MACHINE LEARNING TECHNIQUES BASED EXPLORATION OF VARIOUS TYPES OF CRIMES IN INDIA

Geetika Bhardwaj

PhD Scholar, Department of Computer Applications, Punjabi University NH64 next to Urban Estate Phase II, Patiala, Punjab 147002, India
Inbox.geetika@gmail.com

Dr.R.K.Bawa

Dean, Faculty of Computing Sciences, Punjabi University NH64 next to Urban Estate Phase II, Patiala, Punjab 147002, India
Rajesh.k.bawa@gmail.com

Abstract

The reports of crime has been on a rise especially the cyber crime cases have seen a swift increase since the year 2018. It has been reported that there were two lakh incidents in 2018 which almost became seven times by 2021 i.e 14,02,809 cases in 2021; and 2,12,485 incidents in the starting two months of 2022 as per a renowned newspaper. The pace at which crime rate in digital world is increasing has become a pressing issue. As a result, it is critical to employ various techniques to forecast the rate and timing of digital crime events. Hence, in this paper, data for 2018 and 2019 have been collected from six different datasets to research multiple crimes occurring under cyber crime category. The data has been pre-processed and graphically visualized to assess the crime data appropriately. Using R2 and mean square error metrics, six machine learning algorithms have been used to evaluate the performance of the system and it has been discovered that decision trees had the highest R2 value and the lowest mean square error value of 99.9 and 0.01 for all crimes, respectively. In addition, a report on cognizable crime has also been provided for the years 2018 to 2021.

***Keywords*: Crimes; Cybercrime; Machine Learning; Decision Tree; NCRB.**

1. Introduction

Crime is a socioeconomic issue done by criminals that negatively impacts life quality as well as economic growth of a country, and criminology is a method to determine the behaviour characteristics of such criminals. The detective agencies, police departments, and crime branches use criminology to help them identify the actual nature of a criminal. Since 1800, the criminology bureau has been used in the actions of criminal investigations [Middleton (2021)]. The crime pattern is largely determined by the kind of community one resides in which forms the basis of the various aspects of how the crime is committed. Every second, many crimes occur in various locations, in multiple patterns, and at different times, due to which its number is growing by the day. Furthermore, with the advancement of technology, crime has expanded its reach digitally, which we call cybercrime. Cybercrime is commonly associated with various terminologies such as computer crime, e-crime, Internet crime, etc. Such crimes can be defined as involving any electronic device and a network. Using technological devices to commit a crime or making electronic device a target of crime is quite a common phenomenon these days. However, two types of cybercrimes have been noted down to date: computer-assisted, which includes Online Fraud, Cyber stalking, money laundering and many more are examples of computer-assisted cybercrimes, whereas hacking, defacement of websites, phishing, etc. are examples of computer-focused cybercrimes [Al-Khater *et al.* (2020)].

According to figure 1, Uttar Pradesh had the most cybercrime cases in 2020, with 11,097, followed by Karnataka, which had 10,741 cases. According to the data, Maharashtra (5,496 cases) and Telangana (5,024 cases) came in second and third, respectively. Between 2019 and 2020, the following states saw a significant increase in cybercrime cases: Assam (2,232 to 3,531); Arunachal Pradesh (from 8 to 29); Goa (16 to 41); Chhattisgarh (176 to 298); Manipur (5 to 80); Gujarat (785 to 1,284); and Telangana (4 to 79).

Furthermore, states such as Bihar reported a significant increase in cybercrime, i.e. (2,692 to 5,025) last year compared to 2018. Cases of cybercrime have more than quadrupled in Bihar and Telangana, while they have increased by more than 75 percent in Uttar Pradesh since 2018. Cases have doubled in Odisha, West Bengal, Karnataka, and Chhattisgarh since 2018, while they have tripled in Tamil Nadu and Manipur.

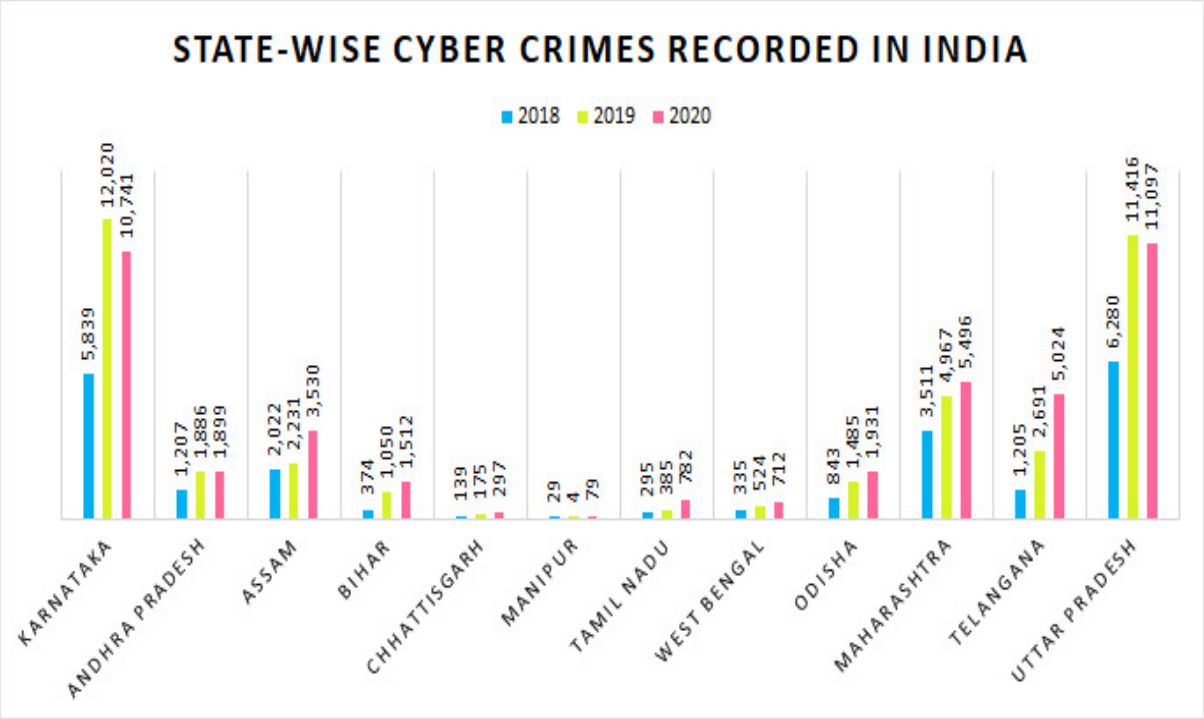


Fig 1: Cybercrimes in India 2018-2020 [NCRB 2020]

Many scholars [Pramanik *et al.* (2021)] feel that solving crime is demanding and requires time as it requires human knowledge and experience. After analyzing a lot of work in this field, it has been finally observed that data mining and AI algorithms, is the technology that can assist us in dealing with this problem. In the mid-1990s, data mining became famous for extracting meaningful data from multiple datasets and discovering the relationships between the features of the data [David and Suruliandi (2017)]. Data collection, pattern identification, classification, visualization, and prediction are all common steps to analyzing crime using machine-learning-based techniques. They have also been used for association analysis, cluster analysis, classification, prediction, and outlier analysis in structured as well as unstructured data [Kim *et al.* (2018)]. Few researchers in [Zhao and Tang (2017)] used a reverse-geo coding technique as well as density-based clustering algorithm to develop a machine-learning model to predict crime. Likewise, a transfer-learning framework was presented for exploiting meteorological data, data relating to geographical movement of people to capture temporal-spatial patterns. In [Budur *et al.* (2015)], the Gradient Boosting Machine (GBM) approach was implemented in a prediction model based on machine-learning. The use of page-rank-based method with weights was effective to weaken and eliminate criminal networks. Machine learning algorithms have also been used to detect data breaches such as man in the middle (MIM), malware attacks, eavesdropping attacks, spear-phishing attacks and many other types of attacks. Hence, to educate and train state and Central Police to predict crimes using AI-based approaches, the Indian government has taken several initiatives by developing programs and software by collaborating with NCRB [Gupta *et al.* (2014)].

In continuing the facts mentioned above, the primary objective of this work is to analyze and accurately predict crime in India based on crime data provenance. In our research, six machine learning algorithms such as decision tree, random forest, Gradient Boosting, and its related algorithms were applied to analyze the crime data that had been compiled from six datasets such as crimes in India, NCRB, cyber crimes in Indian cities, Indian crimes data, the crime rate in India, and Indian crime analysis between 2018 and 2019. The contribution to this research work is as follows:

- Initially, the data has been pre-processed to remove NAN and missing values using the SimpleImputer library.
- In the second stage, data has been visualized graphically for its better understanding, followed by applying feature scaling techniques like principal component analysis, standard scalar and min-max techniques to reduce the dimensions of the dataset to get better-structured data.
- In the next process, the data, after splitting into training and testing phases, is sent to the multiple machine learning models to evaluate their performances using the R2 score and mean square error rate.

2. Background

Researchers have done a lot of work in the field of crime detection using multiple machine learning algorithms. In their paper [Shah *et al.* (2015)], the authors predicted physical or psychological-based crimes by applying machine learning-based algorithms and computer vision in an accurate and timely manner. They used techniques such as kernel density estimation (KDE), support vector machines (SVM) and regression analysis to develop a crime prediction system. The three parts of their strategy were data gathering, statistical analysis between the relationships of crime episodes as well as obtained data, and precise prediction of crime occurrences. In [DeBruin *et al.* (2006)], the authors suggested a method for assessing how criminals are clustered based on their history of crime. Every year for each crime the criminal profile is extracted from the database had been used to construct a profile distance based on which the profile per year distance matrix was generated. The distance matrix, which included the frequency value, had been utilized to construct clusters using the nave clustering technique and developed a criminal profile to portray an offender's criminal history over a year. This knowledge could easily be used to evaluate a big group of criminals, and individual suspects' future behaviour can be anticipated. As per the authors, their technique will aid in the formation of a comprehensive picture of the numerous types of criminal jobs currently available. They also utilized the program to extract characteristics from the Dutch National Criminal Record Database that could be utilized to determine a person's criminal history. As a result, the researchers focused on FP tree, CART algorithms, K-means, Apriori Algorithm and few others along with machine learning for deeper understanding of the criminal pattern and weigh it with actual data.

In the contemporary era, the majority of persons use credit cards to purchase goods, and as a result, they have become victims of fraudsters. Credit card fraud has increased due to emerging technologies, wreaking havoc on the lives of credit card users. In [Shabbir and Kannadasan (2013)], the authors described a generic method for preventing credit card fraud. It was used to build sophisticated systems to lower computation costs over time. It could detect a fraudulent transaction in a matter of seconds. Misrepresentation exchanges are likely to follow credit card exchanges, and hostile extortion mechanisms may be set up to shield banks from massive losses and minimize hazards. Likewise, in [More *et al.* (2021)], the authors proposed an effective fraud detection system based on random forest algorithms to address credit card users' real-time issues to protect users from such frauds. The authors also used a learning-to-rank approach to rank the alerts, which reduced the number of false alerts generated by the fraud detection system and provided more reliable fraud alerts. The authors used a dataset of 100000 cardholder transactions and obtained an accuracy of 97.93 percent. In addition, the model was compared to Nave Bayes and Decision Tree to demonstrate that it was effective at detecting credit card fraud.

Cyber bullying has harmed people's lives, causing severe problems and, in some cases, prompting victims to attempt suicide. This necessitates the development of models to detect cyber bullying to protect individuals' interests. Hence, in their paper [Sandesh *et al.* (2022)], the authors provided insights into cyber bullying and the process of detecting cyber bullying using machine learning algorithms. Their proposed system used labelled data for experimentation and extracted the information related to network, users, and tweet contents from the Twitter platform. Similarly [Gomez *et al.* (2004)], the authors examined outside physical activity and violent crime among inside-city children. In a multiple regression analysis, there was utilization of outdoor physical activities. The purpose of the study was to see if there were any relationships between adolescent outdoor physical activity and violent crime rates, as well as other natural important variables.

The authors in [Tamilarasi and Rani (2020)] explained how the frequency of crimes and crime features in India, such as rape, abetment, sexual assault, and kidnapping, can be investigated using machine learning models. They developed a model to predict crime rates. They tested the accuracy of six machine learning algorithms on crime data, including Linear Regression, KNN, SVM,Nave Bayes, decision trees and CART (Classification and Regression Tree. In [Avila *et al.* (2021], the authors suggested a new classification technique to detect information breaches using GDPR principles. The authors first identified the most used dataset publicly available to detect the threat and later described twenty types of attacks in them. Moreover, they also told the thirty algorithms that had been used to detect whether the data had been leaked or not.

Similarly, the authors in [Izziden *et al.* (2021)] used the golden rule concept in their paper to respond to criticisms and measure the text's fairness using PCA and ALiteBERT techniques. During the experiment, the authors obtained an F1 score of 0.85 and proposed the use of technology along with its implementation so as to steer clear of unjust partiality in word embeddings. In [Cisco], the author demonstrated in 2019 that malware-based encrypted or decrypted software can be identified using the L1 logistic regression model and IP flow data fields. Using this, the author obtained an accuracy of 99.978 percent and a false discovery rate (FDR) of 0.00 percent. Later, he introduced a new concept of encrypted traffic analytics in his paper that included the extra IP flow data fields that captured, extracted, and analyzed the network flow data to solve cybersecurity problems.
In [Fatih and Bekir (2015)], the authors stated that obtaining the suspected profiles and analyzing the same from multiple databases can be done using a facial recognition system. Moreover, getting the suspected vehicle involved in a crime can be done using a license plate reader system. The developers could also use body-attached cameras that capture information more than a human eye to record every detail of crime-happening events. In [Canhoto (2021)], the authors analyzed that there has been a limited scope of using supervised-based machine learning

algorithms because of large datasets, lack of good quality data, etc. Hence, keeping that in mind, the authors used reinforcement and unsupervised machine learning algorithms that could predict terrorist funding by finding bizarre behavior in a financial section other than money laundering. Major headings should be typeset in boldface with the first letter of important words capitalized.

3. Proposed System and its Implementation

In this section, the detailed illustration regarding the flow of research has been shown (Figure 2) such as dataset, libraries, data pre processing, exploratory data analysis, models, and evaluative parameters.

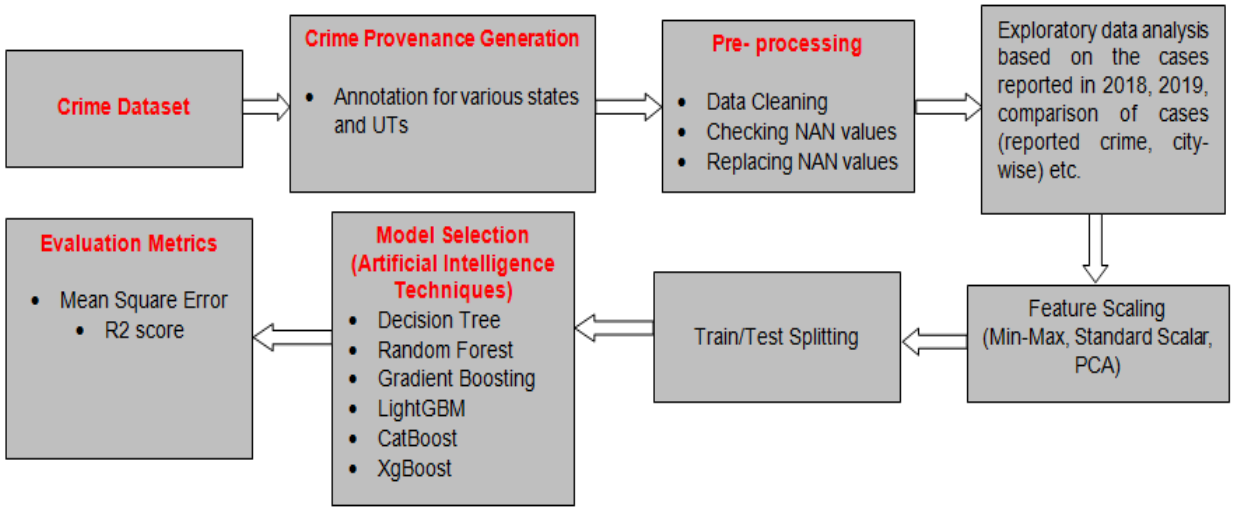


Fig 2: Proposed System

3.1 Dataset

The research work has been conducted by using six datasets which include various crimes such as anger, extortion that causes disrepute, personal revenge, fraud, sexual exploitation, prank, illegal drugs, Sale of Purchase of Illegal Drugs, piracy, stealing information, terrorist funding, Developing Our Business, Psycho or Pervert, Abetment of Suicide, Terrorist Recruitment, Inciting hate against the country, Political Moves, cognizable crime, etc. The data collected from district-wise crimes in India consist of 33 columns which include states, districts, a year from 2001 to 2012, and information related to crimes like murder, kidnapping, etc. [crime in India]. Likewise, the data collected from the National Crime record bureau has a kidnapping and abduction report of 34 metropolitan cities from 2014-2020[crime in Indian cities]. The third dataset has been collected from cybercrime in Indian cities. It contains various types of cybercrime motives along with the cases reported in the particular 34 metropolitan cities of India, including the crime rate in those cities for 2018-2019. In this dataset, the first column contains the name of the cities. The subsequent columns have the motives for cybercrime until the second last column. The second last column contains the total crimes reported in the particular city, and the last column contains the crime rate in the city [India crimes data]. The fourth dataset has been collected from Indian crime data [crime rate in India]. It contains Year wise crime data for each state of India. The fifth dataset has been taken from the crime rate in India [Indian crime analysis]. The dataset includes crimes as per reported between 2019 and 2021 June. Additional information is given alongside the column description. Finally, the last dataset has been taken from the Indian crime analysis, which provides information related to burglary and murders in India in 2013.

3.2 Provenance Generation

The number of cases for different categories of crime reported at various places has also been organized and segregated according to various States and Union territories. As we know, India has 28 States and 8 Union Territories, and the data has been compiled by the number of cases of different crime categories reported in these places. The states and union territories have been annotated to be used as Provenance for crime prediction purposes. Table 1 consists of the name of all the states and Union territories and their annotations.

Table 1: Annotations for different states and Union Territories

Name of the State/Union Territory	State/UT Code(Annotations)
Arunachal Pradesh	SCAR
Assam	SCAS
Bihar	SCBI
Chhattisgarh	SCCH
Goa	SCGO
Gujarat	SCGU
Haryana	SCHA
Himachal Pradesh	SCHP
Jharkhand	SCJH
Karnataka	SCKA
Kerala	SCKE
Madhya Pradesh	SCMP
Maharashtra	SCMH
Manipur	SCMA
Meghalaya	SCME
Mizoram	SCMI
Nagaland	SCNA
Odisha	SCOD
Punjab	SCPU
Rajasthan	SCRA
Sikkim	SCSI
Tamil Nadu	SCTN
Telangana	SCTE
Tripura	SCTR
Uttar Pradesh	SCUP
Uttarakhand	SCUT
West Bengal	SCWB
Andaman and Nicobar Island	UTAN
Chandigarh	UTCH
Dadra and Nagar Haveli and Daman and Diu	UTDNH
Delhi	UTDH
Ladakh	UTLD
Lakshadweep	UTLA
Jammu and Kashmir	UTJK
Puducherry	UTPU

3.3 Libraries used

Several Python libraries have been loaded, and each one of them serves a specific predefined purpose. *Matplotlib* is a Python library that can create any chart. It's a data visualization library that runs on multiple platforms and helps with data visualization. *Pandas* are used to manage and import the datasets. It is an open-source data manipulation and analysis library that provides high-performance data manipulation in Python [Nakib *et al.* (2018)]. *Numpy* is a python package for the computation and processing of any mathematical operation, especially the multidimensional and single-dimensional array elements. *Seaborn* provides the facility of various color palettes as well as default styles for creating statistical plots. It aims to create a more attractive visualization of the central part of understanding and exploring data [18]. *OS is* Python's standard utility modules which provide functions for to interact with the operating system. *Sklearn* is a free machine learning library for Python that includes many practical tools for statistical modeling and machine learning, such as clustering, regression, classification, and dimensionality reduction. In the end, *Maths* is a module consisting of the most famous mathematical functions, which include trigonometric functions, representation functions, logarithmic functions, etc. [Ajagbe *et al.* (2020)].

3.4 Data Pre-processing

Data pre-processing is necessary for data cleaning and for preparing data for machine learning model, which also improves the accuracy and effectiveness of the model. The pre-processing data will be cleaned by Checking nan values and replacing Nan Values with 0 using the SimpleImputer library. The pre-processing of data has been done in three phases, i.e., reading, checking, and cleaning the data. The data has been cleaned by removing the NAN value; its output is shown in table 2.

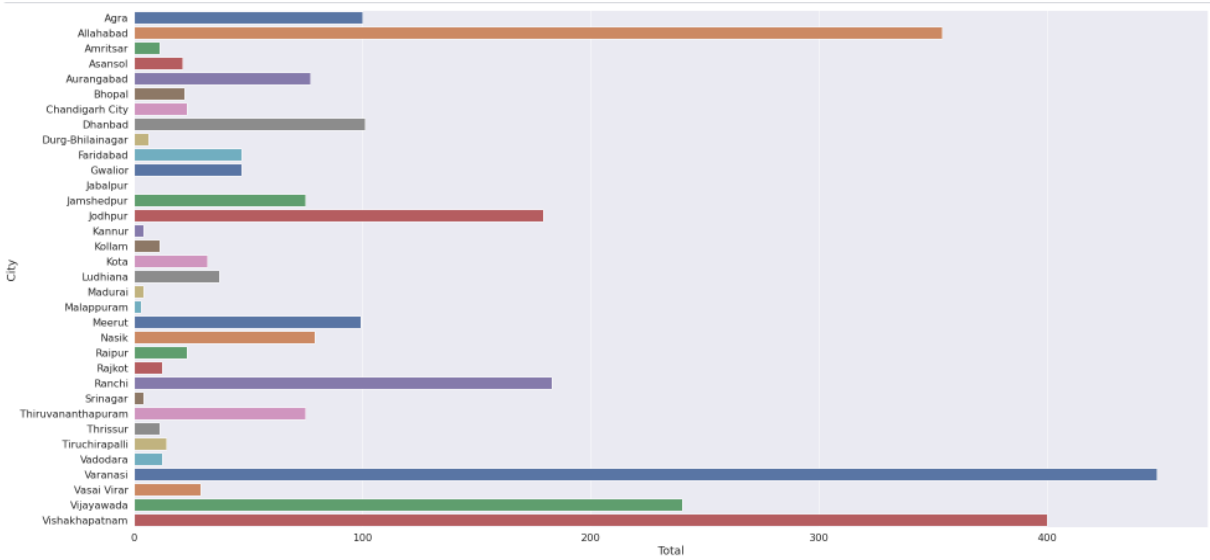
Table 2: Pre-processed dataset

2018		2019	
City	1	City	0
Personal Revenge	1	Personal Revenge	0
Anger	1	Anger	0
Fraud	1	Fraud	0
Extortion	1	Extortion	0
Causing Disrepute	1	Causing Disrepute	0
Prank	1	Prank	0
Sexual Exploitation	1	Sexual Exploitation	0
Political Motives	1	Political Motives	0
Terrorist Activities	1	Terrorist Activities	0
Terrorist Recuit- ment	1	Terrorist Recuit- ment	0
Terrorist Funding	1	Terrorist Funding	0
Inciting Hate against Country	1	Inciting Hate against Country	0
Disrupt Public Service	1	Disrupt Public Service	0
Sale purchase illegal drugs	1	Sale purchase illegal drugs	0
Developing own business	1	Developing own business	0
Spreading Piracy	1	Spreading Piracy	0
Psycho or Pervert	1	Psycho or Pervert	0
Steal Information	1	Steal Information	0
Abetment to Suicide	1	Abetment to Suicide	0
Others	1	Others	0
Total	1	Total	0
Crime Rate	1	Crime Rate	0

3.5 Exploratory Data Analysis

In this part, the pre processed dataset has been visualized graphically for the better understanding of the crime data. It includes cases reported in the year 2020 and 2021, comparison of those cases, and city wise comparison of reported crime cases.

(a)



(b)

Fig 3: Total Cases Reported in (a) 2018 and (b) 2019

Figure 3 depicts the total number of cases reported in 2018 and 2019. Vishakapatnam reported the highest number of crime cases in 2018, with over 400, while Jabalpur and Madurai reported the fewest. Similarly, Varanasi had the highest number of crime cases in 2019, with over 400, while Jabalpur had the lowest number of cases. Let's carefully examine both the 2018 and 2019 cases. We can see that Vishakapatnam and Varanasi have the highest number of cases recorded, ranking first and second compared to the other cities.

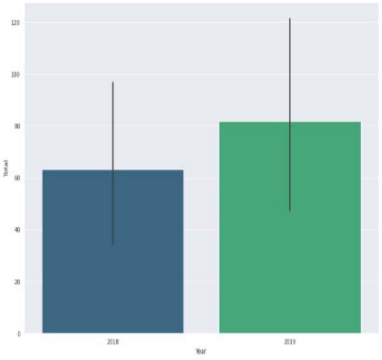


Fig 4: Cases of 2018 vs 2019

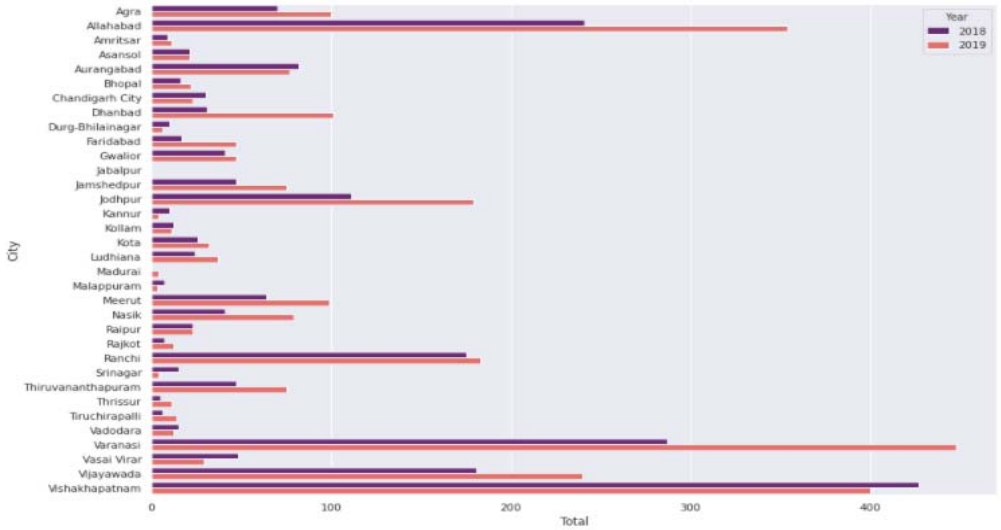


Fig 5: City-wise comparison of cases

Figure 4 compares cases recorded in 2018 and 2019, whereas Figure 5 depicts a city-by-city comparison of criminal cases. Based on the analysis of both graphs, it can be concluded that 2019 set a record for having the highest number of cases in India and that Varanasi had the highest number of cases compared to the other cities. In 2018, Jabalpur was the safest area, with no issues reported, and followed by Madurai. In 2019, Allahabad also had a low crime rate, followed by Vijayawada. The rest of the cities have crime rates, but they are quite common compared to the aforementioned cities.

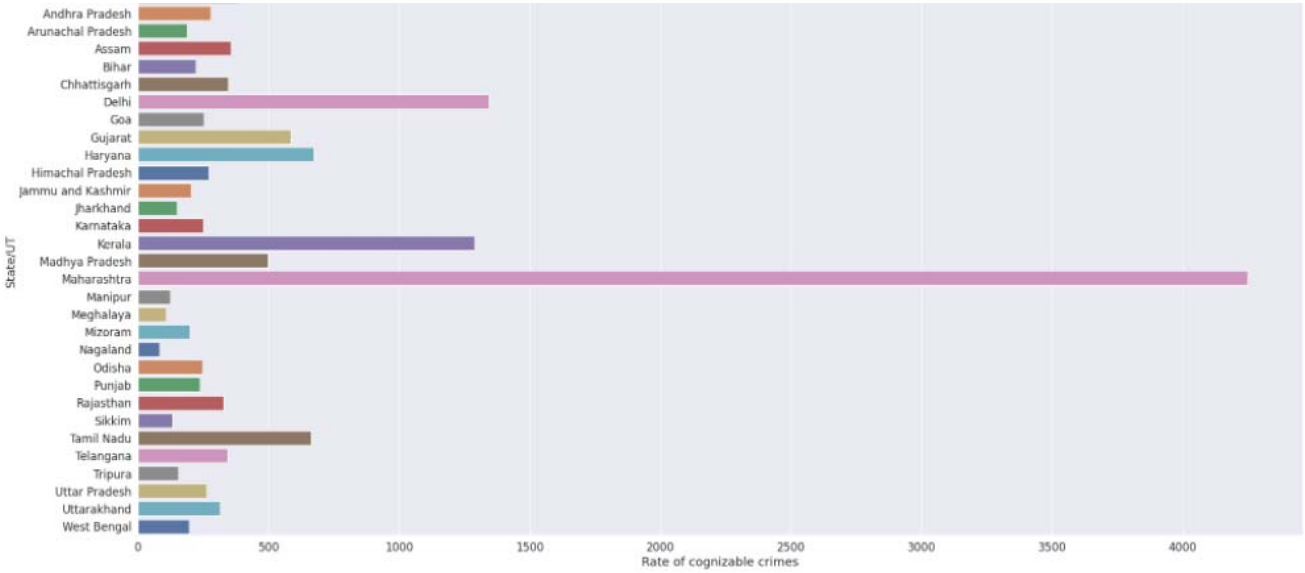


Fig 6: Cases reported for cognizable crime in India

A cognizable crime is one in which a police officer can arrest the convict without a warrant and begin an investigation without the permission of the court, in accordance with the first schedule or any other law in effect at the time. In a nutshell, police can arrest the perpetrator without a warrant in such cases. Figure 6 shows the rate of cognizable crimes in India, with the highest number of over 4000 in Maharashtra and the lowest in Nagaland.

3.6 Feature scaling

Feature scaling is the culminating step of data pre-processing. It is a technique for standardizing the independent variables of the dataset in a specific range. Many scaling techniques can be used, but the one that gives more optimized results after normalizing the data must be prioritized. The scalar techniques that have been used for women's crime data are:-

Min Max: In min-max, all the data is scaled between 0 and 1. Equation (1) shows the formula for calculating the scaled value using the min-max scaling technique [Ghankutkar *et al.* (2019)]

$$x_{scaled} = \frac{(x - x_{\min})}{x_{max} - x\_min} \quad (1)$$

Standard Scalar: Standard Scalar scales the values in a manner that it brings the mean to 0 and the standard deviation to 1(or the variance) [Ivan *et al.* (2017)].

Standardization is calculated by subtracting every value present in the dataset from the mean and then dividing the resultant value by the overall deviation in the data set. It is solved by using eq (2-4)

$$v = \frac{z - \mu}{\sigma} \quad (2)$$

Where,

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(z_i) \quad (3)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \mu)^2} \quad (4)$$

Here, v is standardization, x is a variable value, σ is mean, μ is standard deviation, and n is several observations.

Principal Component Analysis: It is a factor analysis and a statistical method used in image processing to isolate features and apply the characteristics comprised of the reduction of n-dimensions to determine interrelations among variables [Mathew and Asha (2022)]. It is solved by using eq (5-8)

$$z = \frac{value - mean}{standard\ deviation} \quad (5)$$

$$cov\ (X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) \quad (6)$$

$$\det\ (A - \lambda\, I) = 0 \quad (7)$$

Here A is a square matrix, I is an identity matrix, X and Y are the data points and λ is an eigen vector and det is the determinant, cov is a covariance matrix

$$FinalDataset = FeatureVector^T * StandarizedOriginalDataset^T \quad (8)$$

Overall, the Scaling technique that gave the best results after normalizing the data was Standard Scalar. After feature scaling, the dataset has been split into training dataset and testing dataset each having 75% and 25% of the dataset respectively on which machine learning algorithms have been applied.

3.7 Models Applied

Decision tree: It is an effective classification as well as prediction tool. The purpose is to remember simple and clear decision rules inferred from the data features so as to construct a model which is capable of predicting the value of a target variable.

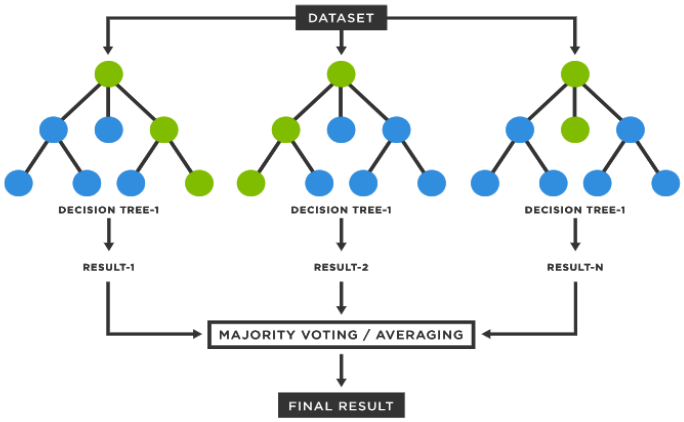


Fig 7. Decision tree architecture [Rawat *et al.* (2021)]

In the decision tree, as shown in figure 7, every leaf node (terminal node) has a class label, every branch constitute the result of the test, and each internal node depicts a test on a particular attribute. The three main criteria used to solve decision tree algorithms are Gini impurity which measures the impurity of a node, and Entropy [Rawat *et al.* (2021)]. They are calculated by the eqs (9-11)

$$I_G\ (n) = \ 1 - \sum_{i=1}^{J}(pi)^2 \quad (9)$$

$$Entropy = \ \sum_{i=1}^{J} - p_i * \ log_2(p_i) \quad (10)$$

where J is the number of classes and p is the distribution of the class in the node [Rawat *et al.* (2021)].

Random forest: It is a machine learning technique to solve regression and classification related problems. Random Forest uses ensemble learning to solve complex problems by combining many classifiers.

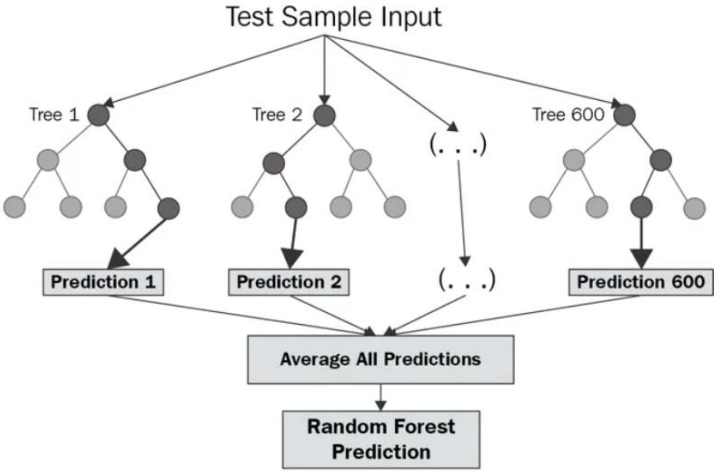


Fig 8. Random forest architecture [Amanoul *et al.* (2021)]

It consists of many tiny decision trees, often known as estimators as shown in figure 8. Each node in the tree is taught to make its predictions using a separate set of observations. The final predictions of random forest is determined by doing the average of the forecasts of each tree [Amanoul *et al.* (2021)]. Eqs solve the Random Forest algorithm. (11-14)

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \tag{11}$$

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \tag{12}$$

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \tag{13}$$

$$ni_j = W_j C_j - W_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)} \tag{14}$$

Here ni_j = node j, W_j = weighted number of samples to reach node j, C_j = node j's impurity value, $left(j)$ = child node from left split on node j, $right(j)$ = child node from right split on node j, fi_i = the importance of feature i, $RFfi_i$ = the importance of feature i calculated from all trees in the Random Forest model, $normfi_i$ = the normalized feature importance for i in tree j, T = total number of trees.

Gradient Boosting: Gradient Boosting is one of the most powerful techniques for constructing predictive models. It trains many models in parallel. Every new model minimizes the loss function using the Gradient Descent Method. It benefits from regularization methods by penalizing different parts of the algorithm and reducing overfitting to improve the algorithm's performance by [Mathew and Asha (2022)].

XgBoost: eXtreme Gradient Boosting is the abbreviation for eXtreme Gradient Boosting. It is based on the concept of gradient boosting. It also makes use of decision trees. It is a custom, parallelized tree-building algorithm which contains several decision trees. It provides features like penalization of trees, efficient handling of missing data, and automatic feature selection [Walteros-Alcázar *et al.* (2021)]. It's an effective method for creating supervised regression models.

XGBoost is a high-speed software library that supports diverse interfaces, including the Command Line Interface (CLI), JVM languages C++, R, Julia, Java and Python [Prabakaran and Mitra (2018)]. Gradient Boosting method, Stochastic Gradient Boosting, and Regularized Gradient Boosting are all supported by XGBoost. The primary goal of using XGBoost is to improve the project's execution speed and model performance. Regression linear is the most frequent loss function in XGBoost for regression issues, while regression logistical is the most used loss function for binary classification [Prabakaran and Mitra (2018)]. Equation (15,16) shows the formula for calculating it:

$$L(\emptyset) = \sum_i l(\hat{y}_i,\ y_i) + \sum_k \Omega(f_k) \tag{15}$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{16}$$

y_i is real value (label) known from the training data set.

CatBoost: "CatBoost" is an open-source algorithm which is combination of two words "Category" and "Boosting." It uses gradient boosting on decision trees. The usage of non-digit factors is allowed in place of pre-

processing data. It combines one-hot encoding and an advanced mean encoding [Sachin *et al.* (2020)]. CatBoost is a recently open-sourced machine learning algorithm created by Yandex. It interacts nicely with well-known deep learning frameworks such as Core ML by Apple as well as TensorFlow from Google. It can work with a range of data types and may aid companies in resolution of a wide range of issues [Sachin *et al.* (2020)]. Additionally, it boasts the highest accuracy in its class. It generates novel insights without needing extensive data training, as other machine learning techniques do, and provides off the wall support for the more descriptive data formats associated with a wide variety of business issues.

LightGBM: Microsoft created LightGBM, a free and open-source distributed gradient boosting platform for machine learning. It's based on decision tree algorithm which improves the efficiency of the model while lowering usage of memory. It produces far more complex trees by using a leaf-wise split strategy which is the primary factor in getting better precision [Rawat *et al.* (2021)].

3.8 Evaluative Parameters

R2 score: It's the total variance elaborated by the model divided by the total variance. The R2 score can range from 0 to 100 percent. If it is 100 percent, the two variables are perfectly correlated, meaning they have no variance [Kumar *et al.* (2020)]. It is shown by eq (17)

$$R^2 = 1 - \frac{sum\ of\ squares\ of\ residuals}{total\ sum\ of\ squares} \quad (17)$$

MSE: The average of all the square of the errors is called the mean square error (MSE) [Avila *et al.* (2021)]. Larger the number, greater the error. It is computed by eq (18)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (18)$$

Where n is total number of observations, Y_i is observed values, $\hat{Y}_i$ is predicted values

4. Background

In this section, results have been computed for various crimes such as personal revenge, anger, fraud, extortion, disrepute, prank, exploitation, developing own business, steal information, psycho or pervert, spreading piracy, abetment to suicide, and others. Multiple machine learning algorithms have been applied like decision tree, gradient boosting, XgBoost, CatBoost, LightGBM, and random forest to detect crime and are evaluated using R2 score and mean square error.

Table 3: Evaluation of models under min-max technique

Crimes	Decision Tree		Random Forest		Gradient Boosting		XgBoost		CatBoost		LightGBM	
	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE
Personal Revenge	99.9	0.01	62.1	0.21	97.6	0.22	99.9	0.02	99.0	0.35	51.5	2.96
Anger	99.9	0.01	89.3	152.3	99.6	4.49	99.9	0.50	99.9	0.78	44.85	20.63
Fraud	99.9	0.01	35.9	16.7	98.9	0.27	99.7	0.52	99.9	0.22	42.3	30.03
Extortion	99.9	0.01	78.2	1.28	98.3	0.09	99.7	0.01	99.3	0.03	80.4	6.53
Disrepute	99.9	0.01	97.6	0.11	81.9	6.87	83.6	5.95	90.5	0.39	93.89	0.25
Prank	99.9	0.01	62.1	0.21	97.6	0.22	99.9	0.02	99.0	0.35	51.5	2.96
Sexual Exploitation	99.9	0.01	99.5	0.01	95.6	0.05	80.65	131	92.6	0.30	69.9	4.11
Political Motives	99.9	0.01	97.5	0.03	85.6	3.31	91.45	0.26	80.63	3.56	99.7	0.02
Terrorist Activities	99.9	0.01	99.4	0.02	80.25	3.35	95.6	0.30	75.65	4.56	99.9	0.01
Terrorist Recruitment	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Terrorist Funding	99.9	0.01	97.6	0.11	81.9	6.87	83.6	5.95	90.5	0.39	93.89	0.25
Inciting Hate Against Country	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Disrupt Public Service	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Sale/Purchase Illegal Drugs	99.9	0.01	98.5	0.02	91.56	3.31	99.8	0.01	95.64	2.36	65.42	21.56
Developing Own Business	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Piracy	99.9	0.01	98.9	0.02	80.56	4.39	89.41	2.44	95.6	1.42	96.57	0.34
Psycho or Pervert	98.9	0.03	97.9	0.05	90.5	4.31	94.4	0.23	78.56	6.64	91.45	1.05
Steal Information	95.9	1.91	85.3	5.64	95.3	2.74	90.4	3.09	96.5	3.23	65.5	28.2
Abetment	95.9	1.01	95.9	1.02	90.5	2.31	90.4	2.03	75.56	6.64	90.45	1.05
Others	95.9	1.91	85.3	5.64	95.3	2.74	90.4	3.09	96.5	3.23	65.5	28.2

After applying min-max in table 3, decision tree has given the best results for all the mentioned crimes in terms of R2 score and mean square error rate while as the least scores have been obtained by lightGBM, random forest, gradient boosting, catboost for different crimes. The highest value achieved so far is 99.9 R2 score and 0.01 mean square error rate while the least is 35.5 R2 score and 48.2 root mean square error.

Table 4**:** Evaluation of models under PCA technique

Crimes	Decision Tree		Random Forest		Gradient Boosting		XgBoost		CatBoost		LightGBM	
	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE
Personal Revenge	99.9	0.01	80.6	0.23	85.5	0.22	98.9	0.02	98.7	0.22	70.5	4.35
Anger	99.9	0.01	78.2	1.28	98.3	0.09	99.7	0.01	99.3	0.03	80.4	6.53
Fraud	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.30	69.9	4.11
Extortion	99.9	0.01	65.5	3.29	97.8	0.20	99.1	0.07	99.5	0.04	78.5	17.05
Disrepute	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.30	69.9	4.11
Prank	99.9	0.01	97.6	0.11	81.9	6.87	83.6	5.95	90.5	0.39	93.89	0.25
Sexual Exploitation	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.31	69.9	4.11
Political Motives	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.30	69.9	4.11
Terrorist Activities	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.30	69.9	4.11
Terrorist Recruitment	99.9	0.01	99.4	0.02	80.25	3.35	95.6	0.30	75.65	4.56	99.9	0.01
Terrorist Funding	99.9	0.01	97.5	0.03	85.6	3.31	91.45	0.26	80.63	3.56	99.7	0.02
Inciting Hate Against Country	99.9	0.01	97.5		0.03	85.6	3.31	91.45	0.26	80.63	99.7	0.02
Disrupt Public Service	99.9	0.01	98.5	0.02	85.6	3.57	99.8	0.01	80.59	4.00	98.6	0.02
Sale/Purchase Illegal Drugs	99.9	0.01	97.5	0.91	70.56	6.82	71.40	6.10	80.56	3.73	91.56	0.08
Developing our Businesses	98.9	0.03	97.9	0.05	90.5	4.31	94.4	0.23	78.56	6.64	91.45	1.05
Piracy	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Psycho or Pervert	99.9	0.01	97.5	0.26	96.45	0.35	8056	3.77	99.4	0.01	94.56	0.41
Steal Information	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41
Abetment	98.9	0.03	97.9	0.05	90.5	4.31	94.4	0.23	78.56	6.64	91.45	1.05
Others	97.9	0.91	80.3	4.65	93.3	1.74	96.4	2.09	95.5	2.23	35.5	48.2

After applying the PCA technique in table 4, the decision tree has given the best results for all the mentioned crimes in terms of R2 score and mean square error rate. At the same time, CatBoost obtained the highest R2 score value for crime of stealing information. The highest value achieved so far is 99.9 R2 and 0.01 mean square error rate, while the least is 65.5 R2 score and 28.2 root mean square error.

Table 5: Evaluation of models under Standard scalar technique

Crimes	Decision Tree		Random Forest		Gradient Boosting		XgBoost		CatBoost		LightGBM	
	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE
Personal Revenge	99.9	0.01	78.2	1.28	98.3	0.09	99.7	0.01	99.3	0.03	80.4	6.53
Anger	99.9	0.01	80.6	0.23	85.5	0.22	98.9	0.02	98.7	0.22	70.5	4.35
Fraud	9.9	0.01	89.3	152.3	99.6	4.49	66.9	0.50	99.9	0.78	44.85	20.63
Extortion	99.9	0.01	62.1	0.21	97.6	0.22	99.9	0.02	99.0	0.35	51.5	2.96
Disrepute	99.9	0.01	35.9	16.7	98.9	0.27	99.7	0.52	99.9	0.22	42.3	30.03
Prank	99.9	0.01	99.5	0.01	95.6	0.05	80.65	1.31	92.6	0.30	69.9	4.11
Sexual Exploitation	99.9	0.01	65.5	3.29	97.8	0.20	99.1	0.07	99.5	0.04	78.5	17.05
Political Motives	99.9	0.01	97.6	0.11	81.9	6.87	83.6	5.95	90.5	0.39	93.89	0.25
Terrorist Activities	99.9	0.01	98.5	0.02	91.56	3.31	99.8	0.01	95.64	2.36	65.42	21.56
Terrorist Recruitment	99.9	0.01	97.5	0.03	85.6	3.31	91.45	0.26	80.63	3.56	99.7	0.02
Terrorist Funding	99.9	0.01	99.4	0.02	80.25	3.35	95.6	0.30	75.5	4.56	99.9	0.01
Inciting Hate Against Country	99.9	0.01	94.56	0.91	62.56	5.89	82.3	2.19	60.23	6.79	97.56	0.03
Disrupt Public Service	99.9	0.01	83.16	0.02	99.9	2.89	99.9	1.02	99.3	0.78	86.5	0.14
Sale/Purchase Illegal Drugs	99.9	0.01	98.5	0.02	85.6	3.57	99.8	0.01	80.59	4.00	98.6	0.02
Developing our Businesses	99.9	0.01	97.5	0.91	70.56	6.82	71.40	6.10	80.56	3.73	91.56	0.08
Piracy	99.9	0.01	97.5	0.26	96.45	0.35	80.56	3.77	99.4	0.01	94.56	0.41

Psycho or Pervert	99.9	0.01	88.7	1.39	99.9	0.01	99.9	9.54	99.9	3.00	80.5	0.13
Steal Information	99.9	0.01	98.9	0.02	80.56	4.39	89.41	2.44	95.6	1.42	96.57	0.34
Abetment	99.9	0.01	98.9	0.02	91.5	3.31	95.4	0.03	79.56	5.64	92.45	0.05
Others	99.9	0.01	83.3	2.65	95.3	0.74	99.4	0.09	98.5	0.23	39.5	38.2

After applying the standard scaling technique in table 5, the decision tree has given the best results for all the mentioned crimes in terms of R2 score and mean square error rate. The highest value achieved so far is 99.9 R2 and 0.01 mean square error rate, while the least is 39.5 R2 score and 38.2 root mean square error by lightGBM. On analyzing the algorithms after applying the three scaling techniques, it can be said that the decision tree has performed very well for the evaluative metrics as compared to the other algorithms.

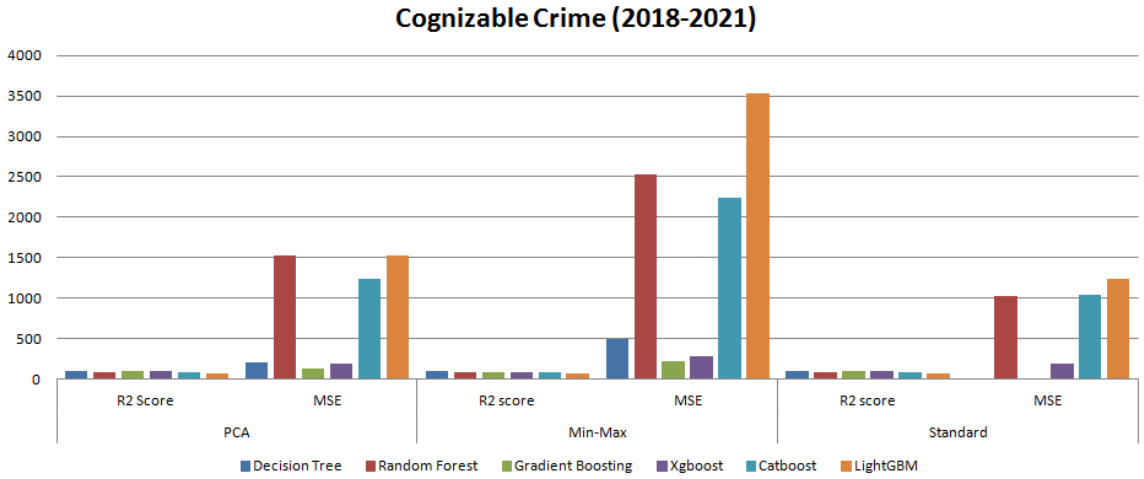


Fig 9: Cognizable crime (2018 to 2021)

In the case of cognizable crime, figure 9 depicts that the decision tree has performed well for PCA, Min-Max, and standard scaling techniques by 98.9, 95.9, and 99.9 R2 scores while 203, 503, and 0 mean square error rates, respectively, on comparing with the other algorithms. LightGBM has shown the least results after applying the min-max scaling technique by 65.4 R2 scores and a 3536.13 mean square error rate.

5. Conclusion and Future Scope

The advancement in technology has revolutionized all aspects of human life but it all has also lead to various social problems with cyber crime being one of the primary concerns. Crime problems in recent years have become more intelligent and diverse than in the past, and violent crimes are on the rise. Many researchers have used varying techniques to detect or predict different types of crime including cybercrime but each has its own set of limitations. Similarly, the datasets that have been used for different types of research on crime have their unique characteristics. The main aim of the research conducted in this paper was to analyze crimes in India using Artificial Intelligence techniques. Data for the years 2018 and 2019 were collected from six different datasets to research various crimes, including crimes in India, NCRB, cyber crimes in Indian cities, Indian crimes data, the crime rate in India, and Indian crime analysis. Six machine learning algorithms have been used ,for instance Gradient Boosting, Decision Tree, CatBoost, Random Forest, Xgboost, and LightGBM ; for evaluation using R2 and mean square error metrics. While in comparison with various algorithms, decision trees had the highest R2 value and the lowest mean square error value of 99.9 and 0.01 for all crimes, respectively, during the research. Furthermore, a report on cognizable crime has been provided for 2018 to 2021. After applying the learning models, we discovered that the performance of decision tree was better than other algorithms.
In the future, a model that particularly predicts the likelihood of a criminal committing cyber crime based on the types of crime already committed should be developed and deployed to web services. The profiling of cyber criminals can also be done keeping in mind the privacy of data and government regulations. The limited cybercrime datasets for Indian cyber crimes poses a restriction on the research work being done on them. New methods can be devised to collect data of cyber criminals which does not breach the privacy concern. Further on, the newly developed system can be reviewed for performance by comparing it with the existing cyber crime techniques used worldwide by various researchers and a separate study can be done on the various cyber crime datasets used internationally to identify the limitations of our own and to make it more efficient.

References

[1] Ajagbe, S. A., Idowu, I. R., Oladosu, J. B., & Adesina, A. O. (2020). Accuracy of machine learning models for mortality rate prediction in a crime dataset. *International Journal of Information Processing and Communication (IJIPC)*, *10*(1), 150-160.
[2] Al-Khater, W. A., Al-Maadeed, S., Ahmed, A. A., Sadiq, A. S., & Khan, M. K. (2020). Comprehensive review of cybercrime detection techniques. *IEEE Access*, *8*, 137293-137311.
[3] Amanoul, S. V., Abdulazeez, A. M., Zeebare, D. Q., & Ahmed, F. Y. (2021, June). Intrusion Detection Systems Based on Machine Learning Algorithms. In *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)* (pp. 282-287). IEEE.
[4] Ávila, R., Khoury, R., Khoury, R., & Petrillo, F. (2021). Use of security logs for data leak detection: a systematic literature review. *Security and communication networks*, *2021*.
[5] Budur, E., Lee, S., & Kong, V. S. (2015). Structural analysis of criminal network and predicting hidden links using machine learning. *arXiv preprint arXiv:1507.05739*.
[6] Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of business research*, *131*, 441-452.
[7] Cisco IOS NetFlow. [Online]. http://www.cisco.com/en/US/products/ps6601/ products-ios-protocol-group-home.html
[8] David, H., & Suruliandi, A. (2017). Survey on crime analysis and on crime analysis and prediction using data mining techniques. *ICTACT journal on soft computing*, *7*(3).
[9] De Bruin, J. S., Cocx, T. K., Kosters, W. A., Laros, J. F., & Kok, J. N. (2006). Onto clustering criminal careers. In *Proceedings of the ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges* (pp. 92-95).
[10] Fatih, T., & Bekir, C. (2015). Police use of technology to fight against crime. *European scientific journal*, *11*(10).
[11] Ghankutkar, S., Sarkar, N., Gajbhiye, P., Yadav, S., Kalbande, D., & Bakereywala, N. (2019, December). Modelling machine learning for analysing crime news. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)* (pp. 1-5). IEEE
[12] Gomez, J. E., Johnson, B. A., Selva, M., & Sallis, J. F. (2004). Violent crime and outdoor physical activity among inner-city youth. *Preventive medicine*, *39*(5), 876-881.
[13] Gupta, M., Chandra, B., & Gupta, M. P. (2014). A framework of intelligent decision support system for Indian police. *Journal of Enterprise Information Management*. (3)
[14] Ivan, N., Ahishakiye, E., Omulo, E. O., & Taremwa, D. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm.
[15] Izzidien, A., Watson, J., Loe, B., Romero, P., Fitz, S., & Stillwell, D. (2021). The Golden Rule as a Heuristic to Measure the Fairness of Texts Using Machine Learning. *arXiv preprint arXiv:2111.00107*.
[16] Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018, November). Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 415-420). IEEE..
[17] Kumar, Y., Kaur, K., & Singh, G. (2020). Machine Learning Aspects and its Applications Towards Different Research Areas. *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 150–156.
[18] Mathew, J. A., & Asha, K. (2022). A Prognostic Approach to Crime Analysis. In *Advances in Machine Learning for Big Data Analysis* (pp. 71-99). Springer, Singapore.
[19] Middleton, S. E. (2021). Use of Artificial Intelligence to Support Cybercrime Research. In *Researching Cybercrimes* (pp. 213-232). Palgrave Macmillan, Cham. (1)
[20] More, R., Awati, C., Shirgave, S., Deshmukh, R., & Patil, S. (2021). Credit Card Fraud Detection Using Supervised Learning Approach. *International Journal of Scientific & Technology Research*, *9*, 216-219.
[21] Nakib, M., Khan, R. T., Hasan, M. S., & Uddin, J. (2018, February). Crime scene prediction by detecting threatening objects using convolutional neural network. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
[22] Prabakaran, S., & Mitra, S. (2018, April). Survey of analysis of crime detection techniques using data mining and machine learning. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012046). IOP Publishing.
[23] Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y., & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *7*(4), e1208. (4)
[24] Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R. N., & Ghosh, A. (2021). Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm. In *Innovations in Electrical and Electronic Engineering* (pp. 671-681). Springer, Singapore.
[25] Sachin Bhardwaj Yogesh Kumar, M. K. P. K. R. (2020). Recent Trends of Data Mining in the Field of Intrusion Detection System. *Journal of Critical Reviews*, *7*, 2360–2365.
[26] Sandesh, A., Asha, H. V., & Supriya, P. (2022). Detection of Cyberbullying on Twitter Data Using Machine Learning. In *Emerging Research in Computing, Information, Communication and Applications* (pp. 703-713). Springer, Singapore.
[27] Shabbir, S. A., & Kannadasan, R. (2013). An Effective Fraud Detection System Using Mining Technique.
[28] Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 1-14.
[29] Tamilarasi, P., & Rani, R. U. (2020, March). Diagnosis of crime rate against women using k-fold cross validation through machine learning. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1034-1038). IEEE.
[30] Walteros-Alcázar, M. A., Aguirre-Yacup, N., Castillo-Landínez, S. P., & Caicedo-Rodríguez, P. E. (2021). General crime from the data mining point of view. A systematic literature review. *International Journal of Business Intelligence and Data Mining*, *19*(3), 371-393.
[31] Zhao, X., & Tang, J. (2017, November). Exploring transfer learning for crime prediction. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 1158-1159). IEEE.
[32] https://ncrb.gov.in/hi/crime-in-india-table-contents?field_date_value%5Bvalue%5D%5Byear%5D=&field_select_additional_table_ti_value=19&items_per_page=All (ncrb Official Government Website)
[33] https://www.kaggle.com/dhruvanurag20/cyber-crime-in-indian-cities-2018-2019 (Serious Crime of year 2018 and 2019)
[34] https://www.kaggle.com/shishir349/indian-crime-analysis?select=murder.csv (2013 Datasets for murders and abduction
[35] https://www.kaggle.com/tanushagupta/crime-rate-in-india (Additional DataBase for Year 2019 2020 and 2021 = This database results will be Considered as the Plus point for Publishing Research Papers)
[36] https://www.kaggle.com/webaccess/india-crimes-data?select=kidnapping_2016.csv (Crimes in india for year 2016)
[37] https://www.nextias.com/current-affairs/17-09-2021/crime-in-india-2020-ncrb
[38] Akerkar, R. A.; Lingras, P. (2008). *An Intelligent Web: Theory and Practice*, 1st edn. Johns and Bartlett, Boston.
[39] Albert, R.; Jeong, H.; Barab´asi, A.-L. (1999): Diameter of the world-wide Web. Nature, **401**, pp. 130–131.

[40] Berry M. W., Dumais S. T., O'Brien G. W. (1995): Using linear algebra for intelligent information retrieval, SIAM Review, **37**, pp. 573-595.
[41] Bharat, K.; Broder, A. (1998): A technique for measuring the relative size and overlap of public Web search engines. Computer Networks, **30**(1–7), pp. 107–117.
[42] Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. (2000): Graph structure in the Web. Computer Networks, **33**(1–6), pp. 309–320.
[43] Chakrabarti, S. (2000): Data mining for hypertext: A tutorial survey. SIGKDD explorations, **1**(2), pp. 1–11.

Conflicts of Interest: - The authors have no conflicts of interest to declare

Authors Profile

Geetika Bhardwaj, a research scholar pursuing Ph.D from Punjabi University, Patiala .Her areas of specialization are Machine learning, Data mining .She has 7 years of academic experience, and published few papers in conferences and esteemed journals.

Dr. Rajesh K. Bawa has done PhD in the area of Numerical Computing from IIT Kanpur. His areas of specializations are Scientific Computing, Computer Graphics, and Digital Image Processing. Presently he is working as a Dean in Faculty of Computing Sciences and Professor in Department of Computer Science at Punjabi University, Patiala. He has 28+ years of academic experience, and more than 70 publications in major International Journals and Conference Proceedings.