CRIME PREDICTION USING MACHINE LEARNING ALGORITHMS

PROJECT REPORT

Submitted by

PRAVEEN ELANGO 2015115099

DIWAKAR 2017115524

submitted to the Faculty of

INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600 025

NOVEMBER 2020

ANNA UNIVERSITY
CHENNAI – 600 025
BONAFIDE CERTIFICATE

Certified that this project report titled CRIME PREDICTION USING MACHINE LEARNING ALGORITHMS was carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE: CHENNAI
DATE: 16/12/2020

Dr.L.SAI RAMESH
TEACHING FELLOW
PROJECT GUIDE
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

Dr. SASWATI MUKHERJEE
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

Crime prediction and identifying criminals is one of the top priority problems to the police department because there is a tremendous amount of data related to crime that exists. There is a need for technology through which case-solving could be faster. The idea behind this project is that crimes can be easily predicted once sorted through a huge amount of data to find patterns that are useful to configuring what is required. In this project, a dataset containing the crime records of a city spanning a particular period of time is given as input for visualization and analysis. The visualization is implemented to identify crime rates and patterns from the data while algorithmic analysis is implemented on the data to be able to make prediction of crime rates in the future. This analysis is implemented using several machine learning algorithms. The prediction, if put to good use, can be used to suppress crimes by installing some safety and security measures if the type of crime that is going to happen is known beforehand, thereby reducing crime rates.

திட்டப்பணி ச்சுருக்கம்

குற்ற முன்கணிப்பு மற்றும் குற்றவாளிகளை அடையாளம் காண்பது பொலிஸ் திணைக்களத்திற்கு முன்னுரிமை அளிக்கும் பிரச்சினைகளில் ஒன்றாகும், ஏனெனில் குற்றம் தொடர்பான தரவுகளின் மிகப்பெரிய அளவு உள்ளது. வழக்குத் தீர்க்கும் வேகமான தொழில்நுட்பத்தின் தேவை உள்ளது. இந்தத் திட்டத்தின் பின்னணியில் உள்ள யோசனை என்னவென்றால், ஒரு பெரிய அளவிலான தரவுகளின் மூலம் வரிசைப்படுத்தப்பட்டால், குற்றங்களை எளிதில் கணிக்க முடியும். இந்த திட்டத்தில் ஒரு குறிப்பிட்ட கால இடைவெளியில் ஒரு நகரத்தின் குற்ற பதிவுகள் அடங்கிய தரவுத்தொகுப்பு காட்சிப்படுத்தல் மற்றும் பகுப்பாய்விற்கான உள்ளீடாக வழங்கப்படுகிறது. தற்போதுள்ள தரவுகளிலிருந்து குற்ற விகிதங்கள் மற்றும் வடிவங்களை அடையாளம் காண காட்சிப்படுத்தல் செயல்படுத்தப்படுகிறது, அதே நேரத்தில் எதிர்காலத்தில் குற்ற விகிதங்களை கணிக்கக்கூடிய வகையில் தரவுகளில் வழிமுறை பகுப்பாய்வு செயல்படுத்தப்படுகிறது. இந்த பகுப்பாய்வு பல இயந்திர கற்றல் வழிமுறைகளைப் பயன்படுத்தி செயல்படுத்தப்படுகிறது. முன்னறிவிப்பு, நல்ல பயன்பாட்டுக்கு வந்தால், சில பாதுகாப்பு மற்றும் பாதுகாப்பு நடவடிக்கைகளை நிறுவுவதன் மூலம் குற்றங்களை அடக்குவதற்கு பயன்படுத்தப்படலாம், நடக்கவிருக்கும் குற்றங்களின் வகை முன்பே தெரிந்தால், குற்ற விகிதங்களைக் குறைக்கும்.

ACKNOWLEDGEMENT

First and foremost, we would like to express our deep sense of gratitude to our guide **DR.L.SAI RAMESH**, Professor Department of Information Science and Technology, Anna University for his excellent guidance, counsel, continuous support and patience. He helped us with this topic and guided us in the development of this project. He gave us the moral support to finish out creative and innovative project in a successful manner.

We express our sincere gratitude to **Dr. SASWATI MUKHERJEE**, Professor and Head, Department of Information Science and Technology, Anna University, for her kind support and for providing necessary facilities to carry out the work.

We also express our gratitude to our project committee coordinator **Dr.S.SWAMYNATHAN**, Professor and the committee members **Dr.J.INDUMATHI**, Professor, **Dr.SELVI RAVINDRAN**, Assistant Professor, **Dr. P.GEETHA**, Assistant Professor, **MS.S.KANIMOZHI**, Teaching Fellow, **Ms.R.L.JASMINE**, Teaching Fellow, Department of Information Science and Technology, Anna University, Chennai, for their valuable guidance and technical support.

PRAVEEN ELANGO DIWAKAR

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Crimes are a significant threat to humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. There has been tremendous increase in machine learning algorithms that have made crime prediction feasible based on past data. The focus of such algorithms is to create a model that can help to detect the number of crimes by its type in a particular place and point of time.

1.1 MOTIVATION

The recent spurt of crime worldwide has put everyone wondering as to what will happen in the future. The need of the hour is to make people realize the gravity of the situation. The motivation for the development of the project is providing automatically an efficient way to classify and predict criminal incidents. The proposed solution is based on machine learning techniques and can make the case-solving and investigation process easier and faster.

1.2 PROBLEM STATEMENT

A research paper may belong to more than one category at the same time. In existing literature, many machine learning classification methods have been applied for predicting criminal incidents. But there is a limitation in identifying patterns and factors of crime incidents over a period of time. Advancements in machine learning algorithms can find new patterns in various datasets and reveal new information.

1.3 OBJECTIVE

The project aims to find an appropriate and efficient method for analysis and prediction of criminal incidents.

1.4 ORGANIZATION OF THE REPORT

The project report is organized as follows,

chapter 2 discusses the existing systems and various methods required for the proposed system.

chapter 3 discusses the various concepts used in the proposed system along with overall system architecture.

chapter 4 discusses the implementation detail of the proposed system along with the necessary algorithms.

CHAPTER 2

LITERATURE SURVEY

A literature survey is done by surveying research papers. The limitations and the knowledge gained from the papers will help us to create a better system.

2.1 CRIME PREDICTION USING K-NEAREST NEIGHBORING ALGORITHM

This paper proposes that crimes can be easily predicted once sorted through a huge amount of data to find patterns that are useful to configuring what is required [1]. The recent developments in machine learning make this task possible. The date, time, location (longitude, latitude) is given as input and the output will be generated which will give information about which crime is likely to happen in that area. It basically gives the hotspots of crime. The data is taken considering the time and type of crime that happened in the past. KNN algorithm then uses its approach which assumes that similar things exist in close proximity and classifies new cases based on similarity measures.

Classes of crimes are:

- Act 379 – Robbery
- Act 13 – Gambling
- Act 279 – Accident

- Act 323 – Violence
- Act 302 – Murder
- Act 363 – Kidnapping

This prediction, if put to good use, can be of great help in investigating cases that have happened.

2.2 CRIME ANALYSIS THROUGH MACHINE LEARNING

This paper proposes to create a prediction model that can accurately predict crime [2]. Two classification algorithms, K-Nearest Neighbor (KNN) and boosted decision tree, were implemented to analyze the VPD crime dataset compiled between 2003 and 2018 with more than 560,000 records. The dataset was processed using two different approaches:

• In the first approach, each neighborhood and crime category were given a unique number when a certain crime happens in a certain neighborhood.

• In the second approach, the neighborhood and the day of the week during which the crime was committed were given a binary number and marked as 1 when the crime happened on that day in that neighborhood, and 0 otherwise.

A prediction accuracy between 39% to 44% is obtained when predicting crime in Vancouver.

2.3 STUDY OF ENSEMBLE LEARNING METHODS FOR CLASSIFICATION IN BIOINFORMATICS

In this research work [3], a novel ensemble learning approach "BBS method" which stands for Bagging, Boosting and Stacking is proposed with appropriate base classifiers for classification of the five UCI datasets taken from the field of Bioinformatics. Experiments were conducted using Weka and Java Eclipse and it has been observed empirically that this approach gives better accuracy with lower root mean square error rate using the technique of ensemble learning. Henceforth it is concluded that the proposed ensemble learning method is more suitable in handling the classification problem in the bioinformatics domain. Such approaches can be efficiently used in related real-life scenarios of classification domain.

2.4 A MACHINE LEARNING APPROACH TO PREDICT CRIME USING TIME AND LOCATION DATA

This paper proposes to use machine learning techniques to classify a criminal incident by type, depending on its occurrence at a given time and location [4]. The experimentation was conducted on a dataset containing San Francisco's crime records from 2003 - 2015. For this supervised classification problem, Decision Tree, Gaussian Naive Bayes, k-NN, Logistic Regression, Adaboost, Random Forest classification models were used. As crime categories in the dataset are imbalanced, oversampling methods, such as SMOTE and under sampling methods such as Edited NN, Neighborhood Cleaning Rule were used. Solving the imbalanced class problem,

the machine learning agent was able to categorize crimes with approximately 81% accuracy.

2.5 CRIME PREDICTION AND FORECASTING IN TAMILNADU USING CLUSTERING APPROACHES

In this work, various clustering approaches of data mining are used to analyze the crime data of Tamilnadu [5]. The crime data is extracted from National Crime Records Bureau (NCRB) of India. K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) algorithms are used to cluster crime activities based on some predefined cases and the results of these clustering are compared to find the best suitable clustering algorithm for crime detection.

CHAPTER 3

SYSTEM DESIGN

3.1 PROPOSED SYSTEM ARCHITECTURE

The proposed system follows a tightly coupled architecture comprising of the following modules.

- Dataset
- Data transformation
- Data visualization
- Model engine
- Preliminary results
- Performance tuning

3.2 ARCHITECTURE DIAGRAM

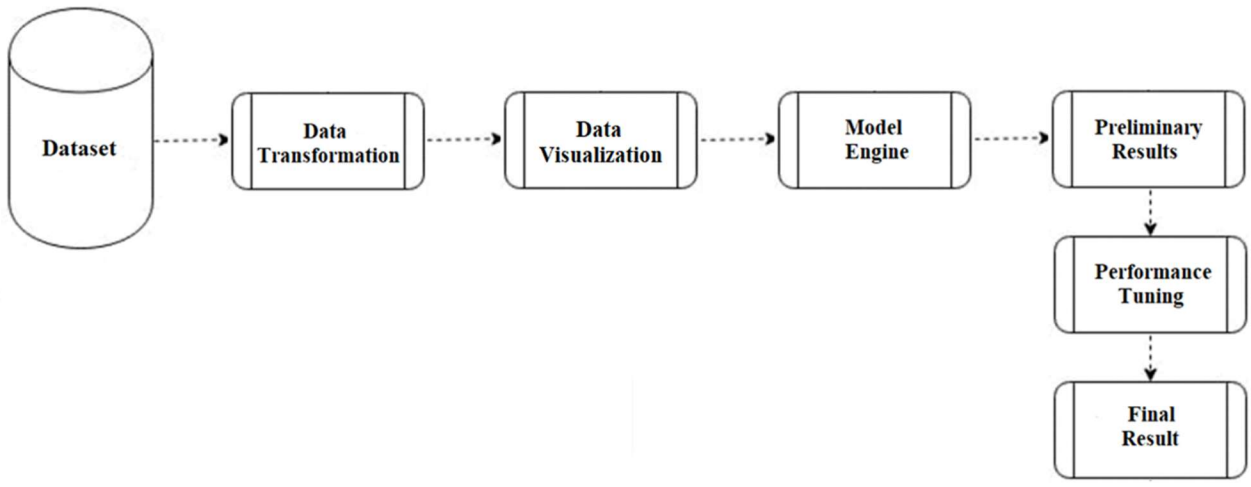


Figure 3.1 System Architecture

3.3 MODULE DESCRIPTION

The description of the modules is mentioned below.

3.3.1 DATASET

The most important part of any machine learning based project is the dataset. The dataset used in this project reflects reported incidents of crime that occurred in the city of Chicago from January 1, 2018 to November 21, 2020. The dataset is obtained from the official Chicago Data Portal (https://data.cityofchicago.org/). It contains the following attributes:

- ID: Unique identifier for the record.
- Case Number: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- Date: Date when the incident occurred.
- Block: address where the incident occurred
- IUCR: The Illinois Uniform Crime Reporting code.
- Primary Type: The primary description of the IUCR code.
- Description: The secondary description of the IUCR code, a subcategory of the primary description.
- Location Description: Description of the location where the incident occurred.
- Arrest: Indicates whether an arrest was made.
- Domestic: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

- Beat: Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car.
- District: Indicates the police district where the incident occurred.
- Ward: The ward (City Council district) where the incident occurred.
- Community Area: Indicates the community area where the incident occurred. Chicago has 77 community areas.
- FBI Code: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
- X Coordinate: The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
- Y Coordinate: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
- Year: Year the incident occurred.
- Updated On: Date and time the record was last updated.
- Latitude: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Longitude: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
- Location: The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

The dataset is imported and sent to the transformation module.

3.3.2 DATA TRANSFORMATION

Data transformation / preprocessing involves the following steps:

i) Creation of data frame

ii) Finding shape, dimensions and data types.

iii) Checking the data for null and duplicate values and dropping them

iv) Removing primary key type attributes that cannot be used for analysis

v) Conversion of the date column to datetime object to get the day of the week, month and time of the crime for better analysis

vi) Mapping of crimes under one group by various parameters

- Type of crime
- Zone where crime has occurred
- Season
- Location

3.3.3 DATA VISUALIZATION

In this module, data is visualized over the time period concerned with the dataset to analyze crime patterns. The data is grouped into years (2018, 2019 and 2020) for separate analysis. The graphics used for visual analysis are bar graph, line graph and pie chart. Finally, a map of Chicago is plotted with the distribution

of various types of crimes happening in districts all over the state. These graphics are analyzed to infer crime patterns in detail and can be used to suppress crimes by installing some safety and security measures in areas plagued with high crime, thereby reducing crime rates.

3.3.4 MODEL ENGINE

In this module, the initial step is to prepare the data for training. The numerical attributes are converted to categorical attributes. This is done as it is better suited for training the model. The attributes that are not required for training are dropped to give more accurate results. The dataset is then split into training set and testing set in the ratio of 75% and 25% with ‘arrest’ as the target variable. Several algorithms are then used for training on the training set. This is done by importing the respective classifiers. After training the model, the classifiers are applied on the testing set. This process is performed individually for each algorithm.

The following classifiers are used for training and testing:

i) Gaussian Naive Bayes

ii) Decision Tree

iii) Random Forest

iv) Logistic Regression

3.3.5 PRELIMINARY RESULTS

After classification is done on the testing set, confusion matrix is computed and plotted to obtain the classification results for each algorithm.

The following classification metrics are used in confusion matrix method:

i) Accuracy

ii) Error

iii) Precision

iv) Recall

v) F-1 Score

3.3.6 PERFORMANCE TUNING AND FINAL RESULT

After implementation of the previous module, the classifier with the highest accuracy value is chosen for hyper tuning. The hyper parameters are tuned to further increase accuracy of selected classifier by cross validation. The final accuracy is obtained after hyper tuning.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 TOOLS USED

The following tools, libraries and environments are used in this project.

4.1.1 Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

4.1.2 NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

4.1.3 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It is used for creating static, animated, and interactive visualizations in Python.

4.1.4 Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

4.1.5 Scikit-learn

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

4.1.6 Mlxtend

Mlxtend (machine learning extensions) is a Python library of useful tools for the day-to-day data science tasks.

4.1.7 Google Colab Notebooks

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

4.2 ALGORITHMS USED

4.2.1 Gaussian Naive Bayes

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution.

4.2.2 Decision Tree

Decision Tree is a supervised learning algorithm whose goal is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

4.2.3 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. Random forest algorithm creates decision trees

on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

4.2.4 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems.

4.3 PROJECT WORK IMPLEMENTATION

4.3.1 Getting Started

In the initial step, the libraries and tools mentioned previously are imported in the Colab notebook. The dataset to be used for the project is then retrieved and imported and a data frame is created as shown in Figure 4.1.

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code
0	12225699	JD436533	11/19/2018 03:00:00 PM	061XX N GLENWOOD AVE	1195	DECEPTIVE PRACTICE	FINANCIAL EXPLOITATION OF AN ELDERLY OR DISABL...	APARTMENT	False	False	2433	24	48.0	77	11
1	11243458	JB168952	02-12-18 14:00	018XX N CICERO AVE	320	ROBBERY	STRONG ARM - NO WEAPON	BANK	False	False	2533	25	36.0	19	3
2	11214298	JB129783	01/25/2018 08:00:00 PM	0000X E MONROE ST	320	ROBBERY	STRONG ARM - NO WEAPON	BANK	False	False	112	1	42.0	32	3
3	11499606	JB504989	11-05-18 2:30	002XX E CHICAGO AVE	281	CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	HOSPITAL BUILDING / GROUNDS	True	False	1833	18	2.0	8	2
4	12179865	JD383631	09/29/2018 12:00:00 AM	048XX N LAWNDALE AVE	1754	OFFENSE INVOLVING CHILDREN	AGGRAVATED SEXUAL ASSAULT OF CHILD BY FAMILY M...	APARTMENT	False	True	1712	17	35.0	14	2

X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
NaN	NaN	2018	11/20/2020 03:50:28 PM	NaN	NaN	NaN
1144104.0	1912042.0	2018	11/20/2020 03:48:09 PM	41.914667	-87.746015	(41.914667382, -87.746015079)
1177003.0	1899949.0	2018	11/20/2020 03:48:09 PM	41.880802	-87.625516	(41.880801792, -87.625515824)
1178058.0	1905778.0	2018	11/19/2020 03:52:48 PM	41.896773	-87.621464	(41.896772914, -87.621464432)
NaN	NaN	2018	11/19/2020 03:50:24 PM	NaN	NaN	NaN

Figure 4.1 Data frame head

4.3.2 Data Preprocessing

The dataset is checked for null and duplicate values and dropped. Inconsistencies in attribute names are handled and primary key type attributes are removed as they have no use for any type of analysis. The date column is converted to datetime object to get the day of the week, month and time of the crime for better analysis. The crimes are mapped under one group by various parameters.

	date	block	iucr	primary_type	description	location_description	arrest	domestic	beat	district	ward	community_area	fbi_code
0	02-12-18 14:00	018XX N CICERO AVE	320	ROBBERY	STRONG ARM - NO WEAPON	BANK	False	False	2533	25	36.0	19	3
1	01/25/2018 08:00:00 PM	0000X E MONROE ST	320	ROBBERY	STRONG ARM - NO WEAPON	BANK	False	False	112	1	42.0	32	3
2	11-05-18 2:30	002XX E CHICAGO AVE	281	CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	HOSPITAL BUILDING / GROUNDS	True	False	1833	18	2.0	8	2
3	05/30/2018 09:08:00 PM	107XX S CHAMPLAIN AVE	041A	BATTERY	AGGRAVATED - HANDGUN	ALLEY	False	False	513	5	9.0	50	04B
4	10/25/2018 08:00:00 PM	065XX S SANGAMON ST	265	CRIMINAL SEXUAL ASSAULT	AGGRAVATED - OTHER	RESIDENCE	True	False	723	7	6.0	68	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
684309	01-01-20 2:45	021XX S OAKLEY AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET	True	False	1234	12	25.0	31	08B
684310	01-01-20 21:23	042XX N BROADWAY	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	True	1915	19	46.0	3	08B
684311	01-01-20 17:00	013XX N LOCKWOOD AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	False	False	2532	25	37.0	25	14
684312	01-01-20 13:00	086XX S PHILLIPS AVE	2825	OTHER OFFENSE	HARASSMENT BY TELEPHONE	APARTMENT	False	False	423	4	7.0	46	26
684313	01-01-20 0:15	104XX S AVENUE H	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE	False	False	432	4	10.0	52	14

x_coordinate	y_coordinate	year	updated_on	latitude	longitude
1144104.0	1912042.0	2018	11/20/2020 03:48:09 PM	41.914667	-87.746015
1177003.0	1899949.0	2018	11/20/2020 03:48:09 PM	41.880802	-87.625516
1178058.0	1905778.0	2018	11/19/2020 03:52:48 PM	41.896773	-87.621464
1182509.0	1833939.0	2018	11/19/2020 03:50:24 PM	41.699538	-87.607345
1171101.0	1861443.0	2018	11/17/2020 03:47:01 PM	41.775269	-87.648315
...	...	...	...	...	...
1161360.0	1889913.0	2020	01-08-20 15:49	41.853602	-87.683235
1169146.0	1928470.0	2020	01-08-20 15:49	41.959239	-87.653536
1140776.0	1908599.0	2020	01-08-20 15:49	41.905281	-87.758327
1193996.0	1848162.0	2020	01-08-20 15:49	41.738294	-87.564821
1202826.0	1836387.0	2020	01-08-20 15:49	41.705762	-87.532871

Figure 4.2 Dataset after preprocessing

4.3.3 Data Visualization and Analysis

The data is visualized over the time period concerned with the dataset to analyze and identify crime patterns. The data is grouped into years (2018, 2019 and 2020) for separate analysis.

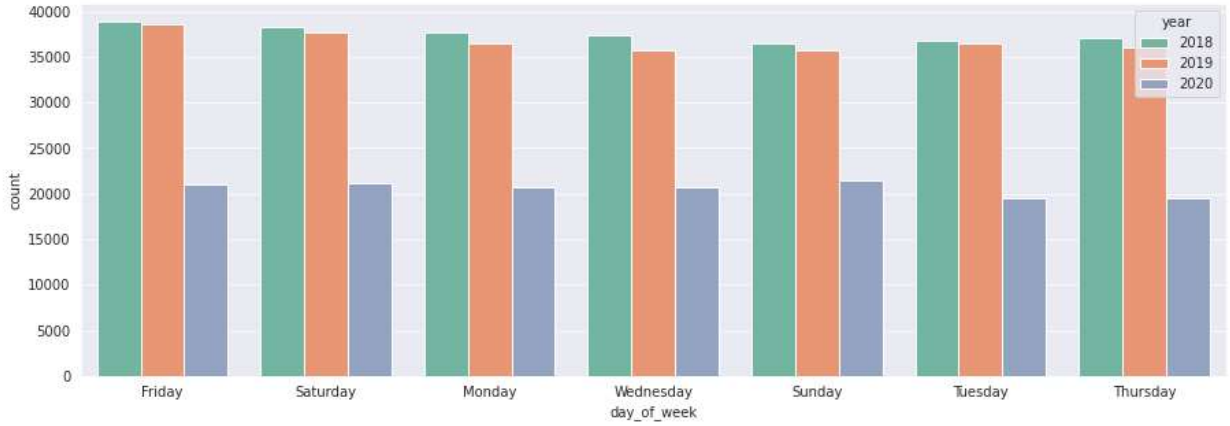


Figure 4.3 Crime rate with respect to days of the week

As shown in Figure 4.3, a bar graph is plotted for the crime count with respect to days of the week for the last three years. It is observed that the day of the week has very little influence on the crime, it seems like almost every day the crime count was almost the same. When comparing 2018 and 2019 it can be seen that the number of crimes is less in 2019 than in 2018 and the number of crimes in 2020 are less than both 2018 and 2019.

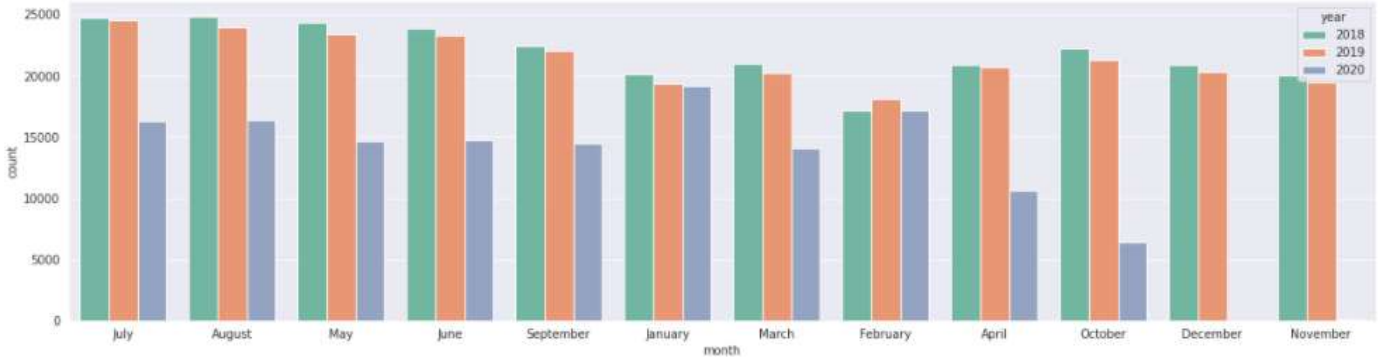


Figure 4.4 Crime rate with respect to months of the year

Similarly, in Figure 4.4, a bar graph is plotted for the crime count with respect to the month for the last three years.

As shown in Figure 4.5, a line graph illustrates the frequency of crimes that happened with respect to hour of the day for the year 2018. Around midnight it is observed that there were a lot of crimes happening and then it decreases gradually. There is a spike at 12 noon. The crime rate was also higher in the evening.

Figure 4.5 Crime rate with respect to hours of the day

Similar graphs are also plotted for 2019 and 2020.

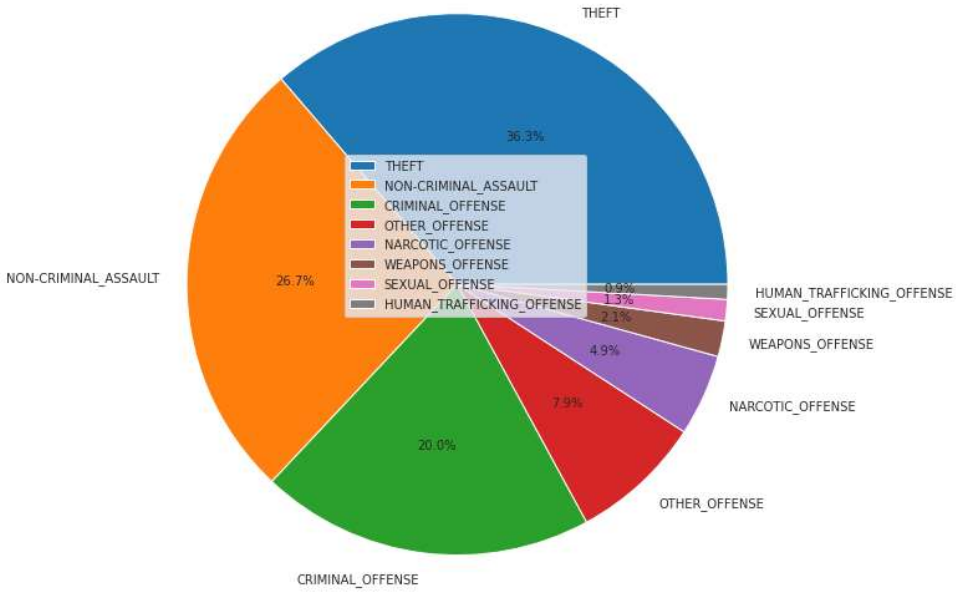


Figure 4.6 Percentage of each type of crime

In Figure 4.6, a pie chart is plotted that reflects the percentage of each type of crime in 2018. It can be seen that there were a lot of Thefts and Non- Criminal Assaults that happened in the year 2018, but there were very less sexual and trafficking offences. Theft and Non-Criminal Offences constituted more than 50% of the crimes committed. In the same manner, graphs are also plotted for 2019 and 2020.

For further detailed analysis, different types of visualizations are also plotted in the project to reflect crime rates with respect to various parameters such as location, zone, season, arrests made etc. in the time period concerned with the data (2018-2020).

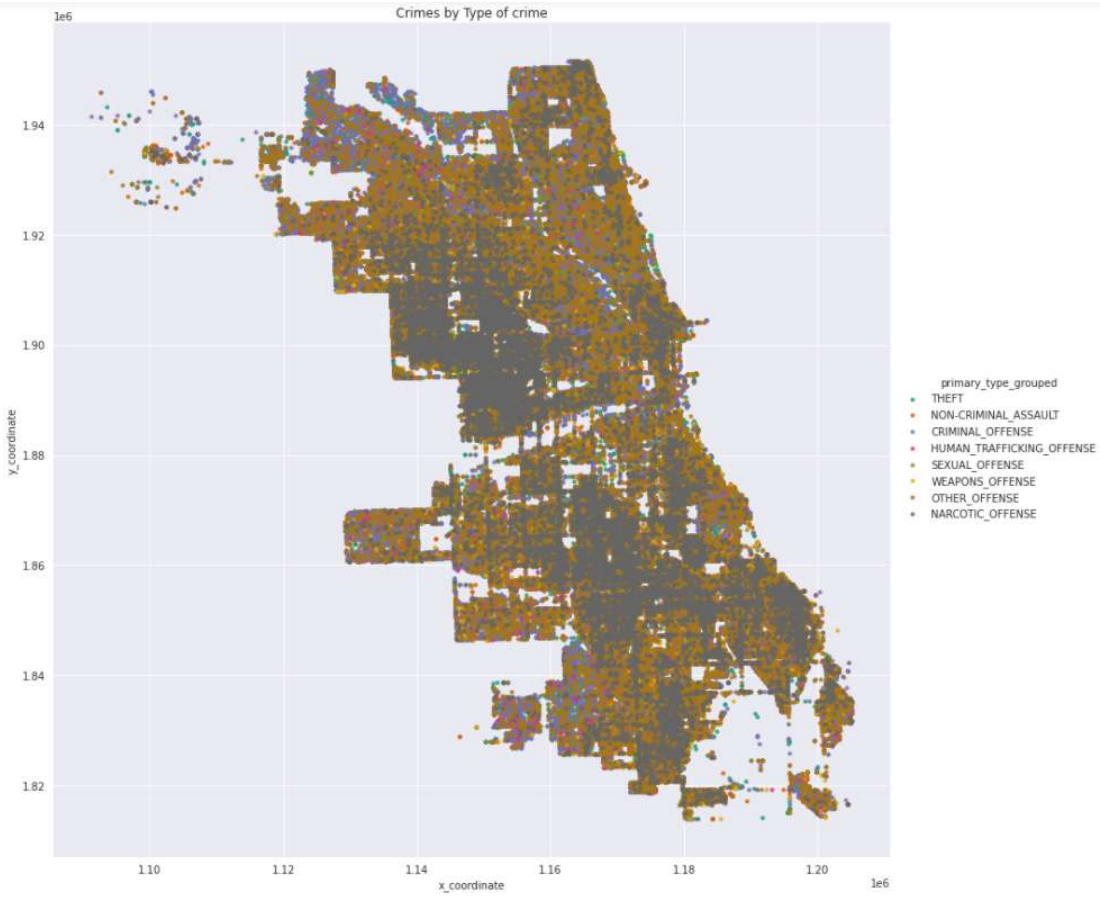


Figure 4.7 Chicago crime map

In Figure 4.7, a map of Chicago is plotted over which the distribution of various types of crimes that happened in districts all over the state with respect to the concerned time period is represented.

4.3.4 Model Engine and Results

The data is prepared for algorithm analysis by converting the numerical attributes to categorical attributes and the attributes that are not required for training are dropped to give more accurate results. The dataset is then split into training set and testing set in the ratio of 75% and 25% with ‘arrest’ as the target variable as shown in Figure 4.8.

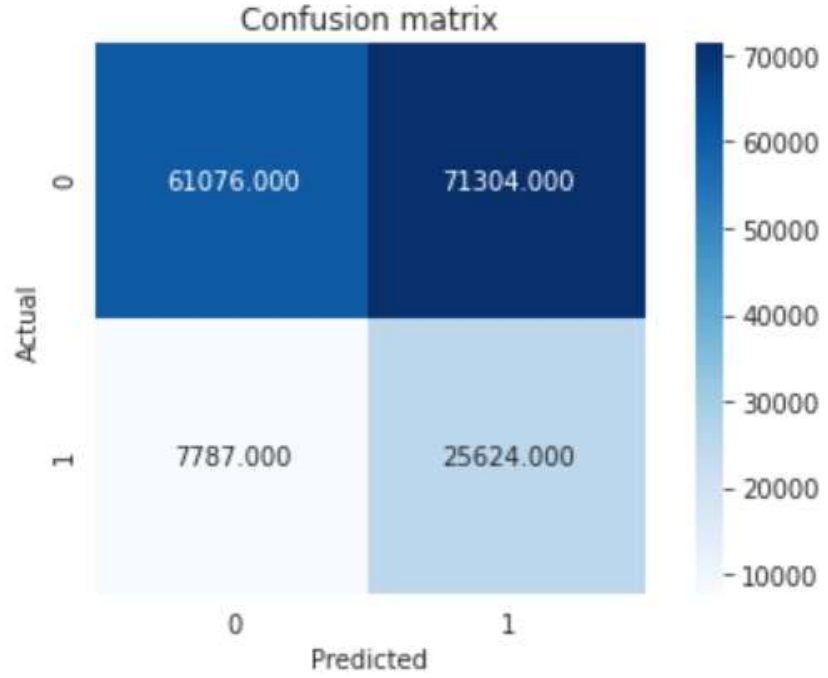
```
X_train, X_test, y_train, y_test = train_test_split(crimes_data_prediction.drop(['arrest_1'],axis=1),crimes_data_prediction['arr
est_1'], test_size=0.25, random_state=42)
```

Figure 4.8 Splitting of dataset

Several algorithms are then used for training and testing which is done by using the respective classifiers. The implementation of each algorithm with the resultant confusion matrix and accuracy results is shown in the following sections.

i) Gaussian Naive Bayes

Using Gaussian Naive Bayes algorithm for training and testing and plotting the confusion matrix yields the following results:



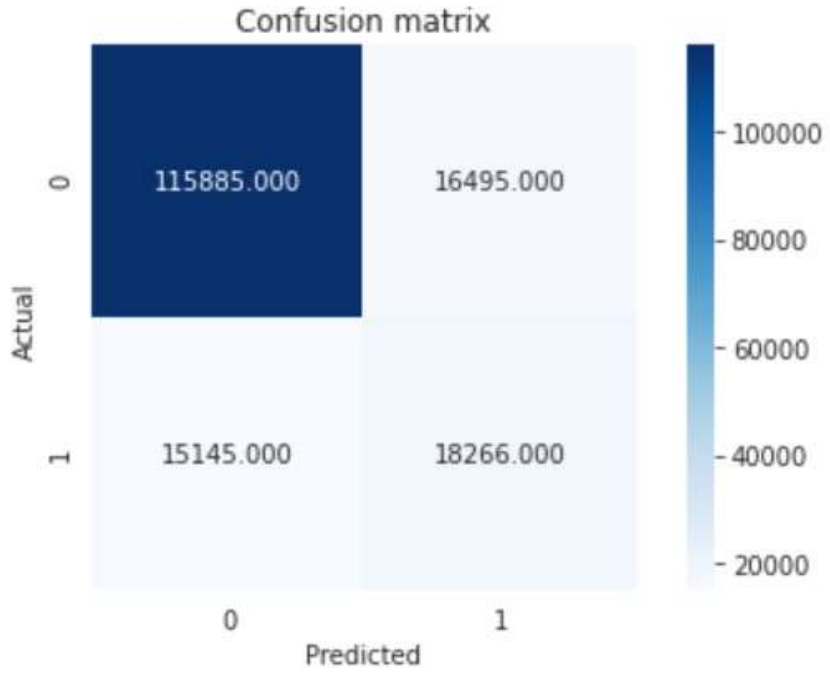

```
Accuracy =  0.522947566514467
Error =     0.47705243348553295
Precision = 0.26436117530538134
Recall =    0.7669330460028134
F-1 Score = 0.39319006590506295
Classification Report
              precision    recall  f1-score   support

           0       0.89      0.46      0.61    132380
           1       0.26      0.77      0.39     33411

    accuracy                           0.52    165791
   macro avg       0.58      0.61      0.50    165791
weighted avg       0.76      0.52      0.56    165791
```

ii) Decision Tree

Using Decision Tree algorithm for training and testing and plotting the confusion matrix yields the following results:



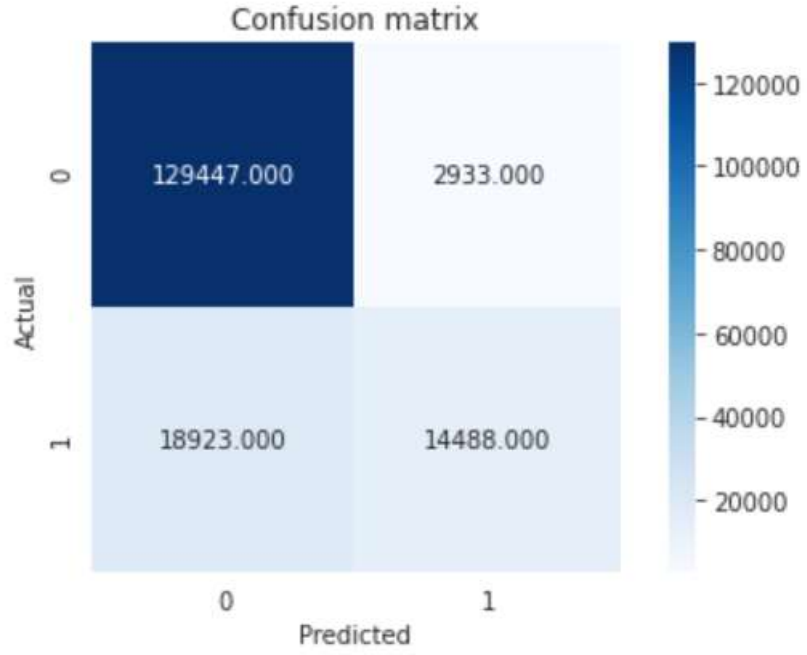

```
Accuracy =  0.80915731251998
Error =  0.19084268748001998
Precision =  0.5254739506918673
Recall =  0.5467061746131514
F-1 Score =  0.5358798333626709
Classification Report
              precision    recall  f1-score   support

           0       0.88      0.88      0.88    132380
           1       0.53      0.55      0.54     33411

    accuracy                           0.81    165791
   macro avg       0.70      0.71      0.71    165791
weighted avg       0.81      0.81      0.81    165791
```

iii) Random Forest

Using Random Forest algorithm for training and testing and plotting the confusion matrix yields the following results:



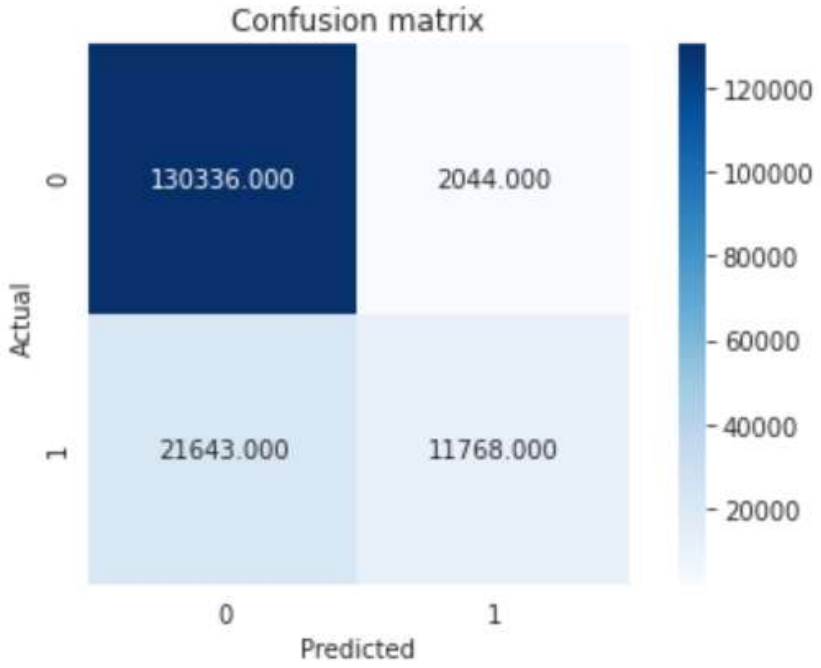

```
Accuracy =  0.8681713723905399
Error =  0.13182862760946013
Precision =  0.8316399747431261
Recall =  0.43362964293196854
F-1 Score =  0.5700346238589865
Classification Report
              precision    recall  f1-score   support

           0       0.87      0.98      0.92    132380
           1       0.83      0.43      0.57     33411

    accuracy                           0.87    165791
   macro avg       0.85      0.71      0.75    165791
weighted avg       0.86      0.87      0.85    165791
```

iv) Logistic Regression

Using Logistic Regression algorithm for training and testing and plotting the confusion matrix yields the following results:




```
Accuracy =  0.8571273470815666
Error =  0.14287265291843343
Precision =  0.8520127425427165
Recall =  0.352219328963515
F-1 Score =  0.49840120280371847
Classification Report
              precision    recall  f1-score   support

           0       0.86      0.98      0.92    132380
           1       0.85      0.35      0.50     33411

    accuracy                           0.86    165791
   macro avg       0.85      0.67      0.71    165791
weighted avg       0.86      0.86      0.83    165791
```

4.3.5 Hyper tuning and Final result

After algorithm analysis on the data, it is observed that the Random Forest classifier gives the highest accuracy on the model. Therefore, the hyper parameters of this classifier are tuned using cross validation to further increase the accuracy as shown in Figure 4.9.

```
In [ ]: feature_scaler = StandardScaler()
        X_train = feature_scaler.fit_transform(X_train)
        X_test = feature_scaler.transform(X_test)

In [ ]: classifier = RandomForestClassifier(n_estimators=10, random_state=42)

In [ ]: from sklearn.model_selection import cross_val_score
        all_accuracies = cross_val_score(estimator=classifier, X=X_train, y=y_train, cv=5)

In [ ]: print(all_accuracies.max())

        0.8697448579528319
```

Figure 4.9 Hyper tuning and Final accuracy

CHAPTER 5

CONCLUSION AND FUTURE WORK

With the help of machine learning algorithms, analysis and prediction on a large amount of crime data has been made easier. The work in this project mainly revolves around crime visualization and prediction. Data visualization helps in the analysis of the dataset. The use of different types of graphs, each having its own characteristics helps in identifying existing crime rates and patterns with respect to various parameters. After using several machine learning algorithms for training and testing, we chose the model that gave the highest accuracy. The hyperparameters of the model were then tuned and now gives a prediction accuracy of 86.97%. This work can help in case investigation and solving and also help to suppress crimes and ensure a safer society.

The future work of the project includes incorporating deep learning models and also incorporating economic and weather data of the region concerned with the dataset. Once the above work is completed, forecasting of crime rates and patterns for upcoming months/years can be implemented.

REFERENCES

[1] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, 2020.

[2] S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, "Crime Analysis Through Machine Learning," *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, 2018.

[3] A. Verma and S. Mehta, "A comparative study of ensemble learning methods for classification in bioinformatics," *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Noida, 2017.

[4] Nishat Shama, "A Machine Learning Approach to Predict Crime Using Time and Location Data," *BRAC University*, Dhaka, 2017.

[5] S. Sivaranjani, S. Sivakumari and M. Aasha, "Crime prediction and forecasting in Tamilnadu using clustering approaches," *2016 International Conference on Emerging Technological Trends (ICETT)*, Kollam, 2016.

[6] S. Agarwal, L. Yadav and M. K. Thakur, "Crime Prediction Based on Statistical Models," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, 2018.

[7] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma and N. Yadav, "Crime pattern detection, analysis & prediction," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, 2017.

[8] Y. Lin, T. Chen and L. Yu, "Using Machine Learning to Assist Crime Prevention," *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, Hamamatsu, 2017.

[9] G. Jha, L. Ahuja and A. Rana, "Criminal Behaviour Analysis and Segmentation using K-Means Clustering," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2020.

[10] A. Almaw and K. Kadam, "Crime Data Analysis and Prediction Using Ensemble Learning," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2018.