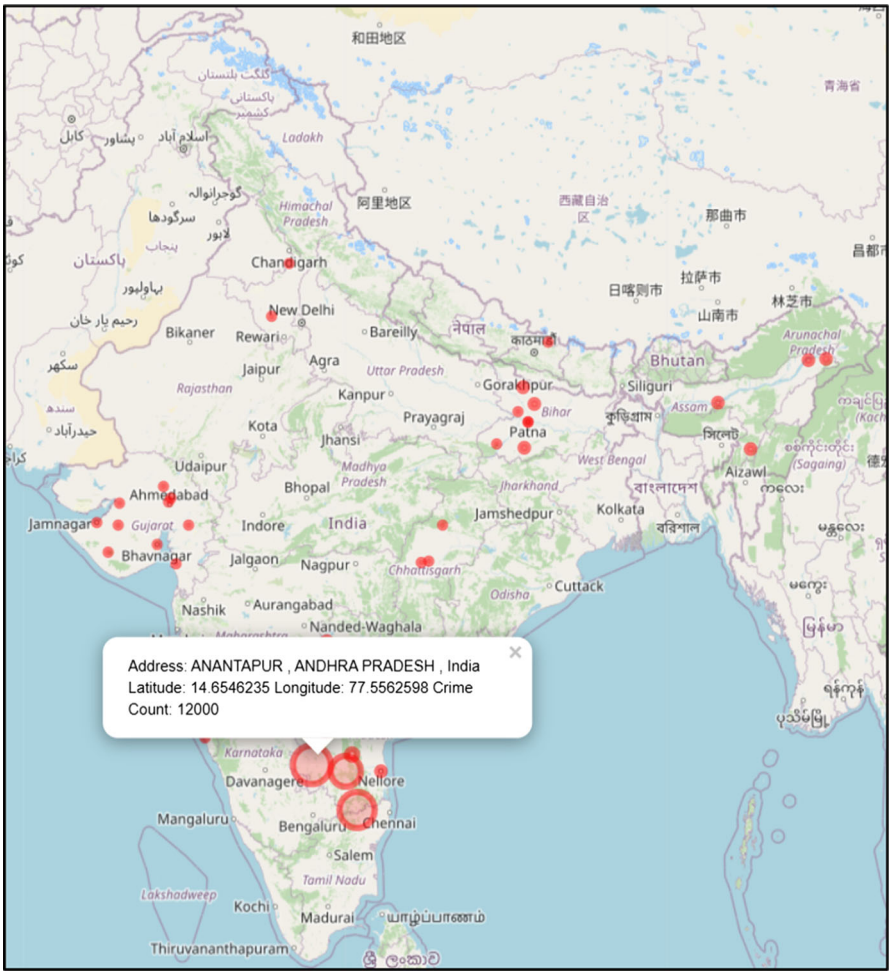

Fig. 10 Leaflet map plot showing top50 districts with highest number of predicted theft crime counts in 2022

4.4 Chord Diagrams

The density of many crime sources, such as residences, eating and drinking facilities, and major commercial shops, are the most important variables for prediction. Although population density is an essential demographic element, the other demographic features are less relevant in comparison to the crime-generating factors. To know the crime count (> 10 k) across the states in India from 2012 to 2020, we use chord diagrams. The primary use of chord diagrams is to show the flows or connections between several entities (called nodes). Each entity is represented by a fragment (often colored or pattered) along the circumference of the circle. Arcs are drawn between entities to show flows. 12 states have a crime count greater than 10,000 (Fig. 12).

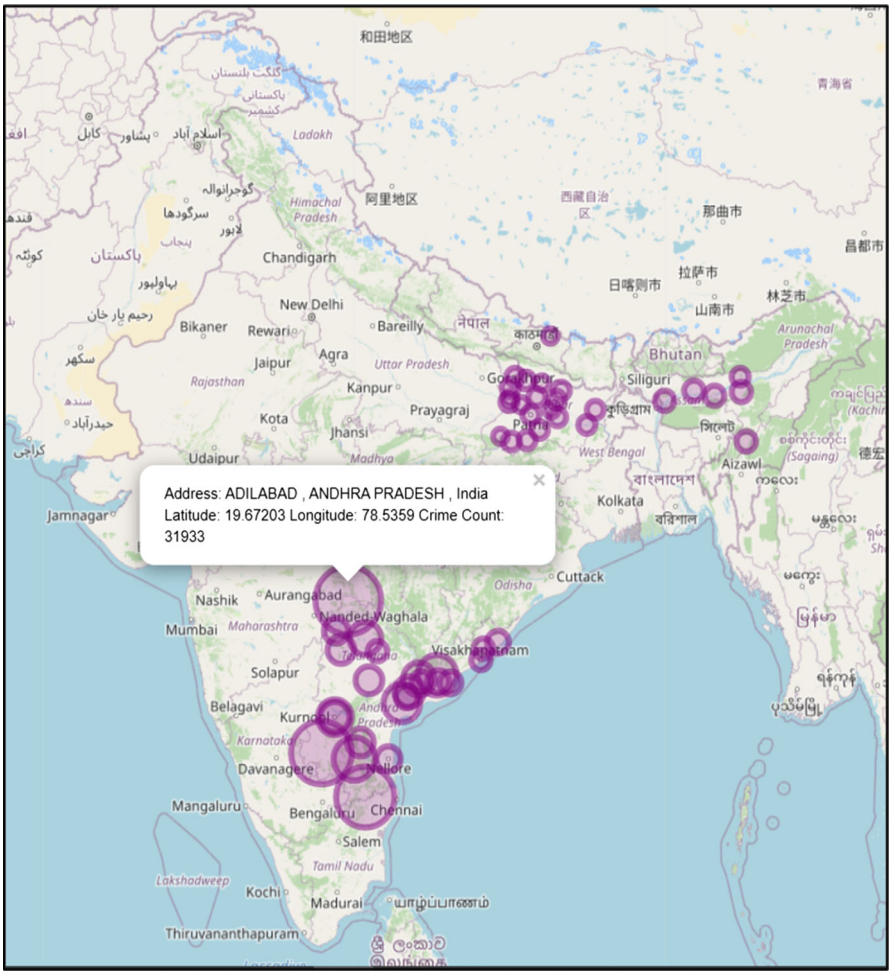

Fig. 11 Leaflet map plot showing top50 districts with highest number of predicted total IPC crime counts in 2022

Figure 13 shows that the total number of crime counts in Andhra Pradesh is 1,234,942 from 2012 to 2020. The line connecting the arc of state and the year is in sequence, with the first line of the arc showing which year has the maximum number of criminal counts. As shown in Fig. 14, we can see that the first line of the arc of the state of Andhra Pradesh shows that the crime count is at its maximum (238,903) in 2020, and the sequence is decreasing, with the last line of the arc having the minimum (84,270) number of crime counts (Fig. 15).

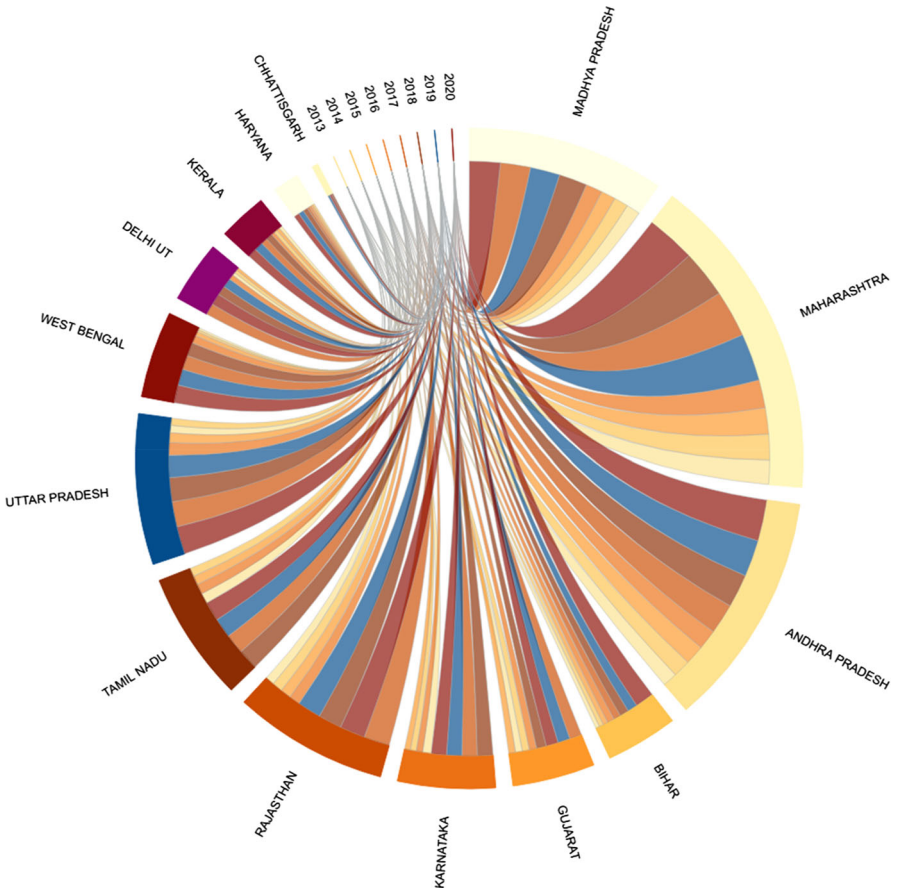

Fig. 12 Flow of crime count in India from 2012 to 2020

4.5 Madhya Pradesh Crime Pattern Analysis

Based on the map plot and chord plot analysis, the Andhra Pradesh state has the greatest theft and overall crime count district by district. Madhya Pradesh, on the other side, has the highest average crime rate state. So, for further analysis, we apply different prediction models for Madhya Pradesh crime data.

Figure 16 depicts the annual frequency of crimes per type and their trend in Madhya Pradesh. The most common types of crimes are theft and hurt grievous hurt. They are irregular in trend, but most of the time they are increasing. Other than other types of crime, Hurt Grievous Hurt is the most common type of crime.

In Fig. 16 most of the crimes are very low, so to visualize it more clearly, we neglect the crime types whose crime count is less than 1 k or 1000. Figure 17 shows the crime kinds that have a crime count of more than 1,000.

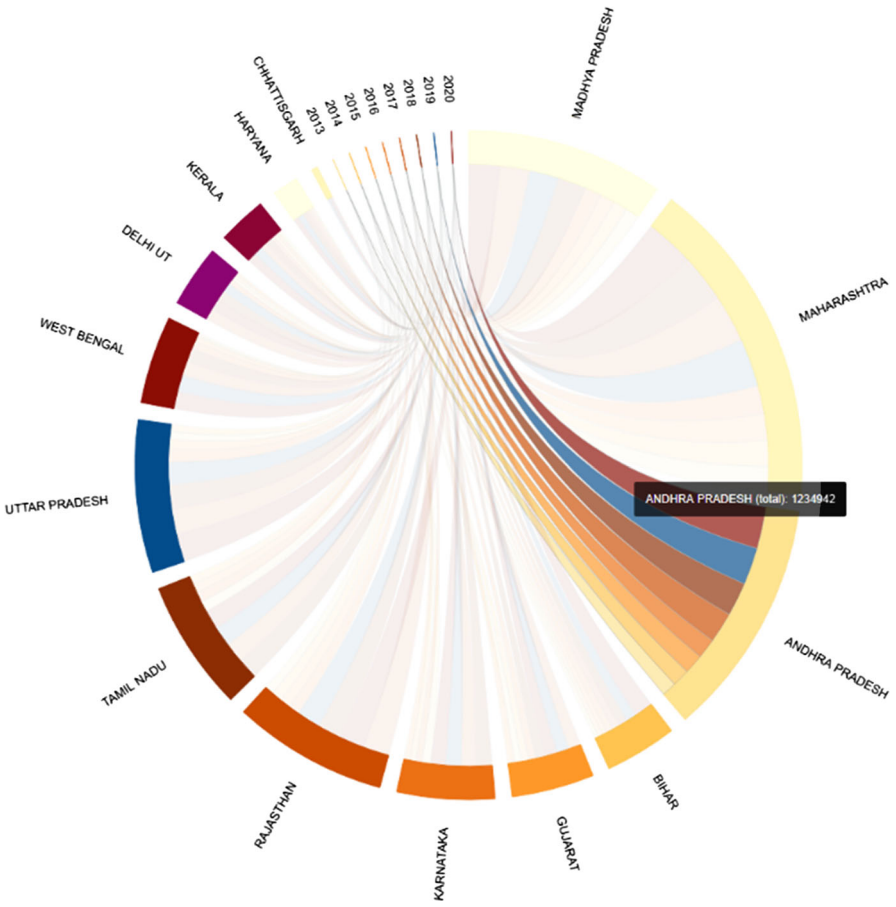

Fig. 13 Total Crime Count of AP

Figure 18 shows the flow of crime count for all the districts of Madhya Pradesh from 2012 to 2020. If we want to see the specific district with the total crime count in 2012–2020, just point to that district. We can also see which year the crime count is at its maximum or minimum in the different districts as explained above.

Figure 19 is the heat map of all of the crime types lying in the range between 0 and 40,000 in 2012–2020. The heat map depicts the distribution of crimes per location; typically occurring in specific districts, and demonstrates that the darker the color, the more specific crime has occurred in a specific district.

There are only four crime types that occur often in all districts. Theft and Hurt Grevious Hurt are the most common crimes, with a rising tendency, whereas burglary offenses decline, then remain steady, and attacks on women are extremely rare, despite their absence from the graph. Figure 20 depicts the trend of these crimes.

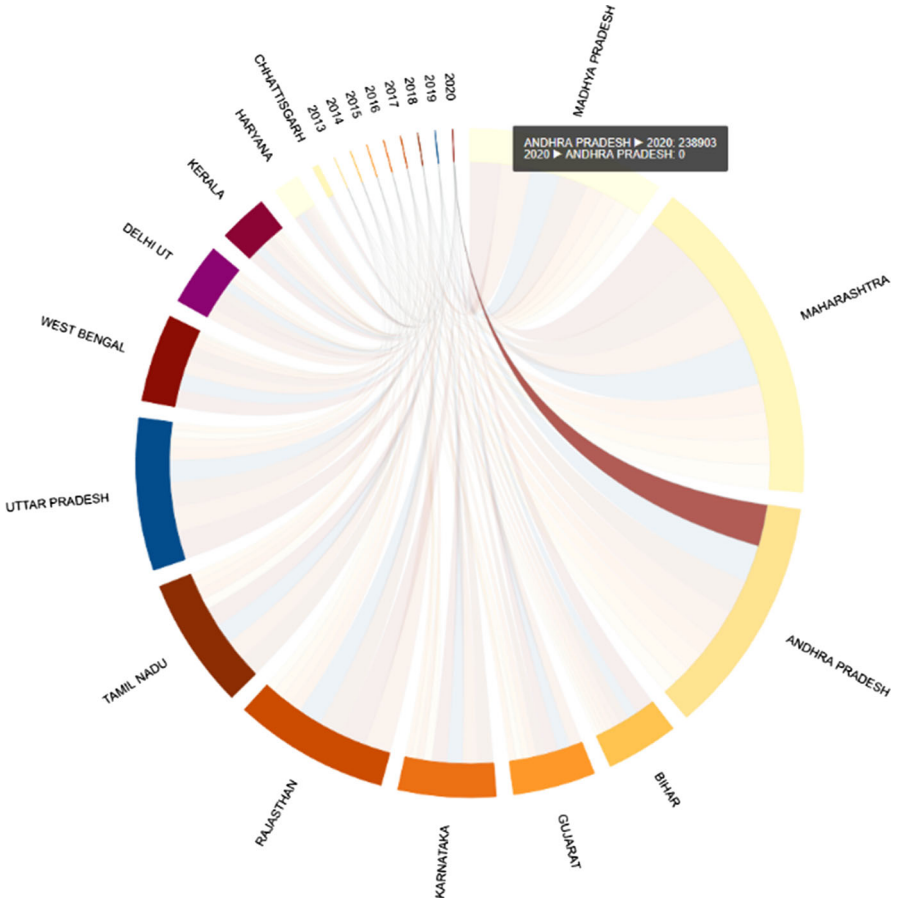

Fig. 14 In 2020 AP has the Maximum crime count

4.6 Regression Analysis Through Neural Network

In this session apply the neural network model as regression for crime analysis tasks. Before applying the neural network model to our dataset, we divide the dataset by training and testing it in a 70:40 ratio and again we divide the testing dataset into test and validation data (The goal of validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset) it in a 50:50 ratio. We define our model as a Sequential class this will be the container that will contain all layers [3]. This model has three main components: the input layer (we give the 500 neurons in it) according to our variables and the input shape must be explicitly designated. The second one is the hidden/encoder layers this layer is all up to the creator we give 3 hidden layers with 500,600 and 700 neurons with activation as "relu", the last one is the output layer with an output_dimension =

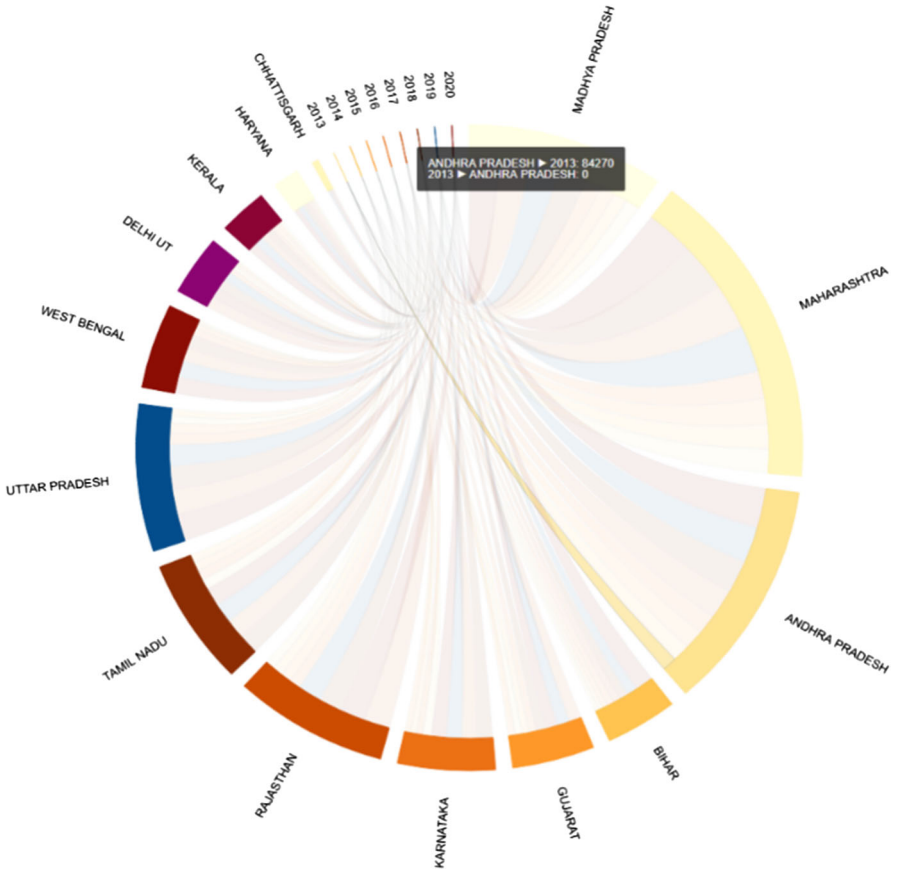

Fig. 15 In 2012 AP has the Minimum crime Count

1 and activation as "linear" since it is a regression problem. After that, all previously learned features are then combined by a fully connected layer to recognize significant patterns from data for the regression. While this is a regression problem, we see the mean_square_error as "mse" and the loss function we give an epoch is 100 (epoch number shows how many times it going to be iterating) and the batch_size is 16 (it's all up to the creator of the data is large or small and batch size shows after how many iterations this model iterating by itself). In every epoch, the "mse" is going to decrease a time comes when the loss and "mse" are moving in a certain range it shows that this amount of epoch is sufficient. When loss is going to increase and "mse" is going to decrease at that point we know that our model is going to be over fitted. The crime pattern that we observed with neural network regression is given below with Fig. 21, 22, and 23.

Figure 21 describes a phenomenon in a certain location and time for example, in our case; we can see that the crime counts increased over the years through the

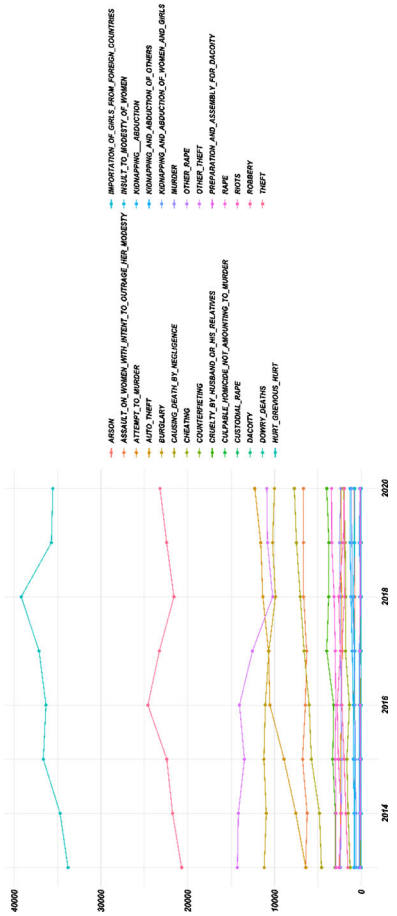

Fig. 16 Crime evolution of Madhya Pradesh per type of crime between 2012–2020

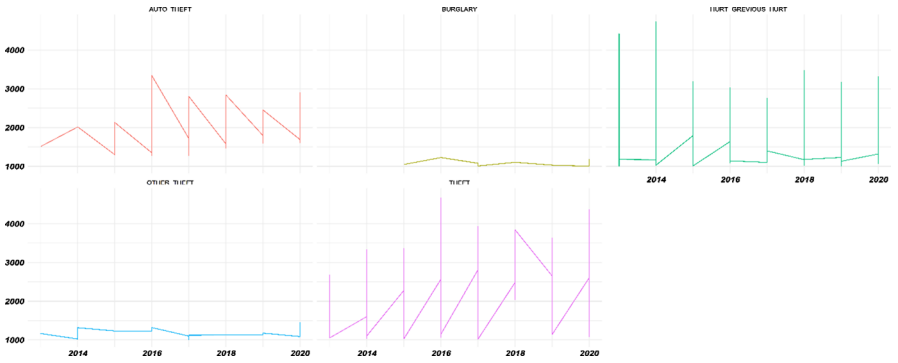

Fig. 17 Crime evolution of MP per type individually and who's crime count $\geq$ 1 k

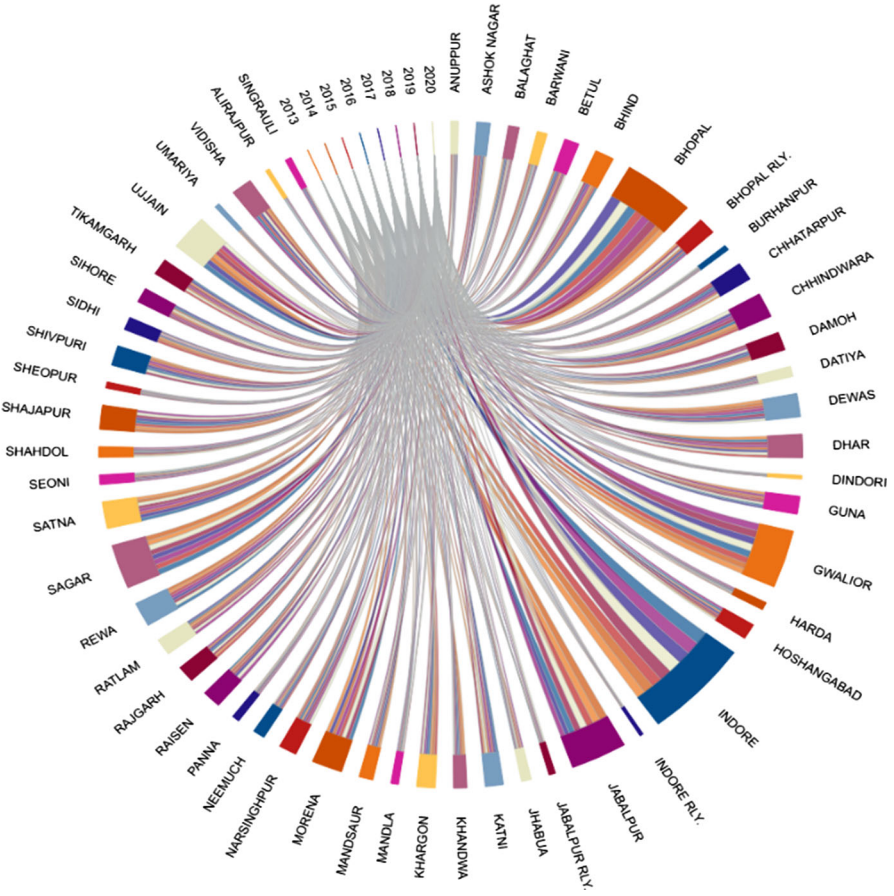

Fig. 18 Flow of crime count in Madhya Pradesh from 2012 to 2020

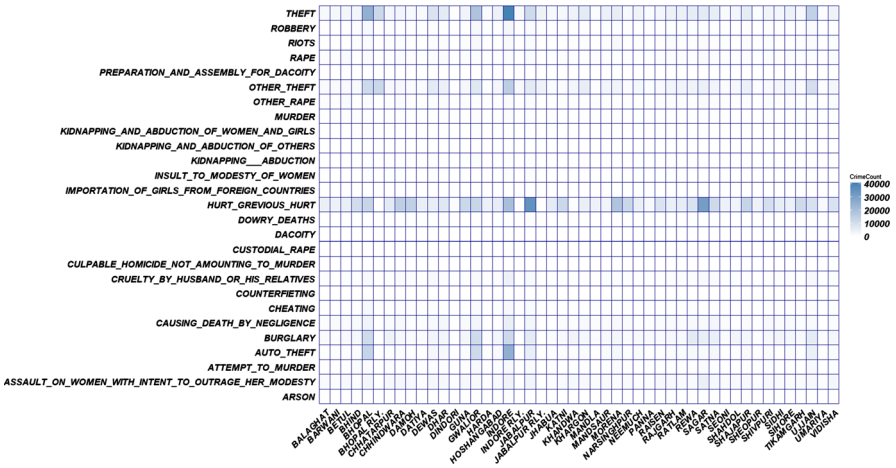

Fig. 19 Crime type in district of MP Vs district of MP

geographical area of the top 50 districts in India. Figure 21 depict that the larger the red area show the more crime in the year 2020.

In Fig. 22, we can see that the crimes have occurred a greater number of times. Figure 23 show that the THEFT is more likely to happen in 2019. By pointing out every time we can see that Crime counts and in which year it is most likely to happen as we see in the figure by pointing out the in the DOWRY_DEATH that in 2019 that crime count is 8618 and it is high in that year and 2011 it is 6208 which is less in that year.

5 Conclusion

For the chosen data, random forest regression, which predicts total IPC cognizable crime, fits relatively best with a 0.96 adjusted r squared value and a MAPE value of 0.2, and among regression models predicting region-wise theft crime count. The random forest regression-based model fits relatively best with a 0.96 adjusted R squared value and a MAPE value of 0.166. According to these regression models, Andhra Pradesh will have the largest crime count in 2022, with Adilabad district leading the way with a forecast crime count of 31,933 and Anantapur district leading the way with a predicted theft crime count of 12,000. Madhya Pradesh, on the other side, has the highest average crime rate state. This paper also shows the crime pattern in MP state. The predicted data by the regressors can be used with data visualizations techniques like leaflet map plot as demonstrated in the paper or other geospatial data visualization techniques like a heat map, choropleth map, dot map, cluster map, bubble map, cartogram map, hexagonal binning, etc. To represent data and visualize the predicted data to draw insightful knowledge from the predicted data. The proposed approach provides a framework for other data analysts to use for Indian crime data prediction and visualization using regression algorithms. Nowadays, when there is

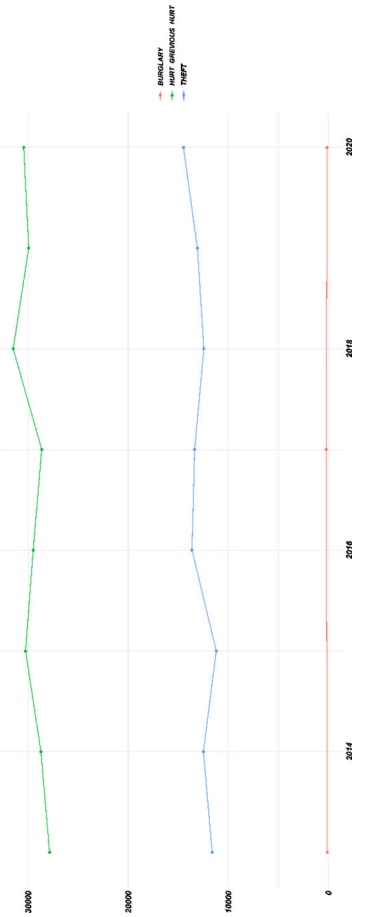

Fig. 20 Max crime Vs district of Madhya Pradesh

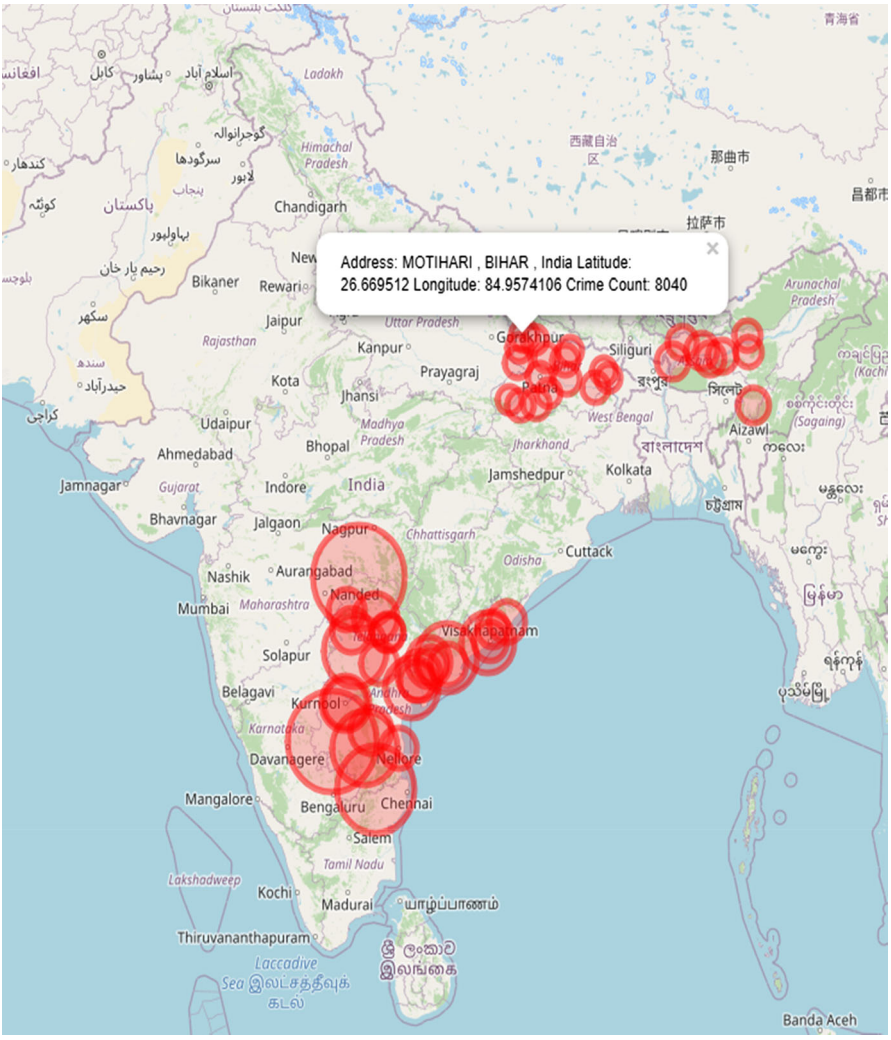

Fig. 21 Top 50 districts in India in 2020

ample data associated with crime maintained out there, such a data-driven approach can help police and other law enforcement organizations control and prevent crime. In this paper, efficient regression models are produced which were able to predict any type of crime count for a given region and year. As the data available was annual and not on a daily basis so the regression model cannot predict crime cunt for a specific day, also the model still predicts crime count for a wide region (area wise), that is it can be narrowed down with a better data set available.

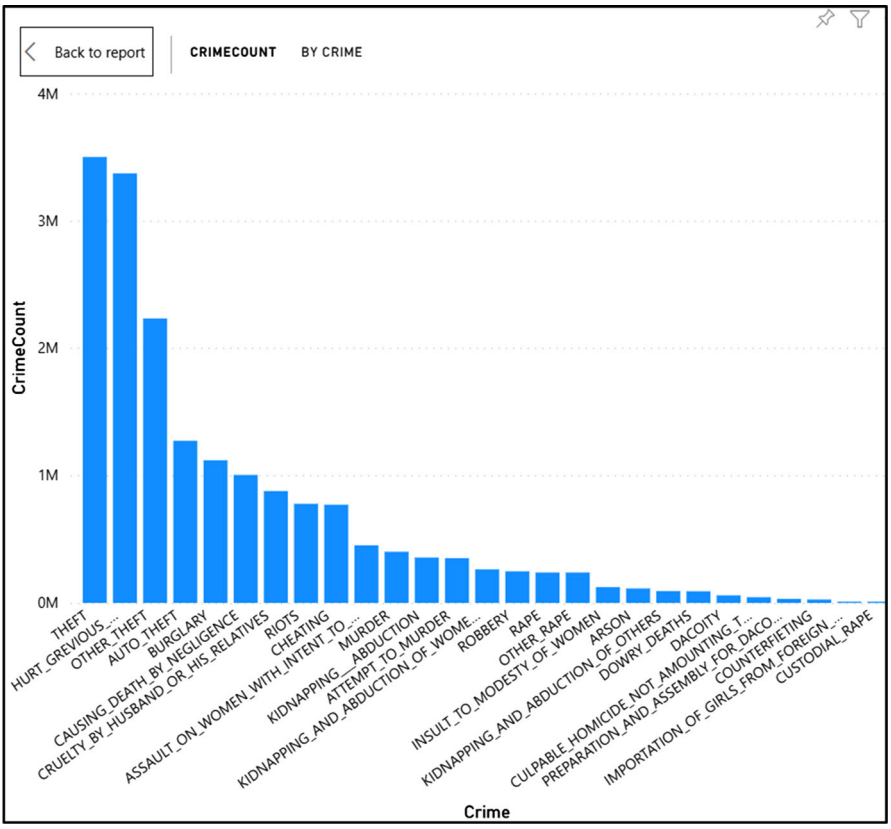

Fig. 22 Crime Vs crime count in decreasing order

5.1 Future Scope

The proposed framework is potential and flexible with any future crime data in terms of building efficient regression models, which can predict crime counts for different types of crimes and total IPC cognizable crime count region-wise, state-wise, and year-wise. For further studies it is suggested to use methods like spatiotemporal and time series analysis to study the recent years crime count data and the factors that contributes to the trends observed, if one can establish strong correlation between such factors and crime count change, the study will vitally contribute in understanding the complex structure of interconnecting factors contributing to crime. Currently in this paper, the challenge is that the data which is used is the best available data from the official NCRB website but still far from enough to make a strong machine learning driven system which can precisely contribute crime counts date wise and pin point the place of predicted crimes by the system, in future if organizations and researchers are able to maintain data with attributes like date, time, gender, crime type, verdict and other useful details then a much power machine learning data driven system can be achieved. The

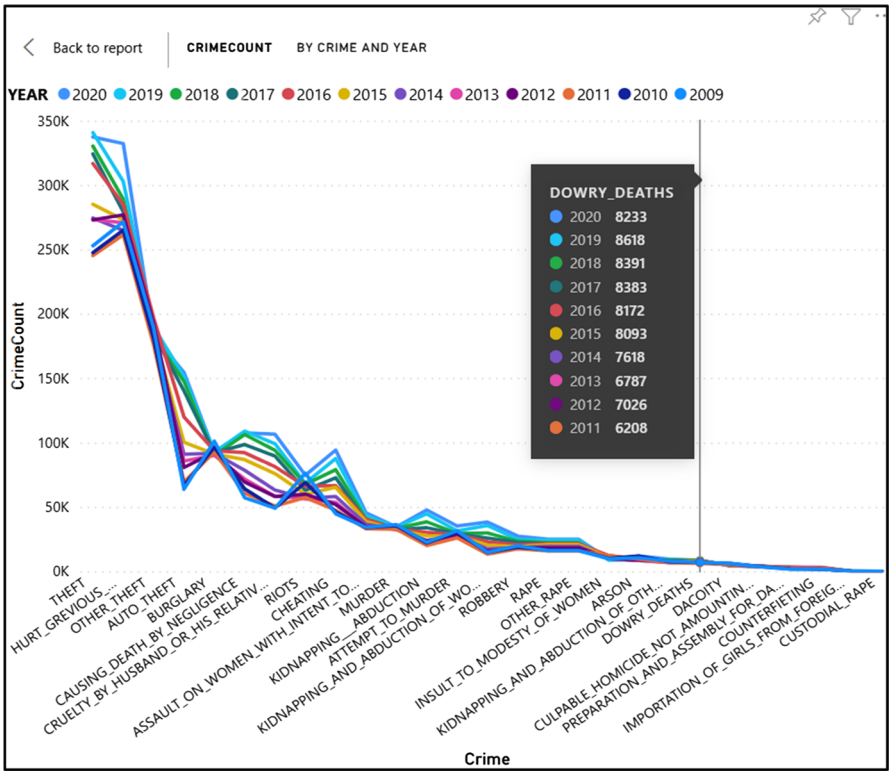

Fig. 23 which crime has more crime count and in which year it happens in more numbers

limitations of this paper are that there is not enough practical data to test the practicality of the regression models, the crime count predicted is still over a wide region area wise so it can be much more useful if it was more precise in terms of location of the predicted crime count. Using predicted data from regression models with efficient data visualization techniques and automated systems producing data analysis reports will help police and other law enforcement organizations to control crime rates and crime prevention. The proposed model will be highly useful with the fully developed (crime and criminal activity tracking systems) CCTNS, an Indian government project which is under development, an integrated system through which police departments distributed all over the country will be able to achieve effective policing through e-governance. It may also be a part of this system and will work wonders once incorporated or used with the system. Other data analysts, researchers, and contributors can use the proposed framework and approach in the future to build efficient regressors using the proposed regression models by making necessary changes.

Author Contributions Material preparation, data collection, and data analysis were performed by Prajwal Sharma and Aftab Hussian. Manuscript writing and all other work performed by Dr. Rabia Musheer Aziz.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability All used data are benchmark and are freely available in repositories.

Code Availability All used code are freely available on net.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical statements This material is the authors' own original work, which has not been previously published elsewhere. The paper is not currently being considered for publication elsewhere. The paper reflects the authors' own research and analysis in a truthful and complete manner.

References

1. Gupta M, Chandra B, Gupta MP (2014) A framework of intelligent decision support system for Indian police. J Enterp Inf Manag 27(5):512–540. https://doi.org/10.1108/JEIM-10-2012-0073
2. Himabindu BL, Arora R, Prashanth NS (2014) Whose problem is it anyway? Crimes against women in India. Glob Health Action 7(1):23718
3. Zavadzki T, de Pauli S, Kleina M, Bonat WH (2020) Comparing artificial neural network architectures for Brazilian stock market prediction. Ann Data Sci 7(4):613–628
4. Aziz R, Verma CK, Srivastava N (2017) A novel approach for dimension reduction of microarray. Comput Biol Chem 71:161–169
5. Misra S (2021) The Police System in India, Global Perspectives in Policing and Law Enforcement
6. Kassem M, Ali A, Audi M (2019) Unemployment rate, population density and crime rate in Punjab (Pakistan): an empirical analysis. Bull Bus Econ 8(2):92–104
7. Shi Y (2022) Advances in big data analytics: theory, algorithms and practices. Springer Nature, Singapore
8. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill/Irwin, New York, pp 2250–2254
9. Shermila AM, Bellarmine AB, Santiago N (2018) Crime data analysis and prediction of perpetrator identity using machine learning approach. In: 2018 2nd international conference on trends in electronics and informatics (ICOEI), 2018. IEEE, pp 107–114
10. Musheer RA, Verma C, Srivastava N (2019) Novel machine learning approach for classification of high-dimensional microarray data. Soft Comput 23(24):13409–13421
11. Aziz RM (2022) Cuckoo search-based optimization for cancer classification: a new hybrid approach. J Comput Biol. https://doi.org/10.1089/cmb.2021.0410
12. Shabat H, Omar N, Rahem K (2014) Named entity recognition in crime using machine learning approach. In Asia information retrieval symposium, 2014. Springer, pp 280–288
13. Goody J (2012) The theft of history. Cambridge University Press, Cambridge
14. Heeramun R, Magnusson C (2017) Gumpert CH, Granath S, Lundberg M, Dalman C, Rai D. Autism and convictions for violent crimes: population-based cohort study in Sweden. J Am Acad Child Adolesc Psychiatry 56(6):491–497
15. McDermott RC, Kilmartin C, McKelvey DK, Kridel MM (2015) College male sexual assault of women and the psychology of men: past, present, and future directions for research. Psychol Men Masc 16(4):355
16. Morewitz S (2019) Kidnapping and Violence: New Research and Clinical Perspectives. Springer, New York
17. van Dijk A, Wolswijk H (2017) Criminal liability for serious traffic offences: essays on causing death, injury and danger in traffic. Eleven International Publishing, Amsterdam
18. ToppiReddy HKR, Saini B, Mahajan G (2018) Crime prediction & monitoring framework based on spatial analysis. Procedia Comput Sci 132:696–705

19. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer, Berlin
20. Liao R, Wang X, Li L, Qin Z (2010) A novel serial crime prediction model based on Bayesian learning theory. In 2010 international conference on machine learning and cybernetics, 2010, vol 4. IEEE, pp 1757–1762
21. Hosseinkhani J, Taherdoost H, Keikhaee S (2021) ANTON framework based on semantic focused crawler to support web crime mining using SVM. Ann Data Sci 8(2):227–240
22. Keyvanpour MR, Javideh M, Ebrahimi MRJPCS (2011) Detecting and investigating crime by means of data mining: a general crime matching framework. Proc Procedia Comput Sci 3:872–880
23. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
24. Tayal et al (2015) (2015) Crime detection and criminal identification in India using data mining techniques. AI Soc 30(1):117–127
25. Awal MA, Rabbi J, Hossain SI, Hashem M (2016) Using linear regression to forecast future trends in crime of Bangladesh. In: 2016 5th international conference on informatics, electronics and vision (ICIEV), 2016. IEEE, pp 333–338
26. Yadav S, Timbadia M, Yadav A, Vishwakarma R, Yadav N (2017) Crime pattern detection, analysis & prediction. In: 2017 International conference of electronics, communication and aerospace technology (ICECA), 2017, vol 1. IEEE, pp 225–230
27. Kim S, Joshi P, Kalsi PS, Taheri P (2018) Crime analysis through machine learning. In 2018 IEEE 9th annual information technology, electronics and mobile communication conference (IEMCON), 2018. IEEE, pp 415–420
28. Kumar H, Sainia B, Mahajana G (2018) Crime prediction & monitoring framework based on spatial analysis. In: International conference on computational intelligence and data science, Jaipur
29. Rastogi I et al (2020) Knowledge discovery in databases for prediction of future crimes. Turk J Physiother Rehabil 32:3
30. Mittal M, Goyal LM, Sethi JK, Hemanth DJ (2019) Monitoring the impact of economic crisis on crime in India using machine learning. Comput Econ 53(4):1467–1485
31. Das P, Das AK (2019) Application of classification techniques for prediction and analysis of crime in India. In: Computational intelligence in data mining. Springer, pp 191–201
32. Hossain S, Abtahee A, Kashem I, Hoque MM, Sarker IH (2020) Crime prediction using spatio-temporal data. In: International conference on computing science, communication and security, 2020. Springer, pp 277–289
33. Pinto M, Wei H, Konate K, Touray I (2020) Delving into factors influencing New York crime data with the tools of machine learning. J Comput Sci Coll 36(2):61–70
34. Wheeler AP, Steenbeek W (2021) Mapping the risk terrain for crime using machine learning. J Quant Criminol 37(2):445–480
35. Hatcher WG, Yu WJIA (2018) A survey of deep learning: platforms, applications and emerging research trend. IEEE Access 6:24411–24432
36. Aziz RM, Baluch MF, Patel S, Kumar P (2022) A machine learning based approach to detect the Ethereum fraud transactions with limited attributes. Karbala Int J Mod Sci 8(2):139–151
37. Aziz RM, Hussain A, Sharma P, Kumar P (2022) Machine learning-based Soft Computing regression analysis approach for crime data prediction. Karb Int J Mod Sci 8(1):1–9
38. Aziz RM, Baluch MF, Patel S, Ganie AH (2022) LGBM: a machine learning approach for Ethereum fraud detection. Int J Inf Technol 29:1–1
39. Safat W, Asghar S, Gillani SA (2021) Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access 9(2021):70080–70094
40. Berger PD, Maurer RE, Cell GB (2018) Multiple linear regression. In: Experimental design. Springer, pp 505–532
41. Aziz RM (2022) Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data. Med Biol Eng Comput 60(6):1627–1646
42. Vural MS, Gök M (2017) Criminal prediction using Naive Bayes theory. Neural Comput Appl 28(9):2581–2592
43. Aziz R, Verma CK, Srivastava N (2018) Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction. Ann Data Sci 5(4):615–635
44. Cootes TF, Ionita MC, Lindner C, Sauer P (2012) Robust and accurate shape model fitting using random forest regression voting. In: European conference on computer vision, 2012. Springer, pp 278–291

45. Xia Z, Stewart K, Fan J (2021) Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major us metropolitan area. Comput Environ Urban Syst 87:101599
46. Aziz RM (2022) Application of nature inspired Soft Comput. techniques for gene selection: a novel frame work for classification of cancer. Soft Comput. https://doi.org/10.1007/s00500-022-07032-9
47. Aziz R, Verma C, Srivastava N (2015) A weighted-SNR feature selection from independent component subspace for NB classification of microarray data. Int J Adv Biotech Res 6(2015):245–255
48. Desai NP, Baluch MF, Makrariya A, MusheerAziz R (2022) Image processing model with deep learning approach for fish species classification. Turk. J. Comput. Math. Educ. 13(1):85–99
49. Lakovic V (2020) Modeling of entrepreneurship activity crisis management by support vector machine. Ann Data Sci 7(4):629–638