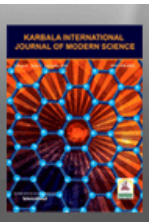


Karbala International Journal of Modern Science

Volume 8 | Issue 1 Article 1

Machine Learning-based Soft Computing Regression Analysis Approach for Crime Data Prediction

Rabia Musheer Aziz
VIT Bhopal University

Aftab Hussain
VIT Bhopal University

Prajwal Sharma
VIT Bhopal University

Pavan Kumar
VIT Bhopal University, pavankmaths@gmail.com

Machine Learning-based Soft Computing Regression Analysis Approach for Crime Data Prediction

Abstract

The crime rate in India is considerably increasing day by day. Consequently, the data associated with crime is also increasing, opening doors for data-driven approaches to these data to extract insightful knowledge, which can help police and other law enforcement organizations of the country in crime control and prevention. Crime prediction using machine learning algorithms on crime data can predict region-wise crime counts. In this paper, a machine learning-based soft computing regression analysis approach for Indian Crime Data Analysis (ICDA) is proposed. Different regression algorithms, namely, Simple Linear Regression (SLR), Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Support Vector Regression (SVR), and Random Forest Regression (RFR) are uses to build regression models. These regression models can predict a total number of Indian Penal Code (IPC) crime counts and crime counts of different types of crime (murder, rape, kidnapping and abduction, riots, to name a few) region-wise and state-wise and all over the country for a given year. Adjusted R squared value and Mean Absolute Percentage Error (MAPE) is used to evaluate and compare proposed regression models. In the proposed approach for ICDA, district-wise spatial-temporal crime data of years 2001 to 2012 is used, collected from the official website of NCRB. For the chosen data, it is concluded that the region-wise total IPC crime prediction RFR model fits the best with an adjusted R squared value of 0.9631551 and an error of 0.2027437. Whereas for region-wise thefts crime count prediction, the RFR model fits the best with an adjusted R squared value of 0.966604 and an error of 0.16571.

Keywords

Indian Penal Code (IPC), Support Vector Regression (SVR), Random Forest Regression (RFR), Decision Tree Regression (DTR), Multiple Linear Regression (MLR), Machine Learning algorithms, Indian Crime data analysis (ICDA)

This research paper is available in Karbala International Journal of Modern Science: https://kijoms.uokerbala.edu.iq/home/vol8/iss1/1

RESEARCH PAPER

Machine Learning-based Soft Computing Regression Analysis Approach for Crime Data Prediction

Rabia M. Aziz [a], Aftab Hussain [b], Prajwal Sharma [b], Pavan Kumar [a,*]

[a] Department of Mathematics, VIT Bhopal University, Sehore, 466116, India
[b] Department of Computer Science, VIT Bhopal University, Sehore, 466116, India

Abstract

The crime rate in India is considerably increasing day by day. Consequently, the data associated with crime is also increasing, opening doors for data-driven approaches to these data to extract insightful knowledge, which can help police and other law enforcement organizations of the country in crime control and prevention. Crime prediction using machine learning algorithms on crime data can predict region-wise crime counts. In this paper, a machine learning-based soft computing regression analysis approach for Indian Crime Data Analysis (ICDA) is proposed. Different regression algorithms, namely, Simple Linear Regression (SLR), Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Support Vector Regression (SVR), and Random Forest Regression (RFR) are uses to build regression models. These regression models can predict a total number of Indian Penal Code (IPC) crime counts and crime counts of different types of crime (murder, rape, kidnapping and abduction, riots, to name a few) region-wise and state-wise and all over the country for a given year. Adjusted R squared value and Mean Absolute Percentage Error (MAPE) is used to evaluate and compare proposed regression models. In the proposed approach for ICDA, district-wise spatial-temporal crime data of years 2001−2012 is used, collected from the official website of NCRB. For the chosen data, it is concluded that the region-wise total IPC crime prediction RFR model fits the best with an adjusted R squared value of 0.9631551 and an error of 0.2027437. Whereas for region-wise thefts crime count prediction, the RFR model fits the best with an adjusted R squared value of 0.966604 and an error of 0.16571.

Keywords: Indian penal code (IPC), Support vector regression (SVR), Random forest regression (RFR), Decision tree regression (DTR), Multiple linear regression (MLR), Machine learning algorithms, Indian crime data analysis (ICDA)

1. Introduction

The crime rate in India is substantially increased over time. As per NCRB data, in 2019, on average, 8837 (per day) IPC cognizable criminal cases were registered in the country. Parallel to this, data associated with the criminal activities registered all over the country also increases day by day in volume [1,2]. NCRB is the responsible authority to maintain all the data associated with the criminal activities registered all over the country. On the one hand, where NCRB maintains the data with the help of Crime and Criminal Tracking Networks and Systems (CCTNS), whereas on the other hand, roles of data analysts and scientists come into play, who can extract knowledge from the data to make valuable insights [3,4]. In the case of crime data analysis, the temporal data can be used to perform time-series analysis on it. Also, it can be used to predict future crime counts for specific regions and all over the country.

Furthermore, this can be used to predict the probable future crime hot spots [5−7]. With the help of this, police and other organizations

Received 17 July 2021; revised 24 October 2021; accepted 26 November 2021.
Available online 1 February 2022

* Corresponding author at: Department of Mathematics, VIT Bhopal University, Sehore, 466116, India.
E-mail addresses: rabia.aziz2010@gmail.com (R.M. Aziz), ah24121999@gmail.com (A. Hussain), prajwalsharma579@gmail.com (P. Sharma), pavankmaths@gmail.com (P. Kumar).

https://doi.org/10.33640/2405-609X.3197

responsible for law enforcement can be more efficient and effective in their job [8]. The crime rate is affected indirectly and directly by many reasons which change from time to time, so it is pretty impossible to achieve such goals without a data-driven approach [9–12]. For crime count prediction, it is used some suitable regression algorithms, like, simple linear regression (SLR), multiple linear regression (MLR), decision tree regression (DTR), support vector regression (SVR), and random forest regression (RFR). In the proposed approach, predictive model templates are created and then fitted into the desired datasets so that these templates can be further used by other aspirants to create predictive models by making necessary changes.

1.1. Types of crime

In the proposed approach, regression models are built for predicting total IPC cognizable crime counts and crime counts for 28 different crimes. These crimes can be classified into some types of crime, which are as follows:

Theft: This crime is said to be committed if an individual treacherously takes possession of another individual's property without his/her consent. Auto theft comes under theft, in which the automobile is the property intended to be stolen. Every country has its punishment for this crime; primarily, this crime is not considered a significant problem, but generally, many cases are registered under this crime [13].

Violent Crimes: This crime is committed when an individual attempts or does harm someone or threaten another individual with violence. Rape, robbery, dacoity, riots, domestic violence, murder, attempt to murder, dowry deaths, active shooters, etc., are considered violent crimes. In this type of crime, as per IPC, the criminal gets a long time punishment or some time penalty of death [14].

Assault on women: The violent crime against women or girls comes under assault on women. Rape, eve-teasing, acid attack, gender-targeted crime, domestic violence, etc., come under assault on women. Such crimes are considered to be hate crimes. The motive in these types of crime is physical pleasure, superiority, entitlement, etc., [15].

Kidnapping: It is a crime in which a person is transported, taken away, and held/confined against his/her consent by threatening him/her or using violence or deception. Kidnapping and abduction of women and girls, kidnapping and abduction of children, etc., come under this crime [16].

Causing death by negligence: This crime is committed when someone dies due to the negligence or careless act of another person who did not intend to kill the individual who died. Sometimes buildings are demolished, significant accidents, patient death, etc., types of events are happened due to negligence. This negligence is one of the serious crimes, and there are very high penalties or punishments in most countries against this crime [17].

1.2. Related works

Tayal et al. [18] proposed a method to design and implement crime detection and identification in Indian cities with the help of data mining techniques. They have 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.

Awal et al. [19] proposed an SLR fitted to the Bangladesh crime data set to forecast future crime trends of Bangladesh. For different regions, they used this model to forecast crime for specific crimes such as robbery, dacoity, women and child repression, murder, etc.

Yadav et al. [20] proposed an SLR 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 on the data set to increase the accuracy of the predictive model. Kim et al. [21] proposed a machine learning-based approach for crime prediction. They used two different data sets obtained by two approaches of data pre-processing. They used K-Nearest Neighbour and boosted decision tree classification algorithms to build classifiers for both the data sets and achieved 39–44% accuracy. The accuracy of the built classifiers is too low, but they can be used as a basic framework for further use. Mittal et al. [22] proposed an approach to monitor the effect of the economic crisis on the crime rate in India. They used SLR, DTR, RFR algorithms, and a neural network algorithm on 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. Das et al. [23] proposed numerous classification

techniques for crime prediction and analysis in India. They used KNN, DTR, RFR, 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, etc., parameters.

Hossain et al. [24] proposed an approach to predict criminal activities with the help of supervised learning algorithms. K- nearest neighbours (KNN) and Decision tree is used on 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 RFR with Adaboost to increase the accuracy. Log–loss is used to measure the performance of classifiers by penalizing false classification. With a random under-sampling method for the RFR, the classifier built gives the best accuracy. The absolute accuracy is 99.16%, with a 0.17% log loss.

Pinto et al. [25] proposed an approach to minimize the adverse action or crime harmful to human society with the help of the model of the machine learning algorithms. KNN, Decision tree, Multivariate linear regression classification algorithms are used with the New York crime data set for 2019. With the help of past crime data, this model shows trends vs patterns in a crime, which helps correlate the factor that helps predict future crimes. The decision tree gives the best prediction (99.95%) to the borough with the correct name. The decision tree deals better in an extensive data set with many nodes with different layers and a small target with only five boroughs. There is a limited number of possible decision points, which is why it gives the best accuracy. The KNN comes in 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 the three models Is that they give a reasonable accuracy rate at a limited target.

Using the machine learning algorithm, Wheeler et al. [26] give the accurate long–term prediction of crime in micro–cities compared to the other popular techniques and how their advanced model is improving their interpretability helps open the "black box" of RFR. Using this model, they estimate future crimes using different measures of predictive accuracy, and their model accurately predicts the crime in micro places. However, they are unable to understand why these places are predicted very riskily.

Safat et al. [27] proposed an approach to estimate the accurate crime rate, types of different crimes, and hot spot places from the past pattern with the help of machine learning and deep learning techniques. They used machine learning algorithms viz., SVM, Naive Bayes, KNN, decision tree, Multilayer perceptron neural net, RFR, 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 valuable inferences based on their experimentation results.

Aytug Onan et al. [28] proposed supervised machine learning algorithms and sentimental analysis. These methods are beneficial for extracting information from text documents online. Some of the researchers applied different classification methods to enhance the predictive preformation of sentiment classification. The proposed ensemble method incorporates Bayesian logistic regression, naive Bayes, linear discriminant analysis, logistic regression, and support vector machine as a base learner for the static classifier. The proposed classification scheme can predict better than conventional ensemble learning methods such as AdaBoost, bagging, random subspace, and majority voting. The best classification accuracy is 98.86% [29–31].

Various researchers proposed a sentimental analysis to express their views, complaints, feelings, and attitudes towards subjects. For that, it uses a sentimental analysis with supervised and unsupervised term weighting schemes. Most researchers used four supervised learning algorithms (i.e., Naïve Bayes, support vector machines, k-nearest neighbour algorithm, and logistic regression) with ensemble learning methods (i.e., AdaBoost, Bagging, and Random Subspace) to explore the predictive efficiency of the term weighting schemes. The experimental results indicate that supervised term weighting models can outperform unsupervised term weighting models [32–36].

Some researchers proposed a sentimental analysis with natural language processing, text mining, and web mining which aims to extract subjective information in source materials. In these methods, feature selection becomes essential in developing a robust and efficient classification method while reducing the training time. The proposed model with these methods indicated that the classification model with feature selection methods is an efficient method. It outperforms individual filter-based feature selection methods for sentiment classification [37–40].

Aytug Onan et al. [37] proposed a text genre classification method to identify text documents'

functional characteristics. The immense quantity of text documents available on the web can be adequately filtered, organized, and retrieved, which may have potential use on several other tasks of natural language processing and information retrieval. Text genre classification is typically performed by supervised machine learning (SVM) algorithm. The highest average predictive performance obtained by the proposed scheme is 94.43%.

M.A. Tocoglu et al. [38] proposed a sentimental analysis method to extract subjective information in the source material. Using this method, it encountered an overwhelming amount of data available. Two essential tasks for achieving scalability in machine learning based on sentimental analysis are instance selection and feature selection. The instance selection methods are evaluated by a decision tree classifier (C4.5 algorithm) and radial basis function networks regarding classification accuracy and data reduction rates. The experimental results indicate that the highest classification accuracy on the C4.5 algorithms is generally obtained by the model class selection method. In contrast, the nearest centroid neighbour edition obtains the highest classification accuracies on radial basis function networks.

Aytug Onan et al. [39] proposed a method that used a sentimental analysis to identify and classify users' views from text documents into different sentiments, such as positive, negative, or neutral. This method is used to extract structured and informative knowledge from unstructured text pieces. In the experimental analysis, three conventional text representation schemes (i.e., term-presence, term-frequency, TF-IDF scheme) and three N-gram models (1-g, 2-g, and 3-g) have been considered in conjunction with four classifiers (i.e., support vector machines, Naïve Bayes, logistic regression, and random forest algorithm).

The predictive performance of four ensemble learners (i.e., AdaBoost, Bagging and Random Subspace, and voting algorithm) have also been evaluated. The empirical results indicate that the machine learning-based approach yields promising results on students' evaluation of higher educational institutions [40].

The objectives of the proposed work are:

i. To create effective predictive model templates in R, which others can further use for performing regression analysis on similar crime data, which can be achieved by making necessary changes.
ii. To build predictive models with Indian district-wise crime data (2001–12), which can predict/forecast total IPC and specific crime counts, region-wise and all over the country.
iii. To use different regression models and choose the best one for predicting different types of crime and total IPC crime counts.

1.3. Proposed work

Fig. 1 shows the flow chart of the proposed work. In the proposed approach for ICDA, the desired crime data set is collected from NCRB records. Afterwards, using MySQL workbench's data import wizard feature, the pre-processed data set (data set after taking care of missing values using R) is imported to the ICDA research database. MySQL workbench smartly creates the table schema as per the data set imported. Then, different 'select' queries are used to generate and export different derived and transformed data set (as per desired future analysis and modelling) in.csv file format. Different regression algorithm-based templates and data pre-processing templates are written in R, and scripts are saved. Then, using desired data pre-processing on different derived data sets, data sets are pre-processed. Then, regression models are created and further evaluated using different regression templates based on their Adjusted R squared value and mean absolute percentage error in predicting the corresponding test set data. After concluding the relatively best regression models, these models are used to predict the crime counts for different regions in India for the year 2022. Leaflet library in R is used to generate map plots to show the spatial-temporal predictions, showing the top 50 predicted regions in India with relatively high crime count.

2. Methods used

2.1. Regression algorithms used

Following is the elaboration of the regression algorithms which has been used in this paper for predictive modelling:

2.1.1. SLR

SLR is the statistical model used to predict the relationship between the independent and dependent variables. It is used to quantify 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.

More the R-square value more the data is fitted to the model. On increasing the data set to measure the model goodness, it is observed an adjusted R-

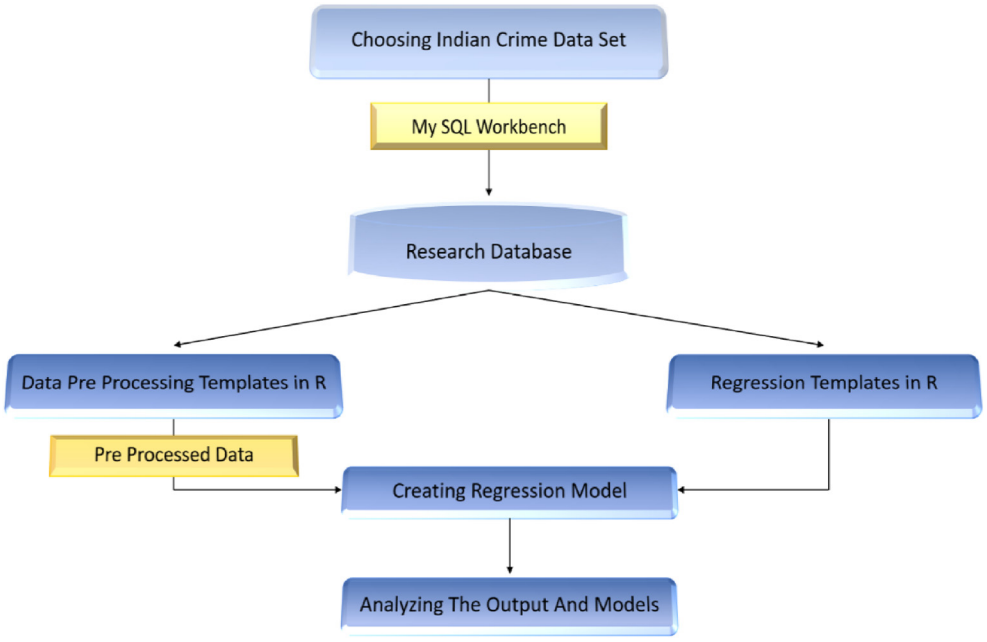


Fig. 1. The flow chart of proposed ICDA Approach.

square value. The representation of the SLR model is given as [19]:

$$Y = a*X + C, \quad (1)$$

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

2.1.2. MLR

MLR is one of the essential 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 continuous, and the independent variable may be continuous or categorical. The representation of the MLR model given as [41,42]:

$$Y = a*X1 + b*X2 + c*X3 + C,$$

where Y: Dependent Variable
X1, X2, X3, ….: Independent Variables
C: Intercept
a,b,c, …: Coefficients of X1, X2, X3, …

2.1.3. DTR

DTR is one of the two encompassed terms in the umbrella term CART, which stands for classification and regression trees. DTR algorithms are generally used to classify the labelled data, but they can also be used for regression analysis. In the training process, initially, the whole data set is considered as the root node. Then different split boundaries are created among the data points at successive levels of the tree. The best split is decided based on the information gain value of the split. Split with the highest information gain is chosen as the best split. There are different DTR algorithms; they all differ in the formula used to calculate the information gain. In the case of decision tree classification, the leaf nodes are the labels of labelled data. In the case of DTR, leaf nodes are the average value of the data points belonging to the region, resulting from splits in the path from the root to the corresponding leaf node [43,44].

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. Therefore,

$$\text{Information Gain (X)} = \text{Entropy(S1) - Entropy(S2)}, \quad (3)$$

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,

$$\text{Entropy}(S1) = \sum_{i=1}^{n} - p_i \log_2(p_i) \quad (4)$$

where n is the total number of class levels, pi is the percentage of data points in class level. The entropy of a set of partitions S2 is given by the weighted sum of the entropy of all the partitions.

$$\text{Entropy}(S2) = \sum_{i=1}^{n} -\text{wi}*\text{Entropy}(P_i)$$

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

2.1.4. RFR

RFR is an ensemble learning-based method. Ensemble learning uses multiple algorithms or the same algorithm multiple times to yield a much more robust algorithm. In the case of the random forest algorithm, DTR is used multiple times to create a forest of decision trees. In the training process of RFR, *k* data points are chosen at random from the training set. Then a decision tree is built with these *k* data points. These two steps are repeated n number of times to create a forest of n decision trees built on n randomly chosen sets of data points. This forest of DTR 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 values is given as the final predicted value of the dependent variable [45–47].

2.1.5. SVR

SVR is a supervised machine learning model that can be used to perform classification and regression analysis on data. For the given sets of vectors in multi-dimensional space, SVM finds a hyperplane (in the case of two-dimensional space, the hyperplane will be a line) that separates the labelled data in space that belongs to different class levels. This hyperplane is often referred to as the maximum margin hyperplane (MMH). SVM can deal with linearly as well as non-linearly separable data with the use of a slack variable. However, 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 which is 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 also when only support vectors are known. 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 [48–51].

3. Experimental setup

This section describes the workflow 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 R squared value is given by,

$$R^2 = 1 - SS_{res}/SS_{tot}, \quad (6)$$

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 worse than the average curve. R squared generally ranges from 0 to 1.

3.1.2. Adjusted R squared

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

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

where n is, the data sample size and p is the number of independent variables. There exist several formulas for calculating adjusted R squared, the McNemar's formula is used.

3.1.3. MAPE (mean absolute percentage error)

MAPE is a measure used for evaluating the prediction accuracy of a predictive model. The formula for MAPE is given by,

$$M = (1/n)\sum_{k=1}^{n} |(\mathrm{Ak} - \mathrm{Fk})/\mathrm{Ak}|, \quad (8)$$

where Ak is the actual value, Fk is the predicted value and the data sample size.

3.2. Data used

The raw data set is taken from the official site of NCRB. The data set contains 9012 records and 33 columns (variables). Each record is distinctly based on 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 28 crimes mentioned above.

TOTAL IPC CRIMES: Count of total cognizable IPC crimes.

Thus, this data set contains information about the total number of cognizable IPC cases and the number of cases registered under different cognizable IPC crimes from 2001 to 2012 for 808 Indian districts and 35 states and union territories. Fig. 2 show the snapshot of the chosen raw data set.

3.2.1. Data pre-processing

Data pre-processing is an essential step of data 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. In our approach for ICDA, NA values (missing values) in a column are replaced by the mean of the values in the corresponding column. Regression algorithms need numerical data, and label encoding will be the wrong approach for the desired regression analysis. The Dummy data frame method from the dummies library in R creates dummy variables for each state/ut and district. It leads to 35-1 state dummy columns and 808-1 district dummy columns. One dummy variable is not created for each categorical variable to avoid the dummy variable trap for a feature containing n class levels n number of dummy variables. A dummy variable stores 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.

The raw data set contains information about district-wise registered cognizable IPC crimes for 2001 to 2012 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 by the mean value of the corresponding column using R. The data set is then imported into a database using MySQL workbench. With the help of structured queries, state-wise and year-wise, crime data was generated. After which, dummy coding is performed on the categorical variables in these data sets before using them to build regression models. After producing desired derived data sets, these data sets are then used to build desired regression models. Derived state-wise, district-wise, and year-wise crime data are then used to build different regression models using different regression algorithms, namely, SLR, MLR, DTR, SVR, and RFR algorithms. These algorithms are used to build regression models for each derived data set by

	A	B	C	D	E	F	G	
1	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMIC	RAPE	CU
2	ANDHRA PR	ADILABAD	2001	101	60	17	50	
3	ANDHRA PR	ANANTAPU	2001	151	125	1	23	
4	ANDHRA PR	CHITTOOR	2001	101	57	2	27	
5	ANDHRA PR	CUDDAPAH	2001	80	53	1	20	
6	ANDHRA PR	EAST GODA	2001	82	67	1	23	
7	ANDHRA PR	GUNTAKAL	2001	3	1	0	0	
8	ANDHRA PR	GUNTUR	2001	182	88	2	54	
9	ANDHRA PR	HYDERABAI	2001	111	113	7	37	
10	ANDHRA PR	KARIMNAG/	2001	162	85	6	56	
11	ANDHRA PR	KHAMMAM	2001	93	60	1	47	
12	ANDHRA PR	KRISHNA	2001	65	51	0	37	
13	ANDHRA PR	KURNOOL	2001	133	72	4	29	
14	ANDHRA PR	MAHABOOE	2001	157	67	26	59	

Fig. 2. District Wise Indian Crime Data (2001–2012) snapshot.

choosing 1 of the 29 dependent variables (28 variables corresponding to different types of cognizable IPC crime and one variable corresponding to total cognizable IPC crimes). After this, regression models predicting crime counts-state, district-wise, and year-wise- are built by making necessary changes in R scripts as per the following chosen dependent variable. Each R script written yields to 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 Table 1 and Table 2. The best model is chosen based on the lowest MAPE value and highest adjusted R squared value among these models. The relatively best model among these is then used to predict district-wise total IPC crime counts and theft crime counts, visualized using leaflet map plots in R software (Figs. 9–10).

4. Experimental results & discussion

This section discusses the results of twelve regression models based on four different regression algorithms, namely, MTR, DTR, RFR, and SVR. These twelve models can be further divided into three groups. The first group consists of models which can predict district wise total cognizable IPC crime counts given district, state, and year data. The second group consists of models which can predict theft crime counts given district, state and year data. At the same time, the third group consists of models which can predict the crime rate per 100k population for a given year. Also, the leaflet map plots based on total IPC crime counts and theft crime predicted by the respective random forest model are discussed in this section.

4.1. Plots to visualize fitness of regressors

Fig. 3(a–b) and Fig. 4(a–b) show the number of total IPC crime cases vs district labels plot, and Fig. 5(a–b) and Fig. 6(a–b) show the number of theft crime cases vs district labels plots for the year 2012. District labels are given to different districts in the training data set and are not used while training the models. They are created to make the data visualization part easy for visualizing the fitness of the regression models. These plots are drawn to visualize the fitness of different regression curves to the training data set belonging to their corresponding regression models. As for each year, there exist more than 800 records in the data set, so the crime counts (total/specific) vs different district plots for a specific year (2012). Although the district label variable is not continuous, considering it continuous only while trying to visualize the fitness of regression models will not do any harm as the model is not compromised in visualizing the fitness.

Different regression templates are used to build different regression models based on the four regression algorithms. Then for each model, the training data set is filtered with records having year equals to 2012 value. The regression curve is plotted using the geom_line () function in R with district labels on the x-axis and predicted total IPC crime counts on the y-axis which is predicted by the corresponding regression model using the previously mentioned filtered training set. Also, the actual data points of the filtered training set are plotted using the geom_point () function in R with district labels on the x-axis, and actual total IPC counts on the y-axis within a common plane using the ggplot () function in R.

In Fig. 3(a–b) and Fig. 4(a–b) blue curve are the regression curves, and red dots are the actual data points. From plots in Fig. 3(a–b) and Fig. 4(a–b), the fitness of different regression models (which can

Table 1. Comparison among regression models predicting district wise total IPC crimes.

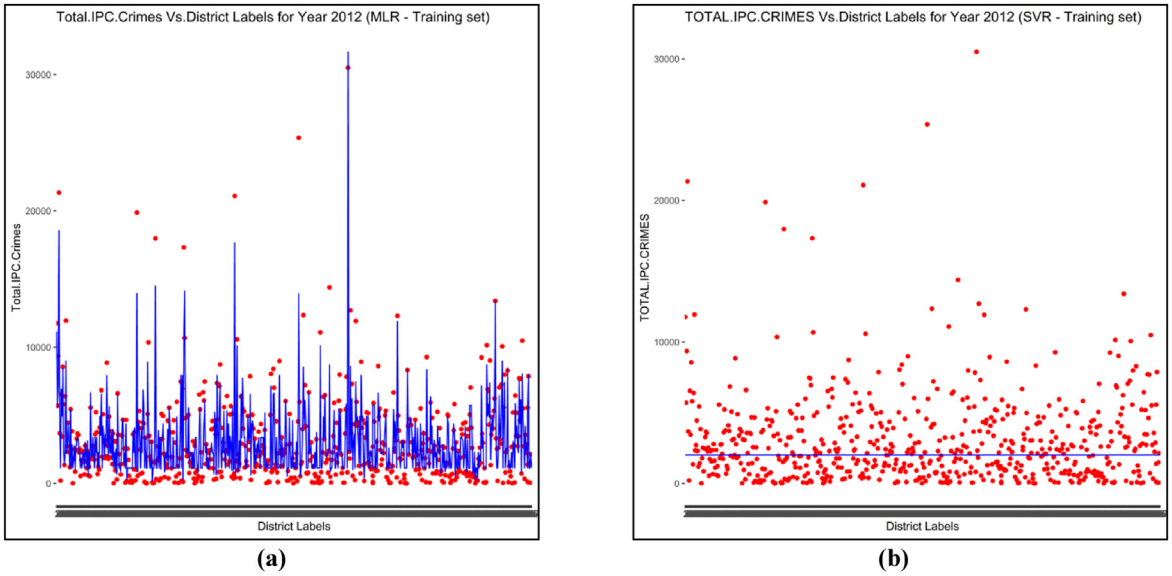
Regression Model Used	R-Squared Value	Adjusted-R squared Value	MAPE
MLR	0.8935085	0.893493	1.99711
DTR	0.5735719	0.5735099	4.543087
RFR	0.9631605	0.9631551	0.2027437

*SVR model is not considered for comparison as it fitted the training data poorer than the average line.

Table 2. Comparison among regression models predicting district wise theft crimes.

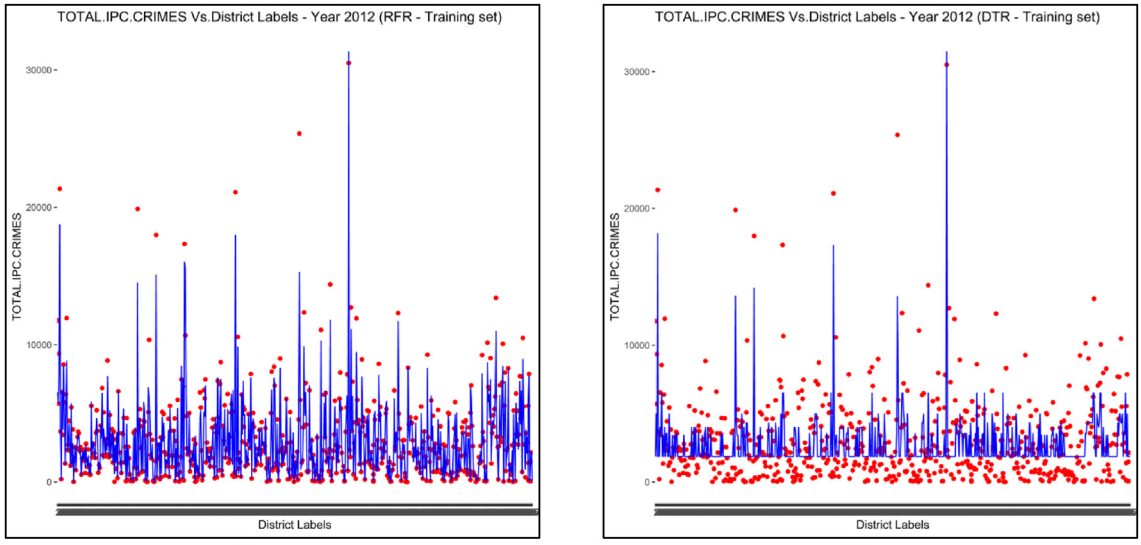
Regression Model Used	R-Squared Value	Adjusted-R squared Value	MAPE
MLR	0.9185906	0.918579	0.8956748
DTR (min. Split = 2)	0.7131494	0.7131085	0.5951368
RFR	0.9666091	0.9666044	0.16571

*SVR model is not considered for comparison as it fitted the training data poorer than the average line.



(a) (b)

Fig. 3. (a–b) Total IPC Crimes Vs. District labels with MLR and SVR model for year 2012.



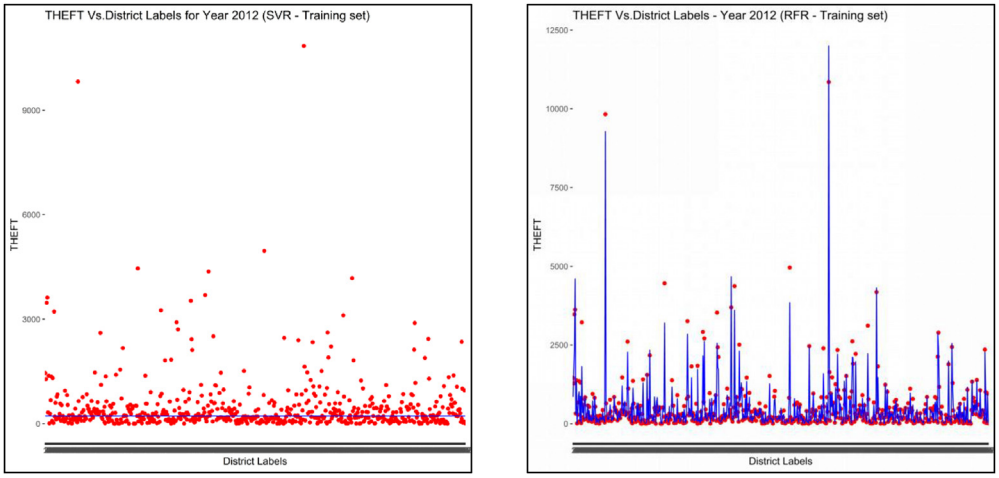
(a) (b)

Fig. 4. (a–b) Total IPC Crimes Vs. District labels with RFR and DTR model for year 2012.

predict total IPC crime counts) on the training data set can be observed. It can be concluded that random forest regression models fit relatively the best for this training data, followed by Multiple linear regression, decision tree regression model, and support vector regression model. To mathematically evaluate and compare these models, adjusted R squared and mean absolute percentage error on test set values are used in Section 4.2.

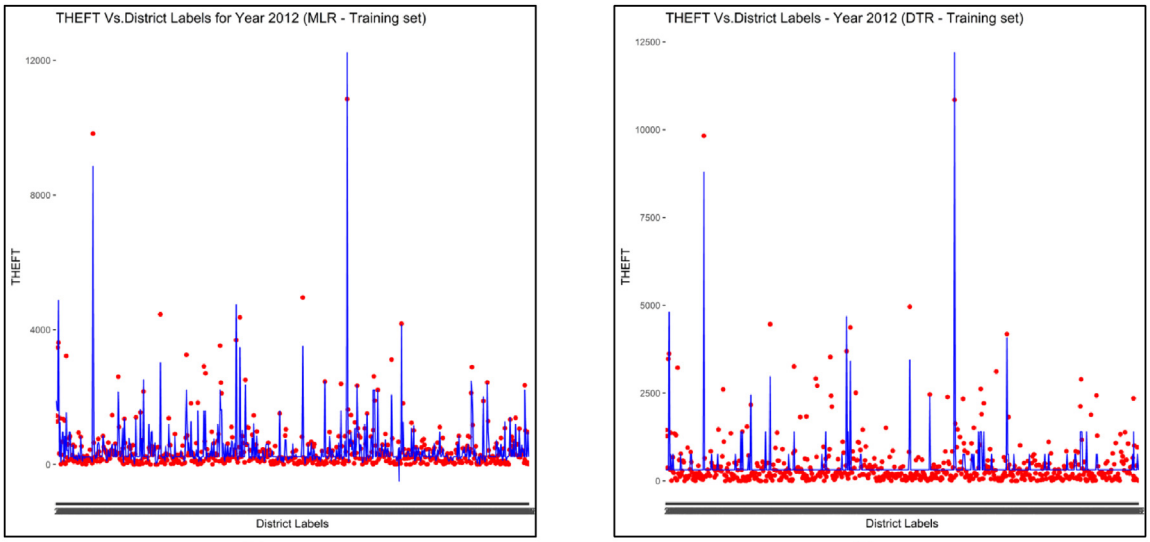
Figures Fig. 5(a–b) and Fig. 6(a–b) shows the fitness plots for different regression models (which can predict theft crime counts) on the training data set. The plots are shown blue curves are the respective regression curves, and red dots are the actual data points. From these plots, it can be concluded that the random forest-based regression model fits the best for the training data followed by MLR based model and DTR based model. SVR based model fits even poorer than the average curve for the training data. These models are mathematically evaluated and compared in Section 4.2.

Figs. 7–10 presents the fitness of regression curves based on different regression models which can predict crime rate per 100k in India for a given year. The plots in Figs. 7–10 have the blue line as the regression curve made by predicted crime rate



(a) (b)

Fig. 5. (a–b) Thefts Vs. District labels with SVR and RFR model for year 2012.



(a) (b)

Fig. 6. (a–b) Thefts Vs. District labels with MLR and DTR model for year 2012.

per 100k population by the regressor on the training data, and red dots are the actual data points from the training set. It can be concluded from the plots in the figures Figs. 7–10 that SLR base regression fits relatively worst as compared to other regression models. To mathematically evaluate and compare these models, the adjusted R squared and MAPE values are used on the test data set in section 4.2.

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 show the comparison of regression models built to forecast total IPC crime counts and Table 2 show the comparison of regression models built to forecast theft crimes, and Table 3 show the comparison among regression models built to predict crime rate for a given year.

Observations:

(i) From Table 1, it is observed that the regression model built using the random forest for district-wise total IPC crime count forecasting is relatively the best model with 0.9631551 adjusted R squared value and

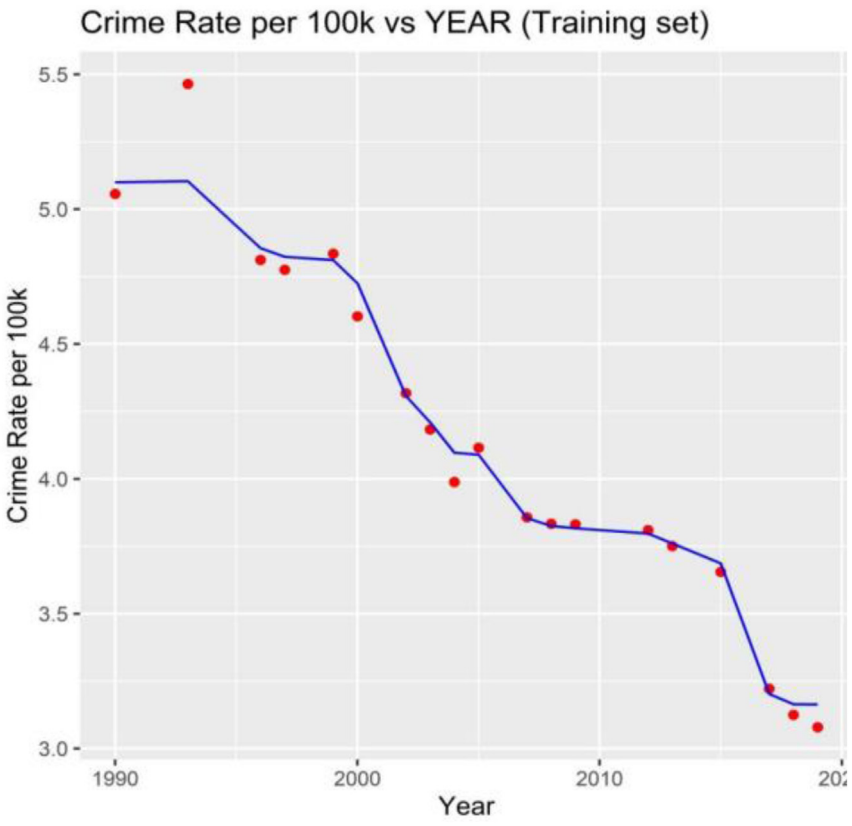


Fig. 7. Crime Rate per 100k population vs.Year (RFR model).

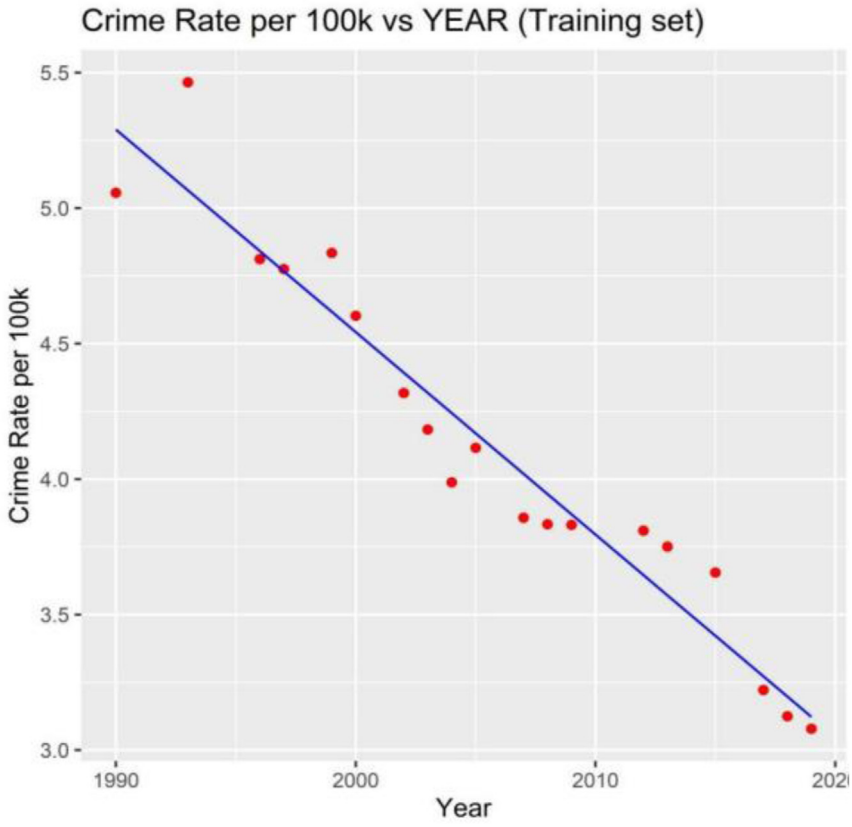


Fig. 8. Crime Rate per 100k population vs.Year (SLR model).

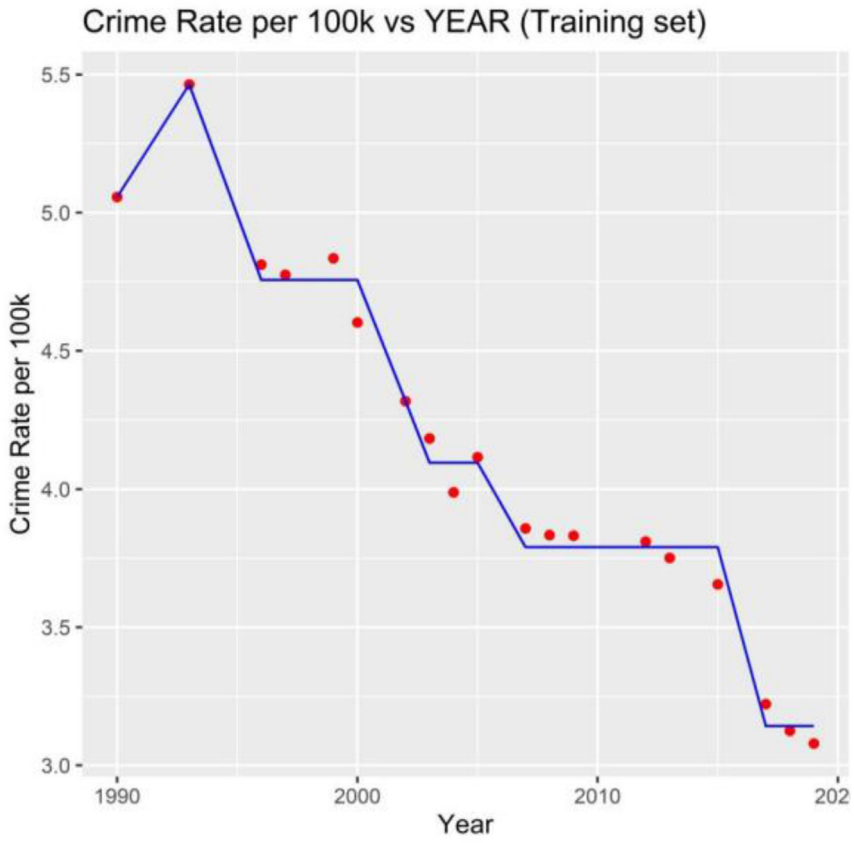


Fig. 9. Crime Rate per 100k population vs.Year (DTR model).

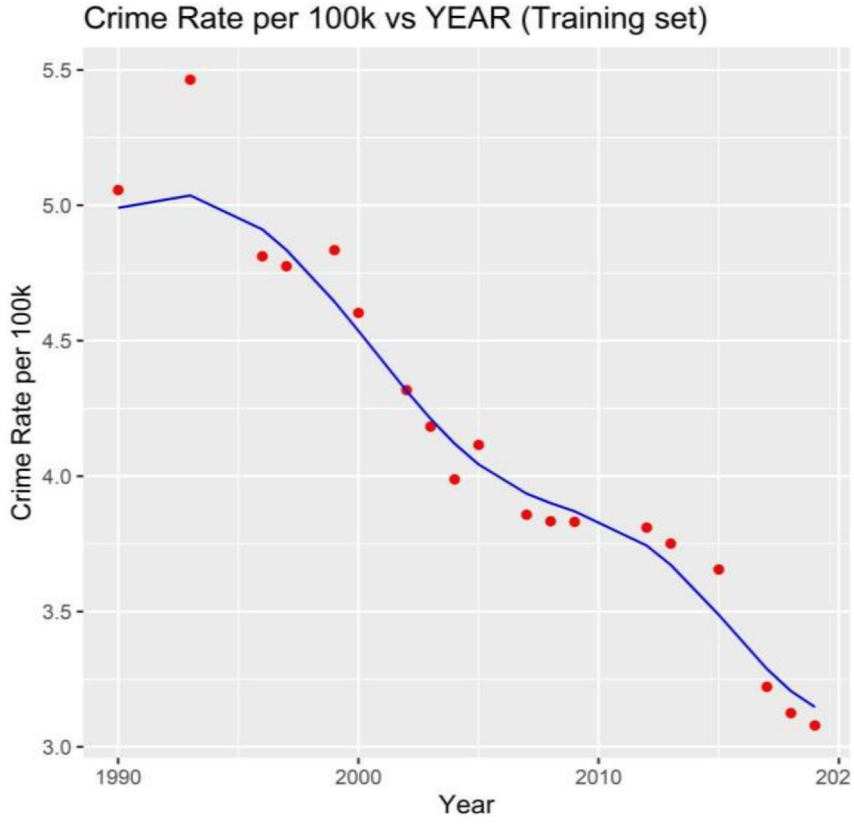


Fig. 10. Crime Rate per 100k population vs.Year (SVR model).

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

Regression Model Used	R-Squared Value	Adjusted-R squared Value	MAPE
SLR	0.9345979	0.9307507	0.0243525
DTR (min. Split = 2)	0.9884693	0.987791	0.03548595
RFR	0.9779519	0.976655	0.02685891
SVR	0.9584286	0.9559832	0.01971657

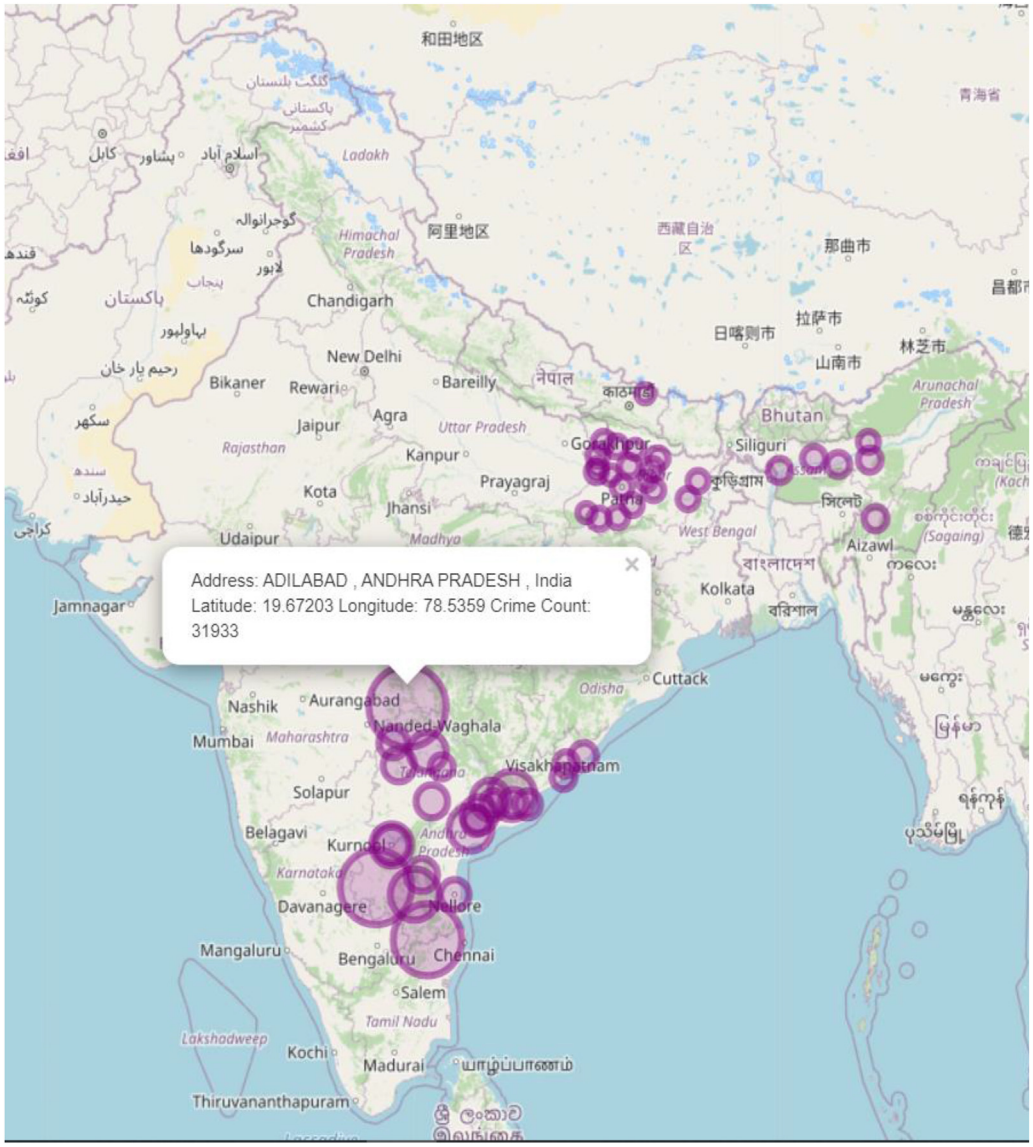


Fig. 11. Leaflet map plot for top 50 districts with highest number of predicted theft crime counts in 2022.

0.2027437 MAPE value on predicting the test set data.

(ii) From Table 2, it is observed that the regression model built using the random forest for district wise theft crime count forecasting is relatively the best model with 0.9666044 adjusted squared value and 0.16571 MAPE value on predicting the test set data.

(iii) Table 3, it can be concluded that the regression model built using decision tree regression is relatively the best for predicting crime rate per 100k for a given year with an adjusted R squared value of 0.987791 and 0.03548595 MAPE value on predicting the test set data.

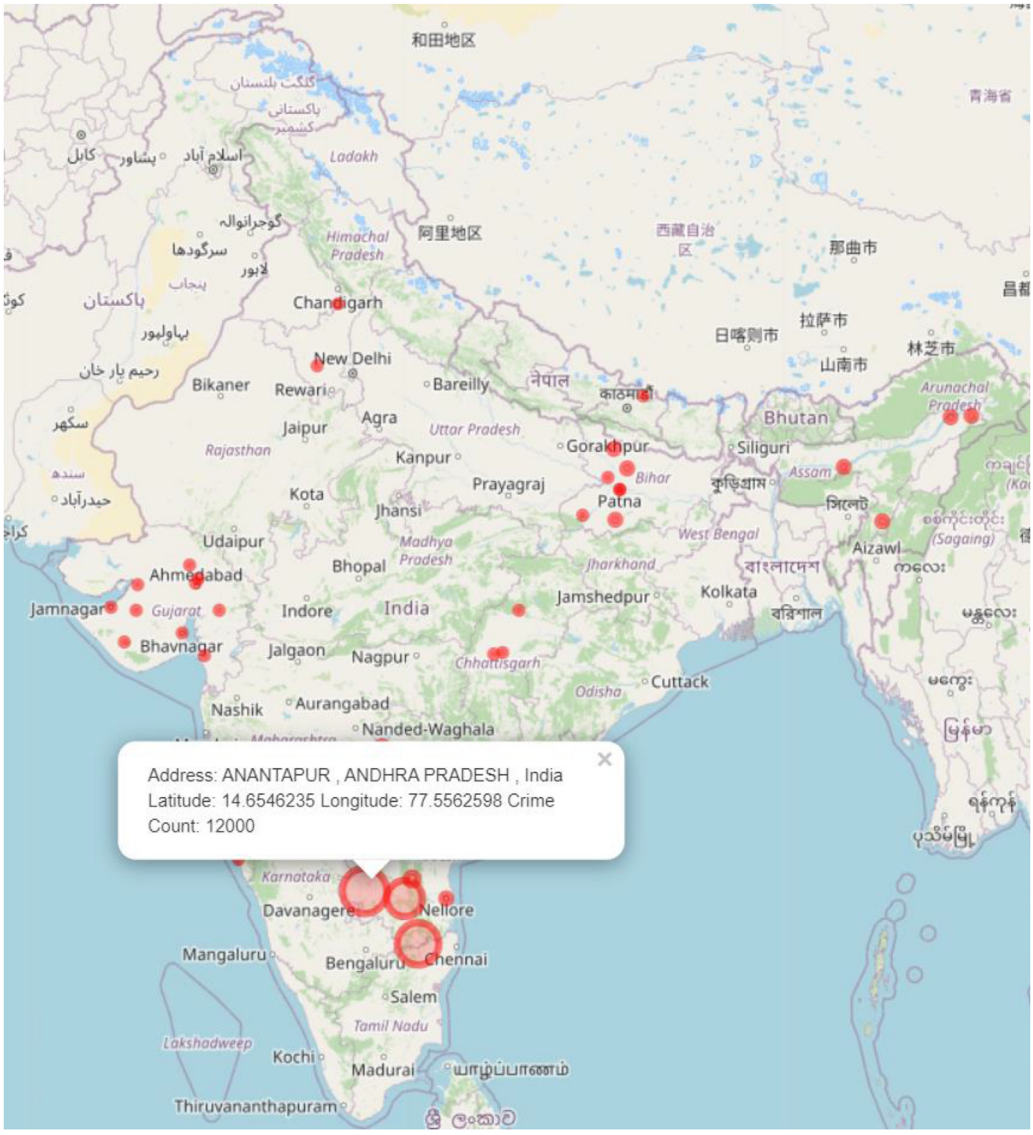


Fig. 12. Leaflet map plot for top 50 districts with highest number of predicted total IPC crime counts in 2022.

4.3. Maps plots

Fig. 11 and Fig. 12 are the map plots plotted using 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 the random forest algorithm.

Observations:

i. Fig. 11 depicts 50 districts with the highest predicted total IPC crime counts 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 district/regions for the year 2022. The larger the radius of the circular mark higher is the value of the predicted crime count.

ii. Fig. 12 depicts 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 higher is the value of the predicted crime count.

iii. These leaflet plots are created so that the predicted crime count can be seen when the

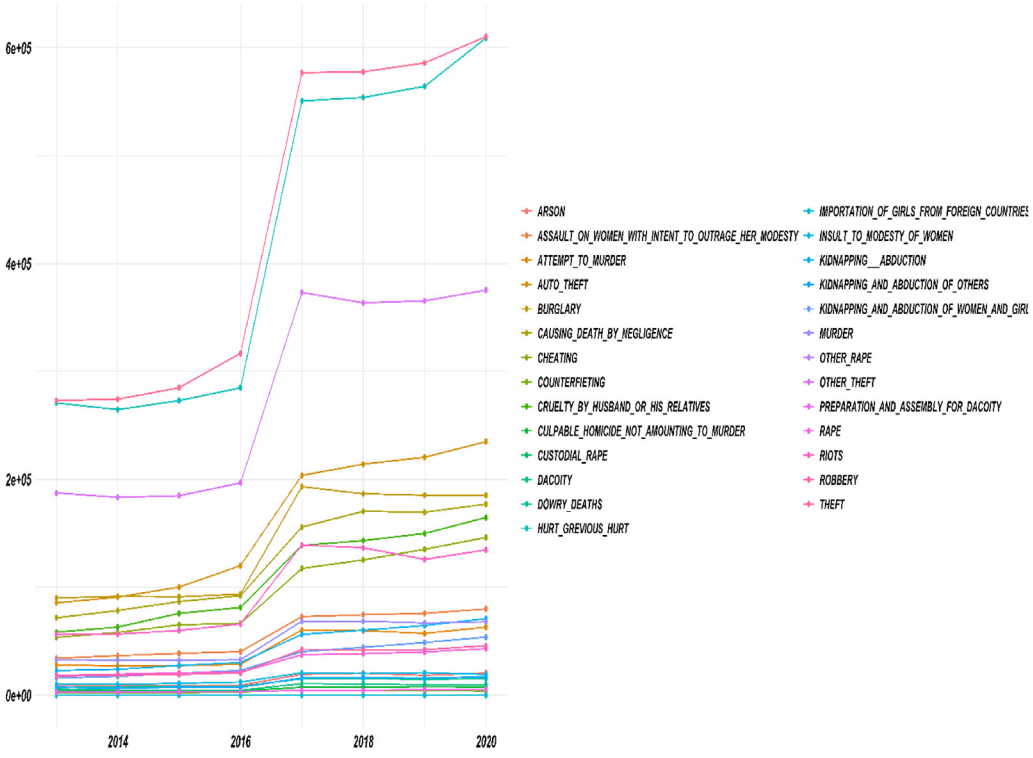


Fig. 13. Crimes evolution per type of crime 2014–2020.

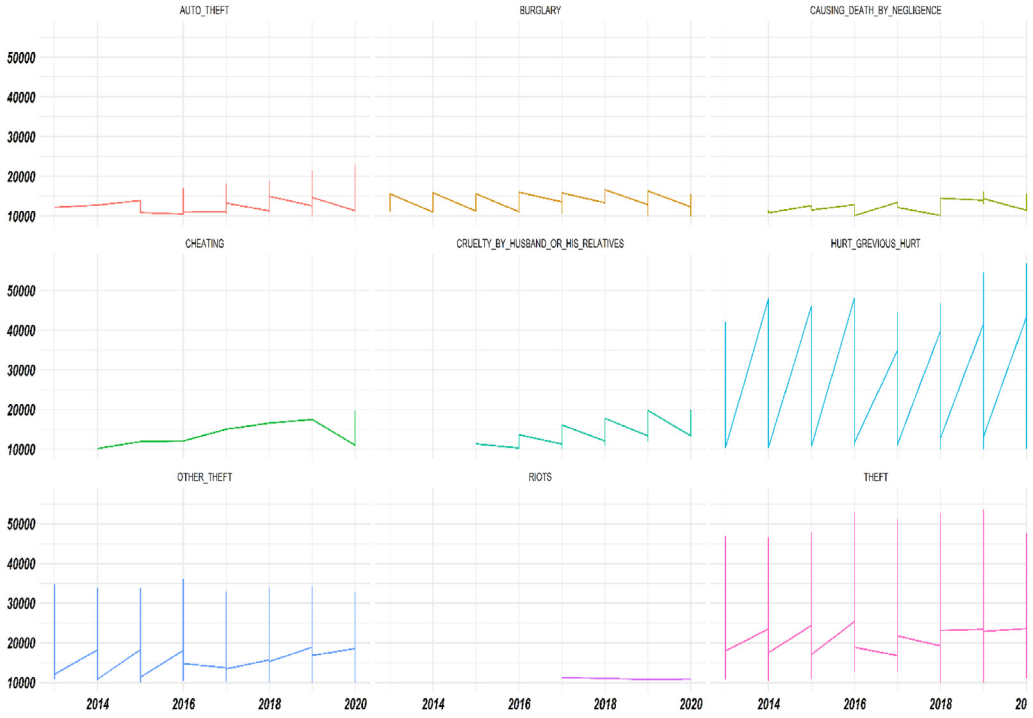


Fig. 14. Crime evolution per type individually and who's crime count ≥ 10k.

viewer hovers the mouse over the circular markers. When markers have clicked, the details about that region are seen viz—address, latitude and longitude coordinates, and crime count as shown in Fig. 12.

iv. Fig. 12 shows that the random forest model predicts Alidabad district/region of Andhra Pradesh state will have the highest crime count in India for 2022 with a predicted crime count of 31933.

v. Fig. 11 shows that the random forest model predicts Anantpur district/region of Andhra Pradesh state will have the highest theft crime count in India for 2022 with a predicted theft crime count of 12000.

vi. Fig. 12 shows that in the year 2022, Andhra Pradesh, Bihar, and Assam will be the states from where the top 50 districts will belong. Also,

vii. Andhra Pradesh state possesses the regions with the highest crime counts as per the model's predictions.

viii. Fig. 11 shows that in the year 2022, Andhra Pradesh state will have regions with the highest theft crime counts per the model's predictions.

4.4. Data visualization on Indian crime data from 2014 to 2020

In this section, different plots based on district-wise Indian crime data are discussed. Fig. 13 depicts the annual frequency of crimes per type and their trend. The most common types of crimes are theft

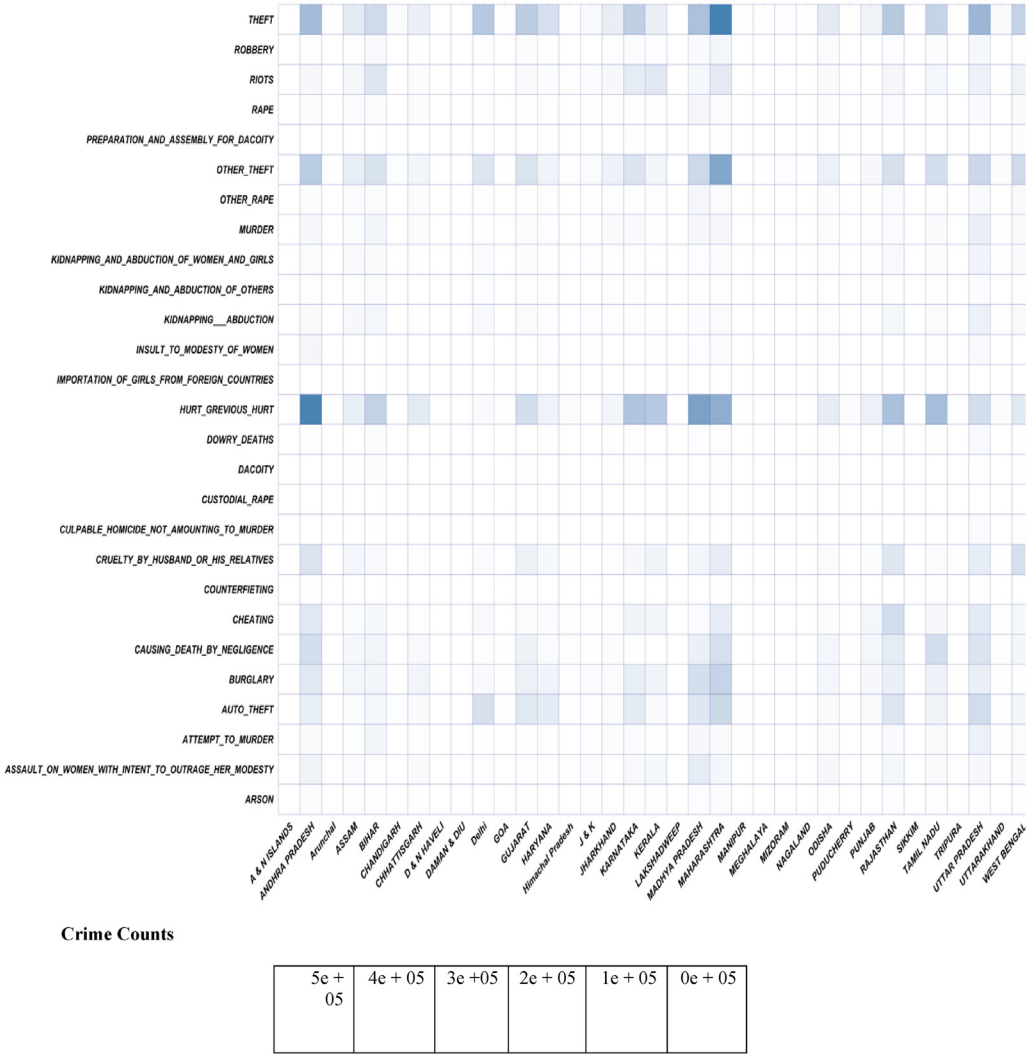


Fig. 15. Types of crime Vs states.

and hurt_grevious_hurt, and after 2016 they are continuously increasing.

In Fig. 13, there are many crimes, and they created a mess because most of them are less in number for clear visualization. It is observed that only those crimes whose crime count is more significant than 10,000 or 10k and to see their trend individually in Fig. 14, where most of them are increasing, and some of the crimes are increasing rapidly after some years like Cruelty_by_husband_or_his_relative.

In Fig. 15, it is observed that some types occur in specific States. The heat map shows that the darker the colour is, the higher the specific crime count is for a specific state.

5. Conclusions

This article investigates the machine learning-based soft computing regression analysis approach for crime data prediction. The predicted data by the five is used with data visualizations techniques like leaflet map plot as demonstrated in the paper or other geospatial data visualization techniques like a heat map, dot map, cluster map, bubble map, etc. For the chosen data, it is concluded that the region-wise total IPC crime prediction random forest regression model fits the best with an adjusted R squared value of 0.9631551 and error of 0.2027437 for region-wise thefts crime count prediction. Also, it is concluded that the RFR algorithm is the best fit with a 0.9666044 adjusted R squared value and an error of 0.16571. The regression model built using SVR is relatively the best for predicting crime rate per 100k for a given year with an adjusted R squared value of 0.9559832 and an error of 0.01971657. Also, the relatively best-fitted regression models predicted that Adilabad district of Andhra Pradesh will have the highest crime count in 2022 with a predicted count of 31933, and Anantpur district Andhra Pradesh will have the highest theft crime count in 2022 with a predicted count of 12000. 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 for Indian crime data prediction and visualization using regression algorithms. When there is much data associated with crime maintained out there, such a data-driven approach can help police and other law enforcement organizations control and prevent crime.

Once well-structured crime data for recent years will be available on the web by NCRB or any other data set offering organizations the proposed framework. It has a solid potential to build powerful regression models that will predict different types of crime count and overall crime count for different regions all over the country. The proposed approach for ICDA is a framework that can be used in the future by other researchers, data analysts, and scientists to build potential predictors using proposed regression models by making necessary changes in their respective approaches or methodology.

Acknowledgments

The authors would like to thank the Editor-in-Chief and anonymous referees for their suggestions and helpful comments that have improved the paper's quality and clarity.

References

[1] M.K. Anser, Z. Yousaf, A.A. Nassani, S.M. Alotaibi, A. Kabbani, K. Zaman, Dynamic linkages between pov-erty, inequality, crime, and social expenditures in a panel of 16 countries: two-step GMM estimates, J. Econ. Struct. 9 (2020) 1–25, http://dx.doi.org/10.1186/s40008-020-00220-6.

[2] M. Kumar, S. Athulya, M. M Mary, M.D.V. Vinodini, K.G.A. Lakshmi, S. Anjana, T.K. Manojkumar, Forecast-ing of annual crime rate in India: a case study, in: 2018 Int. Conf. Adv. Comput. Commun. Inform, IEEE, 2018, pp. 2087–2092. https://ieeexplore.ieee.org/document/8554422.

[3] Manvendra Singh, Danish Gulzar, Role of crime and criminal tracking network system (CCTNS) in crime control: a study of Haryana state, J. Hist. Res. 5 (2019) 982–992. https://thematicsjournals.org/index.php/hrj/article/view/12602.

[4] R. Aziz, C.K. Verma, N. Srivastava, A novel approach for dimension reduction of microarray, Comput. Biol. Chem. 71 (2017) 161–169, http://dx.doi.org/10.1016/j.compbiolchem.2017.10.009.

[5] M. Gupta, B. Chandra, Gupta, A framework of intelligent decision support system for Indian police, J. Enterprise Inf. Manag. 27 (2014) 512–540, http://dx.doi.org/10.1108/JEIM-10-2012-0073.

[6] B. Himabindu, R. Arora, N. Prashanth, Whose problem is it anyway? Crimes against women in India, Glob. Health Action 7 (2014) 23718, http://dx.doi.org/10.3402/gha.v7.23718.

[7] T. Chitra, S. Karunanidhi, The impact of resilience training on occupational stress, resilience, job satisfaction, and psychological well-being of female police officers, J. Police Crim. Psychol. 36 (2021) 8–23, http://dx.doi.org/10.1007/s11896-018-9294-9.

[8] J. Dreze, R. Khera, Crime, gender, and society in India: insights from homicide data, Popul. Dev. Rev. 26 (2000) 335–352, http://dx.doi.org/10.1111/j.1728-4457.2000.00335.x.

[9] A.M. Shermila, A.B. Bellarmine, N. Santiago, Crime data analysis and prediction of perpetrator identity using machine learning approach, in: 2018 2nd Proc. Int. Conf. Trends Electron. Inform, IEEE, 2018, pp. 107–114, http://dx.doi.org/10.1109/ICOEI.2018.8553904.

[10] R. Aziz, C.K. Verma, N. Srivastava, Novel machine learning approach for classification of high-dimensional microarray data, Soft Comput. 23 (2019) 13409–13421, http://dx.doi.org/10.1007/s00500-019-03879-7.

[11] R. Aziz, C.K. Verma, N. Srivastava, Artificial neural net-work classification of high dimensional data with novel optimization approach of dimension reduction, Ann. Data. Sci. 5 (2018) 615–635, http://dx.doi.org/10.1007/s40745-018-0155-2.

[12] Yee Ling Boo, D. Alahakoon, Building multi-modal crime profiles with growing self organising maps, Stud. Comput. Intell. 555 (2014) 97–124, http://dx.doi.org/10.1007/978-3-319-05885-6_6.

[13] V. Srinidhi, P. Saranya, M. Ashok, An affirmative learning techniques to analyse the crime scene in jewel theft murder, Int. Res. J. Multidiscip. Tech. 2 (2020) 1–7, http://dx.doi.org/10.34256/irjmt2051.
[14] R. Heeramun, C. Magnusson, C.H. Gumpert, S. Granath, M. Lundberg, C. Dalman, D. Rai, Autism and convictions for violent crimes: population-based cohort study in Sweden, J. Am. Acad. Child Adolesc. Psychiatry 56 (2017) 491–497, http://dx.doi.org/10.1016/j.jaac.2017.03.011.
[15] R.C. McDermott, C. Kilmartin, D.K. McKelvey, M.M. Kridel, College male sexual assault of women and the psychology of men: past, present, and future directions for research, Psychol. Men Masc. 16 (2015) 355–366, http://dx.doi.org/10.1037/a0039544.
[16] M. Skubak Tillyer, R. Tillyer, J. Kelsay, The nature and influence of the victim-offender relationship in kidnapping incidents, J. Crim. Justice 43 (2015) 377–385, http://dx.doi.org/10.1016/j.jcrimjus.2015.07.002.
[17] Kairat A. Bakishev, Aleksander V. Bashirov, Alikjan K. Fetkulov, Analysis and prediction of the state of road accidents and traffic crimes in the Republic of Kazakhstan, J. Advanced Res. L. & Econ. 8 (2017) 1456, http://dx.doi.org/10.14505/jarle.v8.5(27).08.
[18] D. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, N. Tyagi, Crime detection and criminal identification in India using data mining techniques, AI Soc. 30 (2015) 117–127, http://dx.doi.org/10.1007/s00146-014-0539-6.
[19] S. Jha, E. Yang, A.O. Almagrabi, A.K. Bashir, G.P. Joshi, Comparative analysis of time series model and machine testing systems for crime forecasting, Neural Comput. Appl. 17 (2021) 10621–10636, http://dx.doi.org/10.1007/s00521-020-04998-1.
[20] R. Kumar, B. Nagpal, Analysis and prediction of crime patterns using big data, Int. J. Inf. Technol. 11 (2019) 799–805, http://dx.doi.org/10.1007/s41870-018-0260-7.
[21] S. Kim, P. Joshi, P.S. Kalsi, P. Taheri, Crime analysis through machine learning, in: 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf., IEEE, 2018, pp. 415–420, http://dx.doi.org/10.1109/IEMCON.2018.8614828.
[22] M. Mittal, L.M. Goyal, J.K. Sethi, D.J. Hemanth, Monitoring the impact of economic crisis on crime in India using machine learning, Comput. Econ. 53 (2019) 1467–1485, http://dx.doi.org/10.1007/s10614-018-9821.
[23] P. Das, A.K. Das, Application of classification techniques for prediction and analysis of crime in India, Comput. Intell. Data Min. 711 (2019) 191–201, http://dx.doi.org/10.1007/978-981-10-8055-5_18.
[24] S. Hossain, A. Abtahee, I. Kashem, M. Hoque, I.H. Sarker, Crime prediction using spatio-temporal data, in: Int. Conf. Comput. Sci. Commun. Secur., vol. 1235, Springer, 2020, pp. 277–289, http://dx.doi.org/10.1007/978-981-15-6648-6_22.
[25] M. Pinto, H. Wei, K. Konate, I. Touray, Delving into factors influencing New York crime data with the tools of machine learning, J. Comput. Sci. Coll. 36 (2020) 61–70.
[26] A.P. Wheeler, W. Steenbeek, Mapping the risk terrain for crime using machine learning, J. Quant. Criminol. 37 (2021) 445–480, http://dx.doi.org/10.1007/s10940-020-09457-7.
[27] W. Safat, S. Asghar, S. Gillani, Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques, IEEE Access 9 (2021) 70080–70094, http://dx.doi.org/10.1109/ACESS.2021.3078117.
[28] A. Onan, S. Korukoğlu, H. Bulut, A multi-objective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification, Expert Syst. Appl. 62 (2016) 1–16, http://dx.doi.org/10.1016/j.eswa.2016.06.005.
[29] A. Onan, S. Korukoğlu, H. Bulut, A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification, Inf. Process. Manag. 53 (2017) 814–833, http://dx.doi.org/10.1016/j.ipm.2017.02.008.
[30] R. Aziz, C.K. Verma, M. Jha, N. Srivastava, Artificial neural network classification of microarray data using new hybrid gene selection method, Int. J. Data Min. Bioinf. 17 (2017) 42–65.
[31] A. Onan, M.A. Toçoğlu, Satire identification in Turkish news articles based on ensemble of classifiers, Turk. J. Electr. Eng. Comput. Sci. 28 (2020) 1086–1106, http://dx.doi.org/10.3906/elk-1907-11.
[32] A. Onan, Ensemble of classifiers and term weighting schemes for sentiment analysis in Turkish, Commun. Res. 1 (2021) 1–12, http://dx.doi.org/10.52460/src.2021.004.
[33] A. Onan, An ensemble scheme based on language function analysis and feature engineering for text genre classification, J. Inf. Sci. 44 (2018) 28–47, http://dx.doi.org/10.1177/0165551516677911.
[34] A. Onan, M.A. Toçoğlu, A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification, IEEE Access 9 (2021) 7701–7722, http://dx.doi.org/10.1109/ACCESS.2021.3049734.
[35] A. Onan, S. Korukoğlu, H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification, Expert Syst. Appl. 57 (2016) 232–247, http://dx.doi.org/10.1016/j.eswa.2016.03.045.
[36] A. Onan, Enriched word embeddings for sarcasm identification, in: Advances in Intelligent Systems and Computing Software Engineering Methods in Intelligent Algorithms, Springer, 2019, pp. 293–304, http://dx.doi.org/10.1007/978-3-030-19807-7_29.
[37] A. Onan, S. Korukoğlu, A feature selection model based on genetic rank aggregation for text sentiment classification, J. Inf. Sci. 43 (2017) 25–38, http://dx.doi.org/10.1177/0165551515613226.
[38] I. Goni, J.M. Gumpy, T.U. Maigari, M. Muhammad, A. Saidu, Cybersecurity and cyber forensics: machine learning approach, Mach. Learn. Res. 5 (2020) 46–50.
[39] A. Onan, S. Korukoğlu, Exploring performance of instance selection methods in text sentiment classification, in: R. Silhavy, R. Senkerik, Z. Oplatkova, P. Silhavy, Z. Prokopova (Eds.), Artificial Intelligence Perspectives in Intelligent Systems, vol. 464, Springer, 2016, pp. 167–179, http://dx.doi.org/10.1007/978-3-319-33625-1_16.
[40] A. Onan, Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach, Comput. Appl. Eng. Educ. 29 (2021) 572–589, http://dx.doi.org/10.1002/cae.22253.
[41] D. Weisburd, C. Gill, A. Wooditch, W. Barritt, J. Murphy, Building collective action at crime hot spots: findings from a randomized field experiment, J. Exp. Criminol. 17 (2021) 161–191, http://dx.doi.org/10.1007/s11292-019-09401-1.
[42] M. Suhail, I. Babar, Y.A. Khan, M. Imran, Z. Nawaz, Quantile-based estimation of liu parameter in the linear regression model: applications to portland cement and US crime data, Math. Probl Eng. 1 (2021) 1–11, http://dx.doi.org/10.1155/2021/1772328.
[43] A. Emmanuel, O. Elisha, T. Danison, N. Ivan, Crime prediction using decision tree (J48) classification algorithm. International journal of computer and information technology, Karbala Int. J. Mod. Sci 6 (2017) 188–195. http://hdl.handle.net/20.500.12493/113.
[44] C. Wu, F. Yang, Y. Wu, R. Han, Prediction of crime tendency of high-risk personnel using C5. 0 decision tree empowered by particle swarm optimization, Math. Biosci. Eng. 16 (2019) 4135–4150.
[45] J.M. Sadler, J.L. Goodall, M.M. Morsy, K. Spencer, Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest, J. Hydrol. 559 (2018) 43–55, http://dx.doi.org/10.1016/j.jhydrol.2018.01.044.
[46] T.F. Cootes, M.C. Ionita, C. Lindner, P. Sauer, Robust and accurate shape model fitting using random forest regression voting, in: European Conference on Computer Vision,

Springer, 2012, pp. 278–291, http://dx.doi.org/10.1007/978-3-642-33786-4_21.
[47] Z. Xia, K. Stewart, J. Fan, 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 (2021) 101599, http://dx.doi.org/10.1016/j.compenvurbsys.2021.101599.
[48] R. Aziz, C.K. Verma, N. Srivastava, A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data, Genom. Data. 8 (2016) 4–15, http://dx.doi.org/10.1016/j.gdata.2016.02.012.
[49] R. Aziz, C.K. Verma, N. Srivastava, A weighted-SNR feature selection from independent component subspace for NB classification of microarray data, Int. J. Adv. Biotechnol. Res. 6 (2015) 245–255.
[50] R. Aziz, N. Srivastava, C.K. Verma, T-independent component analysis for SVM classification of DNA-microarray data, Int. J. Bioinf. Res. 6 (2015) 305–312.
[51] M. Rao, N.K. Kamila, Bayesian network based energy efficient ship motion monitoring, Karbala Int. J. Mod. Sci. 4 (2018) 69–85, http://dx.doi.org/10.1016/j.kijoms.2017.11.001.