Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 8, July 2021: 2415 – 2435

Analysing Crimes of Indian Datasets Based on Machine Learning Methods

Jessica Sarah[a], Amisha Michelle Danny[b], Juan Mark Deen[c], Lovesh Dongre[d], Chitransh S. V.[e], Harshita Ramchandani[f]

a UG scholar, Department of computer science and Engineering ,Vellore Institute of Technology, Kotri Kalan, Ashta, Near, Indore Road, Bhopal, Madhya Pradesh 466114, India, Email: jessica.sarah2020@vitbhopal.ac.in[a]
b UG scholar, Department of Computer Science and Engineering Kalinga Institute of Industrial Technology, KIIT Road, Patia, Bhubaneswar, Odisha 751024,India, Email: amisha.danny@gmail.com[b]
c UG scholar, Department of computer science and Engineering and Bioinformatics, Vellore Institute of Technology, Vellore Campus, Tiruvalam Rd, Katpadi, Vellore, Tamil Nadu 632014, India, Email:juanmark0521@gmail.com[c]
[d,e, f] UG scholars, Department of Computer Science and Engineering ,University Institute of Technology, RGPV ,Bhopal 462033,India, Email: loveshdongre57@gmail.com[d],chitrakarma@gmail.com[e],harshitaramchandani123@gmail.com[f]

Abstract: Present scenario problems were facing the crime rate increases day by day, due to crimes being a typical social issue influencing personal satisfaction and the economic development of the general public. It is a fundamental factor that decides if individuals move to another city and what spots should be kept away from when they travel. With the expansion of violations, law requirement organizations are proceeding to request progressed geographic data frameworks and new information mining to deal with enhanced crime examination and better secure their groups. Even though violations could happen all over, offenders must deal with wrongdoing openings they look in most commonplace territories for them. Therefore, it becomes vital to study the factors that impact the crime rate. To be better prepared to respond to criminal activity, it is important to understand the patterns in crime; this study analyzes crime data from all over the nation, available publicly on websites such as data.gov.in or Kaggle. This study investigates the relationship between various factors and the crime rate in India and focuses on the extent of effects of various factors on crime registered under IPC or SLL in all Indian states and the union territories. The study covers the data of all the Indian states and the union territories of periods ranging from 2001 to 2014. However, different classes of crimes have a slightly different range of years. The data for district-wise crimes in India is from 2001 to 2014, the data for crimes against SC, ST & Children is from 2001 to 2012, and the data for murder victims by age and gender is from 2001 to 2010. The findings show that these factors are crucial determinants of the rate of criminal cases registered in India. We unequivocally believe that discovering connections between wrongdoing components could profoundly help foresee the potentially risky hotspots later on.

Keywords: Crime datasets, Machine Learning methods, KNN, Decision Tree, Random Forest

1. Introduction

Machine learning techniques are one of the numerous technologies used for crime detection and prevention. They provide light on the common goal of minimizing and eliminating this harmful activity that affects a human civilization. Authors [1] studies New York City crime data and give a comparative examination of them. they employed the machine learning based classification models for this supervised classification problem. Like Decision Tree, Multivariate Linear Regression, and kNN they used. The methodologies used by them shows based on the historical crime data trends[1]. Crimes are a typical social issue influencing the personal satisfaction and the monetary development of the general public. It is a fundamental factor that decides if individuals move to another city and what spots should be kept away from when they travel. With the expansion of violations, law requirement organizations are proceeding to request progressed geographic data frameworks and new information mining ways to deal with enhanced crime examination and better secure their groups. Machine learning is increasingly being used by researchers in the social and computer sciences to analyze and answer global development concerns. Data analytics is having an ever-increasing impact on addressing a wide range of societal issues. In article [4] focuses at how data from various disparate web sources can be used to get insights and generate predictions regarding the spatial distribution of crime in big urban areas. Following a purely data-driven strategy and methodology, several crucial research problems are addressed. First, they look at how different relevant forms of data are for predicting crime levels, with a particular focus on how prediction accuracy may be enhanced by combining data from multiple sources. Their aim, to look at prediction accuracy across all individuals and at prediction accuracy across all groups[4]. Crime has been an evident technique to disturb individuals and society in these recent times. An increase in crime causes an imbalance in a country's constituency. The crime patterns must be understood in order to analyze and answer this type of criminal activity. Authors [5] are investigating such crimes by employing data acquired from open-source Kaggle utilized in turn to forecast recent crimes. The key feature of this project is to estimate which crime types contribute most to the timeframe and location. Researchers[6] examines the prediction of crime based on machine learning. Due to two alternative data processing methodologies, crime statistics from Vancouver have been evaluated for the last 15 years. An accurate and effective analysis of the expanding volume of crime data poses a severe problem for all law enforcement and intelligence institutions. It can also be challenging to detect cybercrime since heavy network traffic and frequent online transactions generate large amounts of data, with little connection with unlawful operations. Database extraction is an excellent method to swiftly and efficiently investigate enormous databases for criminals lacking considerable experience as data analysts. Authors[9] provides a framework of criminal data mining based on experience obtained with the University of Arizona research project Coplink. In order to this many articles published based on investigate to the crimes via machine learning prediction model. Machine learning can play a significant role in revealing Bangladesh's criminal tendencies and patterns. Various machine learning regression models are utilized here for the prediction of crimes trends and patterns in Bangladesh, such as linear regression, polynomial regression, and random forest [11]. It is critical to conduct thorough and accurate investigations into crime situations. Computer programs can significantly assist law enforcement officials with the increased use of computerized systems to track crime and violence. Article [12] used machine learning techniques on a datasets to predict some parameters such as criminal age, gender, race, crime manner, etc. Their study showed that four distinct algorithms, K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest Classifier (RFC), and Decision Tree Classifier (DTC). Article[13] work based on Bangla Crime Type Classification. They have analyzed word vectors like a bag of words, TF-IDF in state-of-art machine learning algorithms, and most good semantic and syntactic word embeddings like Word2Vec, GloVe, fast-Text in both shallow and deep CNN, and RNN to select the best word embeddings their classifier module. In recent years, there has been a considerable increase in the number of crimes committed against women. Every year, a massive amount of data is collected based on crime reporting. This data can be highly useful for assessing and forecasting crime and helping us stop it to some extent. Data analysis is the act of reviewing, cleansing, transforming, and modelling data to provide meaningful information, report conclusions, and support decision-making. One of the essential strategies for standardizing the independent characteristics and placing the data in a set range is feature scaling. It is carried out during data pre-processing. K-fold cross-validation is a re-sampling approach that is used to calculate machine learning models on a limited sample of data. It is a popular strategy because it is straightforward to grasp and produces a model deftness calculation that is less biased or negative than other approaches, such as a simple train or test divide. Machine learning is vital in the data processing. Article [14] discussed various machine learning algorithms, including KNN and decision trees, Nave Bayes and Linear Regression, CART (Classification and Regression Tree), and SVM. In Article[15], historical data on public property crime are used as research data for the estimation of predictive power amongst several machine learning algorithms from 2015 to 2018 from a section of a big coastal town in southeast China. Results based solely on historical crime data reveal that KNN, the random forests, vector support, the Naive Bays, and convolution neural networks were over performed by the LSTM model. In the most recent couple of decades, the innovation made spatial information-digging a handy answer for wide gatherings of people of law implementation authorities that is moderate and accessible. This immense information is utilized as a record for making a criminal record database. The article[3] investigates the developing world's emerging field of machine learning (ML4D). They provided an overview of relevant literature. Following that, for best practices gleaned from the literature for ensuring that ML4D projects contribute to the achievement of development goals. Even though violations could happen all over, it is basic that offenders deal with wrongdoing openings they look in most commonplace territories for them. By providing an information mining way to decide the most criminal hotspots and discover the sort, area, and time of perpetrated crimes, this paper try to improve individuals' mindfulness regarding the hazardous areas in specific eras. Hence, our proposed model can enable crime rate in Indian datasets using machine learning methods.

2.Significance Of The Study

The fundamental difficulties we are confronting are:

- Increase in crime data that must be put away and dissected.
- Analysis of information is troublesome since information is inadequate and conflicting.
- Limitation in getting crime information records from the Law Enforcement division.
- The accuracy of the program relies upon the exactness of the preparation set.

These are the steps involved in Crime Analysis:

- Data Collection
- Visualization
- Data Pre-processing
- Feature Extraction
- Classification
- Prediction

Rationale

“Madhya Pradesh's commercial capital Indore has topped the crime record in the country in 2008 followed by Bhopal and Jaipur. Crime rate of Indore was 941.4, which is the highest in the country, according to National Crime Record Bureau's (NCRB) report - "Crime in India 2008". Bhopal followed with crime rate of 791.4 in the same year as per the report.[18]”

With the rapid urbanization and development of big cities and towns, the graph of crimes is also increasing. This phenomenal rise in offenses and crime in cities is a matter of great concern and alarm to all of us. There are robberies, murders, rapes, and whatnot. The frequent and repeated thefts, burglaries, robberies, murders, killings, rapes, shoplifting, pick-pocketing, drug- abuse, illegal trafficking, smuggling, theft of vehicles, etc., have made the everyday citizens have sleepless nights and restless days. They feel very insecure and vulnerable in the presence of anti-social and evil elements. The criminals have been operating organized and sometimes even have nationwide and international connections and links.

3.Review Of Related Studies

Crime is a social nuisance that has been on the rise in almost all parts of the world[16]. Including India, therefore, it becomes vital to study the factors that impact the crime rate. This study aims to investigate the relationship between various factors and the crime rate in India. The study focuses on the extent of various factors on crime registered under IPC in all Indian states and major union territories. Criminologists analyze the data with varying degrees of success. However, with the increasing crime rate, human skills fail when provided with massive data sets. Application of data analysis techniques can be used to facilitate the task that can extract the hidden knowledge from the massive data sets and provide the crime investigation department a new edge for crime analysis. [11] Collecting crime information from government portals employs data analysis techniques to predict or avoid future crime trends. Past crime records accumulated from government portals constitute the crime type, time, location, information about the victims, genders, ages, social status, and many more. Thus, crime prediction, a subtask of crime analysis, consider all the past criminal records, classifies the crime categories, and predicts the future crime. Numerous research works exist in the literature that employs different data mining techniques for crime analysis of different countries and cities[12]. Analysis of crime has been done by mapping. And similarities have been found with the past records of crime trends compared to the present scenario. This task was an approach to determine places where maximum numbers of crime incidents take place. The application of data analysis techniques has proven to be crucial for crime detection and prevention tasks. A comparative study was conducted for different crime patterns that exhibited better linear regression than other classification methods. An architecture was implemented to collect the raw data and categorize the data into crime types, locations, and places. Then, existing classification algorithms were used, and the most effective technique was chosen, resulting in crime prediction. The present work demonstrates crime prediction for 28 states (Andhra Pradesh inclusive of Telangana) and 7 union territories of India[18]. It has considered collecting crime records from 2001–2014 containing information about different types of crime like juvenile offense, murder, etc. Classification techniques involve the assignment of any object to one of the multiple predefined classes. Here, the predictive modelling is separately used for each crime type for all the states. Three different classification techniques decision tree, K-Nearest Neighbour and random forest have been used in this work.

4.Objectives Of The Study

The objective of this study work is to:
1.Understanding the crime pattern
2.Analysis of crime in India
3.Understanding the accuracy of various predictive machine learning algorithms.

Goal

Study of the this work focuses on three directions

- Understanding patterns of criminal behaviour that could help in solving criminal investigations
- Visualizing and analyzing the data about various crimes throughout the nation
- Comparison of performance of the following predictive machine learning algorithms
  - KNN (K-Nearest Neighbours)
  - Decision tree
  - Random forest

5. **Problem Description**

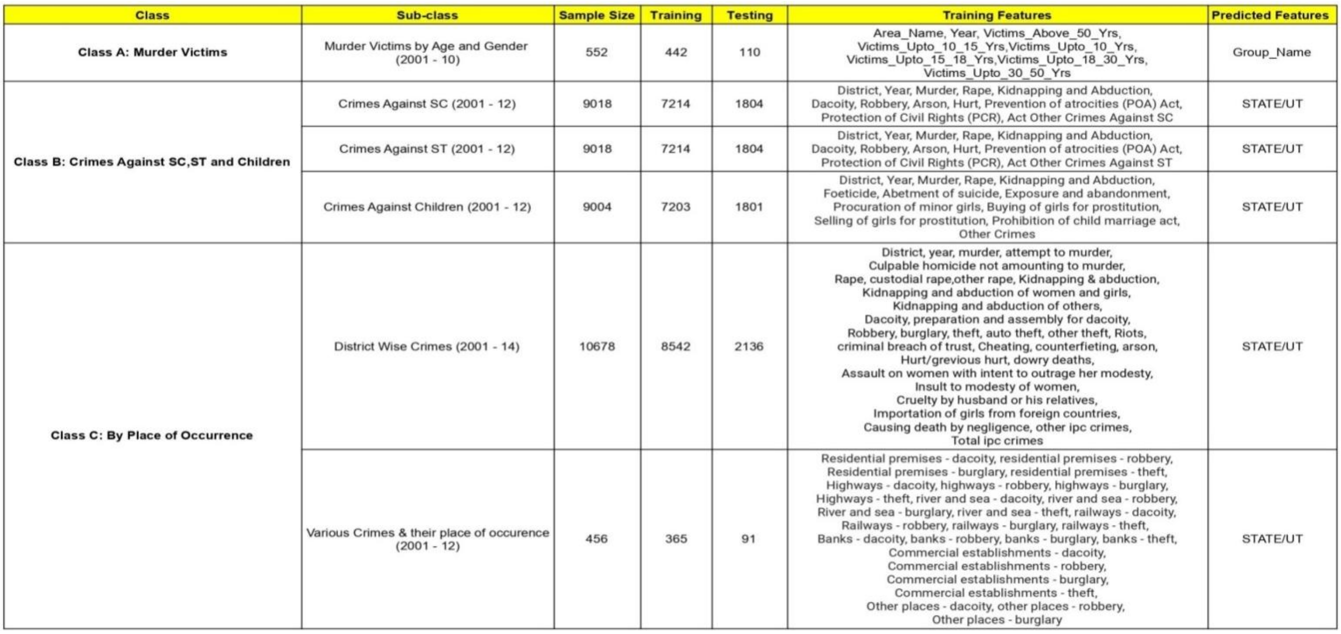
This study investigation plans based to discover various factors affecting crimes in India, utilizing an arrangement of valid datasets of crimes. We will attempt to find them in all probability wrongdoing areas. Also, we will foresee what sort of wrongdoing may happen next in a particular area. At last, we expect to give an insight into the performance of various algorithms in predicting various factors related to crimes.

6. Population And Sample

6.1 Dataset

Datasets that we are using are downloaded from www. data.gov.in and www.Kaggle.com[17], but these data samples are not used directly as input to the machine learning algorithms. Table 1 shown the snapshot of one of the files present in our datasets.

Table 1.Dataset Classes and their description



Class	Sub-class	Sample Size	Training	Testing	Training Features	Predicted Features
Class A: Murder Victims	Murder Victims by Age and Gender (2001 - 10)	552	442	110	Area_Name, Year, Victims_Above_50_Yrs, Victims_Upto_10_15_Yrs,Victims_Upto_10_Yrs, Victims_Upto_15_18_Yrs,Victims_Upto_18_30_Yrs, Victims_Upto_30_50_Yrs	Group_Name
Class B: Crimes Against SC,ST and Children	Crimes Against SC (2001 - 12)	9018	7214	1804	District, Year, Murder, Rape, Kidnapping and Abduction, Dacoity, Robbery, Arson, Hurt, Prevention of atrocities (POA) Act, Protection of Civil Rights (PCR), Act Other Crimes Against SC	STATE/UT
	Crimes Against ST (2001 - 12)	9018	7214	1804	District, Year, Murder, Rape, Kidnapping and Abduction, Dacoity, Robbery, Arson, Hurt, Prevention of atrocities (POA) Act, Protection of Civil Rights (PCR), Act Other Crimes Against ST	STATE/UT
	Crimes Against Children (2001 - 12)	9004	7203	1801	District, Year, Murder, Rape, Kidnapping and Abduction, Foeticide, Abetment of suicide, Exposure and abandonment, Procuration of minor girls, Buying of girls for prostitution, Selling of girls for prostitution, Prohibition of child marriage act, Other Crimes	STATE/UT
Class C: By Place of Occurrence	District Wise Crimes (2001 - 14)	10678	8542	2136	District, year, murder, attempt to murder, Culpable homicide not amounting to murder, Rape, custodial rape,other rape, Kidnapping & abduction, Kidnapping and abduction of women and girls, Kidnapping and abduction of others, Dacoity, preparation and assembly for dacoity, Robbery, burglary, theft, auto theft, other theft, Riots, criminal breach of trust, Cheating, counterfieting, arson, Hurt/grevious hurt, dowry deaths, Assault on women with intent to outrage her modesty, Insult to modesty of women, Cruelty by husband or his relatives, Importation of girls from foreign countries, Causing death by negligence, other ipc crimes, Total ipc crimes	STATE/UT
	Various Crimes & their place of occurence (2001 - 12)	456	365	91	Residential premises - dacoity, residential premises - robbery, Residential premises - burglary, residential premises - theft, Highways - dacoity, highways - robbery, highways - burglary, Highways - theft, river and sea - dacoity, river and sea - robbery, River and sea - burglary, river and sea - theft, railways - dacoity, Railways - robbery, railways - burglary, railways - theft, Banks - dacoity, banks - robbery, banks - burglary, banks - theft, Commercial establishments - dacoity, Commercial establishments - robbery, Commercial establishments - burglary, Commercial establishments - theft, Other places - dacoity, other places - robbery, Other places - burglary	STATE/UT

1. Region of occurrence of crimes
2. State/UTs where the crime had occurred

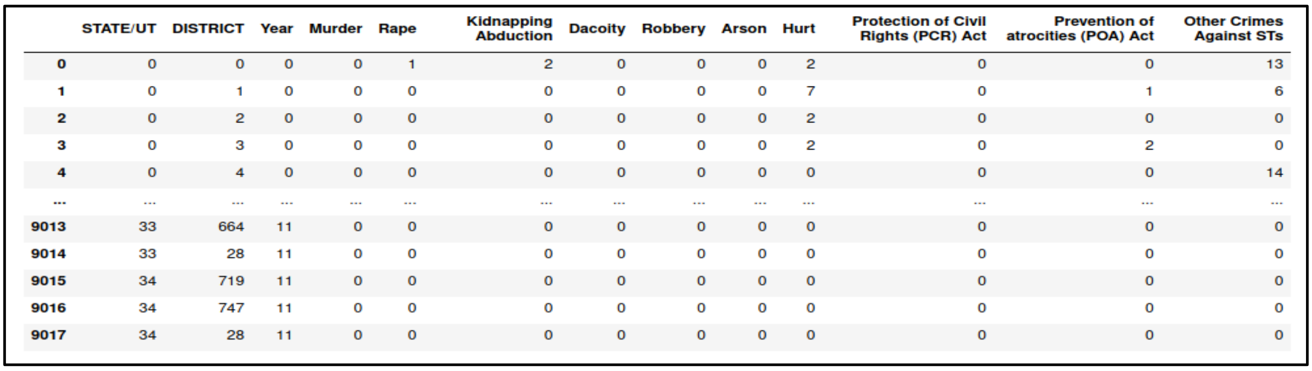
3. Type of crime committed
4. Age group of murder victims

6.2 Data Preprocessing

Data preprocessing is an essential step in any machine learning method, and it plays a significant role in classification to reach maximum correct prediction. In order to these this study went through a series of pre-processing steps to convert crime datasets into executable form.

- We encoded the categorical features such as Region or State/UTs where the crime has occurred during the implementation.
- Encoding categorical data is important because the machine learning models completely works on mathematics and numbers, but if our dataset would have a categorical variable in it, then it may create trouble while building the model. So, it is necessary to encode these categorical variables into numbers .Table 2 shows the files after pre-processing is done{ SC stand for Scheduled castes, which are one of the lower communities in India. The Indian constitution, in Constitution (Scheduled Castes) Order, 1950, lists 1,108 castes across different states}.

Table 2 State-wise crimes against SC over the period of 2001 – 12



	STATE/UT	DISTRICT	Year	Murder	Rape	Kidnapping Abduction	Dacoity	Robbery	Arson	Hurt	Protection of Civil Rights (PCR) Act	Prevention of atrocities (POA) Act	Other Crimes Against STs
0	0	0	0	0	1	2	0	0	0	2	0	0	13
1	0	1	0	0	0	0	0	0	0	7	0	1	6
2	0	2	0	0	0	0	0	0	0	2	0	0	0
3	0	3	0	0	0	0	0	0	0	2	0	2	0
4	0	4	0	0	0	0	0	0	0	0	0	0	14
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9013	33	664	11	0	0	0	0	0	0	0	0	0	0
9014	33	28	11	0	0	0	0	0	0	0	0	0	0
9015	34	719	11	0	0	0	0	0	0	0	0	0	0
9016	34	747	11	0	0	0	0	0	0	0	0	0	0
9017	34	28	11	0	0	0	0	0	0	0	0	0	0

Here, 28 States and 7 Union Territories have been taken into account as the data is from 2001 to 2012 that was before the creation of Telengana as a new State.

7. Proposed Methodology

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades, it has become a standard tool in almost any task that requires information extraction from large data sets. Here proposed machine learning model trained datasets to learn particular feature sets as shown in fig. 2.

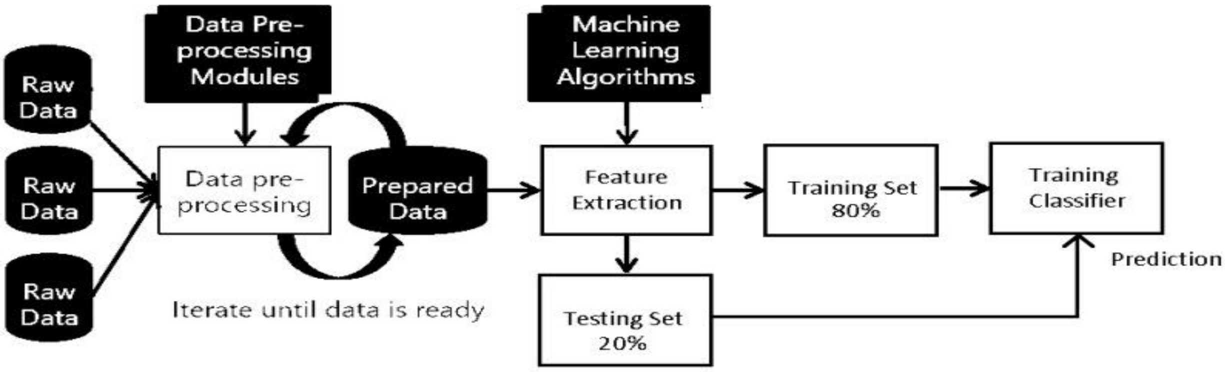


Fig. 2 Proposed Machine Learning Process

7.1 Feature sets : The Following inputs class features are train or learn by proposed algorithms on the basis of year, place, category of crime.

1. Class A (Murder Victims, IPC Section 302)
2. Class B (Crimes Against SC, ST and Children, IPC Sections 376 & 317)
3. Class C (Crimes by Place of Occurrence)

The output is State/UTs (predicted feature) where the class of crime is likely to occur. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbours), Decision Tree, and Random Forest. [10]

7.2 Proposed Algorithm

For the purpose of proper implementation and functioning several Algorithms and techniques were used. Following are the algorithms used:

7.2.1 KNN (K-Nearest Neighbours)

A powerful classification algorithm used in pattern recognition K nearest neighbours stores all available cases and classifies new cases based on a similarity measure (example: distance function).One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method)[21].

KNN: Classification Approach

- An object (a new instance) is classified by a majority votes for its neighbour classes as shown in fig. 3.
- The object is assigned to the most common class amongst its K nearest neighbours(measured by distance function).

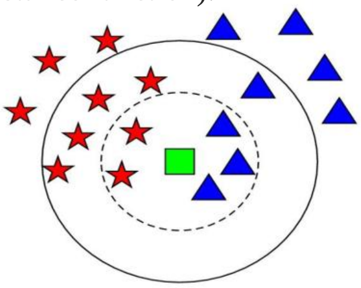


Fig. 3 Principal Diagram of KNN

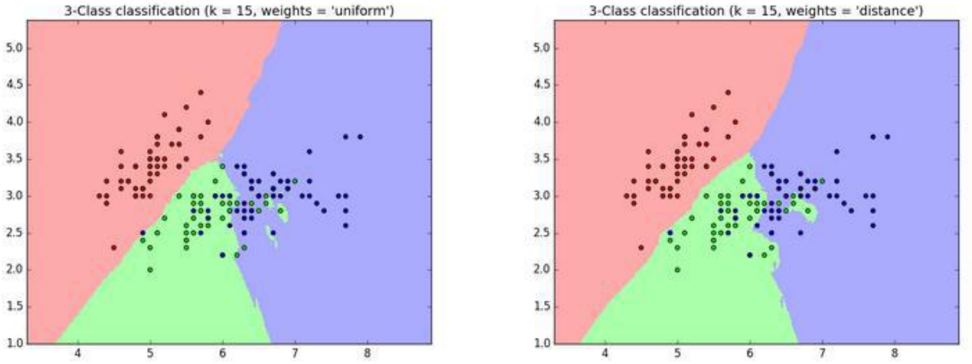


Fig. 4 Shows graphical representation of KNN

In this study have used the default distance function provided in the sklearn library which is the Euclidean Distance and following is the formula eq.(1) for the same and fig.4 represent the graphical representation of KNN[22].

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad \textbf{eq.(1)}$$

7.2.2 Decision Tree

As the name says all about it, it is a tree which helps us by assisting us in decision-making. Used for both classification and regression, it is a very basic and important predictive learning algorithm[23].

- It is different from others because it works intuitively i.e., taking decisions one-by-one.
- Non-parametric: Fast and efficient.

It consists of nodes which have parent-child relationships as shown in fig.5

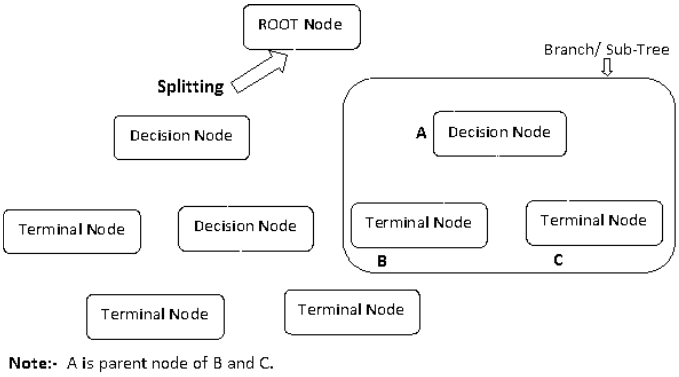


Fig. 5 Data Splitting into in Decision Tree

Decision tree considers the most important variable using some fancy criterion and splits dataset based on it. It is done to reach a stage where we have **homogenous subsets** that are giving predictions with utmost surety. We have used the default type of Decision Tree (entropy) which is inbuilt in the sklearn library.

7.2.3 Random Forest

Random Forests is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data. Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making split decision to help avoid overfitting on the training data. A random forests classifier is an ensemble classifier, which aggregates a family of classifiers $h(x|\theta_1)$,$h(x|\theta_2)$,…,$h(x|\theta_j)$. Each member of the family, $h(x|\theta)$, is a classification tree and k is the number of trees chosen from a model random vector. Also, each θ_j is a randomly chosen parameter vector. If D(x,y) denotes the training dataset, each classification tree in the ensemble is built using a different subset $D\theta_j(x,y) \subset D(x,y)$ of the training dataset. Thus, $h(x|\theta_j)$ is the $(x)_{kth}$ classification tree which uses a subset of features $x\theta_j \subset x$ to build a classification model. Each tree then works like regular decision trees[24]: it partitions the data based on the value of a particular feature (which is selected randomly from the subset), until the data is fully partitioned, or the maximum allowed depth is reached. The final output y is obtained by eq.(2) aggregation of the results.

$$y = \operatorname{argmax}_{p \in \{h(x_1)..h(x_\kappa)\}} \{ \sum_{j=1}^{\kappa} (I(h(x|\theta_j) = p)) \} \qquad \textbf{eq.(2)}$$

where I denote the indicator function.

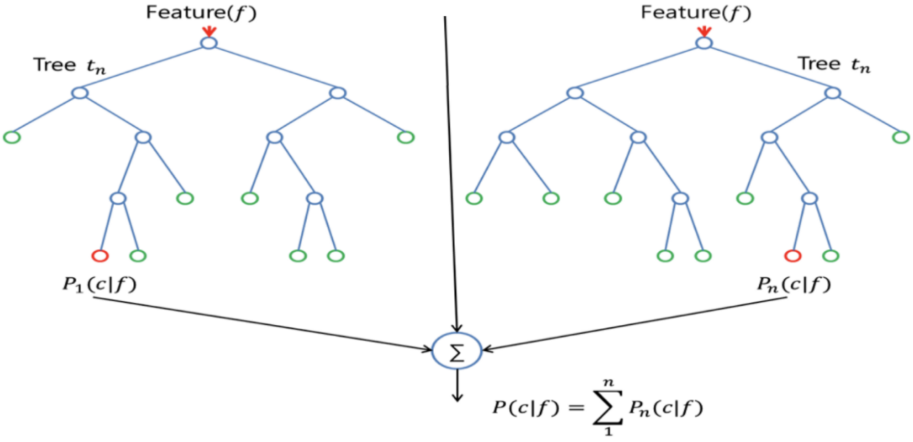


Fig. 6 Tree depth search in random forest for sample

8. Proposed Work flow and Result Outcomes

This investigation focuses on the analysis of crime data of India, primarily for bringing out the factors responsible for various crimes in India and preventive measures that could be taken to prevent or reduce them. According to the problem statement, data collection would be done from various websites such as Kaggle.com, data.gov.in, etc., after collecting data, and it would be cleaned and arranged in our required formats. Analysis of the factors present in the data for feature extraction according to the problem statement. Python programming language used for visualization of results in tabular data, also for proposed model prediction using python modules like matplotlib, NumPy, pandas, etc. This experiment works on an Intel Core2Duo processor with 2GB RAM[19],[20].

Visualization and Analysis

8.1 (a) Class A (Murder Victims, IPC Section 302)

8.1.1 Percentage Breakdown of Male and Female murder victims over the period of 2001 to 2010.The percentage of male murder victims is significantly greater than that of female murder victims, almost three times as many murder victims are males as compared to female murder victims as shown in fig.7.

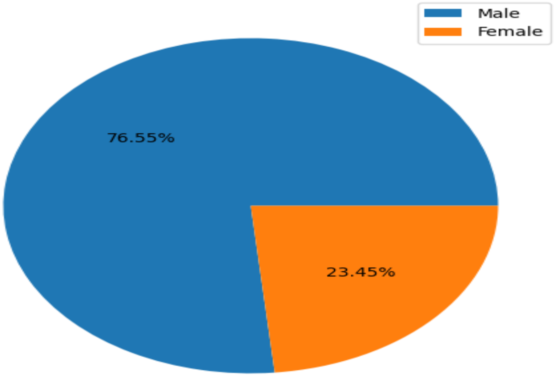


Fig. 7 Percentage breakdown of male and female murder victims over the period of 2001 – 10

The same trend is seen in various age groups of murder victims. The number of male murder victims stands high when observed in different age groups such as less than ten years, between 10 to 15 years, between 15 to 18 years, between 18 to 30 years, between 30 to 50 years, and above 50 years. In this case, the maximum amount of murder victims is found in the age gap of 18 to 30 years in both male and female victims. The second rate of victims found in the age gap of 30 to 50 years, with the lowest rate found in the age group of victims between 10 to 15. Following fig.7 and fig.8 shown a bar graphs for better understanding.

8.1.2 Statistics of Murder Victims by age group over the period of 2001 to 2010

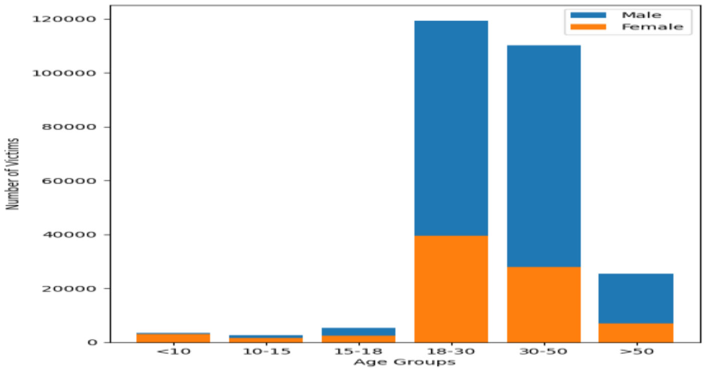


Fig. 8 Murder victims in different age-groups over the period of 2001 – 10

The number of murder victims state-wise we see that Uttar Pradesh tops the list following which are Bihar, Maharashtra, Andhra Pradesh (data is from a period of time when this state wasn’t divided) and Madhya Pradesh. The amount of murder victims from Uttar Pradesh is almost 1.5 times the number in Bihar which is second to UP in this crime, which is shocking. Lowest being in the state of Sikkim (151) and among UTs, Lakshadweep has the lowest reported murder victims (4)[19],[20] as shown in fig.9.

8.1.3 Statistics of Murder Victims in 28 States and 7 Union territories over the period of 2001 to 2010

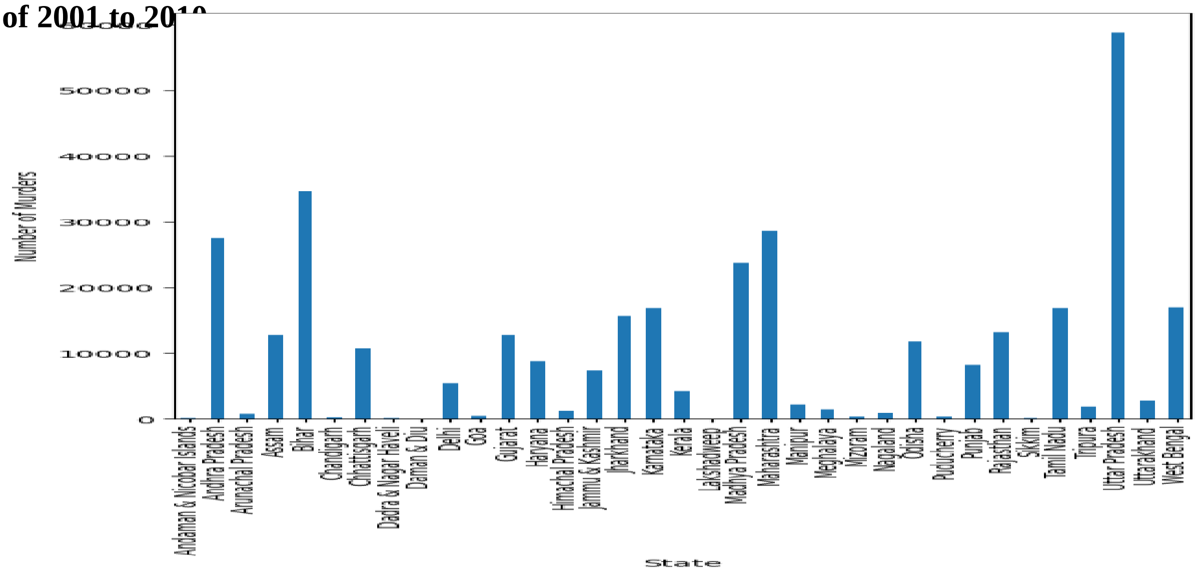


Fig. 9 Murder victims in 28 States and 7 Union Territories of India from 2001 – 10

8.1.4 Trend of Male, Female & Total Murder Victims over the period of 2001 to 2010

Finally, looking at the trend of murders over the period of 2001 – 10 we observe that although there has been an overall decline in the number of male murder victims but in case of female victims the line seems to remain in a steadily but slowly increasing state as shown in fig. 10 and fig.11.

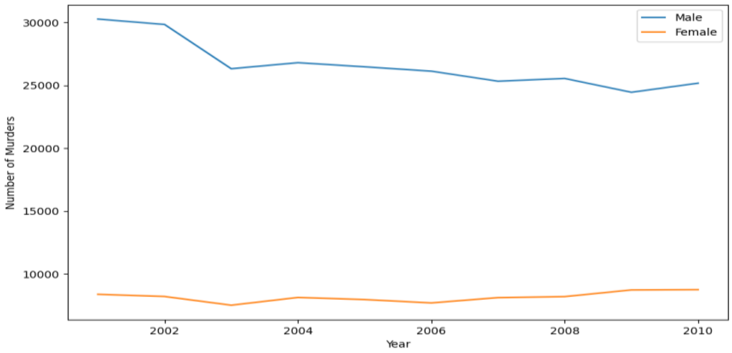


Fig. 10 Trend of gender-wise murder victims from 2001-10

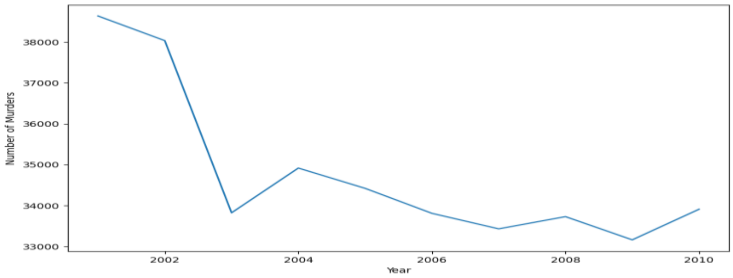


Fig. 11 Trend of total murder victims from 2001-10

8.2 (b) Class B (Crime Against SC, ST & Children, IPC Sections 376 & 317)

8.2.1 Statistics of State-wise crimes committed against SC in India, and in UP over the period of 2001 to 2012

Uttar Pradesh has the maximum count of reported crimes against SC over the period of 2001 – 12, following which is the state of Rajasthan which is surprising as this state is not really known to be a crime spot, following Rajasthan is the notorious state of Madhya Pradesh which has been known for having a high rate of crime in almost every type of crime over the period of 2001 – 12 as shown in fig12 and fig.13.

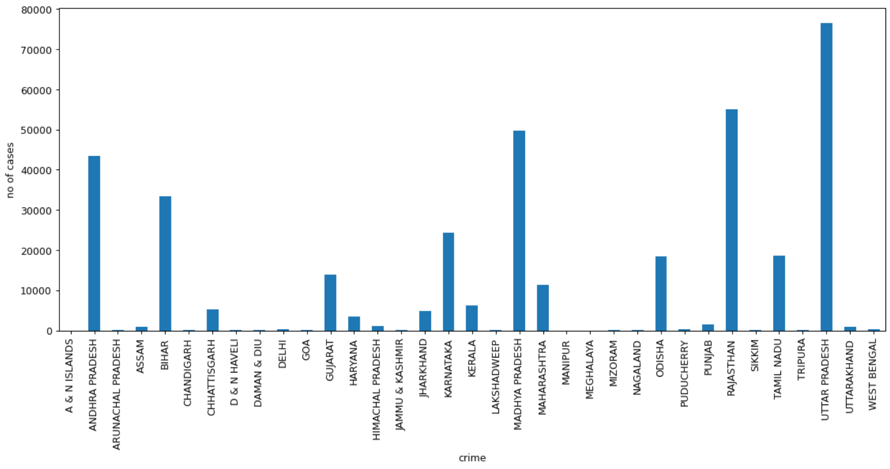


Fig. 12 Statistics of state-wise crime against SC over the period of 2001 – 12

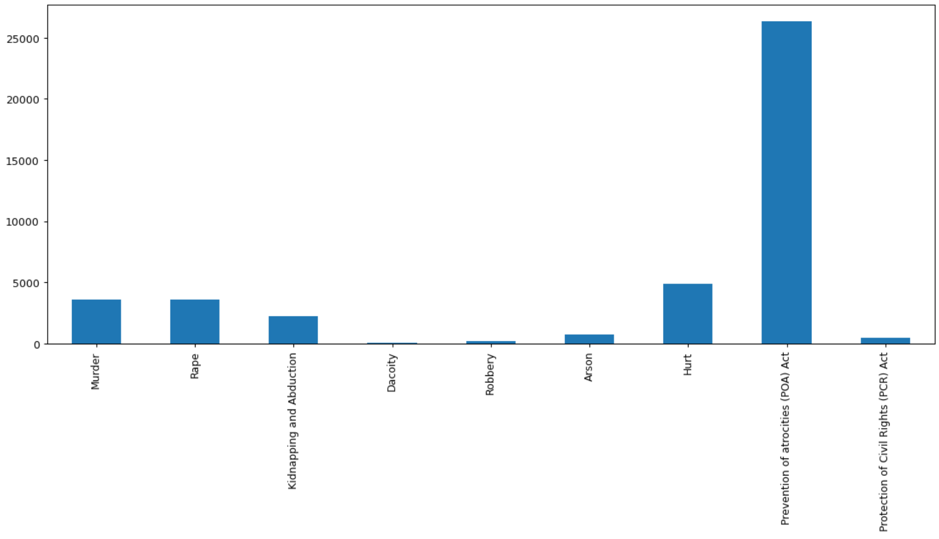


Fig. 13 Crime Against SC in Uttar Pradesh over the period of 2001 – 12

8.2.2 Statistics of State-wise crimes committed against ST over the period of 2001 to 2012

As we can clearly see in the above plot that in the state of UP the POA Act (Prevention of Atrocities, Act)[10] has the highest count of reported cases which is surprisingly higher than the one following it, i.e., Hurt or Grievous Hurt, the graph clearly indicated the amount of hate and the kind of mindset that the people of UP had during the period of 2001 – 12 as shown in fig.14.

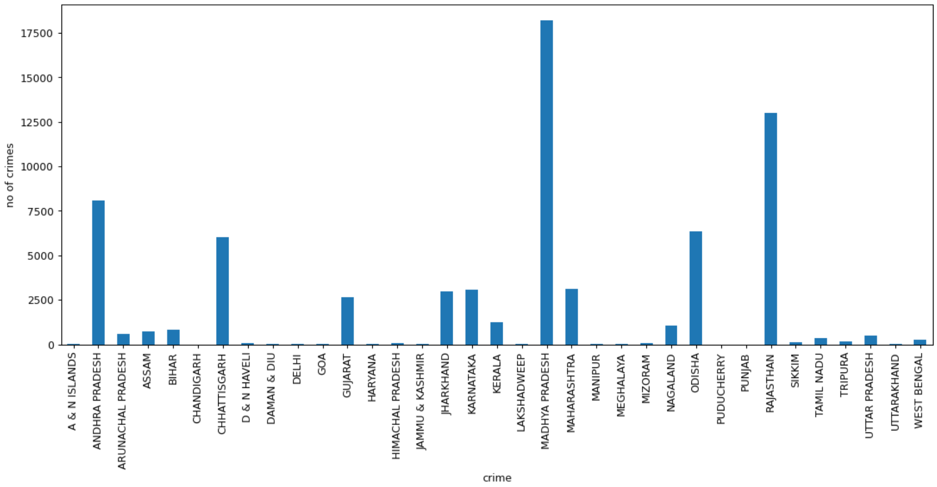


Fig. 14 Statistics of state-wise crime against ST over the period of 2001 – 12

Madhya Pradesh has the maximum count of reported crimes against ST over the period of 2001 – 12, following which is the state of Rajasthan which is surprising as this state is not really known to be a crime spot, following Rajasthan is the state of Andhra Pradesh.

8.2.3 Statistics of crimes committed against ST in Madhya Pradesh over the period of 2001 – 12

Another Sample of the fact stated earlier that the state of Madhya Pradesh has almost always been known for being a notorious crime state with being the highest in the case of crimes such as Rapes and Murders. The following graph adds to the conviction that this state indeed has had the highest number of reported rape cases against ST over the period of 2001 – 12, following fig.15 shows the crime of Hurt or Grievous Hurt against ST(tribal community), following which is the violation of the Prevention of Atrocities (POA), Act, followed by Murder, Kidnapping and Arson[10].

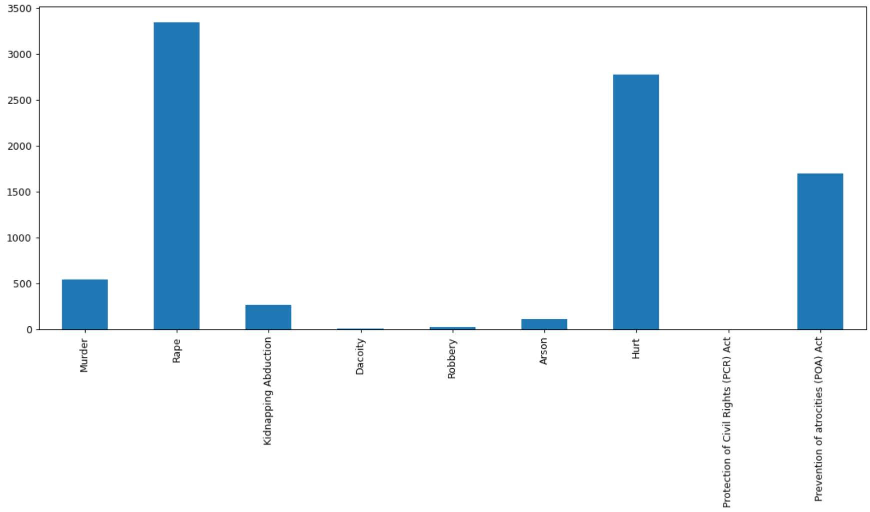


Fig. 15 Crimes Against ST in Madhya Pradesh over the period of 2001 – 12

8.2.4 Statistics of State-wise crimes committed against Children over the period of 2001 – 12

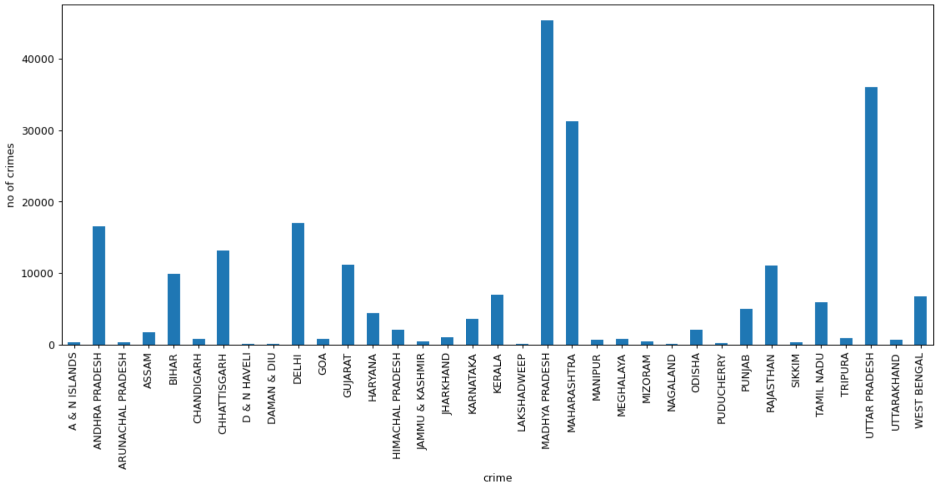


Fig. 16 Statistics of state-wise crime against children

8.2.5 Statistics of crimes committed against Children in Madhya Pradesh over the period of 2001 – 12

As we can observer from the above visual that the state having the highest count of reported crimes against children is the state of Madhya Pradesh, standing for its reputation for being one of the worst states in India to live in, MP has once again proved to be high-crime zone but this time it was for the poor children of this state during the period of 2001 – 12 as shown in fig 17.

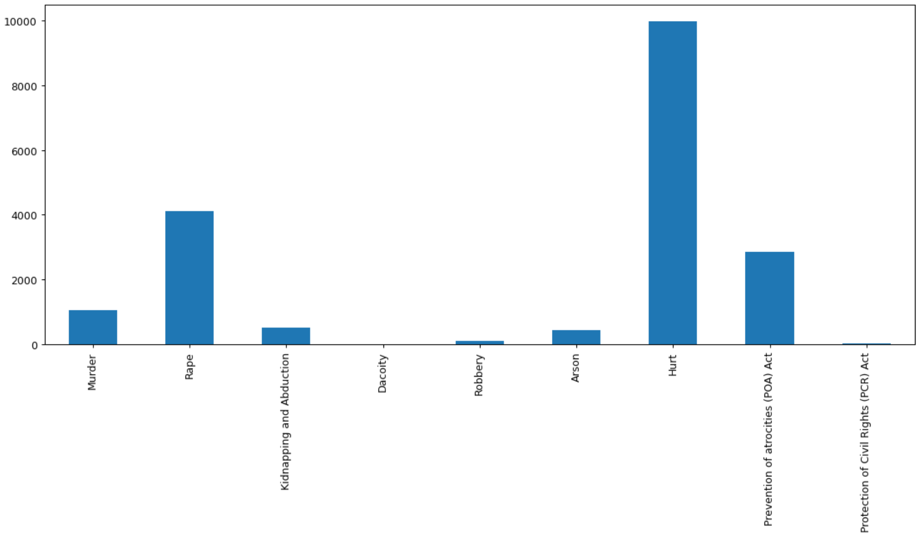


Fig. 17 Crime against children in Madhya Pradesh over the period of 2001 – 12

As we can clearly observe from the above graph that the crime of Hurt or Grievous Hurt against Children is the highest reported crime in the notorious state of Madhya Pradesh following which are Rape and violation of POA act[10].

8.3 Class C (Crime by Place of Occurrence)

8.3.1 Statistics of State-wise total crimes committed over the period of 2001 – 14

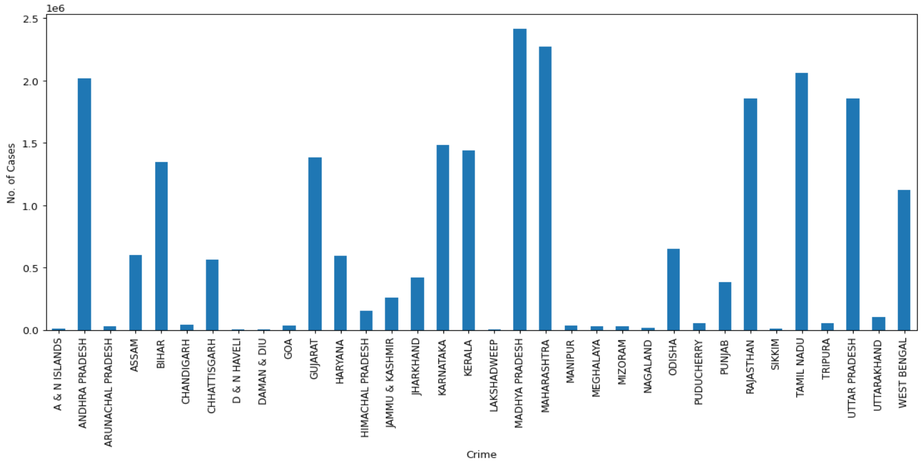


Fig. 18 Statistics of state-wise crime (2001 – 14)

As we observe in the above bar graph that the state of Madhya Pradesh has the highest count of reported crimes over the period of 2001 – 14, second to which is the state of Maharashtra, following which are Andhra Pradesh, Tamil Nadu, Uttar Pradesh and Rajasthan. The lowest being in the state of Sikkim and in the Union Territory of Lakshadweep as shown in fig.18.

8.3.2 Statistics of different crimes committed in Madhya Pradesh over the period of 2001 – 14

In the below Fig. 19 we observe that the crime of Hurt or Greivous Hurt has bagged the highest count of officially reported cases in the state of Madhya Pradesh over the period of 2001 – 14, following which is the crime of theft and other theft committed over the same duration, burglary stands next to these crimes following which is auto theft (theft of vehicles), the least amount of cases reported are for Custodial Rape, Counterfeiting and Import of girls from foreign countries.

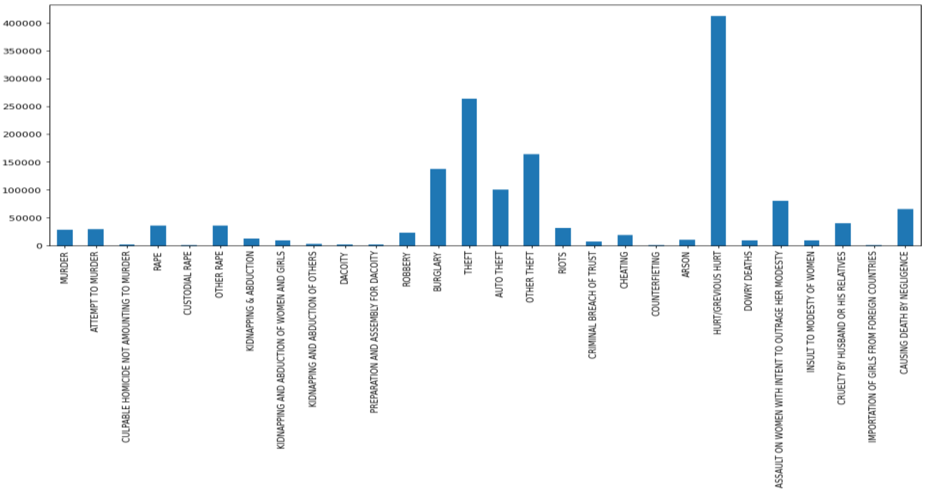


Fig.19 Statistics of various crimes in MP (2001 – 14)

8.3.3 Statistics of occurrence of crimes such as theft, burglary, robbery and dacoity by place over the period of 2001 – 14

Below is the graph plotted for the assessment of the trend of burglary, robbery, theft and dacoity that take place according to various places such as residential, commercial, highways, etc. The fig.20 graph clearly indicates that the maximum theft and burglary incidences are reported from residential premises, also that second to this the next place where these two crimes occur very frequently are the commercial establishments such as Jewellery shops, hardware shops, warehouses, etc.

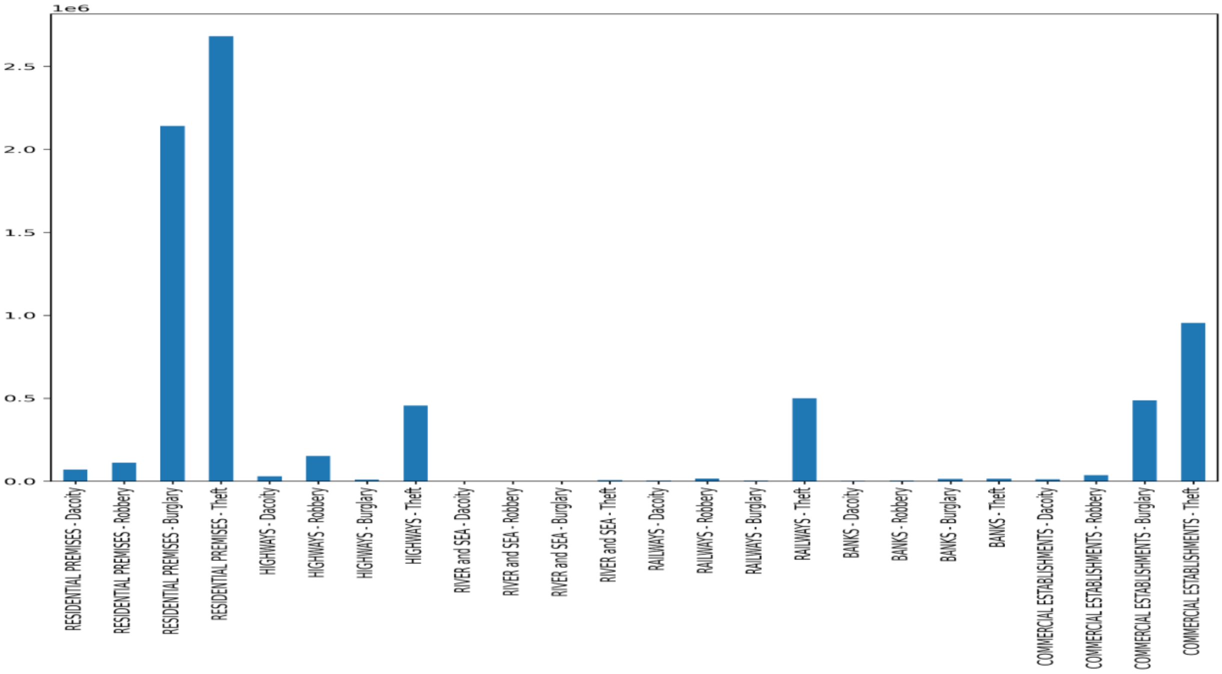


Fig. 20 Stats of Burglary, Theft, Robbery and Dacoity by place of occurrence over the period of 2001 – 12

9 Experimental Results

The performance of the classifiers has been measured by evaluation techniques, namely, precision, recall, F-measure and the accuracy score as evaluated by eq.(3) to (5). Confusion matrix denotes the number of classified and misclassified instances in the crime data. True positive (TP) and True negative (TN) represent correctly classified instances, whereas False positi N) denote the incorrectly classified instances.

$$Precision\,(P) = \frac{TP}{TP + FP}$$

eq.(3)

$$Recall\,(R) = \frac{TP}{TP + FN}$$

eq.(4)

$$F - measure = \frac{2PR}{P + R}$$

eq.(5)

The training data for crime related to murder contain all the instances, and it has been learnt to the predictive model which classifies the gender of the murdered victims. Table 3 to table 8 shows the performance measures for different class of crime cases. Table 3 shows the classification accuracy for predicting the crimes against children over the period of 2001 – 12. decision tree perform well with 95.17% accuracy rate. The following table shows the classification accuracy for predicting various crimes that had occurred by different place of occurrence throughout India.

Table 3 Results of predicting the State or UT for Crimes Committed Against Children (2001 – 12)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	91.49	91.28	91.26	91.28
Decision Tree	95.41	95.17	95.16	95.17
Random Forest	86.78	87.23	86.81	87.23

Table 4 Results for predicting the State or UT for Crime by Place of Occurrence (2001 – 12)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	86.50	80.43	82.00	80.43
Decision Tree	76.87	71.74	70.38	71.74
Random Forest	93.93	90.22	90.47	90.22

Table 5 Results for predicting the Group Name for Age and Gender of Murder Victims (2001 – 12)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	82.90	82.88	82.88	82.88
Decision Tree	81.18	81.08	81.07	81.08
Random Forest	86.77	86.49	86.45	86.49

Table 6 Results for predicting the State or UT for District Wise Total Crimes committed (2001 – 14)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	83.36	84.08	82.57	84.08
Decision Tree	95.9	95.45	95.5	95.45
Random Forest	94.36	94.47	94.01	94.47

Table 7 Results of predicting the State or UT for Crimes Committed Against SC (2001 – 12)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	90.32	90.02	89.93	90.02
Decision Tree	95.17	94.4	94.51	94.4
Random Forest	86.26	86.64	86.17	86.64

Table 8 Results of predicting the State or UT for Crime Committed Against ST (2001 – 12)

Classifier/Metric	Precision	Recall	F-1	Accuracy
KNN	91.49	91.01	91.11	91.01
Decision Tree	95.82	95.23	95.22	95.23
Random Forest	88.62	88.69	88.49	88.69

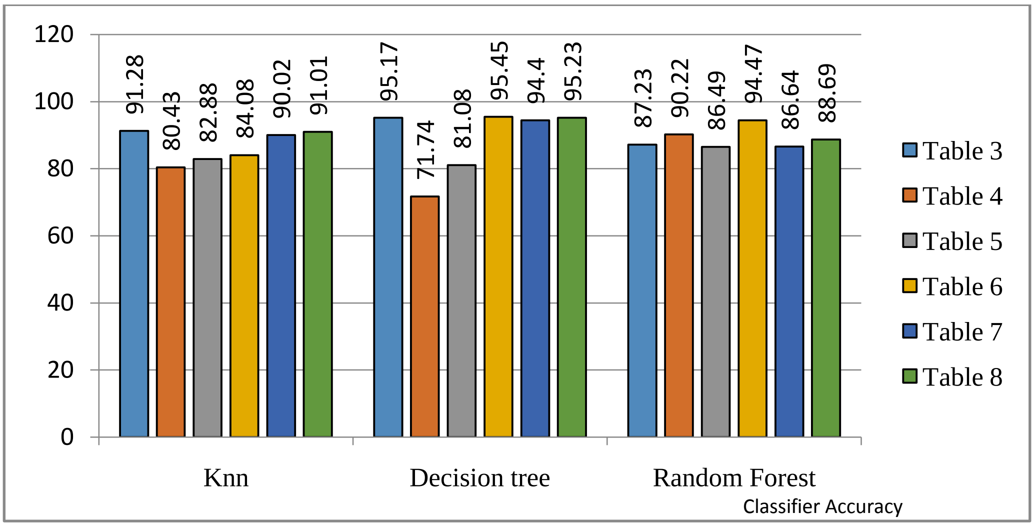


Fig.21 Classifier Accuracy Chart form as result observed in all above Tables content

Random forest performs well in predicting for the State, or UT for Crime by Place of Occurrence (2001 – 12) give 90.22%, as shown in table 4 and indicating for the group name for age and gender of murder victims (2001 – 12) gets 86.49% as shown in table 5. Decision tree performs well in data samples of the State or UT for district Wise total crimes committed (2001 – 14) prediction accuracy results get 95.45% as shown in table 6. And in data samples, the State or UT for crimes committed against SC (2001 – 12) prediction accuracy results get 94.4%, as shown in table 7. And also, in the data samples, the State or UT for crime committed against ST (2001 – 12) decision tree performs well resultant prediction accuracy gets 95.23%, as shown in table 8.Knn perform well and works as moderate in all cases.

10.Conclusion

In crimes against children, the Decision Tree classifier has performed well following the KNN algorithm and the Random Forest algorithm. In case of crimes by place and murder victims of occurrence, the best performing algorithm is the Random Forest algorithm, KNN follows it, and Decision Tree. In the case of district-wise crimes committed over the period of 2001 – 14, Decision Tree performed the best, following it is Random Forest and KNN. In the case of crimes committed against SC and ST, the Decision Tree algorithm performed the best following: the KNN algorithm and the last is Random Forest. Application of the three machine learning algorithms we used in our project viz. Random Forest, Decision Tree, and K-Nearest Neighbours can be beneficial for achieving insights on the crime patterns, which will help the law enforcement prevent the crime with proper crime prevention strategies. Machine learning algorithms are powerful in predicting the different crime rates across the country. In this study, via results observed and data analysis, the category-wise classification is helpful for frontline police workers. According to the results, it is easy to find a more efficient manner to identify and stop crimes. However, thus decreasing the overall crime rate in every part of Indian states. Using such machine learning algorithms are enables stopping crimes and creating a better world for future generations.

References

[1] Marcus Pinto,Hsinrong Wei,Kiyatou Konate, Ida Touray," Delving into factors influencing New York crime data with the tools of machine learning" ,Journal of Computing Sciences in Colleges Volume 36 Issue 2 October 2020 pp 61–70

[2] Frank Anechiarico and James B. Jacobs. 1996. The Pursuit of Absolute Integrity: How Corruption Control Makes Government Ineffective. University of Chicago Press.

[3] Daniel B. Neill,Maria De-Arteaga,William Herlands,Artur Dubrawski, "Machine Learning for the Developing World", ACM Transactions on Management Information Systems Volume 9 Issue ,Article No.(9), pp 1–14 ,https://doi.org/10.1145/3210548, 24 August 2018

[4] Alexandros Belesiotis,George Papadakis, Dimitrios Skoutas,"Analyzing and Predicting Spatial Crime Distribution Using Crowdsourced and Open Data", ACM Transactions on Spatial Algorithms and SystemsVolume 3Issue 4 May 2018 Article No.: 12pp 1–31https://doi.org/10.1145/3190345

[5] N Kanimozhi; N V Keerthana; G S Pavithra; G Ranjitha; S Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm ,"2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 25-27 March 2021, Coimbatore, India

[6] Suhong Kim, Param Joshi, Parminder Singh Kalsi and Pooya Taheri, "Crime Analysis Through Machine Learning", IEEE Transactions on, November 2018.

[7] H Benjamin Fredrick David and A. Suruliandi, "Survey on Crime Analysis and Prediction using Data mining techniques", ICTACT Journal on Soft Computing on, April 2012.

[8] S.Gosavi Shruti and Shraddha S. Kavathekar, "A Survey on Crime Occurrence Detection and prediction Techniques", International Journal of Management Technology And Engineering, vol. 8, December 2018.

[9] H. Chen; W. Chung; J.J. Xu; G. Wang; Y. Qin; M. Chau," Crime data mining: a general framework and some examples", Journals & Magazines Computer,Volume(37) Issue: 4, pp**:** 50 - 56, 4, April 2004, DOI: 10.1109/MC.2004.1297301

[10] Nikita Sonavan(2017), "RAPE AS AN ATROCITY: ANALYSIS OF JUDGMENTS DELIVERED BY THE DISTRICT COURT OF BILASPUR,CHHATTISGARH ", http://glcmumbai.com/lawreview/volume9/04NikitaSonavaneArticle.pdf

[11] Al Amin Biswas;Sarnali Basak,"Al Amin Biswas;Sarnali Basak,"Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model", 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT).pub: IEEE, 28-29 Sept. Jaipur, India 2019

[12] Saqueeb Abdullah,Farah Idid Nibir,Suraiya Salam,Akash Dey,Md Ashraful Alam,Md Tanzim Reza, "Intelligent Crime Investigation Assistance Using Machine Learning Classifiers on Crime and Victim Information" 2020 23rd International Conference on Computer and Information Technology (ICCIT) pub.IEEE,19-21 Dec. 2020

[13] Salma Tabashum,Md. Mamun Hossain,Ariful Islam,Mun Yea Mahafi Taz Zahara,Fahmida Naznin Fami , "Performance Analysis of Most Prominent Machine Learning and Deep Learning Algorithms In Classifying Bangla Crime News Articles",2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020

[14] P. Tamilarasi,R.Uma Rani, "Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning" 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) pub:IEEE,11-13 March 2020

[15] Xu Zhang;Lin Liu;Luzi Xiao; Jiakai Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots", pub: IEEE access Journal, Volume(8),02 October 2020

[16] https://www.ojp.gov/pdffiles1/nij/194616.pdf

[17] https://www.kaggle.com/rajanand/crime-in-india.

[18] Crime in India 2019 Statistics Volume (I) National Crime Records Bureau (Ministry of Home Affairs) Government of India National Highway – 8, Mahipalpur, New Delhi - 110 037. Published By: National Crime Records Bureau (Ministry of Home Affairs) Government of India National Highway – 8, Mahipalpur, New Delhi - 110 037. Phone: 011-26735450 Email: stat@ncrb.nic.in Website: https://ncrb.gov.in

[19] राष्ट्रीय अपराध रिकॉर्ड ब्यूरो NATIONAL CRIME RECORDS BUREAU Empowering Indian Police with Information Technology GOVERNMENT OF INDIA MINISTRY OF HOME AFFAIRS https://ncrb.gov.in/en/crime-india

[20] https://towardsdatascience.com/indian-crime-data-analysis-85d3afdc0ceb

[21] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression", The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

[22] Cover, Thomas M.; Hart, Peter E. (1967). "Nearest neighbor pattern classification"(PDF). IEEE Transactions on Information Theory. 13 (1): 21–27. doi:10.1109/TIT.1967.1053964.

[23] Quinlan, J. R. (1987). "Simplifying decision trees", International Journal of Man-Machine Studies. 27 (3): 221–234. doi:10.1016/S0020-7373(87)80053-6.

[24] Piryonesi, S. Madeh, El-Diraby, Tamer E. (2021), "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling", Journal of Infrastructure Systems. 27 (2): 04021005. doi:10.1061/(ASCE). ISSN 1076-0342.