Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777680

CRIME IDENTIFICATION AND DETECTION USING MACHINE LEARNING

GD Makkar[1], Pradeep Semwal[2], Harish Chandra Sharma[3], Archana Kero[4], Minit Arora[5], Vaibhav Sharma[6]*

[1,2,3,4,5,6]Associate Professor, School of CA & IT, SGRR University, Dehradun, Uttarakhand, India,

Email: [1]gdmakkar@gmail.com, [2]psemwal2222@gmail.com, [3]hcs19@ yahoo.com, [4]archanakero@gmail.com, [5]minitarora@gmail.com, [6]*vsdeveloper10@gmail.com

Orcid ID: 30000-0002-1263-9325, 6*0000- 0002-1404-2012

*Corresponding Author

ABSTRACT

Recognizing crime patterns is crucial for being better prepared to respond to criminal behaviour. Assessed crime statistics from the city of Indore that has been collected from the Indore Polices publicly available website in this study. The goal is to estimate which type of crime is most likely to occur in Indore at a particular time and location. Artificial intelligence and machine learning have become more significant in crime detection and prevention. Applied the Nave Bayes, K-nearest neighbour, and linear regression approaches to analyse crime data using the same limited set of variables on the Communities and Crime Dataset from Indore. In terms of overall performance, the Nave Bayes approach performs well in compare to the other machine learning algorithms.

INTRODUCTION

Crime identification using machine learning is a field of study that aims to leverage the power of machine learning algorithms and techniques to analyze and interpret crime-related data for the purpose of identifying criminal activities, predicting crime patterns, and improving law enforcement strategies. It involves the use of statistical models, data mining, and pattern recognition algorithms to analyze various types of crime data, such as incident reports, criminal records, geographical data, and social media data. Here are some common approaches and techniques used in crime identification using machine learning:

Predictive Modelling: Machine learning algorithms can be trained on historical crime data to create predictive models that can forecast the likelihood of future crimes. These models can consider various factors such as time, location, demographics, and environmental conditions to estimate the probability of criminal activities occurring in a specific area.

Anomaly Detection: Machine learning algorithms can be used to identify abnormal or suspicious patterns in crime data. By learning from historical data, these algorithms can detect outliers and unusual behaviour that may indicate criminal activities or anomalies in the dataset.

Natural Language Processing (NLP): NLP techniques can be applied to analyze unstructured data sources such as social media posts, news articles, and police reports to extract useful information related to crime. Sentiment analysis, entity recognition, and topic modelling can be used to identify crime-related keywords, locations, and individuals involved.

Image and Video Analysis: Machine learning algorithms can analyze images and videos from surveillance cameras or crime scenes to detect and recognize objects, faces, and other visual elements relevant to criminal activities. This can assist in identifying suspects, vehicles, or other objects associated with crimes.

Network Analysis: By analyzing connections and relationships between individuals or entities involved in criminal activities, machine learning algorithms can identify hidden networks, criminal organizations, or patterns of criminal behavior. Social network analysis and graph algorithms can be applied to uncover key individuals and their roles in criminal networks.

Predictive Policing: Machine learning models can be used to allocate police resources more effectively by predicting crime hotspots and identifying areas at higher risk of criminal activities. This can help law enforcement agencies optimize patrol routes and allocate resources based on data-driven insights.

It's important to note that crime identification using machine learning is a complex and evolving field. Ethical considerations and potential biases in the data and algorithms used should be carefully addressed to ensure fair and accurate results. Furthermore, the implementation of machine learning models should always be complemented with human expertise and judgement for effective crime prevention and law enforcement.

Data mining can therefore be very helpful in analysing, visualising, and predicting crime utilising crime data sets from various Indian states. Dataset is classified based on some predefined condition. Here, crimes against women are grouped according to the numerous forms that occur in the various Indian states and localities. Planning measures to prevent crimes against women and taking effective action to reduce crime would be much easier with the aid of crime prediction.

A deeper understanding of crime is helpful in several ways, including by enabling law enforcement to reduce crime through targeted and sensitive actions, as well as by encouraging greater cooperation between citizens and the state in order to foster safe neighbourhoods. Understanding trends in crime from data is an active and expanding area of research because to the Big Data era and the accessibility of quick, effective algorithms for data analysis. Our algorithms' inputs include location (latitude and longitude), kind of crime, and time (hour, day, month, and year):

To analyse the data of previous year’s crime record and to be better prepared for the unforeseen crimes happening in a particular area and to find out the hotspots of crimes likely to happen. The creation of an accurate prediction model for crime is the main goal of this endeavour. In order to assess the crime dataset collected between with more records, two classification techniques, Linear Regression, K-Nearest Neighbour (KNN), and Nave Bayes, were implemented.

We can forecast the locations of future crimes by using historical data and looking at where current crimes have occurred. For instance, a sudden spike in burglaries in one location may indicate that nearby areas may soon experience a similar increase. The system flags potential hotspots, and the police should think about increasing their patrols.

The science of machine learning involves letting computers make judgements on their own. Self-driving cars, speech recognition, web search, and a better knowledge of the human genome have all benefited recently from the use of machine learning. Additionally, it has made it possible to predict crime using the cited data. The prediction method KNearest Neighbour (KNN) Classification supports nominal class labels. Numerous industries have employed classification, including corporate intelligence, healthcare, finance and banking, homeland security, and weather forecasting.

LITERATURE SURVEY

Accurately and effectively processing the expanding volumes of crime data is a significant challenge for all law enforcement and intelligence gathering organisations. Cybercrime detection can also be challenging due to the high volume of data generated by frequent online transactions and active network traffic, of which only a small part is related to illicit activity. Criminal investigators who may not have considerable experience as data analysts can swiftly and effectively investigate enormous datasets with the use of the sophisticated tool known as data mining. With the Cop link project, which University of Arizona researchers have been working on with the Tucson and Phoenix police departments, we propose a generic framework for crime data mining. This framework is based on the knowledge we have learned from this project.

Any social order that includes criminal activity has been around since the dawn of time. Even within a single community, it varies from place to place and from one method of occurrence to another. Additionally, it sometimes increases, sometimes drops, etc., and is concentrated in some areas more than others. Previous studies have shown that the rate of crime is significantly correlated with a variety of social characteristics, including education levels, poverty rates, and the absence of social organisation. Others have called attention to the association between the built environment and the rate of crime. They suggested that crime happens in areas where there are both opportunities and criminals. The purpose of this essay is to pinpoint urban factors that contribute to crime in the Greater Cairo Region and to offer several solutions for lowering these crimes. The primary areas of the agglomeration were further examined in light of socioeconomic analysis, street network pattern, and land usage.

Public safety personnel now have the opportunity to prioritise the deployment of limited resources based on anticipated crime trends thanks to the convergence of public data and statistical models. Observed crime data and details about numerous criminogenic factors are used to train current crime prediction techniques. Due to a dearth of evidence at smaller resolutions (such as ZIP codes), researchers have favoured global models (such as those of entire cities). These global models and their presumptions are in conflict with data showing that there are regional differences in the link between crime and criminogenic factors. We provide area-specific crime prediction models based on hierarchical and multi-task statistical learning in response to this gap. By sharing data across ZIP codes, our models reduce sparsity while retaining the benefits of localised models for tackling non-homogeneous crime trends. Actual crime data used in out-of-sample testing reveals predictive improvements over a number of cutting-edge worldwide models.

Road traffic accidents (RTAs) cause an estimated 1.2 million fatalities and 50 million injuries annually, which raises serious public health issues. RTAs are among the top causes of mortality and injury in developing countries, with Ethiopia having the greatest rate of these mishaps.

Therefore, both traffic agencies and the general public are very interested in strategies to lessen accident severity. In this study, we used data mining techniques to establish a connection between observed road characteristics and the severity of accidents in Ethiopia. We also created a set of guidelines that the Ethiopian Traffic Agency might employ to increase safety.

Naive Bayes can be used for crime identification and detection by classifying instances or incidents as either criminal or non-criminal based on the available features or attributes. Here's how Naive Bayes can be applied in this context:

In this article, we gather a labeled dataset consisting of historical crime incidents with associated attributes/features. These attributes could include time of occurrence, location, type of crime, demographic information, and any other relevant information. Then clean the dataset by removing any inconsistencies, missing values, or irrelevant attributes. And convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding. Identify the features that are most relevant for crime identification and detection. This can be done by analyzing the dataset and consulting with domain experts. Apply the Naive Bayes algorithm to the training set. The algorithm will estimate the probabilities and build a model based on the assumptions of feature independence. Calculate the prior probabilities of criminal and non-criminal incidents. Calculate the posterior probabilities of the incident belonging to each class using Bayes' theorem and the conditional probabilities estimated during training. Evaluate the performance of the Naive Bayes model by comparing the predicted class labels with the true labels in the test set.

PROPOSED SYSTEM

Only if our approach to combating crime is clear and steadfast can crime be reduced. Only by analysing historical and current data will this be possible. In this case, the approach we take makes use of a dataset that details crime against women across our nation's states. Using the data set that is now accessible, we collect the relevant information from prior years, and then we use the naive Bayes classification method to estimate the number of crimes that are likely to occur in the years to come. Time series technique is then applied to the supplied set of data after bayes classification. Additionally, we will explore and highlight the hotspots of locations where a specific crime will occur. The module of the proposed system can be divided into distinct modules.

Pre-processing
Data Source
Pre-Processing
Statistical Analysis
Classification

Data Source

A deeper understanding of crime is helpful in a number of ways, including by enabling law enforcement to reduce crime through targeted and sensitive actions, as well as by encouraging greater cooperation between citizens and the state in order to foster safe neighbourhoods. Understanding trends in crime using data from the city of Indore is an active and expanding subject of research due to the advent of the Big Data era and the availability of quick, effective algorithms for data analysis. Our algorithms' inputs include location (latitude and longitude), kind of crime, and time (hour, day, month, and year):

Accident governs Act 279
Robbery governs Act 379 Ø Gambling governs Act 13
Kidnapping governs Act 363
Violence governs Act 323 Ø Murder governs Act 302

Pre-processing

To add several pertinent features to the original and pre-processed datasets, fill the vacant cells, remove superfluous columns, and pre-process the original dataset.

Statistical Analysis

The crime dataset's distribution is based on the year, month, and day. The yearly average of criminal events is displayed. As the time intervals grow longer, the dataset tends to exhibit a normal distribution. The graph of each day, however, shows an unusual maximum value of a greater number of events, which is thought to be an aberration but actually signals a riot.

Classification

After statistical analysis we classify the prediction values of the crime rate in the taken data set. model predictions from machine learning. To achieve crime-prediction accuracy between 75% and 80%, naive bayes classification and boosted decision trees were used. For various techniques and algorithms, there were minor differences in the accuracy, complexity, and training time of the algorithms. Finally, utilising the earlier information, we can anticipate the crime rate.

ALGORITHM

The Naive Bayes algorithm is a straightforward but effective machine learning method that relies on the Bayes theorem and the assumption of feature independence. The classification of texts, the detection of spam, sentiment analysis, and recommendation systems are all common classification tasks that use it. The algorithm is referred to as "naive" because it assumes that all features are independent of one another, meaning that the presence or absence of one feature has no bearing on the likelihood that another feature would also be present or absent. Naive Bayes still works well in many practical applications and frequently serves as a baseline model, despite the possibility that this assumption may not be accurate in real-world settings. The Naive Bayes method operates as follows:

The algorithm needs a training dataset with labels, where each instance is linked to a class label. The likelihood that each class will appear in the dataset is determined by calculating the prior probabilities for each class. The likelihood or conditional probability of each feature, given each class, is computed for each feature. The approach determines the most likely class for a new instance during the testing phase given feature values but no class label.

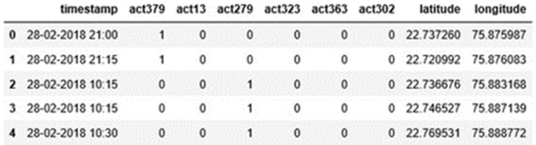
It calculates the posterior probability of each class for the given instance using Bayes' theorem and the assumptions of feature independence. The class with the highest posterior probability is assigned as the predicted class for the instance. The Naive Bayes algorithm uses probability theory to make predictions. It assumes that the probability distribution of each feature given the class is known. The algorithm usually assumes different probability distributions for different types of features.

The Naive Bayes algorithm is computationally efficient, requires relatively small amounts of training data, and can handle high-dimensional feature spaces. However, its strong independence assumption may lead to suboptimal performance in cases where the features are correlated. Additionally, it assumes that all features are equally important, which may not be true in certain scenarios. Despite these limitations, Naive Bayes can be a useful and effective algorithm, especially in situations where the independence assumption holds reasonably well and a fast and interpretable model is desired. Bayes' theorem is mathematically expressed by the following equation:

$$P(A|B) = {}^{P(B|A)\,P(A)}\!\big/_{P(B)} \qquad (1)$$

RESULTS and Discussion

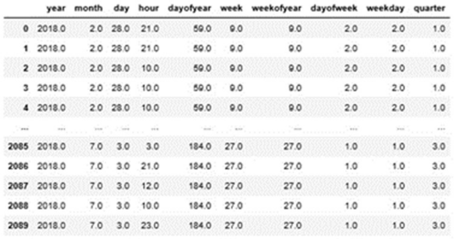
In this below figure, original data which is taken from the database is shown, in which time stamp, acts and the landmark is also shown. If the crime is done then it will be represented as 1 and if that crime is not happened in that particular time, then it will be represented as 0. Figure 1 shows the original data that contains distinct attributes with longitude and latitude.



	timestamp	act379	act13	act279	act323	act363	act302	latitude	longitude
0	28-02-2018 21:00	1	0	0	0	0	0	22.737260	75.875987
1	28-02-2018 21:15	1	0	0	0	0	0	22.720992	75.876083
2	28-02-2018 10:15	0	0	1	0	0	0	22.736676	75.883168
3	28-02-2018 10:15	0	0	1	0	0	0	22.746527	75.887139
4	28-02-2018 10:30	0	0	1	0	0	0	22.769531	75.888772

Fig 1: Original Data

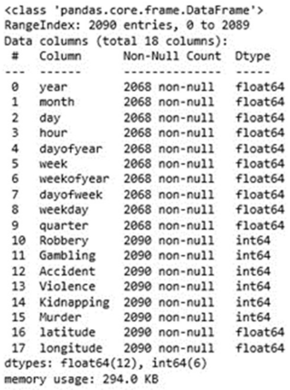
Data will be pre-processed as shown in the figure 2, all the data will be divided into different categories and if any data is not present then it will be filled with null values in this stage.



	year	month	day	hour	dayofyear	week	weekofyear	dayofweek	weekday	quarter
0	2018.0	2.0	28.0	21.0	59.0	9.0	9.0	2.0	2.0	1.0
1	2018.0	2.0	28.0	21.0	59.0	9.0	9.0	2.0	2.0	1.0
2	2018.0	2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0
3	2018.0	2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0
4	2018.0	2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0
...	...	...	...	...	...	...	...	...	...	...
2085	2018.0	7.0	3.0	3.0	184.0	27.0	27.0	1.0	1.0	3.0
2086	2018.0	7.0	3.0	21.0	184.0	27.0	27.0	1.0	1.0	3.0
2087	2018.0	7.0	3.0	12.0	184.0	27.0	27.0	1.0	1.0	3.0
2088	2018.0	7.0	3.0	10.0	184.0	27.0	27.0	1.0	1.0	3.0
2089	2018.0	7.0	3.0	23.0	184.0	27.0	27.0	1.0	1.0	3.0

Fig 2: Pre-processed data

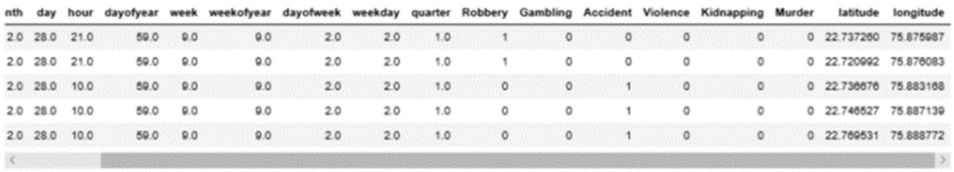
All the pre-processed data will be checked in this below figure 3, where all the null entries will be validated in this stage.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2090 entries, 0 to 2089
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   year        2068 non-null   float64
 1   month       2068 non-null   float64
 2   day         2068 non-null   float64
 3   hour        2068 non-null   float64
 4   dayofyear   2068 non-null   float64
 5   week        2068 non-null   float64
 6   weekofyear  2068 non-null   float64
 7   dayofweek   2068 non-null   float64
 8   weekday     2068 non-null   float64
 9   quarter     2068 non-null   float64
 10  Robbery     2090 non-null   int64
 11  Gambling    2090 non-null   int64
 12  Accident    2090 non-null   int64
 13  Violence    2090 non-null   int64
 14  Kidnapping  2090 non-null   int64
 15  Murder      2090 non-null   int64
 16  latitude    2090 non-null   float64
 17  longitude   2090 non-null   float64
dtypes: float64(12), int64(6)
memory usage: 294.0 KB
```

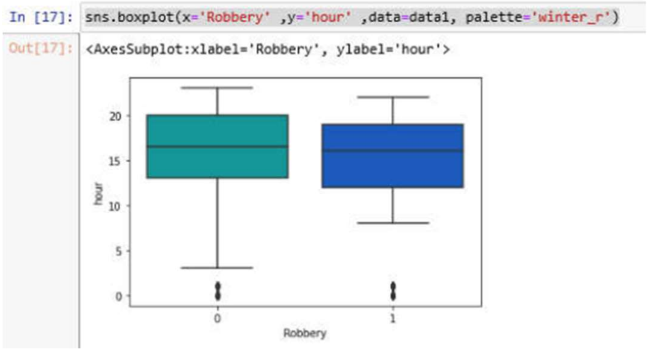
Fig 3: Null Entry Checking

A final dataset will be prepared from the available dataset, all the data will be processed as shown below (Figure 4).



nth	day	hour	dayofyear	week	weekofyear	dayofweek	weekday	quarter	Robbery	Gambling	Accident	Violence	Kidnapping	Murder	latitude	longitude
2.0	28.0	21.0	59.0	9.0	9.0	2.0	2.0	1.0	1	0	0	0	0	0	22.737260	75.875987
2.0	28.0	21.0	59.0	9.0	9.0	2.0	2.0	1.0	1	0	0	0	0	0	22.720992	75.876083
2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0	0	0	1	0	0	0	22.736676	75.883168
2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0	0	0	1	0	0	0	22.746527	75.887139
2.0	28.0	10.0	59.0	9.0	9.0	2.0	2.0	1.0	0	0	1	0	0	0	22.769531	75.888772

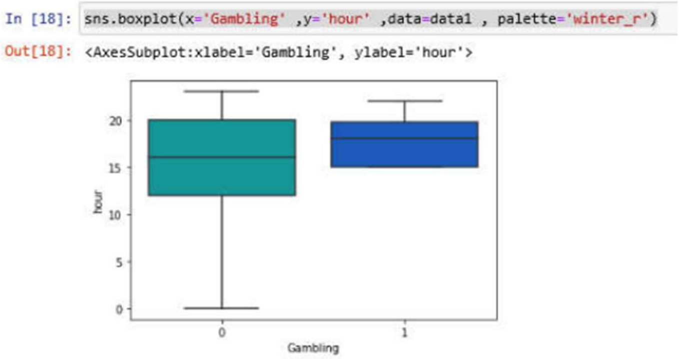
Fig 4: Removal of Null Values



```
In [17]: sns.boxplot(x='Robbery' ,y='hour' ,data=data1, palette='winter_r')
Out[17]: <AxesSubplot:xlabel='Robbery', ylabel='hour'>
```

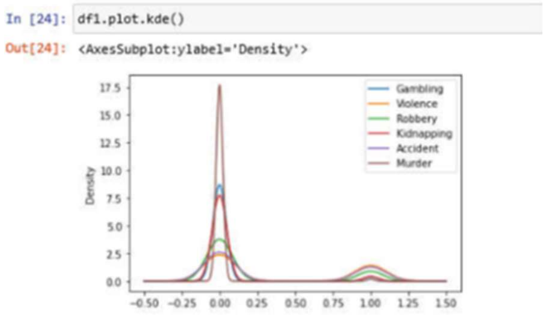
Fig 5: Robbery Related Crime

Box plot for gambling is shown in Figure 5, as the graph shows data in hours.



```
In [18]: sns.boxplot(x='Gambling' ,y='hour' ,data=data1 , palette='winter_r')
Out[18]: <AxesSubplot:xlabel='Gambling', ylabel='hour'>
```

Fig 6: Box plot for crime Gambling

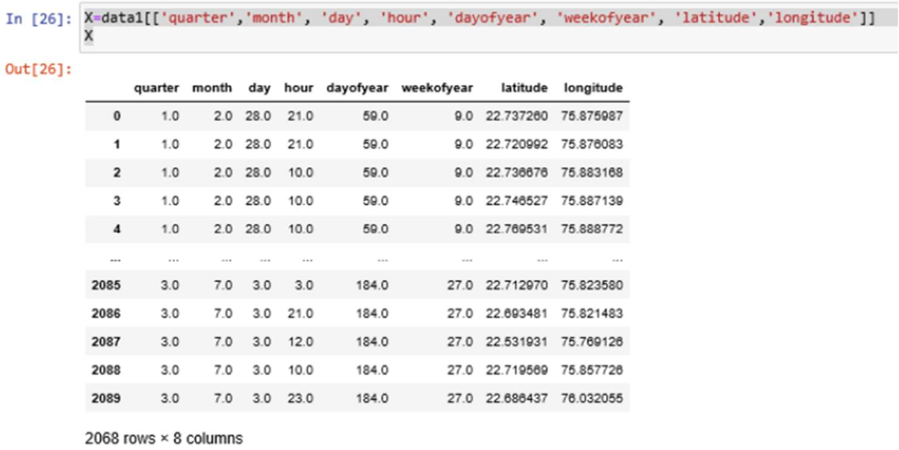


```
In [24]: df1.plot.kde()

Out[24]: <AxesSubplot:ylabel='Density'>
```

Fig 7: Plotting density among all the crimes

Dataset will be trained to predict the crimes that likely to happen with the location. All the data which contains of month, day, year, time and location will be shown.



```
In [26]: X=data1[['quarter','month', 'day', 'hour', 'dayofyear', 'weekofyear', 'latitude','longitude']]
         X

Out[26]:
```

	quarter	month	day	hour	dayofyear	weekofyear	latitude	longitude
0	1.0	2.0	28.0	21.0	59.0	9.0	22.737260	75.875987
1	1.0	2.0	28.0	21.0	59.0	9.0	22.720992	75.876083
2	1.0	2.0	28.0	10.0	59.0	9.0	22.736676	75.883168
3	1.0	2.0	28.0	10.0	59.0	9.0	22.746527	75.887139
4	1.0	2.0	28.0	10.0	59.0	9.0	22.769531	75.888772
...	...	...	...	...	...	...	...	...
2085	3.0	7.0	3.0	3.0	184.0	27.0	22.712970	75.823580
2086	3.0	7.0	3.0	21.0	184.0	27.0	22.693481	75.821483
2087	3.0	7.0	3.0	12.0	184.0	27.0	22.531931	75.769126
2088	3.0	7.0	3.0	10.0	184.0	27.0	22.719569	75.857726
2089	3.0	7.0	3.0	23.0	184.0	27.0	22.686437	76.032055

2068 rows × 8 columns

Fig 8: Training Set

Line plot of crime analysis is shown below, this figure shows all the crimes in this data set. It shows the relation between, rate of crimes per hour. Each crime data can also be separated by doing respective changes in the code.

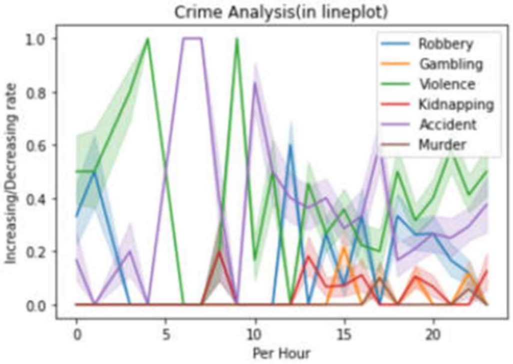
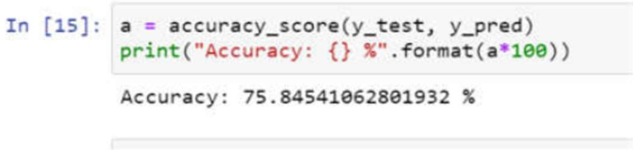


Fig 9: Plotting data distribution

Output of the accuracy can be seen by executing the program.



```
In [15]: a = accuracy_score(y_test, y_pred)
         print("Accuracy: {} %".format(a*100))

         Accuracy: 75.84541062801932 %
```

Fig 10: Output of the Accuracy

Hotspots of the crimes that are likely to happen can be seen in the google maps.

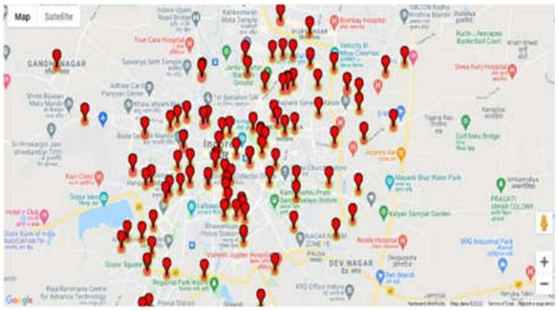


Fig 11: Hotspots on Google map

CONCLUSION

Two alternative dataset techniques were employed in this study to analyse crime data from the Indore police department that was obtained from a publicly accessible website during the previous years. To achieve a crime prediction accuracy of between 70 and 80%, machine learning predictive models such as linear regression, KNN, and naive bayes were applied. For various techniques and algorithms, the accuracy, complexity, and training time varied slightly. By fine-tuning the method and the data for applications, the prediction accuracy can be increased. Despite having poor prediction accuracy, this model nonetheless offers a foundation for additional research.

A method of law enforcement known as "crime prediction" identifies the crimes that are most likely to happen in the future using data and statistical analysis. There has been ongoing research in this area throughout the world. Using face recognition, one may foretell who will commit a crime and whether they will do so before it really happens. The report poses the question of whether more advanced AI/ML will ultimately result in more accurate predictions or if it will only exacerbate current issues. Any system will be based on real-world data, but if those data are produced by biassed police officers, the AI/ML may also be biassed.

REFERENCES

A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," Proc. of the 16th Intl. Conf. on Multimodal Interaction, pp. 427-434, 2014.

H. Adel, M. Salheen, and R. Mahmoud, "Crime in relation to urban design. Case study: the greater Cairo region," Ain Shams Eng. J., vol. 7, no. 3, pp. 925-938, 2016.

"Overall crime rate in Vancouver went down in 2017, VPD says," CBC News, Feb. 15, 2018. [Online] Available: https://www.cbc.ca/news/canada/britishcolumbia/crime-rate-vancouver2017-1.4537831. [Accessed: 09- Apr- 2023].

J. Kerr, "Vancouver police go high tech to predict and prevent crime before it happens," Vancouver Courier,
July 03, 2023. [Online] Available: https://www.vancourier.com/news/vancouver-policego-high-tech-topredict-and-prevent-crime-before-ithappens-1.21295288. [Accessed: 09- Apr- 2023]
[5] J. Han, Data mining: concepts and techniques, Morgan Kaufmann, 2012.

R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219- 4225, Mar. 2013.

H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," IEEE Computer, vol. 37, no. 4, pp. 5056, Apr. 2004.

T. Beshah and S. Hill, "Mining Road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia," Proc. of Artificial Intell. for Develop. (AID 2010), pp. 14-19, 2010.

M. Al Boni and M. S. Gerber, "Area-specific crime prediction models," 15th IEEE Intl. Conf. on Mach. Learn. and Appl., Anaheim, CA, USA, Dec. 2016.