Comput Econ (2019) 53:1467–1485
https://doi.org/10.1007/s10614-018-9821-x

Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning

Mamta Mittal[1] · **Lalit Mohan Goyal**[2] · **Jasleen Kaur Sethi**[3] · **D. Jude Hemanth**[4]

Accepted: 27 April 2018 / Published online: 7 May 2018

Abstract Trends of crimes in India keep changing with the growing population and rapid development of towns and cities. The rise in crimes at any place especially crimes against women, children and weaker sections of the society is a worrying factor for the Indian Government. In India, the crime data is maintained by National Crime Records Bureau as well as an application called Crime Criminal Information System is available to make inquiry and generate reports for the crime data. To curb crime, the Police need countless hours to go through the crime data and determine the various factors that affect it. Therefore, there is necessity of tools which can automatically predict the factors that effects the crimes effectively and efficiently. The field of machine learning has emerged in the recent years for this purpose. In this paper, various machine learning techniques have been applied on crime data to monitor the impact of economic crisis on the crime in India. The effect of unemployment rates and Gross District Domestic Product on theft, robbery and burglary has been monitored across districts of various states in India. Further, Granger causality between crime rates and economic indicators has also been calculated. It has been observed from the experimental work that unemployment rate is the major economic factor which affects the crime rate, thus paving the path to control the crime rate by raising more opportunities for the employment.

✉ D. Jude Hemanth
jude_hemanth@rediffmail.com

1 Department of CSE, G B Pant Government Engineering College, New Delhi, India

2 Department of CSE, Bharti Vidyapeeth's College of Engineering, New Delhi, India

3 USICT, GGSIPU, New Delhi, India

4 Department of ECE, Karunya University, Coimbatore, India

Keywords Crime · Machine learning · Linear regression · Correlation · Coefficient of determination

1 Introduction

Crime is defined as any harmful act against the public which the State wishes to prevent and which, upon conviction, is punishable by fine, imprisonment, and/or death. Historically solving crimes has been the privilege of legal enforcement specialists. But lately computer analysts have aided the legal enforcement in solving crimes by discovering crime patterns. Crime rate is affected by number of Economic factors such as Income level, unemployment rate and Gross Domestic Product (GDP), Consumer Price Index (CPI) etc. in India (Nath 2006).

An economic crisis is a situation where the economy of a country experiences a sudden downturn brought on by financial crisis. Whether there is economic crisis or not, the economic indicators like unemployment have a direct impact on the Indian population and thus affect crime rates. Due to financial crisis, there is an increase in unemployment and large number of population suffers sudden reduction in their income.

In the past, Analysts have used number of techniques of data mining and Machine Learning to study crime data to discover crime patterns. Iqbal (2013) applied Naïve Bayes and Decision trees classification algorithms for predicting crime categories for various states of USA and from the results it was observed that Naïve Bayes out performed Decision Tree classification. Nath (2006) used data mining for identifying crime patterns using the clustering techniques to support police detectives in solving crimes. The significant attributes are identified using expert based semi-supervised learning method and scheme for weighting the significant attributes is also developed. Tayal (2015) proposed an approach for crime detection and criminal identification (CDCI) using data mining techniques for seven Indian cities based on their crime rates. K-means clustering is used for crime detection and KNN classification for crime identification and prediction. These techniques have been used to extract the unstructured crime dataset from various crime Web sources for the year 2000 to 2012 and then the extracted crime data is converted into structured crime instances. Ku and Leroy (2014) developed a Decision Support System (DSS) that uses document similarity algorithm with Naïve Bayes Classifier and binary Logistic Regression for crime analysis and classification of the various crime reports on similar crimes. It was observed that binary Logistic Regression had better classification accuracy than Naïve Bayes Classifier. Kaikhah and Doddameti (2006) used neural network to discover and predict the hidden trends in four types of crimes—murder, rape, Robbery and auto theft in US cities. Gerber (2014) used Twitter messages for the prediction of crime. Out of 25 crime types studied, for the 19 crime types the performance was higher when tweets were used as compared to a standard approach based on kernel density estimation. Cohen et al. (2007) developed an indicator forecast model for the support of police resources to forecast property and violent crimes. This model is used to assist police for deploying resources to prevent increase of crime. Renushe et al. (2012) used data mining techniques for the effective investigation of crimes under Faraskhana

police station of Pune city and designed the actions required for preventing the crime. McClendon and Meghanathan (2015) carried out the analysis of crime data taken from UCI machine learning repository and data for the state of Mississippi using Weka. The various techniques like Linear Regression, Additive Regression and Decision Stump algorithm were performed on the crime dataset where the linear regression had the best performance. Saeed et al. (2015) used the communities and crime dataset from UCI machine learning repository to predict criminal activity using various classification techniques like Naïve Bayes and decision trees. Out of them, it was observed that Naïve Bayes classifier is more accurate for Crime analysis.

Big Data Analytics (BDA) on crime data has been covered in some of the papers. Jain and Bhatnagar (2016) used BDA to take well informed decisions so that law and order can be maintained properly like if the number of complaints from a state is found to be very high, extra security must be provided there. This work has been carried out using Pig in Hadoop. Chauhan and Aluvalu (2016) proposed a system where clustering is applied on the crime data based on the geographical locations and then Artificial Neural Network has been used to detect the prediction pattern for different crimes.

Though a lot of work has been done on the crime data, the effect of financial crisis has not been monitored on different types of crimes at the national level. In this paper, various machine learning algorithms like Decision Trees, Random Forest, Linear Regression and Neural networks have been used for analysing crime dataset on economic indicators like unemployment rate and GDDP (Gross District Domestic Product). Major focus of this paper is based on monitoring the crimes: theft, Robbery and burglary across various districts of India. The data for crimes and economic indicators have been taken from various Government websites http://ncrb.nic.in/StatPublications/CII/PrevPublications.htm, http://www.censusindia.gov.in, https://data.gov.in, http://mospi.nic.in/data, www.labourbureaunew.gov.in.

The remaining paper has been organized as follows: Sects. 2 and 3 presents an overview of Machine Learning and Linear Regression respectively. In Sect. 4, the methodology for the analysis of crime dataset has been presented. The results analysis of Crime data on economic indicator has been presented in Sect. 5. Further, Causal Relationship between Crime Rates and Economic Indicators has been given in Sect. 6. Some actions to control the crime have been briefly explained in Sect. 7. In the last section, work carried out in this study has been concluded.

2 Machine Learning

According to Mitchell (1997), "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." Machine learning is the process of executing a program to optimize performance criterion based on either the experience or training data. The objectives of machine learning include Task Oriented Studies to improve the performance of the model, Cognitive Simulation to enable simulation of the model and theoretical analysis of various learning algorithms (Mitchell 1997; Alpaydin 2014).

The various types of learning include Supervised learning, Unsupervised learning and Reinforcement learning. In the Supervised learning mapping to the output values is done based on labelled training data. This type of learning is shown in Fig. 1a. In this Teacher can be thought as the one which has knowledge of the environment. This knowledge is represented by input output examples. The desired response is provided by the teacher. The learning system gives the actual response. The error signal is the difference between desired response and actual response. Unsupervised learning is learning without a teacher and it aims to find hidden patterns in unlabelled data. There is no external teacher or critic present. The system develops the ability to form internal representations for the input to create classes from the unlabelled data. The block diagram of Unsupervised Learning is shown in Fig. 1b. In Reinforcement learning, the learning is performed by continuous interaction with the environment. These learning systems have a critic which converts a primary reinforcement signal into heuristic reinforcement signal. Thus, these systems learn under delayed reinforcement. Reinforcement learning algorithms learn from past experiences to generate output which consists of sequence of actions. This type of learning is shown in Fig. 1c (Haykin 2001).

There are different types of machine learning algorithms that can be used in data mining (McClendon and Meghanathan 2015). The various machine learning algorithms have been summarized in Fig. 2.

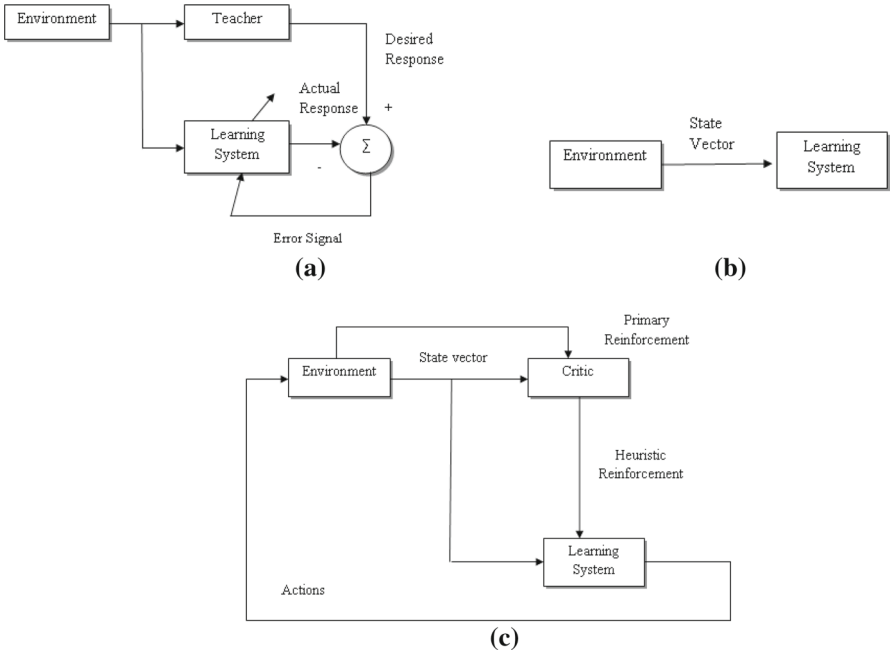

Fig. 1 **a** Block diagram of supervised learning. **b** Block diagram of unsupervised learning. **c** Block diagram of reinforcement learning

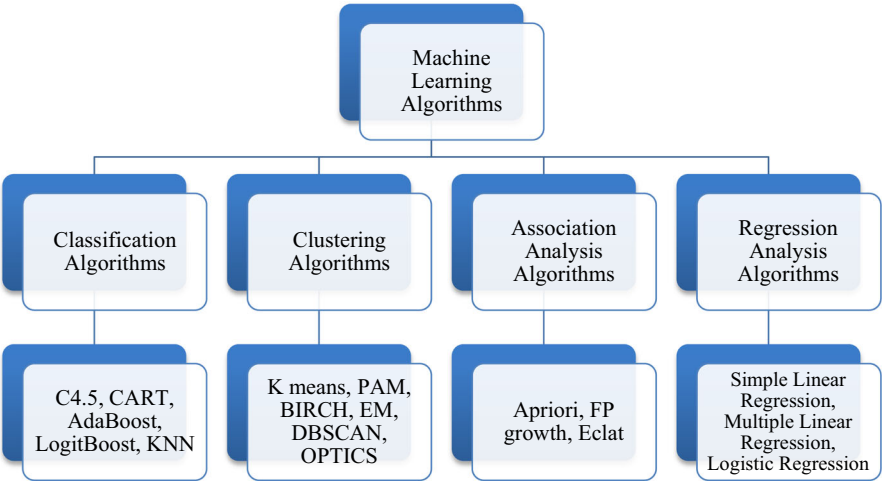

Fig. 2 Various machine learning algorithms

2.1 Classification Algorithms

Classification is supervised learning technique that is used to predict a categorical attribute that take discrete values. The various classification algorithms are Statistical classification which classifies the data into groups based on the training set of various data items. These includes linear discriminant analysis and nearest neighbour algorithms. Decision Tree classifies an instance based on the condition from the root of the tree to its leaf. These algorithms include C4.5 and CART. Rule Induction is a classification algorithm that extracts rules from the input data. In Fuzzy rule induction, fuzzy logic is applied to the input data to extract fuzzy rules. Some fuzzy learning techniques are AdaBoost and LogitBoost. Neural networks are massively parallel distributed processing network modelled on the human brain structure that consists of processing elements called neurons. Multilayer perceptron (MLP), Radial Basis Function Network (RBFN) are examples of neural network algorithms (Romero et al. 2008).

2.2 Clustering Algorithms

Clustering is an unsupervised learning technique that partitions the dataset into groups or clusters based on similarity measures. There are various types of clustering algorithms. Partitioning based algorithms divide the dataset into partitions where each partition is a cluster. K-means and FCM are the examples of partitioning based algorithms. Hierarchical based algorithms partition the data based on medium of proximity in the form of a dendogram. Some Hierarchical based algorithms are BIRCH and AGNES. Density based clustering algorithms can make separation between the data based on the regions of density, connectivity and boundary. DBSCAN, OPTICS, and DENCLUE are density based clustering algorithms. Grid based clustering algorithms

divides the data sets into grids for fast processing time. Wave-Cluster and STING are typical examples of Grid based algorithms. Model based clustering algorithm automatically determines the number of clusters and is based on mathematical model. EM, COBWEB and SOM are examples of model based clustering algorithms (Mittal et al. 2011, 2015).

2.3 Association Analysis Algorithms

Association analysis algorithms find the correlation between the different attributes in a dataset to create association rules. These kinds of algorithms can be used in market basket analysis. Association analysis algorithms partition data as per Horizontal Data Layout or Vertical Data Layout. In Horizontal Data Layout, the dataset is divided into many transactions with a transaction identifier which increases the computation overhead. Apriori and FP growth algorithm are based on Horizontal data layout. Vertical data Layout is comprised of a dataset with number of items and each item is followed by all the transactions identifiers which contain that item. Eclat algorithm follows Vertical data layout.

2.4 Regression Analysis Algorithms

Regression Analysis algorithms are used for prediction of variables that take continuous values. There are various types of Regression analysis algorithms. Simple Linear Regression evaluates the impact of an independent variable on a dependent variable. When the impact of multiple independent variables is evaluated on a dependent variable, then it is known as Multiple linear regression. In logistic regression, dependent variable is represented by either presence or absence of a feature.

3 Linear Regression

Linear Regression is used to model an output variable as a linear function of one or more input variables. These input variables are called independent variables and the response variable is called dependent variable. If the dependent variable is modeled by using a linear function of one independent variable, it is called Simple Linear Regression. Multiple Linear Regression is modeling a dependent variable based on more than one independent variable.

3.1 Simple Linear Regression

Using Simple Linear Regression, a dependent variable Z can be modeled as a linear function of independent variable X using Eq. (1) as:

$$Z = a + bX \tag{1}$$

where the regression parameter a is the intercept and the regression parameter b is the slope of the regression line.

3.2 Least Squares Method

Linear Regression aims to fit a linear line between dependent variable based on independent variable. Least Squares Method is used to estimate the intercept and slope parameters to determine the regression line (Zou et al. 2003). In this method, these regression parameters are found out such that sum of squares residuals (SSR) is minimized. SSR is the difference between observed value of dependent variable and the estimated or predicted value of dependent variable and is given below by Eq. (2):

$$SSR = \sum_{i=1}^{m} (Z_i - Z_i^{''})^2 \tag{2}$$

where Z_i is the observed value of dependent variable Z and $Z_i^{''}$ is the predicted value of dependent variable Z.

3.3 Steps for Regression Model Analysis

Linear regression is to predict the value of a dependent variable based on one or more input independent variables so that a linear relationship is established between the output target variable and the input variable. This formula can be used to estimate the value of the output variable, when only the values of input variables are known. Regression Model Analysis consists of number of steps.

3.3.1 Creation of Scatter Plot

A scatter plot is created between the independent variable on the X axis and dependent variable on the Y axis. The aim to create a scatter plot is look for a visual linear line between the two variables. In case of multiple input independent variables, a scatter plot is created for each of the independent variable against the dependent variable.

3.3.2 Determination of Correlation Coefficient

The purpose of correlation analysis is to determine the strength of a linear or nonlinear relationship between two variables. The correlation coefficients take the values between -1 and $+1$. The value -1 corresponds to negative correlation, 0 corresponds to unrelated and $+1$ corresponds to positive correlation. A low correlation corresponds that the variation of the output variable cannot be explained by the input variable and there is a need to look for more explanatory input variables. The value of correlation can be found out by Pearson Correlation coefficient.

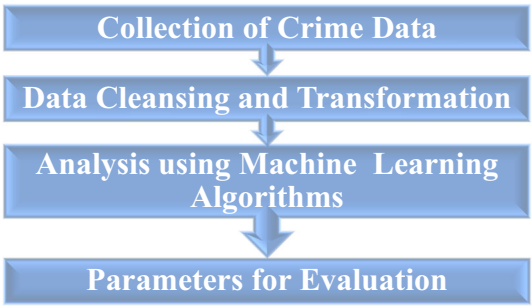

Fig. 3 Steps for analysis of crime data

3.3.3 Calculation of Regression Coefficients

The regression coefficients a and b in Eq. (1) are then calculated to obtain the regression line. The intercept coefficient is calculated using Eq. (3) as:

$$a = Z' - bX'. \tag{3}$$

The slope coefficient b is given by Eq. (4) as:

$$b = \frac{\sum_{i=1}^{m} \left(\mathrm{X_i} - \mathrm{X'}\right)\left(\mathrm{Z_i} - \mathrm{Z'}\right)}{\sum_{i=1}^{m} \left(\mathrm{X_i} - \mathrm{X'}\right)^2} \tag{4}$$

where X' and $Z^{'}$ are the mean values of input variable X and output variable Z respectively.

3.3.4 Calculation of Equation for Regression line

The equation for regression line (1) is completed by using values of the slope and the intercept coefficients from Eq. (4) and (3) respectively. This regression equation determines the predicted value of the output variable $Z_i^{''}$. This can be used to evaluate SSR using Eq. (2).

4 Methodology for the Analysis of Crime Dataset

To analysis crime dataset, firstly data from various government sites has been taken. After data cleansing and transformation, the analysis of Crime dataset is carried out using machine learning algorithms. The results are evaluated using numbers of parameters such as Correlation coefficient, Accuracy etc. These various steps are shown in Fig. 3 and elaborated in this section.

4.1 Collection of Crime Dataset

In India, the responsibilities of the Police include the maintenance of law and order, checking of Militant and Organised Criminal Gangs, maintenance of VIP Security and handling Dharna and Rallies by Trade Union and Political Party. The rate of crime is increasing and there is no time for maintenance of such crime records. Exchanging criminal information amongst different Police Stations is very difficult. This led to the computerization of Criminal Records and hence the development of NCRB (National Crime Records Bureau). Some of the objectives of NCRB include collection and processing of crime statistics at National level and to use fingerprint search for identification of criminals. In this paper, the crime dataset has been collected from NCRB website.

To perform the experimental work, Gross State Domestic Product (GSDP), Net State Domestic Product (NSDP), Per capita Net State Domestic Product, unemployment rate and Consumer Price Index (CPI) have been considered from 2004 to 2013 on the all states to monitor the crimes: Theft, Robbery and Burglary. The data for these variables is available on various Government websites http://ncrb.nic.in/StatPublications/CII/PrevPublications.htm, http://mospi.nic.in/data, www.labourbureaunew.gov.in. In this, Gross State Domestic Product (GSDP) is the volume of all goods and services produced within the State during a given period of time; Net State Domestic Product (NSDP) is the volume of all goods and services produced within the State during a given period of time after deducting the depreciation; NDP per capita is the state's economic output per person. NDP per capita is calculated as state's NDP divided by the state's population; Unemployment rate in a state is defined as the percentage of the number of unemployed individuals divided by the total number of individuals in labour force; The Consumer Price Index (CPI) is the weighted average of prices of a number of consumer goods and services; Theft is the act of taking of property; Robbery is taking of property through fear of by force that involves person to person interaction; Burglary is the act of entering a building with the intention of committing a theft.

The Pearson correlation coefficient of the various economic indicators with the target variables: theft, burglary and Robbery has been calculated. These results have been shown in Table 1.

From this Table 1, it has been observed that GSDP and NSDP have high correlation with all the target variables. Whereas, Per Capita NSDP, Unemployment rate and CPI have low values for the correlation coefficients. Based on these values, the input variables GSDP (the one with highest correlation values) and Unemployment rate

Table 1 Correlation between economic indicators and target variables

Economic indicator	Theft	Burglary	Robbery
GSDP	0.7993	0.6723	0.8065
NSDP	0.8013	0.6734	0.8075
Per capita NSDP	− 0.04	− 0.032	− 0.0058
Unemployment rate	− 0.0918	− 0.0946	− 0.045
CPI	0.2153	0.1431	0.2404

District	Unemployed	GDDP	BURGLARY	THEFT	DACOITY
Anantapur	2044982	7178.59	178	439	17
Chittoor	2240707	7656.74	281	1269	9
East Godavari	3060615	12929.6	403	1286	3
Guntur	2506189	10733.34	293	1266	14
Krishna	2468518	12579.88	155	653	3
Kurnool	2024038	7636.36	179	373	16
Prakasam	1696881	7535.17	181	406	10
Potti Sriramulu Nellc	1648996	6329.53	205	759	6
Srikakulam	1412826	4197.16	101	211	1
Visakhapatnam	2400710	14643.93	393	987	6
Vizianagaram	1186512	3911.08	87	217	6
West Godavari	2163854	9763.69	300	1204	2
Lohit	84781	274.97	20	22	1
Lower Dibang Valley	31754	131.62	13	37	3
Lower Subansiri	52828	177.94	19	12	0
Tawang	21416	124.28	8	18	0
Tirap	61806	177.82	17	17	4
Upper Siang	17676	80.37	8	16	0
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....

Fig. 4 Crime dataset

(the one with lowest correlation values) have been chosen to construct the crime dataset. These input variables have been recorded on the districts of various states for which crime data is available on NCRB website http://ncrb.nic.in/StatPublications/CII/PrevPublications.htm, Census India website http://www.censusindia.gov.in and Open Government website https://data.gov.in. Small chunk of this dataset has been depicted in Fig. 4. For the experimental work, Gross District Domestic Product (GDDP) and the unemployment rate has been considered as independent variables and theft, burglary and Robbery are the dependent variables. This data need to be cleaned and transformed before the applying the machine learning techniques. In the next section, data cleansing and transformation have been discussed.

4.2 Data Cleansing and Transformation

In this phase, to clean the data duplicates and missing values of some variables have ignored. Further, as the units of the independent and dependent variables are different, data is transformed by normalizing the values of variables between 0 and 1. Normalization intends to have scaled values that allow the comparison of corresponding normalized values for different datasets in a way that eliminates the effects of certain gross influences. A variable Y can be normalized as shown in Eq. (5)

$$Y' = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \tag{5}$$

where Y_{min} and Y_{max} are the minimum and the maximum values of variable Y respectively.

4.3 Analysis using Machine Learning Algorithms

For modelling the crime dataset, the analysis has been done using various machine learning algorithms: Decision Trees, Random Forest, Linear Regression and Neural Network. Decision Trees are the most commonly used algorithms, based on recursive partitioning approach; Random Forest is the collection of decision trees and it repeatedly partitions the given dataset; Linear Regression is based on dependent and independent variables; Neural Network creates a model based on the concept of multiple layers of neurons connected to each other. Crime dataset has been analysed using these four algorithms.

4.4 Parameters for Evaluation of Crime Dataset

In this section, the various parameters for the evaluation of crime dataset have been discussed.

4.4.1 Correlation Coefficient

The correlation coefficient is used to determine the strength of linear or nonlinear relationship between two variables. The Pearson correlation coefficient r for variables $\mathrm{X_i}$ and $\mathrm{Z_i}$ for $i = 1, 2, 3, \ldots m$ is calculated using Eq. (6) as:

$$r = \frac{\sum_{i=1}^{m} \left(\mathrm{X_i} - \mathrm{X'}\right)\left(\mathrm{Z_i} - \mathrm{Z'}\right)}{\sqrt{\sum_{i=1}^{m} \left(\mathrm{X_i} - \mathrm{X'}\right)^2 \sum_{i=1}^{m} \left(\mathrm{Z_i} - \mathrm{Z'}\right)^2}} \tag{6}$$

where X' and Z' are the mean values of input variable X and output variable Z respectively.

4.4.2 Coefficient of Determination, R^2

The value of the Pearson correlation coefficient r can be squared to obtain the term R-square (R^2) or coefficient of determination. This measure ranges between 0 and 1, where value of 1 represents good fit. It is calculated as the fraction of the variability in Z by the variability in X through their linear relationship given as by Eq. (7)

$$R^2 = {}^{SSR}/_{SST} \tag{7}$$

where sum of squares of total errors $SST = \sum_{i=1}^{m} \left(Z_i - Z_i^{'}\right)^2$ where Z_i is the observed value and $Z_i^{'}$ is the mean value of the dependent variable Z.

4.4.3 Absolute Mean Error

The Absolute Mean Error (AME) is used to measure how close predicted value is to the actual outcomes. The absolute mean error is defined as:

$$AME = \frac{1}{m} \sum_{i=1}^{m} \left(Z_i - Z_i^{''}\right) \tag{8}$$

where Z_i is the observed value and $Z_i^{''}$ is the predicted value of the dependent variable Z.

4.4.4 Accuracy

Accuracy is the percentage change of predicted value and actual value with the acceptable error.

$$Accuracy = \frac{100}{m} \sum_{i=1}^{m} e_i \tag{9}$$

$$e_i = \begin{cases} 1 & if abs\left(Z_i - Z_i^{''}\right) \leq err \\ 0 & otherwise \end{cases}$$

where Z_i is the observed value and $Z_i^{''}$ is the predicted value of the dependent variable Z respectively and err is for the acceptable error.

5 Result Analysis of Crime Dataset using Machine Learning Algorithms

Analysis of Crime Dataset has been carried out using machine learning algorithms: Decision Tree, Random Forest, Linear Regression and Neural Network. These four algorithms have been applied to the crime dataset with each Theft, Burglary and Robbery as the target variables and with GDDP and unemployment rate as the input variables. In each case, Correlation coefficient, Coefficient of determination, Absolute Mean Error and Accuracy have been calculated.

5.1 Theft as Target Variable

The results of analysis of crime data using different machine learning algorithms for theft as a target variable have been depicted in Table 2.

Table 2 Result of analysis of crime dataset for target variable theft

Machine learning algorithm	Correlation coefficient	Coefficient of determination	Absolute mean error	Accuracy
Decision tree	0.544076	0.296	0.032863	88.31169
Random forest	0.466951	0.218	0.03025	90.90909
Linear regression	0.876262	0.7678	0.026459	93.50649
Neural network	−0.6104	0.3726	0.036965	92.20779

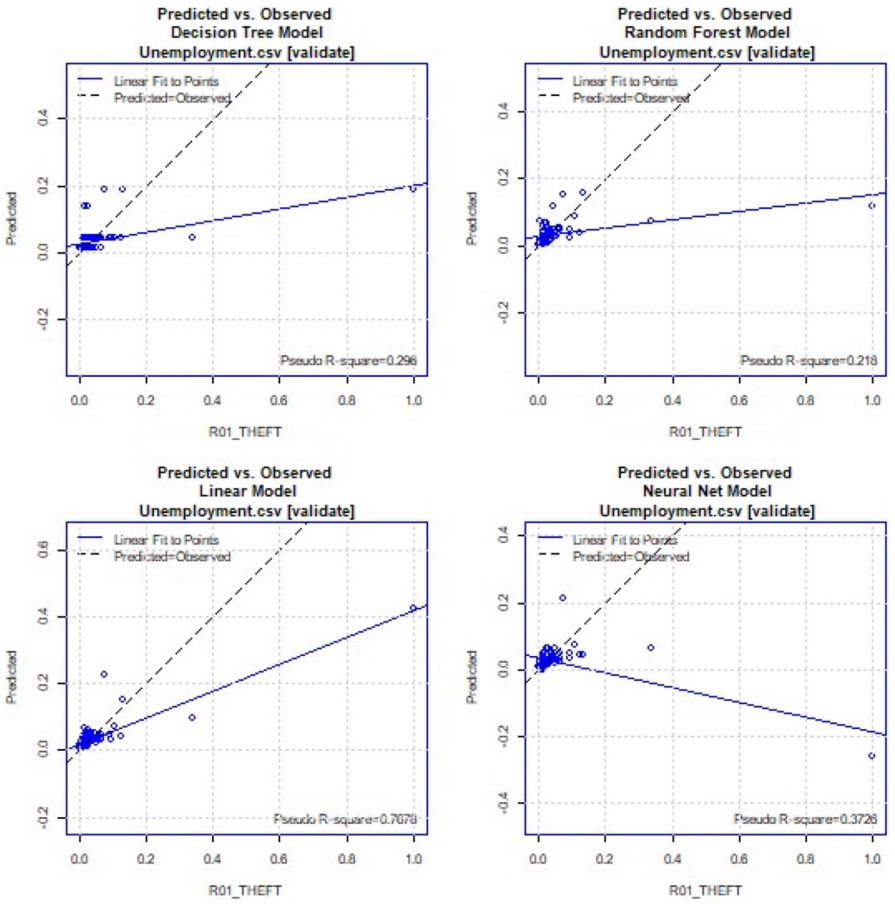

Fig. 5 Predicted versus observed values of theft

From this Table 2, it has been observed that Linear Regression model has the highest Accuracy and least value of Absolute Mean Error compared to all other algorithms. Further, the Correlation coefficient and Coefficient of determination is also highest in Linear Regression as compared to all other algorithms. The plot between observed and predicted values of Theft based on various machine learning algorithms has been depicted in Fig. 5.

Table 3 Result of analysis of crime dataset for target variable burglary

Machine learning algorithm	Correlation coefficient	Coefficient of determination	Absolute mean error	Accuracy
Decision tree	0.501399	0.2514	0.038257	85.71429
Random forest	0.461937	0.2134	0.03621	88.61038
Linear regression	0.862117	0.7432	0.03259	89.61039
Neural network	− 0.17314	0.02998	0.038444	88.31169

5.2 Burglary as Target Variable

The results of analysis of crime data using different machine learning algorithms for Burglary as a target variable have been depicted in Table 3.

From this Table 3, it has been observed that Linear Regression model has the highest accuracy and least value of absolute mean error compared to all other algorithms. Further, the correlation coefficient and coefficient of determination is also highest in Linear Regression as compared to all other algorithms. The plot between observed and predicted values of Burglary based on various machine learning algorithms has been depicted in Fig. 6.

5.3 Robbery as Target Variable

The results of analysis of crime data using different machine learning algorithms for Robbery as a target variable have been depicted in Table 4.

From this Table 4, it has been observed that Linear Regression model has good accuracy and less absolute mean error. It can also be noted that Neural Network also has accuracy same as the linear regression model. Further, the values for the correlation coefficient and coefficient of determination are low for all the models. The plot between observed and predicted values of Robbery based on various machine learning algorithms has been depicted in Fig. 7.

6 Causal Relationship between Crime Rates and Economic Indicators

Causality is the association between simultaneous or successive events where the one event is the cause and the other event is its effect. A variable a_t can granger cause another variable b_t if b_t is predicted using values of a_t and b_t with a lag value. The Granger causality can be either unidirectional or bidirectional (Granger 1988).

To assess the causal relationship between crime rates and economic indicators, all India datasets have been taken from various Government websites for 1971 to 2013 http://ncrb.nic.in/StatPublications/CII/PrevPublications.htm, https://data.gov.in, http://mospi.nic.in/data, www.labourbureaunew.gov.in.

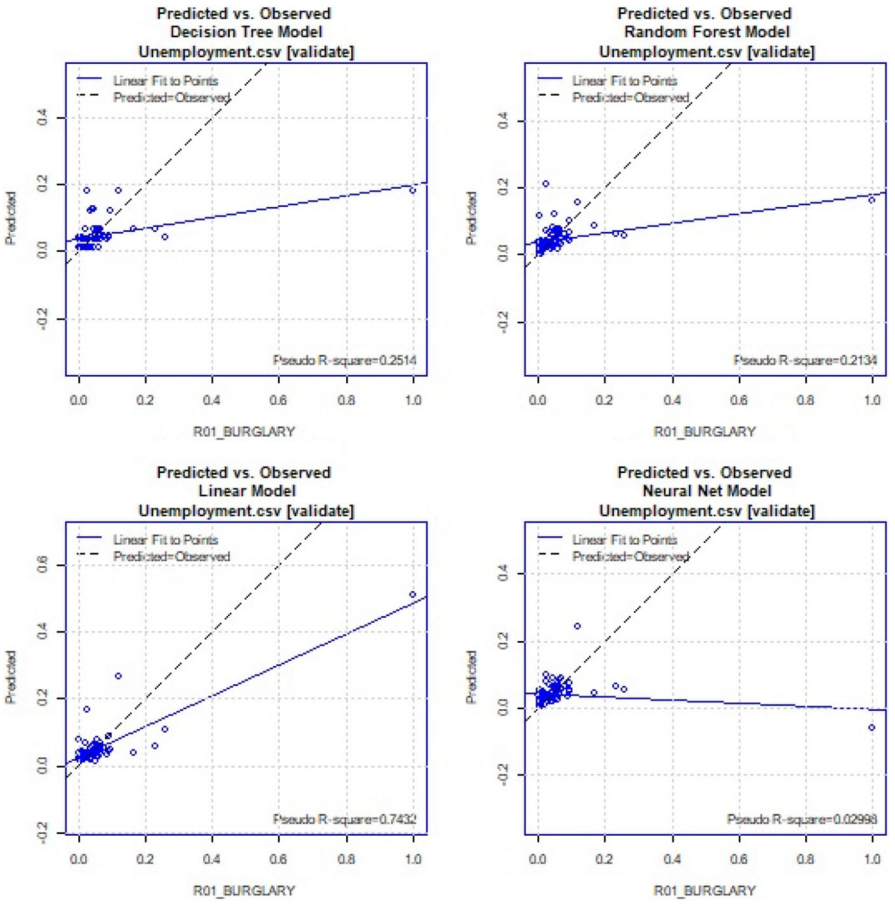

Fig. 6 Predicted versus observed values of burglary

Table 4 Result of analysis of crime dataset for target variable robbery

Machine learning algorithm	Correlation coefficient	Coefficient of determination	Absolute mean error	Accuracy
Decision tree	0.444962	0.198	0.029548	84.41558
Random forest	0.558711	0.3122	0.027503	85.71429
Linear regression	0.578914	0.3351	0.026965	88.31169
Neural network	0.509931	0.26	0.02538	88.31169

6.1 Granger Causality Results between Robbery and Economic Indicators

The results of granger causality between Robbery and economic indicators: GDP, NDP, Unemployment rate, Per Capita Income and Consumer Price Index have been depicted in Table 5. The results of F statistic and probability for the variables a/b have

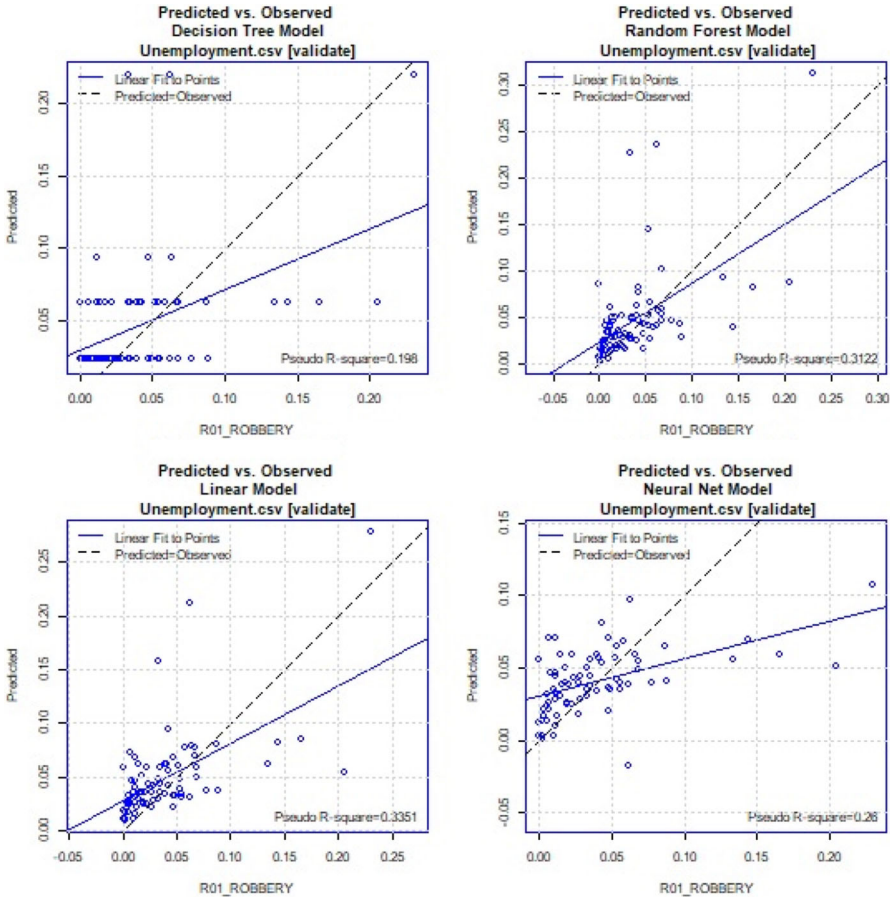

Fig. 7 Predicted versus observed values of robbery

been displayed where *a* is the dependent variable and *b* is the independent variable respectively.

From this Table 5, it has been observed that there is a unidirectional causation between Robbery and GDP that is Robbery granger cause GDP. Granger causality also exists between Robbery and NDP and Unemployment rate and Robbery. Unidirectional causation exists between Robbery and Per Capita Income.

6.2 Granger Causality Results between Burglary and Economic Indicators

The granger causality results between burglary and GDP, NDP, Unemployment rate, Per Capita Income and Consumer Price Index have been presented in Table 6.

From this Table 6, it has been observed that the causation only exists between Burglary and Unemployment rate.

Table 5 Granger causality results between robbery and economic indicators

Variables	F-Statistic	Probability
(Robbery|GDP)	2.934	0.066
(GDP|Robbery)	3.7475	0.03323*
(Robbery|NDP)	2.7435	0.07781
(NDP|robbery)	3.483	0.04142*
(Robbery|unemployment rate)	6.2029	0.004845**
(Unemployment rate|robbery)	0.723	0.4922
(Robbery|Per capita income)	2.4968	0.09651
(Per capita income|robbery)	3.2963	0.04847*
(Robbery|CPI)	2.6584	0.08379
(CPI|Robbery)	0.0555	0.9461

*, **are the statistical significance as 5% and 1% levels respectively

Table 6 Granger causality results between burglary and economic indicators

Variables	F-Statistic	Probability
(Burglary|GDP)	2.3686	0.108
(GDP|Burglary)	2.1705	0.1288
(Burglary| NDP)	2.4195	0.1033
(NDP|Burglary)	2.004	0.1496
(Burglary|Unemployment rate)	0.8111	0.4523
(Unemployment rate|Burglary)	4.9011	0.01311*
(Burglary|Per capita income)	2.63	0.08588
(Per capita income|Burglary)	1.8237	0.176
(Burglary|CPI)	0.2385	0.789
(CPI|Burglary)	0.6012	0.5536

*, ** are the statistical significance as 5 and 1% levels respectively

6.3 Granger Causality Results between Theft and Economic Indicators

The results of granger causality for Theft and GDP, NDP, Unemployment rate, Per Capita Income and Consumer Price Index have been shown in Table 7.

From the Table 7, it has been observed that unidirectional causation exists between Theft and unemployment rate.

It has been observed from the experiments conducted above that when the target variable is considered as crime rates, a unidirectional causation between Unemployment rate and Robbery exists at 1% level. This implies that as Unemployment rate increases, the crime rates (Robbery) also tend to increase.

Table 7 Granger causality results between theft and economic indicators

Variables	F-Statistic	Probability
F(Theft|GDP)	2.5138	0.09507
F(GDP|Theft)	0.6131	0.5472
F(Theft|NDP)	2.731	0.07866
F(NDP|Theft)	0.3407	0.7136
F(Theft|Unemployment rate)	0.7742	0.4686
F(Unemployment rate|Theft)	6.5122	0.003855**
F(Theft|Per capita income)	2.8087	0.07354
F(Per capita income|Theft)	0.1419	0.8682
F(Theft|CPI)	0.4113	0.6659
F(CPI|Theft)	0.6067	0.5506

*, **are the statistical significance as 5 and 1% levels respectively

7 Actions Taken to Control Crime

If the impact of economic stress on crime trends can be determined properly then the crime rate can be controlled. To accomplish this, Government can form numerous policies as some of them are mentioned below:

- As unemployment rate has a unidirectional causal relationship with crime, then the reduction in unemployment reduces crime rates also. Government should raise more employment opportunities by providing more funds to establish Small Scale Business, education organisations and Start-ups.
- Software such as Crime Mapping, Analytics and Predictive System(CMAPS) have been used by Police to monitor the crime http://delhipolice.nic.in/CP%20Forword2015.pdf. This crime forecasting software involves statistical models and algorithms to prevent crime. If data of the economic indicators is included in this software, it would be helpful in predicting the crime rate based on the economic indicators also.

8 Conclusion

In this paper, machine learning techniques have been used to analyse crime data and to discover the economic factors affecting crime in India. Experimental work using various machine learning algorithms like Decision Trees, Random Forest, Linear Regression and Neural Networks have been carried out by taking unemployment rate and GDDP as the independent input variables and Theft, Burglary and Robbery as the target variables. Further, the performance of the regression model has been evaluated in detail based on correlation coefficient r, the coefficient of determination R^2, Absolute Mean Error and Accuracy. It has been found that out of all the four machine learning algorithms, Linear Regression outperformed the other algorithms

and has the highest accuracy. Granger causality for crime rates and economic indicators has been calculated in the paper. It has been found that a unidirectional causation exists between Unemployment rate and Robbery. Therefore, Government should provide more employment and education opportunities to control the crime rate. In future, Android based applications akin to CMAPS can be developed using machine learning algorithms which would be helpful for Government and Police departments.

References

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge: MIT Press.

Chauhan, T., & Aluvalu, R. (2016) Using big data analytics for developing crime predictive model. In: Paper presented at RK University's First International Conference on Research & Entrepreneurship, pp 1–6.

Cohen, J., Gorr, W. L., & Olligschlaeger, A. M. (2007). Leading indicators and spatial interactions: A crime-forecasting model for proactive police deployment. *Geographical Analysis, 39*(1), 105–127.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61,* 115–125.

Granger, C. W. (1988). Causality, cointegration, and control. *Journal of Economic Dynamics and Control, 12*(2–3), 551–559.

Haykin, S. S. (2001). *Neural networks: a comprehensive foundation*. London: Pearson Education.

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms prediction. *Indian Journal of Science and Technology, 6*(3), 4219–4225.

Jain, A., & Bhatnagar, V. (2016). crime data analysis using pig with hadoop. *Procedia Computer Science, 78,* 571–578.

Kaikhah, K., & Doddameti, S. (2006). Discovering trends in large datasets using neural networks. *Applied Intelligence, 24*(1), 51–60.

Ku, C. H., & Leroy, G. (2014). A decision support system: Automated crime report analysis and classification for e-government. *Government Information Quarterly, 31*(4), 534–544.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ), 2*(1), 1–12.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill. p. 2. ISBN 0-07-042807-7.

Mittal, M., Sharma, R. K., & Singh, V. P. (2015). Modified single pass clustering with variable threshold approach. *International Journal of Innovative Computing Information and Control, 11*(1), 375–386.

Mittal, M., Singh, V. P., & Sharma, R. K. (2011). Random automatic detection of clusters. In *2011 International conference on image information processing (ICIIP)* (pp. 1–6). IEEE.

Nath, S. V. (2006). Crime pattern detection using data mining. In Web intelligence and intelligent agent technology workshops, 2006. WI-IAT 2006 workshops. IEEE/WIC/ACM international conference for crime on 2006 (pp. 41–44). IEEE.

Renushe, H. N., Rasal, P. R., & Desai, A. S. (2012). Data mining practices for effective investigation of crime. *International Journal of Computer Applications in Technology, 3,* 865–870.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In: *Educational Data Mining*.

Saeed, U., Sarim, M., Usmani, A., Mukhtar, A., Shaikh, A., & Raffat, S. (2015). Application of machine learning algorithms in crime classification and classification rule mining. *Research Journal of Recent Sciences, 4*(3), 106–114.

Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., & Tyagi, N. (2015). Crime detection and criminal identification in India using data mining techniques. *AI & Society, 30*(1), 117–127.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression 1. *Radiology, 227*(3), 617–628.