CRIME RATE PREDICTION BY USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the

award of

Bachelor of Engineering degree in Computer Science and Engineering

By

P.HARIPAVAN(REG. NO. 39110755)

P.VARUN ACHYUTH RAM(REG. NO. 39110751)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI

– 600119

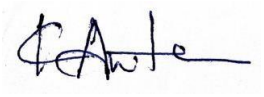
APRIL - 2023

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **P.HARIPAVAN (39110755) and P.VARUN ACHYUTH RAM (39110751)** who carried out the Project Phase-1 entitled "**CRIME RATE PREDICTION BY USING MACHINE LEARNING"** under my supervision from January 2023 to April 2023.



Internal Guide

Ms. K. ANITA DAVMANI, M.E., (Ph.D).

Head of the Department

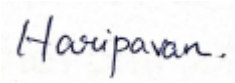
Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on_____

InternalExaminer **ExternalExaminer**

DECLARATION

I,**P.HARIPAVAN (Reg.No- 39110755),** here by declare that the Project Phase-2 Report entitled **"CRIME RATE PREDICTION BY USING MACHINE LEARNING"** done by me under the guidance of **Ms. K. ANITA DAVMANI, M.E., (Ph.D).** is submitted in partial fullfilment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.



DATE:

PLACE: Chennai **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA INSTITUTE OF SCIENCE AN TECHNOOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D**, **Dean**, School of Computing, **Dr. L. Lakshmanan, M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms. K. ANITA DAVMANI, M.E., (Ph.D).** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

The criminal cases in India are increasing rapidly due to which number of cases pending are also piling up. This continuous increase in the criminal cases is proving to be difficult to be classified and to be solved. Recognizing the criminal activity patterns of a place is important in order to prevent it from happening. The crime solving agencies can do a better work if they have a good idea of the pattern of criminal activities that are happening in a particular area. To be better prepared to respond to criminal activity, it is important to understand patterns in crime. In our project, we analyze crime data from the city of Indore, scraped from publicly available website of Indore Police. At the outset, the task is to predict which category of crime is most likely to occur given a time and place in Indore. The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand. The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology. We also attempt to make our classification task more meaningful by merging multiple classes into larger classes. Finally, we report and reflect on our results with different classifiers, and dwell on avenues for future work.

TABLE OF CONTENTS

CHAPTER-1
INTRODUCTION

1.1 General

Paasbaan which is an Urdu word meaning protector, many important questions in public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, in and more concerted efforts by citizens and authorities to create healthy neighbourhood environments. With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

The inputs to our algorithms are time (hour, day, month, and year), place (latitude and longitude),and class of crime:

Act 379 – Robbery Act
13 - Gambling Act 279
- Accident Act 323 -
Violence Act 302 -
Murder
Act 363 - Kidnapping

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbors), Decision Trees, and Random Forests.

We also perform multiple classification tasks – we first try to predict which of 6 classes of crimes are likely to have occurred, and later try to differentiate between the violent and non-violent crimes.
.

Madhya Pradesh's commercial capital Indore has topped the crime record in the country in 2008 followed by Bhopal and Jaipur. Crime rate of India was 941.4, which is the highest in the country according to National Crime Record Bureau's (NCRB) report - "Crime in India 2002 to 2010".

With the rapid urbanization and development of big cities and towns, the graph of crimes is also on the increase. This phenomenal rise in offences and crime in cities is a matter of great concern and alarm to all of us. There are robberies, murders, rapes and what not. 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 common citizens to 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 in an organized way and sometimes even have nationwide and international connections and links.

The objective of our work is to:

1. Predicting crime before it takes place.
2. Understanding crime pattern.
3. Classify crime based on location.
4. Analysis of crime in Indore.

After the preprocessing described in the previous sections, we had three different classifications problems to solve, which we proceeded to attack with an assortment of classification algorithmsThe following are the algorithms which we are using:

KNN(K- Nearest neighbors)

Decision Tree

Random Forests

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets.We are surrounded by a machine learning based technology: search engines learn how to bring us the best results (while placing pro_table ads), anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms.

Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, finedetailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired or re_ned through learning from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to learn and adapt

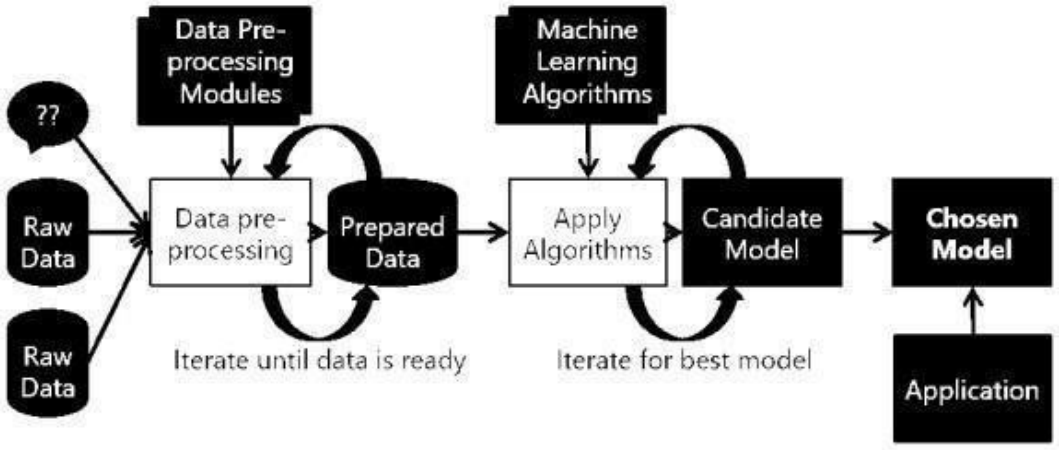


Fig 1.1-Machine learning process

The inputs to our algorithms are time (hour, day, month, year), place (latitude and longitude), class of crime

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

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbors), Decision Trees, and Random Forests.

Dataset which we are using is scraped daily from website of Indore police which is publically available. But the dataset is Hindi and in order to perform machine learning this data cannot be used as it is. Hence the data needs to be processed.


crime_data.csv

Features of this dataset :

1. Police Station
2. Police Station identification number
3.I.P.C. act number 4.Complainant name & address 5.Accused name & address
6.Incident place
7.Incident date & time
8. Reporting date & time

Reason of Time delay in reporting to police Incident information in brief

Before implementing machine learning algorithms on our data, we went through a series of preprocessing steps with our classification task in mind. These included:

Dropping features such police station, station number, Complainant name & address ,Accused name & address

Dropping features such as Resolution, Description and Address: The resolution and description of a crime are only known once the crime has occurred, and have limited significance in a practical, real-world scenario where one is trying to predict what kind of crime has occurred, and so, these were omitted. The address was dropped because we had information about the latitude and longitude, and, in that context, the address did not add much marginal value.

The timestamp contained the year, date and time of occurrence of each crime. This was decomposed into five features: Year (2018), Month (1-12), Date (1-31), Hour (0- 23) and Minute (0-59).

Following these preprocessing steps, we ran some out-of-the box learning algorithms as a part of our initial exploratory steps. Our new feature set consisted of 9 features, all of which were now numeric in nature.

timestamp	act379	act13	act279	act323	act363	act302	latitude	longitude
28-02-2018 21:00	1	0	0	0	0	0	22.73726	75.87599
28-02-2018 21:15	1	0	0	0	0	0	22.72099	75.87608
28-02-2018 10:15	0	0	1	0	0	0	22.73668	75.88317
28-02-2018 10:15	0	0	1	0	0	0	22.74653	75.88714

Table 1.2: Dataset after Preprocessing

The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand. The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

Possible avenues through which to extend this work include time-series modelling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes.

For Example: it could be the case that two or more classes of crimes surge and sink together,which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement. This kind of data analysis was technologically impossible a few decades ago, but the hope is that recent developments in machine learning are up to the task.

Public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and more concerted efforts by citizens and authorities to create healthy neighborhood environments. With the advent of the Big Data era and the availability of fast,efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

The remaining section of the report is structured as follows:

- **Chapter 2** provides detailed business and technical requirements
- **Chapter 3** provides analysis and design of this project
- **Chapter 4** provides Construction, implementation details of this project
- **Chapter 5** provides Conclusion and future scope as well as future application of this project

CHAPTER-2
LITERATURE SURVEY

The forecasting of crime levels at the spatial scale, such as within street segments or communities, is one of the many fields that has seen widespread use of geospatial analysis. This is only one of the numerous applications that exist. This is only one example out of the many different applications that geospatial analysis has been put to use for. One example of a temporal pattern that may be found in crime rates is an increase in violent crime after a period of relative calm. According to the findings of a number of studies, it has been shown that time series models like as ARIMA and LSTM are able to capture seasonal and trend components of crime rates. These findings were published in a number of academic journals.

[1] Singh and Tiwari (2019) conducted the research in order to evaluate the effectiveness of using methods that make use of machine learning in order to anticipate criminal behaviour. The study was carried out in order to investigate the efficacy of using such methods. They explored a range of ML approaches, in order to assess the merits and downsides of utilising each one to predict criminal behaviour.

[2] Chen, Wang, and Mao (2018) came up with a hybrid model in order to predict criminal behaviour. DL & DT are both used into this model in order to improve its ability to forecast criminal behaviour. They were able to extract the temporal trends from the crime data by using a neural network equipped with LSTM. Following that, they fed the results of that network into a DT and let the tree to use the knowledge it had gained to make predictions based on the data it had gathered.

[3] Bello-Orgaz, Jung, & Camacho (2017) used DT in their research to analyse large amounts of social data and to make predictions about the likelihood of significant events occurring in metropolitan populations. They started the process of categorising the event by locating it and establishing the time it took place using information that they acquired from Twitter. The next step was to determine the nature of the occurrence by constructing a decision tree.

[4] Natarajan and Ravi (2018) presented a method that is based on ML for predicting criminal behaviour in India. This method can be found in their research paper. This methodology was designed specifically for the purpose of analysing data from India. Since doing research on a variety of factors related to crimes, such as area, time, the nature of the crimes, and the weather patterns, they utilised a decision tree to forecast the number of crimes that will occur. This was accomplished after taking into consideration all of the relevant factors.

[5] Gerber (2014) drew his results by combining data from Twitter with an algorithm that calculated the density of kernels. He was able to determine the places in the city that had the

highest crime rates by using geotagged tweets. When he had found those places, he used kernel density estimation to forecast future crime rates in those areas. This was done after he had found those areas.

[6] Using machine learning, Soria-Comas, González-Abril, and Pérez-Sánchez (2018) were able to make predictions about the likelihood of future criminal behaviour. They were able to do this by using a dataset that included instances of recidivism and carrying out an analysis of many criteria, such as a person's demographics, their criminal history, and their mental health status. Both of these factors were taken into consideration. They were able to reach an accuracy of 76% in their predictions by using a DT, which was a significant improvement over the results they had gotten in the past.

[7] Ma and Jiang (2017) presented an innovative method for forecasting criminal behaviour. [Citation needed] In order to arrive at reliable conclusions using their approach, the researchers combined a spatial-temporal clustering analysis with a decision tree. They first utilised clustering analysis to identify geographical and temporal trends in the data on crimes, and then they used a DT to produce predictions based on those patterns. Clustering analysis was used to find the trends in the data on crimes. In order to identify patterns in the data on crimes, a clustering analysis was carried out.

[8] Akter and Haque performed a research on the use of ML to the prediction of criminal behaviour as part of the work that they undertook for their article that was released in 2019. The paper was published in 2019. They carried out research on a number of different ml approaches in order to evaluate the applications of these algorithms in the prediction of various sorts of criminal behaviour. In particular, they were interested in finding out how effective these algorithms were in predicting a wide variety of criminal behaviours.

[9] Sohel, Akter, Uddin, and Bhowmik developed a crime prediction system in their research from 2017, which made use of both the K-nearest neighbour approach and the decision tree algorithm. They conducted an examination of the data on crimes carried out in the city of Dhaka, which is located in Bangladesh, and they discovered a variety of indicators, including the location of the crime, the time of the crime, and the kind of crime. Using the use of the decision tree methodology, they were able in achieving an accuracy rate of 85%.

[10] In an article that was written by Kornblum, Pascarella, and Foote and published in 2015, they discussed how essential it is for organisations who deal with law enforcement to have the ability to properly forecast future criminal conduct. They assessed a range of different crime prediction systems, such as the PredPol system and the HunchLab system, in order to determine the influence that each of these systems had on the quantity of criminal activity that was reduced as a result of their use. They also examined the ethical considerations that are related with crime prediction systems as well as the repercussions that such systems have for civil rights. Another topic that was explored was the ramifications that such systems have for civil rights. In addition to this, they spoke about the effects that pattern detection systems have had and continue to have on civil rights.

CHAPTER-3

ANALYSIS AND DESIGN

Use case diagram represent the overall scenario of the system. A scenario is nothing but a sequence of steps describing an interaction between a user and a system. Thus use case is a set of scenario tied together by some goal. The use case diagram are drawn for exposing the functionalities of the system.

A Use case diagram is a unified modelling language (UML) is a type of behavioral diagram defined by and created from a Use case diagram analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms cases. And any dependencies between those use cases. The main purpose of use case diagram to show what system function are performed by crimes

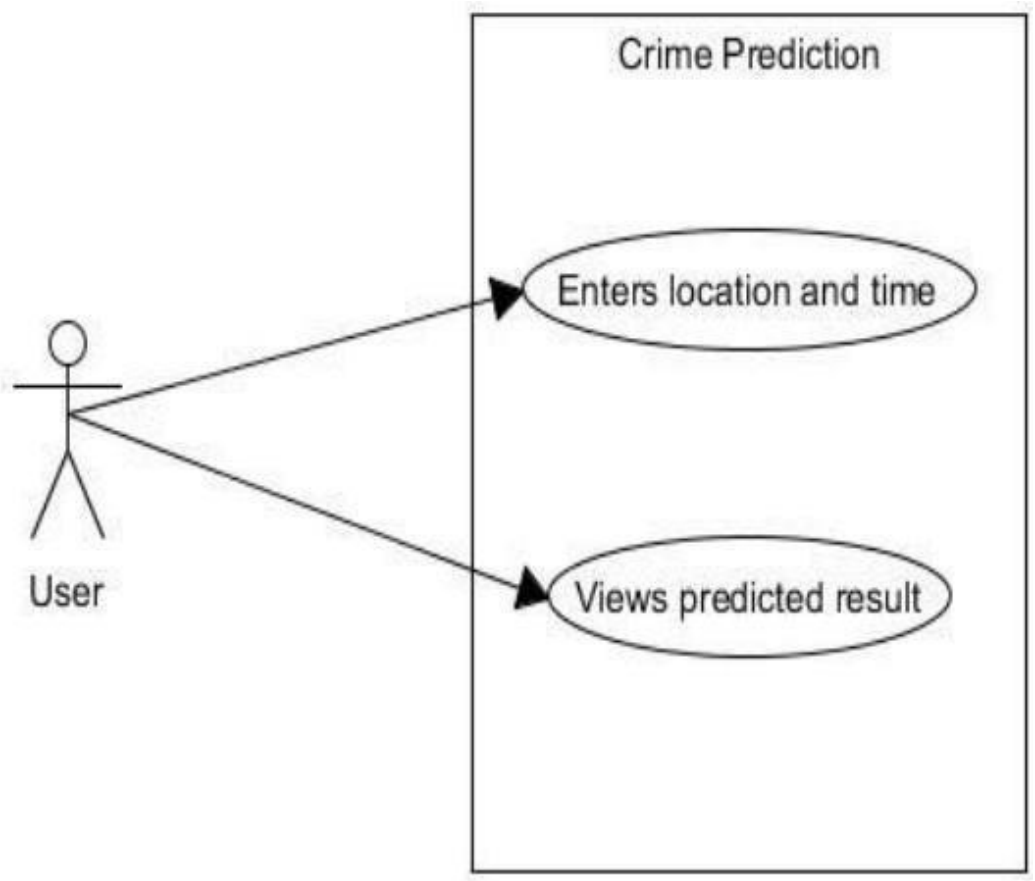


Fig 3.1-Use case diagram of Paasbaan

The activity diagram is a graphical representation for representing the flow of interaction within specific scenatios. It is similar to a flowchart in which various activities that can be performed in the system are represented.

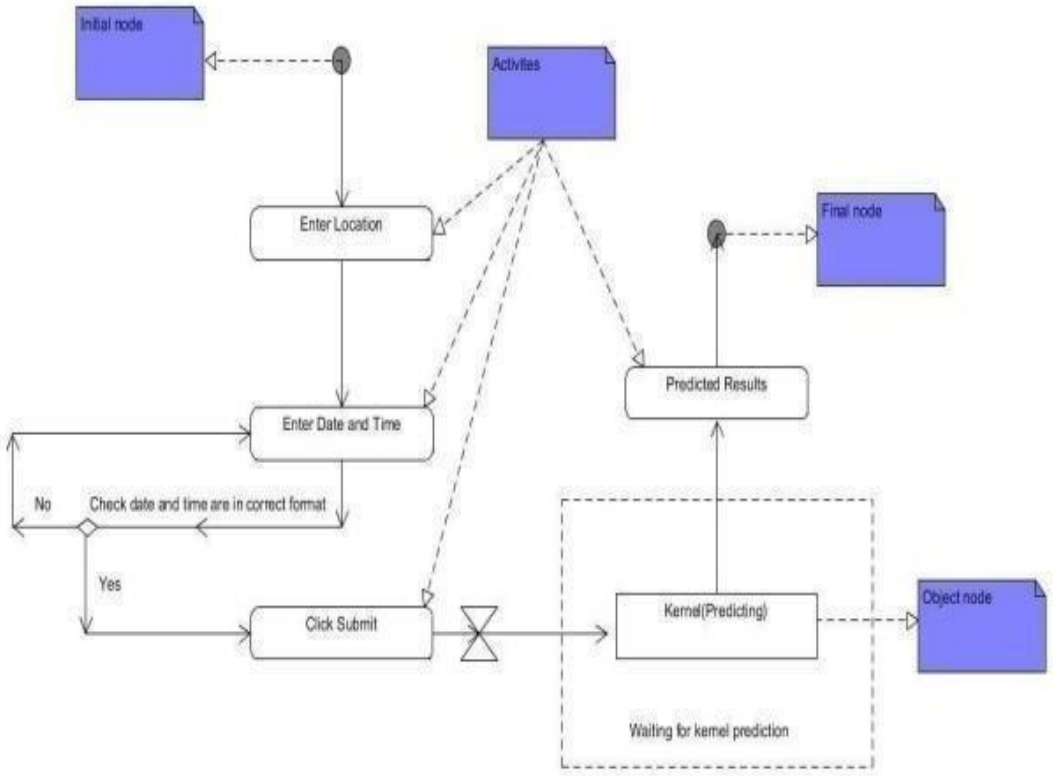


Fig 3.2-Activity diagram of Paasbaan

3.2.1 Example of Activity diagram :

An activity diagram visuvally presents a series of actions to control in a system similar to a flow chart or a data flow diagram. Activity diagram are often used in business process modelling. They can also describe the steps in a use case diagram. Activities modelled can be sequential and concurrent

In the sequence diagram how the object interacts with the other object is shown. There are sequence of events that are represented by a sequence diagram. It is a time oriented view of the interation between objects to accomplish a behavioural goal of the system.

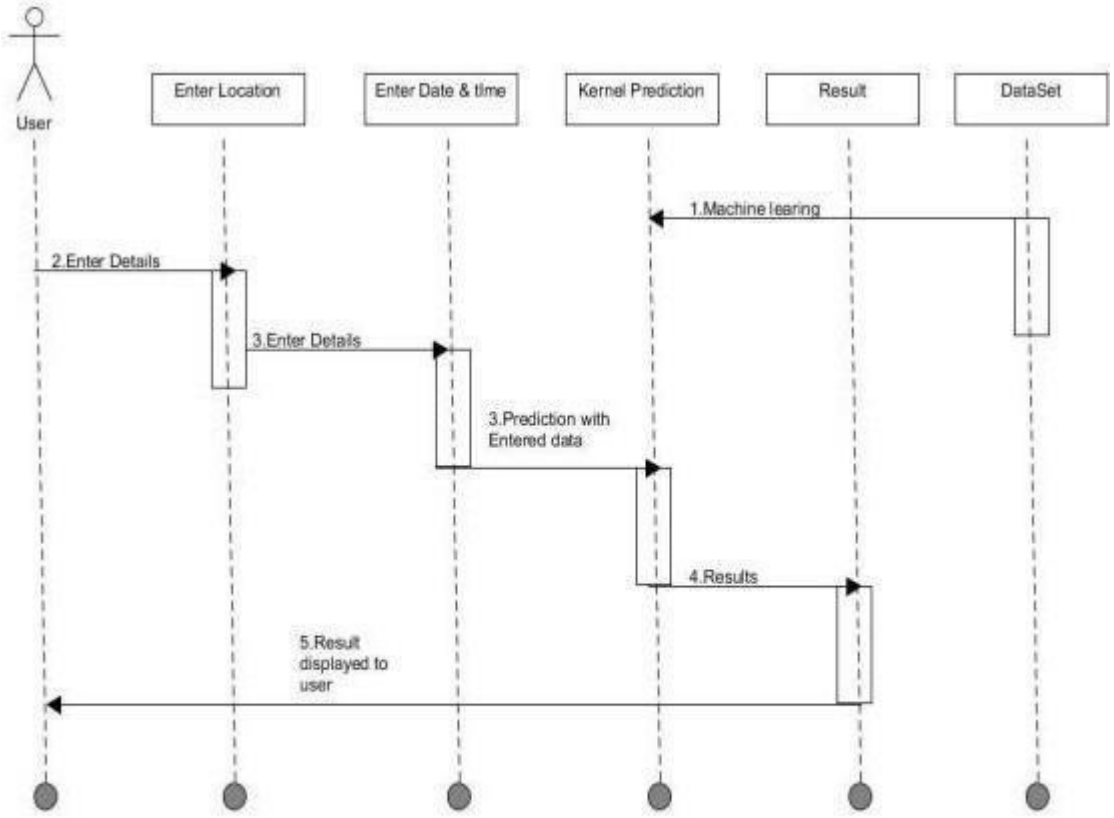


Fig 3.3-Sequence diagram of Paasbaan

3.3.1 Sequence diagram Example:

Sequence diagrams, commonly used by developers, model the interactions between objects in a single use case. They illustrate how the different parts of a system interact with each other to carry out a function,and the order in which the interactions occur when a particular use case is executed. In simpler words, a sequence diagram shows how different parts of a system work in a ‘sequence’ to get something done. shows different parts of a system work in a ‘sequence’ to get something done.

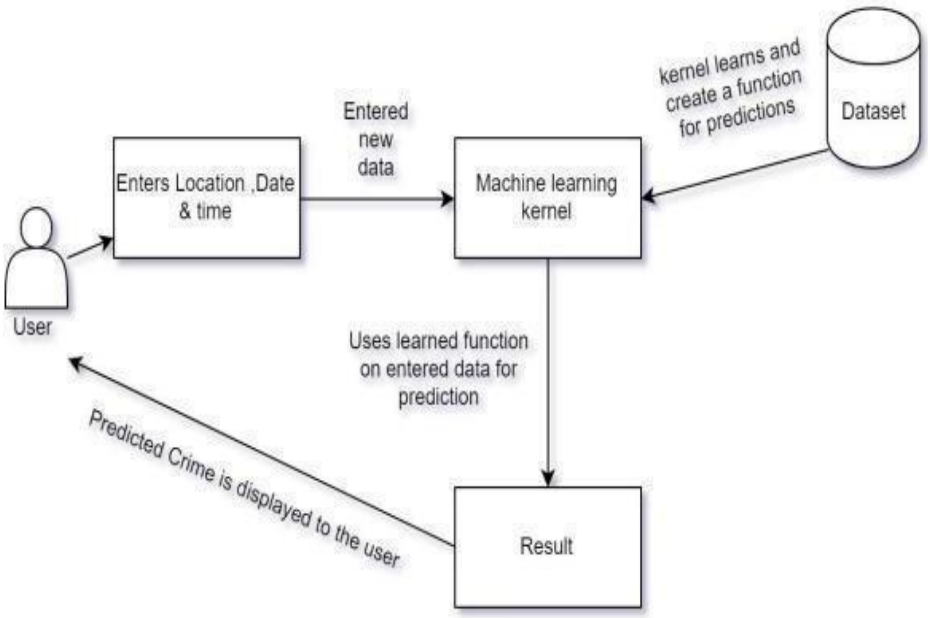


Fig 3.4-System architecture of Paasbaan

The system architectural design is the design process for identifying the subsystems making up the system and framework for subsystem control and communication. The goal of the architectural design is to establish the overall structure of software system. By outlining the specifics of how the application should be constructed, the software design will be utilised to assist in the software development of an android application. Use case models sequence diagrams, and other supplementary requirement data are included in the software design specifications, which are narrative and graphical documentation of the software design

All the images are first pre processed. Then it goes through feature extraction where H ear cascade is used. The video is captured from the surveillance camera which are converted into frames. When a face is detected in a frame, it is pre processed. Then it goes through feature extraction where Hear cascade is used. he features of the processed real-time image is compared with the features of processed images which are stored in the citizen database. If a match is found, it is further compared with the features of images stored in a local watch list database to identify if the person is criminal or not. If he is criminal a notification is sent to the police personnel with all the details and the time for which he was under the surveillance of the camera. If he is not a citizen of that country, it is then compared with the features of images stored in the international watch list database.

CHAPTER-4

IMPLEMENTATION

The implementation of the project is done with the help of python language.To be particular,for the purpose of machine learning Anaconda is being used. Anaconda is one of several Python distributions. Anaconda is a new distribution of the Python. It was formerly known as Continuum Analytics. Anaconda has more than 100 new packages. Anaconda is used for scientific computing, data science, statistical analysis, and machine learning.

On Python technology, we found out Anaconda to be easier. Since it helps with the following problems:

1. Installing Python on multiple platforms.
2. Separating out different environments.
3. Dealing with not having correct privileges.
4. Getting up and running with specific packages and libraries.

This data was scraped from the publically available data from Indore police website which had been made by people in police station of different areas. Implementation of the idea started from the Indore city itself so as to limit an area for the prediction and making it less complex. The data was sorted and converted into a new format of timestamp, longitude, latitude, which was the input that machine would be taking so as to predict the crime rate in particular location or city

The entries was done just to make the machine learn what all it has to do with the data and what actually the output is being demanded. As soon as the machine learnt the algorithms and the process, accuracy of different algorithms were measured & the algorithm with the most accuracy is used for the prediction kernel i.e. Random forest.

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

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

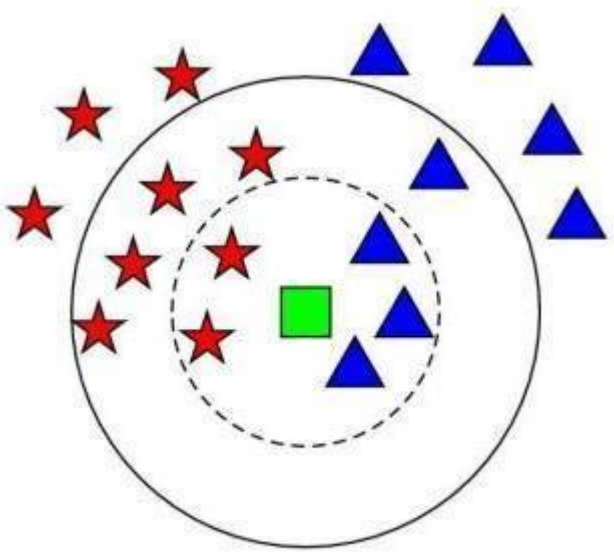


Fig 4.1.1 Principle diagram of KNN

KNN: Classification Approach

An object (a new instance) is classified by a majority votes for its neighbor classes.

The object is assigned to the most common class amongst its K nearest neighbors.(measured)

A powerful classification algorithm used in pattern recognition K nearest neighbors stores all available cases and classifies new cases based on a similarity measure (e.g. distance function).One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method).The classification method of estimating the likehood that the data point will become a member of one grouo or another based the group data point nearest to it belong to. The k-nearest neighbour algorithm is type of supervised machine learning algorithm used to solve classification and regression problems. However its mainly used for classification problems.

Its considered as non-parametric method because it doesn't make any assumptions about under lying data distribution. Simply put KNN tries to determines what group a data point belongs to by looking at the data point around it. The k-nearest neighbours algorithm is highly susceptible to over fitting due to the curse of dimensionality. However this problem can be resolved with the brute force implementation of KNN algorithm. but it isn't practical for large datasets.

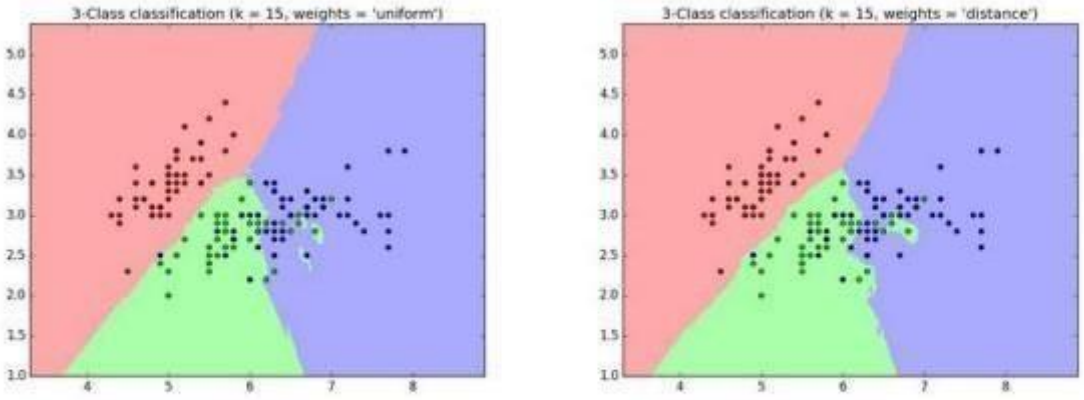


Fig 4.1.2 Shows graphical representation of KNN

Some frequently used distance functions.	
Camberra : $d(x,y)=\sum_{i=1}^{m}\frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$ (2)	Euclidean : $d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$ (5)
Minkowsky : $d(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^r\right)^{1/r}$ (3)	Manhattan / city - block : $d(x,y)=\sum_{i=1}^{m}\lvert x_i-y_i\rvert$ (6)
Chebychev : $d(x,y)=\max_{i=1}^{m}\lvert x_i-y_i\rvert$ (4)	

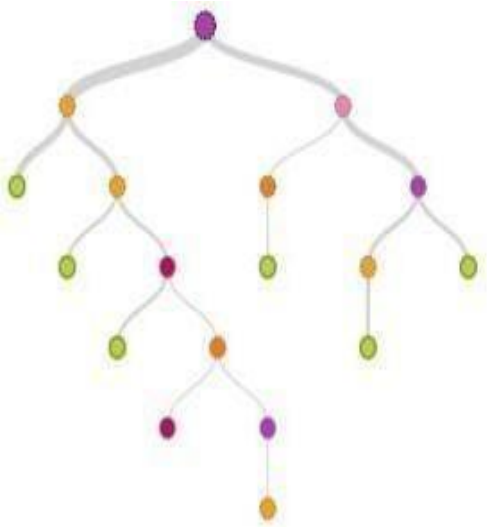
Fig 4.1.3 Distance functions

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.

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

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.



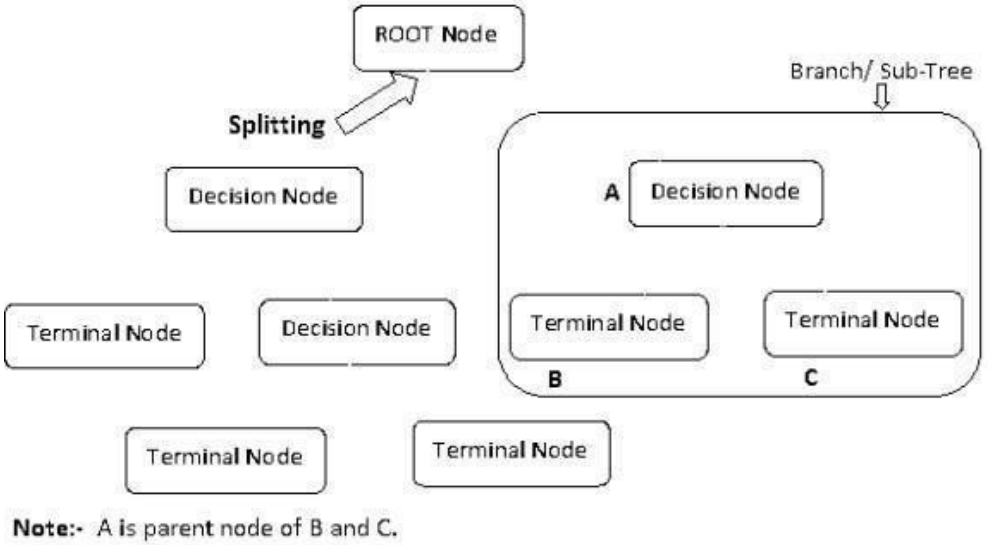


Fig 4.2.2 Decision Tree example

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|θ1),h(x|θ2),..h(x|θk). Each member of the family, h(x|θ), is a classification tree and k is the number of trees chosen from a model random vector.

Also, each θk 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 Dθk(x,y) D(x,y) of the training dataset. ⊂

Thus, h(x|θk) is the kth classification tree which uses a subset of features xθk x to build a classification model. Each tree then works like regular decision trees: 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 aggregating the results thus:

4.2.3 Random forest Formula :

$$y = \text{argmax}_{p \in \{h(x_1)..h(x_k)\}} \{ \sum_{j=1}^{\kappa} (I(h(x|\theta_j) = p)) \}$$

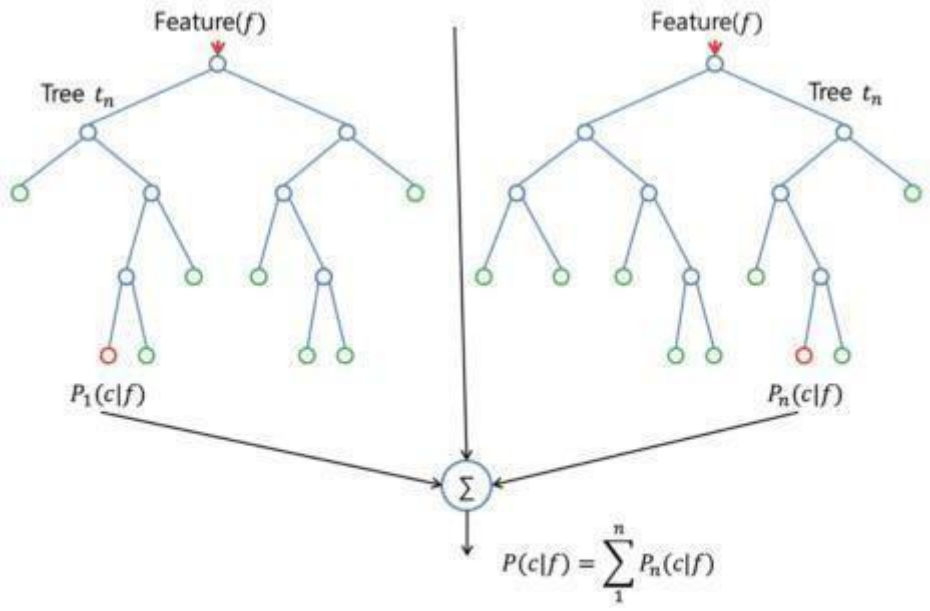


Fig 4.3.1 Random Forest Example

4.3.1 Random forest Example

A Random forest Algorithm is a supervised language machine learning algorithm that is extremely popular and it is used for classification and regression problems in machine learning. we know that the crimes are comprises different crimes, and the more crimes more it will be robust.

Similarly, the greater the number of Crimes in random forest algorithm, the higher its accuracy and problem-solving ability. Random forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the prediction of accuracy dataset. It is based on concept of ensemble learning which is process of combining multiple classifiers to solve a complex of problem to improve the performance of the model

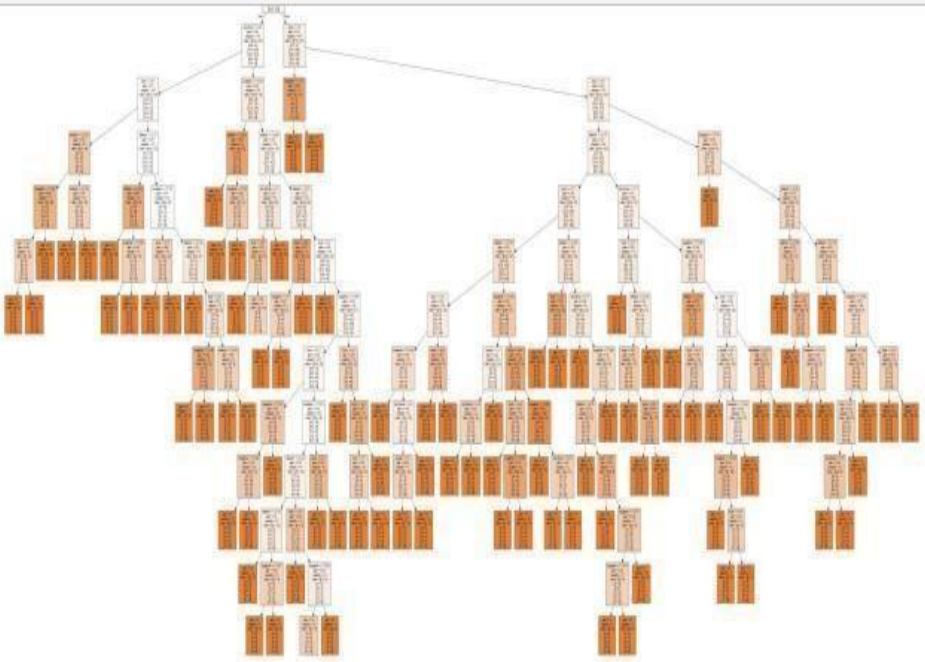


Fig 4.3.2 Decision Tree of Paasbaan

4.2.4 Data Visualization

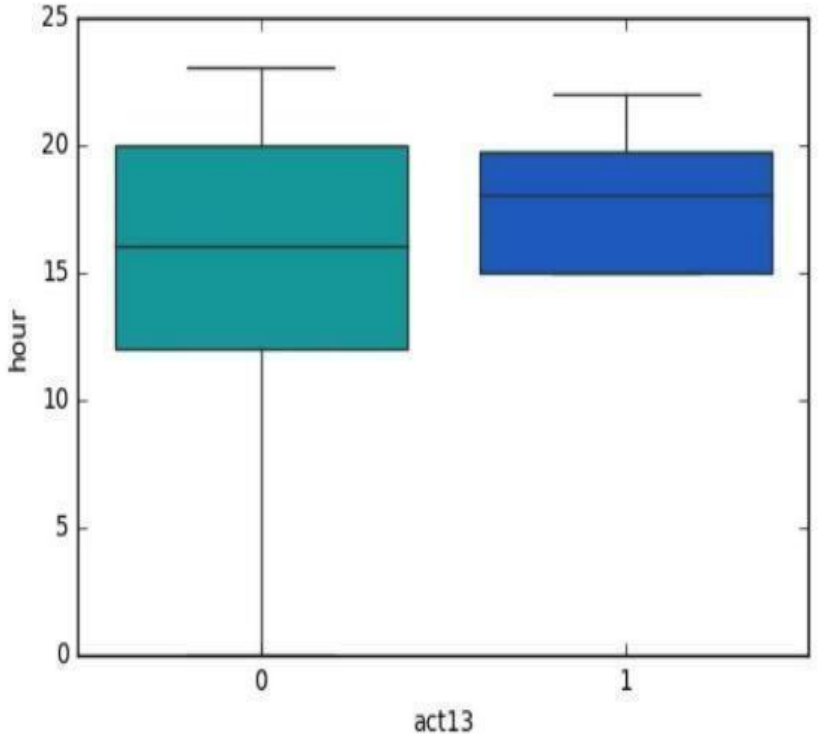


Fig 4.4.1 Act13(Gambling vs Hour)

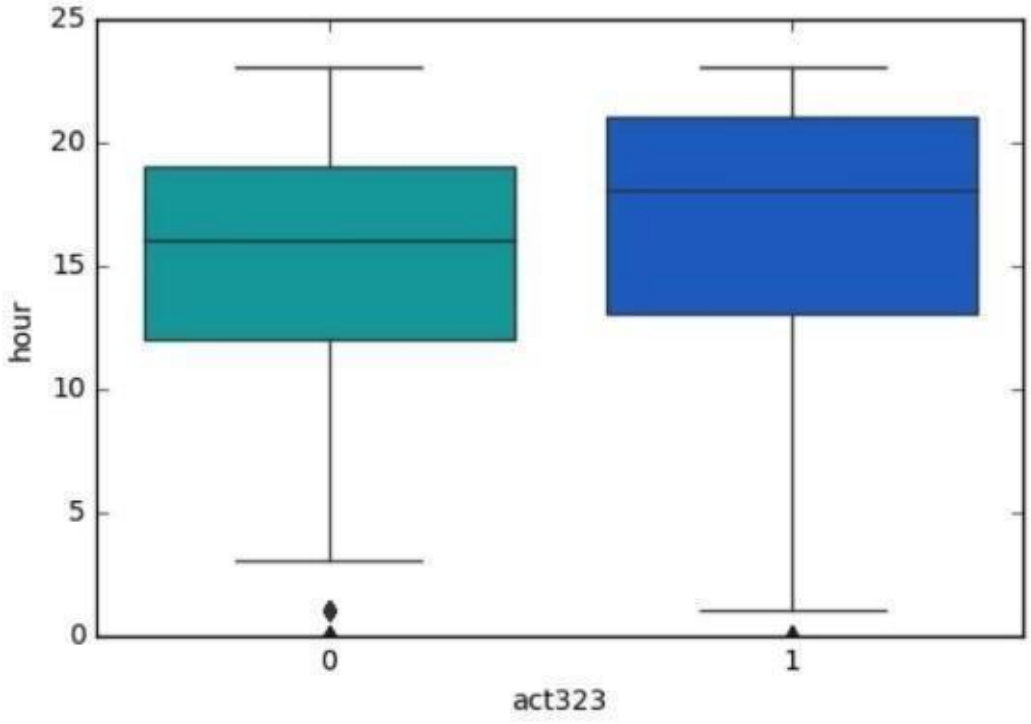


Fig 4.4.2 Act323(Violence vs Hour)

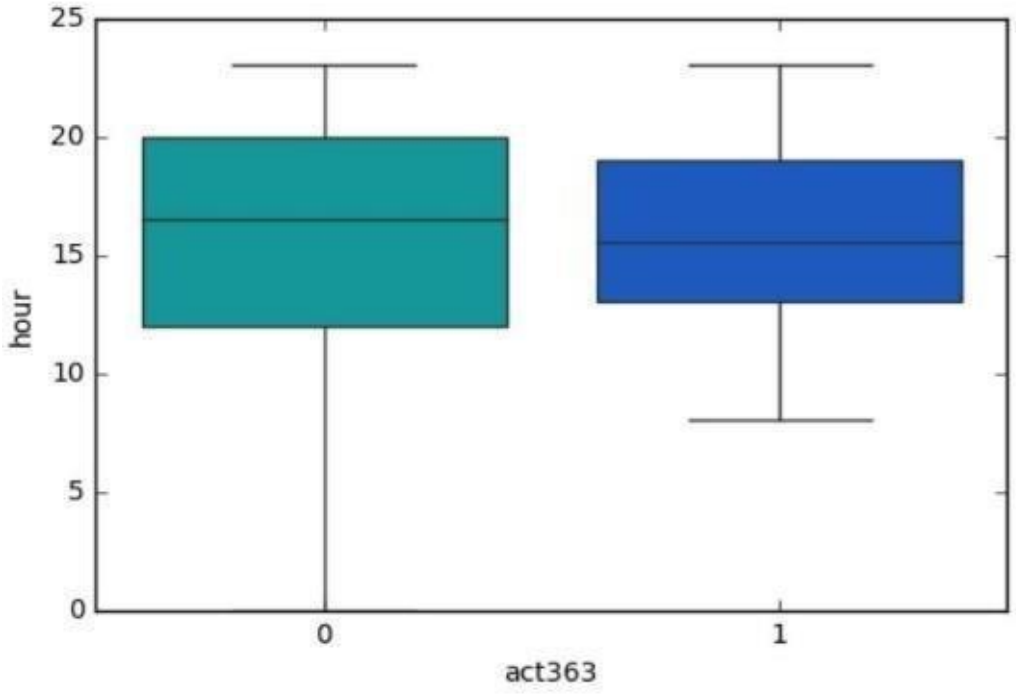


Fig 4.4.3 Act363(Kidnapping vs Hour)

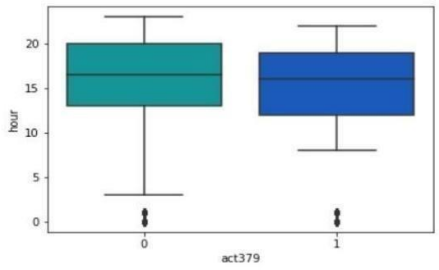


Fig 4.4.4 Act379(Robbery vs Hour)

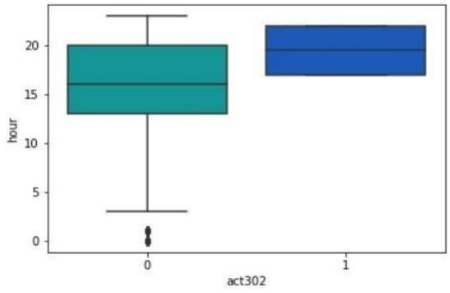


Fig 4.4.5 Act302(Murder vs Hour)

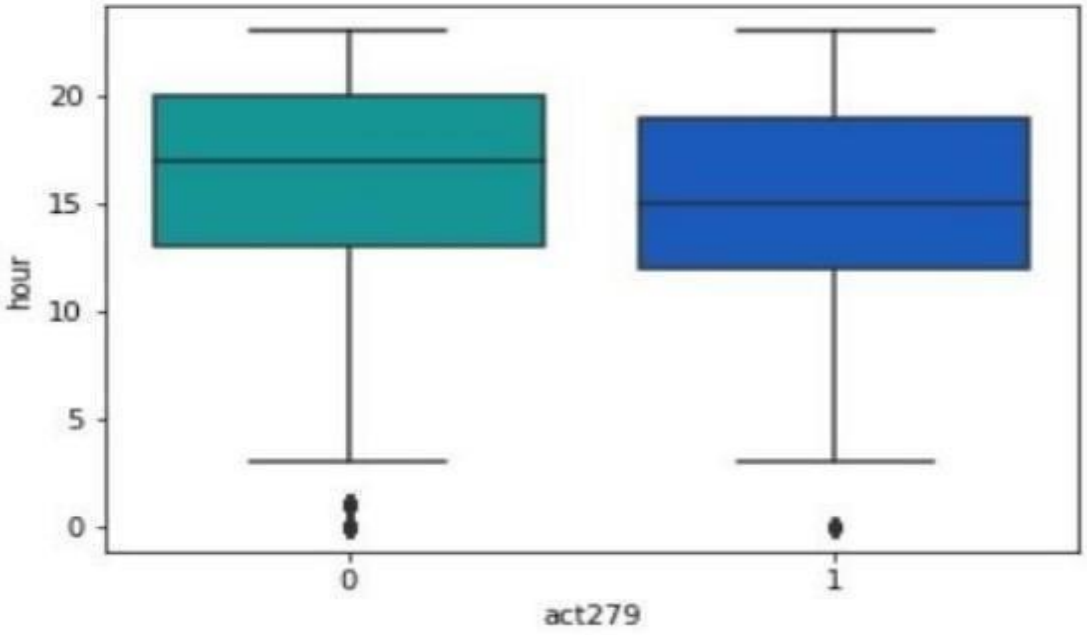


Fig4.4.6 Act279(Accident vs Hour)

\

.

The development of software involves a series of production activities were opportunities for injection of human fallibilities are enormous.

Error may begin to occur at very inspection of the process where the objective may be enormously or imperfectly specified as well as in lateral design and development stage. Because of human inability to perform and communicate with perfection, software development quality assurance activities.

Software testing is a crucial element of software quality assurances and represents ultimate review of specification, design and coding.

4.5.1 White box testing

It focuses on the program control structure. Here all statement in the project have been executed at least once during testing and all logical condition have been exercised.

This is designed to uncover the error in functional requirements without regard to the internal working of

the project. This testing focuses on the information domain of the project , deriving test case by partitioning the input and output domain of programming – A manner that provides through test coverage

The testing can be done at system, integration and unit levels of software development. One of the basic goals of white box testing is to verify a working flow for an application.it involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output

In white box testing we have divided into two basic steps to give a simplified explanation

1. Understand the source code
2. Create test case and execute

CHAPTER-5

CONCLUSION AND FUTURE

SCOPE

The initial problem of classifying 6 different crime categories was a challenging multi-class classification problem, and there was not enough predictability in our initial data-set to obtain very high accuracy on it. We found that a more meaningful approach was to collapse the crime categories into fewer, larger groups, in order to find structure in the data. We got high accuracy and precision on Prediction. However, the Violent/Non-violent crime classification did not yield remarkable results with the same classifiers – this was a significantly harder classification problem. Thus, collapsing crime categories is not an obvious task and requires careful choice and consideration.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes – for example, it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

The goal of any society shouldn't be to just catch criminals but to prevent crimes from happening in the first place

5.2.1 Predicting Future Crime Spots:

By using historical data and observing where recent crimes took place we can predict where future crimes will likely happen. For example a rash of burglaries in one area could correlate with

more burglaries in surrounding areas in the near future. System highlights possible hotspots on a map the police should consider patrolling more heavily

Fig 5.1 Predicting Surges

5.2.2 Predicting Who Will Commit a Crime:

Using Face Recognition to predict if a individual will commit a crime before it happens. The system will detect if there are any suspicious changes in their behavior or unusual movements. For example if an individual seems to be walking back and forth in a certain area over and over indicating they might be a pickpocket or casing the area for a future crime. It will also track individual over time.

5.2.3 Pretrial Release and Parole:

After being charged with a crime, most individuals are released until they actually stand trial. In the past deciding who should be released pretrial or what an individual's bail should be set at is mainly now done by judges using their best judgment. In just a few minutes, judges had to attempt to determine if someone is a flight risk, a serious danger to society, or at risk to harm a witness if released. It is an imperfect system open to bias. The media organization's analysis indicated the system might indirectly contain a strong racial bias. They found, "That black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent)." The report raises the question of whether better AI/ML can eventually produce more accurate predictions or if it would reinforce

existing problems. Any system will be based off of real world data, but if the real world data is generated by biased police officers, it can make the AI/ML biased.

The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement. This kind of data analysis was technologically impossible a few decades ago, but the hope is that recent developments in machine learning are up to the task.

The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand. The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative.

Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes – for example, it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

REFERENCES

Bogomolov, Andrey and Lepri, Bruno and Staiano, Jacopo and Oliver, Nuria and Pianesi, Fabio and Pentland, Alex.2014. Once upon a crime: Towards crime prediction from demographics and mobile data, Proceedings of the 16th International Conference on Multimodal Interaction.

Yu, Chung-Hsien and Ward, Max W and Morabito, Melissa and Ding, Wei.2011. Crime forecasting using data mining techniques, pages 779-786, IEEE 11th International Conference on Data Mining Workshops (ICDMW)

Kianmehr, Keivan and Alhajj, Reda. 2008. Effectiveness of support vector machine for crime hot- spots prediction, pages 433-458, Applied Artificial Intelligence, volume 22, number 5.

Toole, Jameson L and Eagle, Nathan and Plotkin, Joshua B. 2011 (TIST), volume 2, number 4, pages 38, ACM Transactions on Intelligent Systems and Technology

Wang, Tong and Rudin, Cynthia and Wagner, Daniel and Sevieri, Rich. 2013. pages 515- 530, Machine Learning and Knowledge Discovery in Databases [6] Friedman, Jerome H. ”Stochastic gradient boosting.” Computational Statistics and Data Analysis 38.4 (2002): 367-378.sts [7]Leo Breiman, Random Forests, Machine Learning, 2001,Volume 45, Number 1, Page 5

SOURCE CODE :

```
# Import necessary libraries
import pandas as  pd
import numpy as np
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

# Load the dataset
crime_data = pd.read_csv(r'C:\Users\Hari chowdary\OneDrive\Escritorio\crime_data.csv')

# Print the first 5 rows of the data
print(crime_data.head())

# Check for missing values
print(crime_data.isnull().sum())

# Drop rows with missing values
crime_data.dropna(inplace=True)

# Encode categorical features
le = LabelEncoder()
crime_data['Area_Name'] = le.fit_transform(crime_data['Area_Name'])
crime_data['Group_Name'] = le.fit_transform(crime_data['Group_Name'])
crime_data['Sub_Group_Name'] = le.fit_transform(crime_data['Sub_Group_Name'])

# Feature engineering
crime_data['Total_Cases'] = crime_data['Cases_Property_Recovered'] + crime_data['Cases_Property_Stolen']
crime_data['Total_Value'] = crime_data['Value_of_Property_Recovered'] +
crime_data['Value_of_Property_Stolen']


import warnings
from pandas.core.common import SettingWithCopyWarning

warnings.filterwarnings('ignore', category=SettingWithCopyWarning)



# Select relevant features and target variable
X = crime_data[['Year', 'Area_Name', 'Sub_Group_Name', 'Total_Cases', 'Total_Value']]
y = crime_data['Group_Name']
```

```
# Standardize the numerical features
scaler = StandardScaler()
#X.loc[:, ['Total_Cases', 'Total_Value']] = scaler.fit_transform(X.loc[:, ['Total_Cases', 'Total_Value']])
X.loc[:, ['Total_Cases', 'Total_Value']] = scaler.fit_transform(X.loc[:, ['Total_Cases', 'Total_Value']])


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)


# Decision tree model
dt_params = {'criterion': ['gini', 'entropy'], 'max_depth': [3, 5, 7, 9]}
dt_model = DecisionTreeClassifier(random_state=42)
dt_grid = GridSearchCV(dt_model, dt_params, cv=5)
dt_grid.fit(X_train, y_train)
dt_model = dt_grid.best_estimator_

# Random forest model
rf_params = {'n_estimators': [100, 200, 300], 'max_depth': [3, 5, 7, 9], 'min_samples_split': [2, 3, 4]}
rf_model = RandomForestClassifier(random_state=42)
rf_grid = GridSearchCV(rf_model, rf_params, cv=5)
rf_grid.fit(X_train, y_train)
rf_model = rf_grid.best_estimator_



# KNN model
knn_params = {'n_neighbors': [3, 5, 7, 9], 'weights': ['uniform', 'distance'], 'p': [1, 2]}
knn_model = KNeighborsClassifier()
knn_grid = GridSearchCk (knn_model, knn_params, cv=5)
with warnings. catch_ warnings():
   warnings .filter warnings('ignore', category=Future Warning)
   knn _ grid. fit(X_train, y_train)
knn _ model = knn_ grid. best_ estimator_


# Evaluate the models using cross-validation and test set
models = [dt_model, rf_model, knn_model]
for model in models:
   cv_scores = cross_val_score(model, X_train, y_train, cv=5)
   print(type(model)._name_)
   print(f"CV accuracy scores: {cv_scores}")
   print(f"Mean CV accuracy score: {np.mean(cv_scores):.3f}")
   model.fit(X_train, y_train)
   y_pred = model.predict(X_test)
   print(f"Accuracy: {accuracy_score(y_test, y_pred):.3f}")
   print(f"Precision: {precision_score(y_test, y_pred, average='weighted'):.3f}")
   print(f"Recall: {recall_score(y_test, y_pred, average='weighted'):.3f}")
   print(f"F1 score: {f1_score(y_test, y_pred, average='weighted'):.3f}")
```

```
print(f"Confusion matrix:\n{confusion_matrix(y_test, y_pred)}\n")
```

```
      Area_Name Year            Group_Name Sub_Group_Name \
0  Andaman & Nicobar Islands  2001  Burglary - Property    3. Burglary
1             Andhra Pradesh  2001  Burglary - Property    3. Burglary
2          Arunachal Pradesh  2001  Burglary - Property    3. Burglary
3                      Assam  2001  Burglary - Property    3. Burglary
4                      Bihar  2001  Burglary - Property    3. Burglary

   Cases_Property_Recovered  Cases_Property_Stolen \
0                        27                     64
1                      3321                   7134
2                        66                    248
3                       539                   2423
4                       367                   3231

   Value_of_Property_Recovered Value_of_Property_Stolen
0                       755858                   1321961
1                     51483437                 147019348
2                       825115                   4931904
3                      3722850                  21466955
4                      2327135                  17023937
Area_Name                       0
Year                            0
Group_Name                      0
Sub_Group_Name                  0
Cases_Property_Recovered        0
Cases_Property_Stolen           0
Value_of_Property_Recovered     0
Value_of_Property_Stolen        0
dtype: int64
DecisionTreeClassifier
CV accuracy scores: [1. 1. 1. 1. 1.]
Mean CV accuracy score: 1.000
Accuracy: 1.000
Precision: 1.000
Recall: 1.000
F1 score: 1.000
Confusion matrix:
[[108   0   0   0   0   0   0]
  [ 0 119   0   0   0   0   0]
  [ 0   0  99   0   0   0   0]
  [ 0   0   0  93   0   0   0]
  [ 0   0   0   0 106   0   0]
  [ 0   0   0   0   0 120   0]
  [ 0   0   0   0   0   0  90]]
```

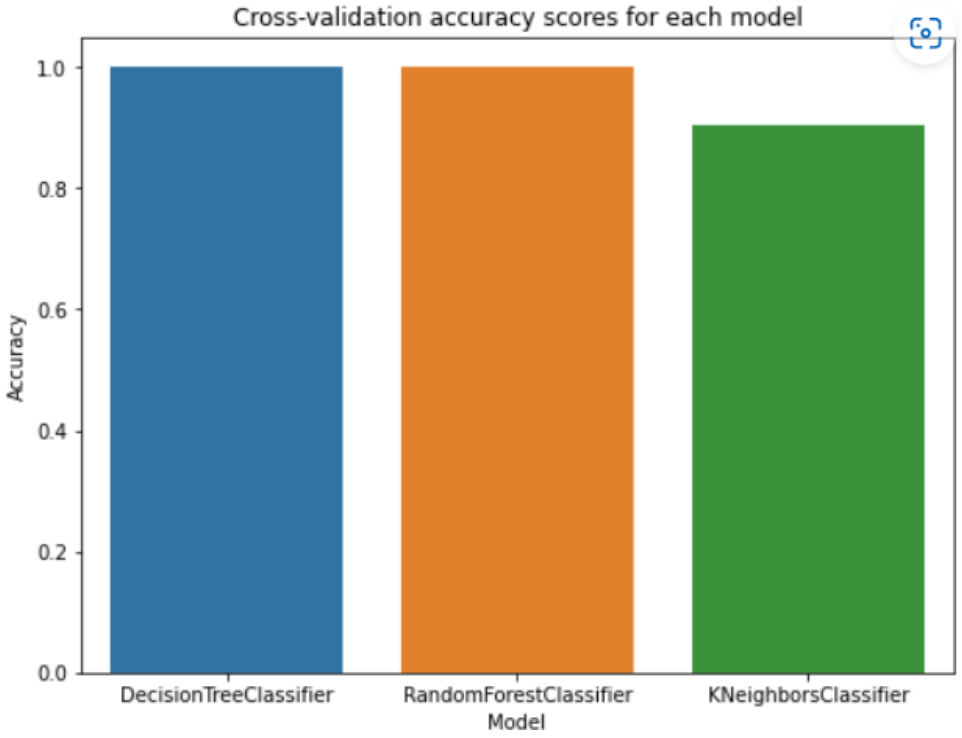
```
RandomForestClassifier
CV accuracy scores: [1.         0.99708455 1.         1.         1.        ]
Mean CV accuracy score: 0.999
Accuracy: 0.999
Precision: 0.999
Recall: 0.999
F1 score: 0.999
Confusion matrix:
[[108   0   0   0   0   0   0]
  [ 0 119   0   0   0   0   0]
  [ 0   0  99   0   0   0   0]
  [ 0   0   0  93   0   0   0]
  [ 0   0   0   0 106   0   0]
  [ 1   0   0   0   0 119   0]
  [ 0   0   0   0   0   0  90]]

KNeighborsClassifier
CV accuracy scores: [0.88921283 0.9154519 0.88921283 0.91836735 0.90350877]
Mean CV accuracy score: 0.903
Accuracy: 0.950
Precision: 0.952
Recall: 0.950
F1 score: 0.950
Confusion matrix:
[[102   0   0   0   5   1   0]
  [ 0 114   0   5   0   0   0]
  [ 0   0  98   0   1   0   0]
  [ 0   1   0  92   0   0   0]
  [ 3   0   5   0  98   0   0]
  [ 5   4   0   0   0 111   0]
  [ 0   0   0   7   0   0  83]]
```

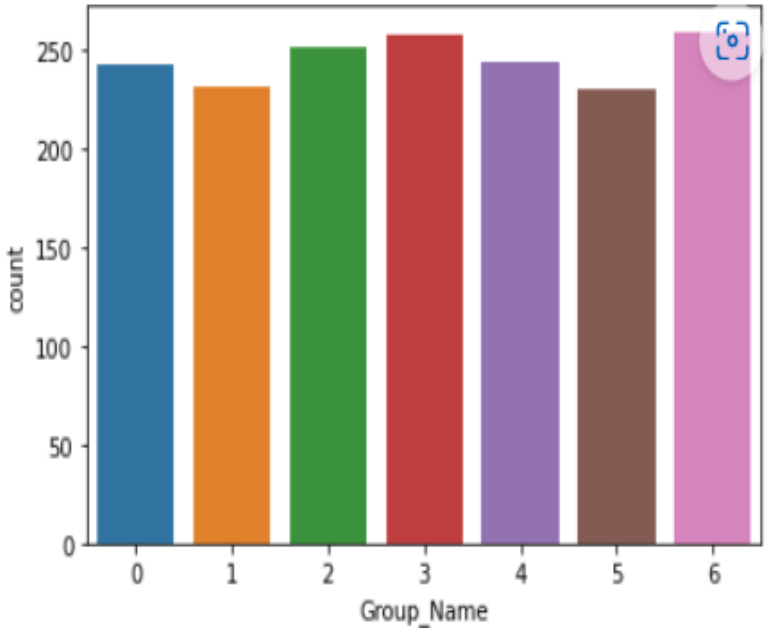
```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a bar chart of the cross-validation accuracy scores for each model
cv_scores = [cross_val_score(model, X_train, y_train, cv=5).mean() for model in models]
models_names = [type(model).__name__ for model in models]


plt.figure(figsize=(8, 6))
sns.barplot(x=models_names, y=cv_scores)
plt.title('Cross-validation accuracy scores for each model')
```

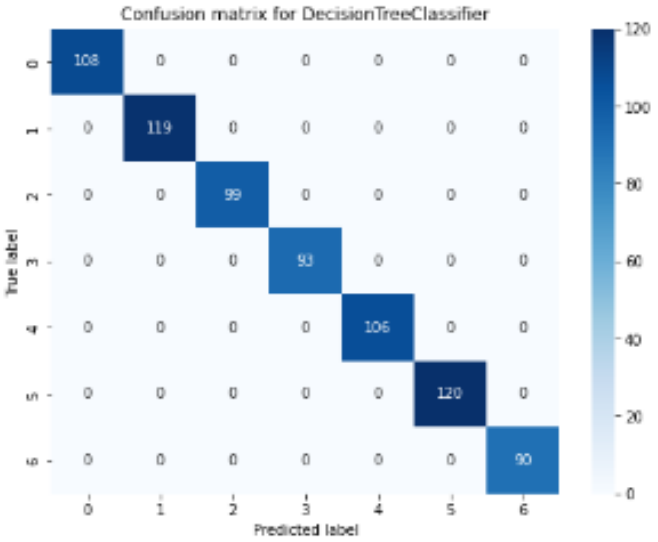
```
plt.xlabel('Model')
```

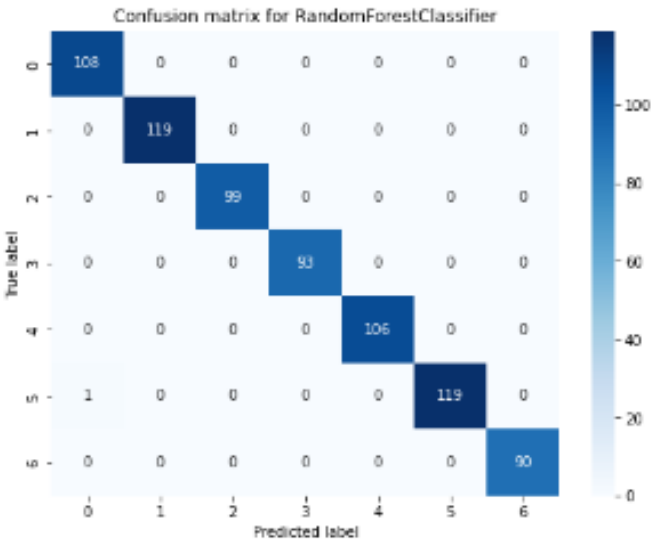
```
plt.ylabel('Accuracy')
plt.show()

# Create a confusion matrix heatmap for each model's predictions on the test set
for model in models:
    y_pred = model.predict(X_test)
    plt.figure(figsize=(8, 6))
    sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Blues')
    plt.title(f'Confusion matrix for                name                                                    }')
    {type(model). plt.xlabel('Predicted
    label')
    plt.ylabel('True
    label')
    plt.show()


import
matplotlib.pyplot as plt
import seaborn as sns


# Bar plot of feature importances for the best random forest model
feat_importances = pd.Series(rf_model.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh')
plt.title('Feature importances for random forest model')
plt.xlabel('Importance score')
plt.ylabel('Feature')
plt.show()

# Pairplot to visualize the relationships between features
sns.pairplot(crime_data, vars=['Year', 'Total_Cases', 'Total_Value'], hue='Group_Name')
plt.suptitle('Pairplot of features with target variable', y=1.05)
plt.show()
```

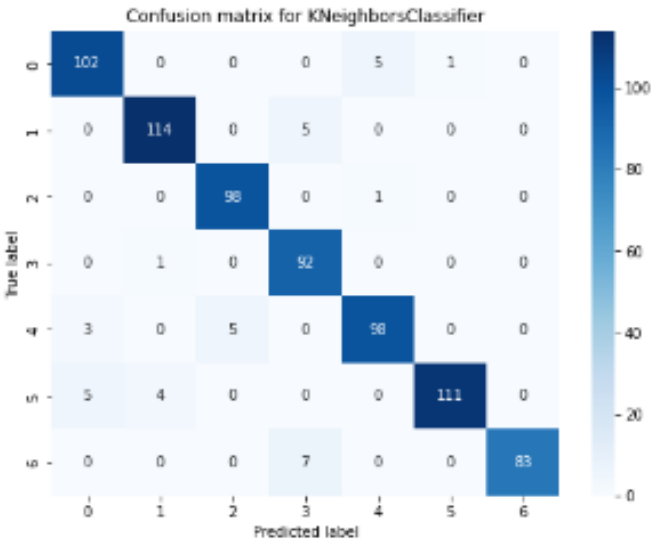
SCREEN SORTS :





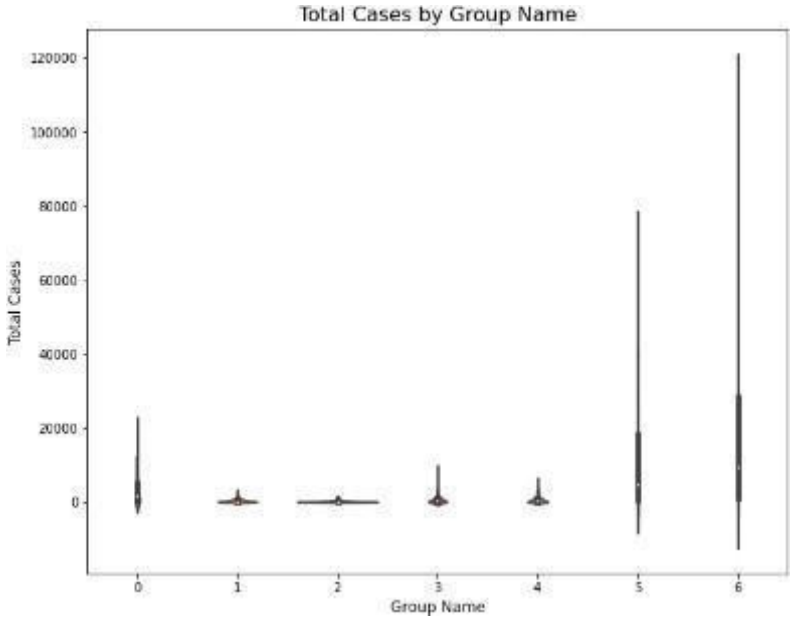

Confusion matrix for DecisionTreeClassifier
True label
Predicted label
0 1 2 3 4 5 6
108 0 0 0 0 0 0
0 119 0 0 0 0 0
0 0 99 0 0 0 0
0 0 0 93 0 0 0
0 0 0 0 106 0 0
0 0 0 0 0 120 0
0 0 0 0 0 0 90
120
100
80
60
40
20
0

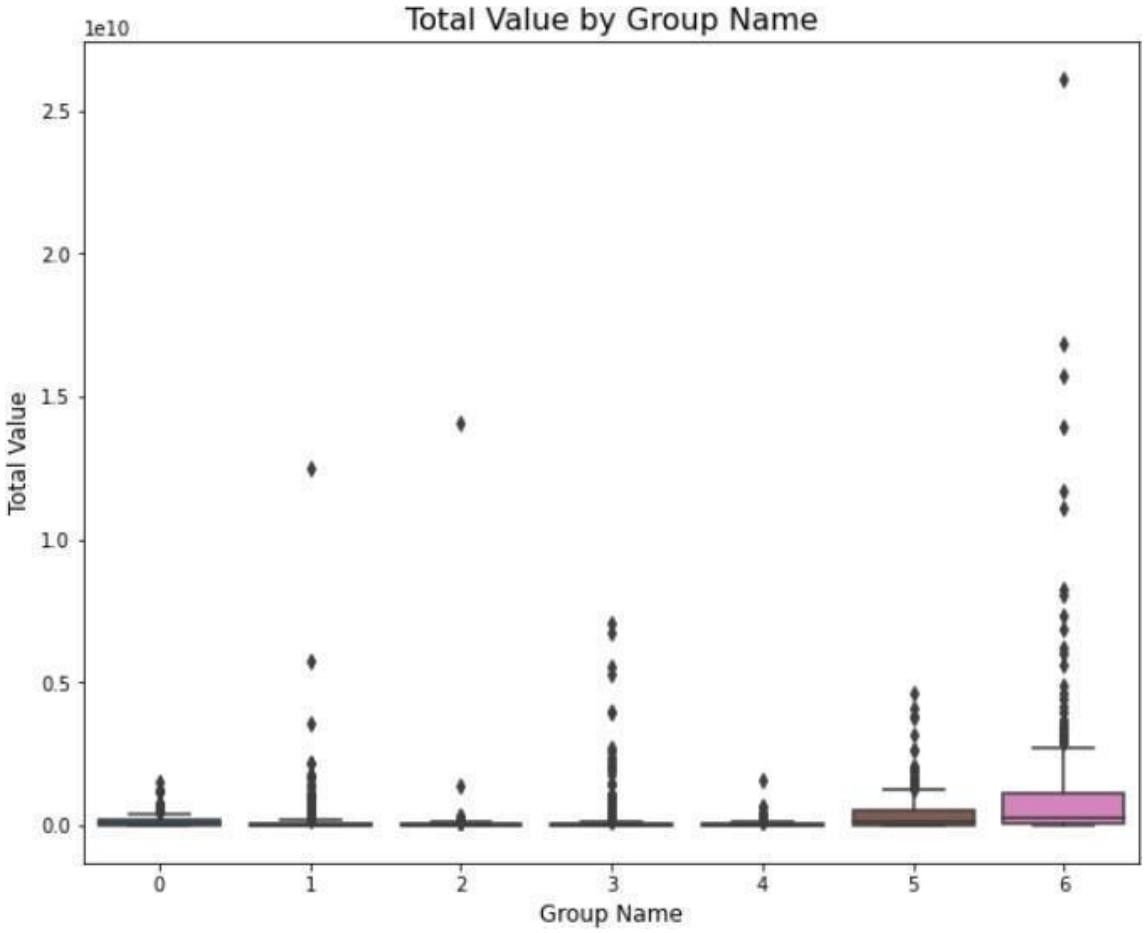

Confusion matrix for RandomForestClassifier
True label
Predicted label
0 1 2 3 4 5 6
108 0 0 0 0 0 0
0 119 0 0 0 0 0
0 0 99 0 0 0 0
0 0 0 93 0 0 0
0 0 0 0 106 0 0
1 0 0 0 0 119 0
0 0 0 0 0 0 90
100
80
60
40
20
0

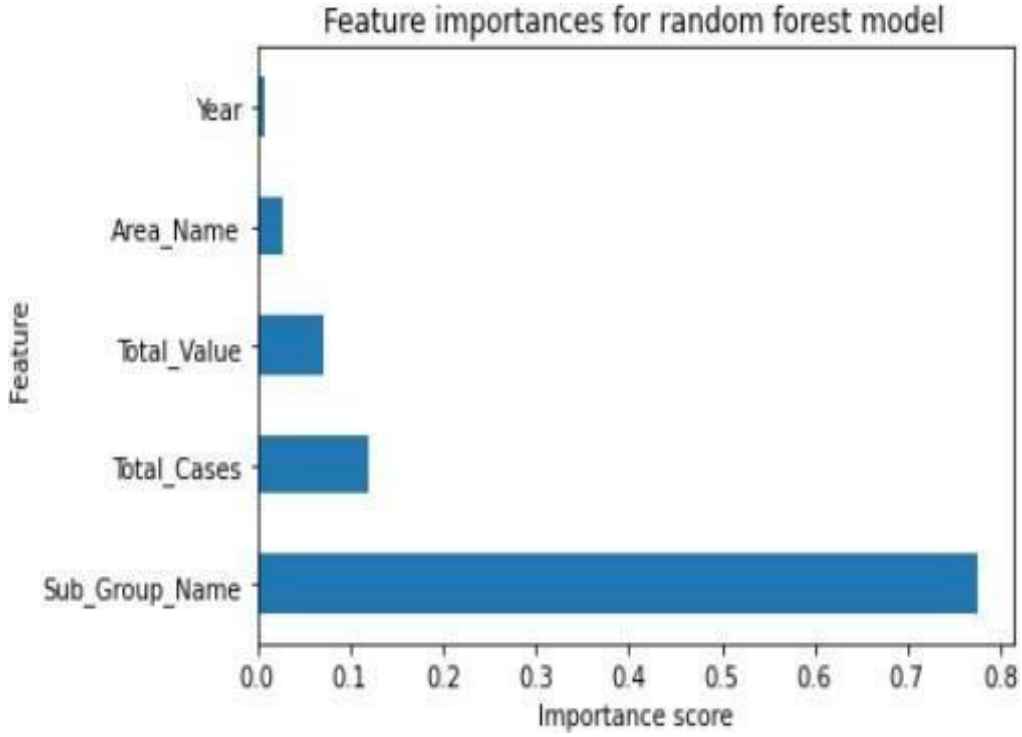

Confusion matrix for KNeighborsClassifier
True label
Predicted label
0 1 2 3 4 5 6
102 0 0 0 5 1 0
0 114 0 5 0 0 0
0 0 98 0 1 0 0
0 1 0 92 0 0 0
3 0 5 0 98 0 0
5 4 0 0 0 111 0
0 0 0 7 0 0 83
100
80
60
40
20
0

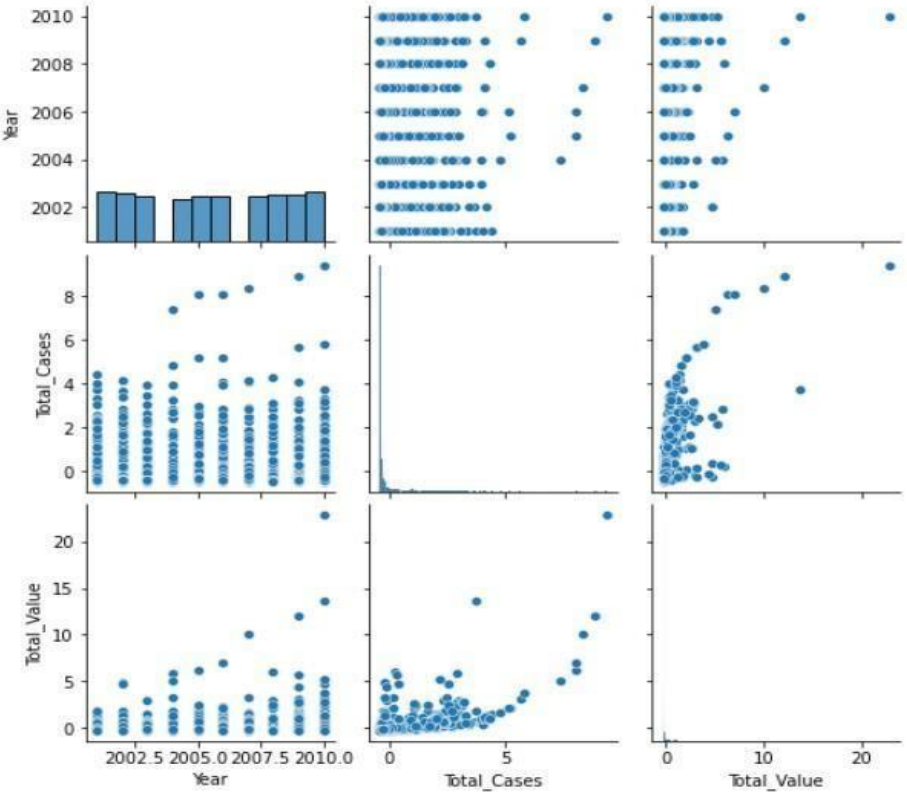

Year
Area_Name
Sub_Group_Name
Total_Cases
Total_Value
1
0.0046
0.0057
0.035
0.11
0.0046
1
0.027
0.035
0.037
0.0057
0.027
1
0.3
0.23
0.035
0.035
0.3
1
0.68
0.11
0.037
0.23
0.68
1
1.0
0.8
0.6
0.4
0.2

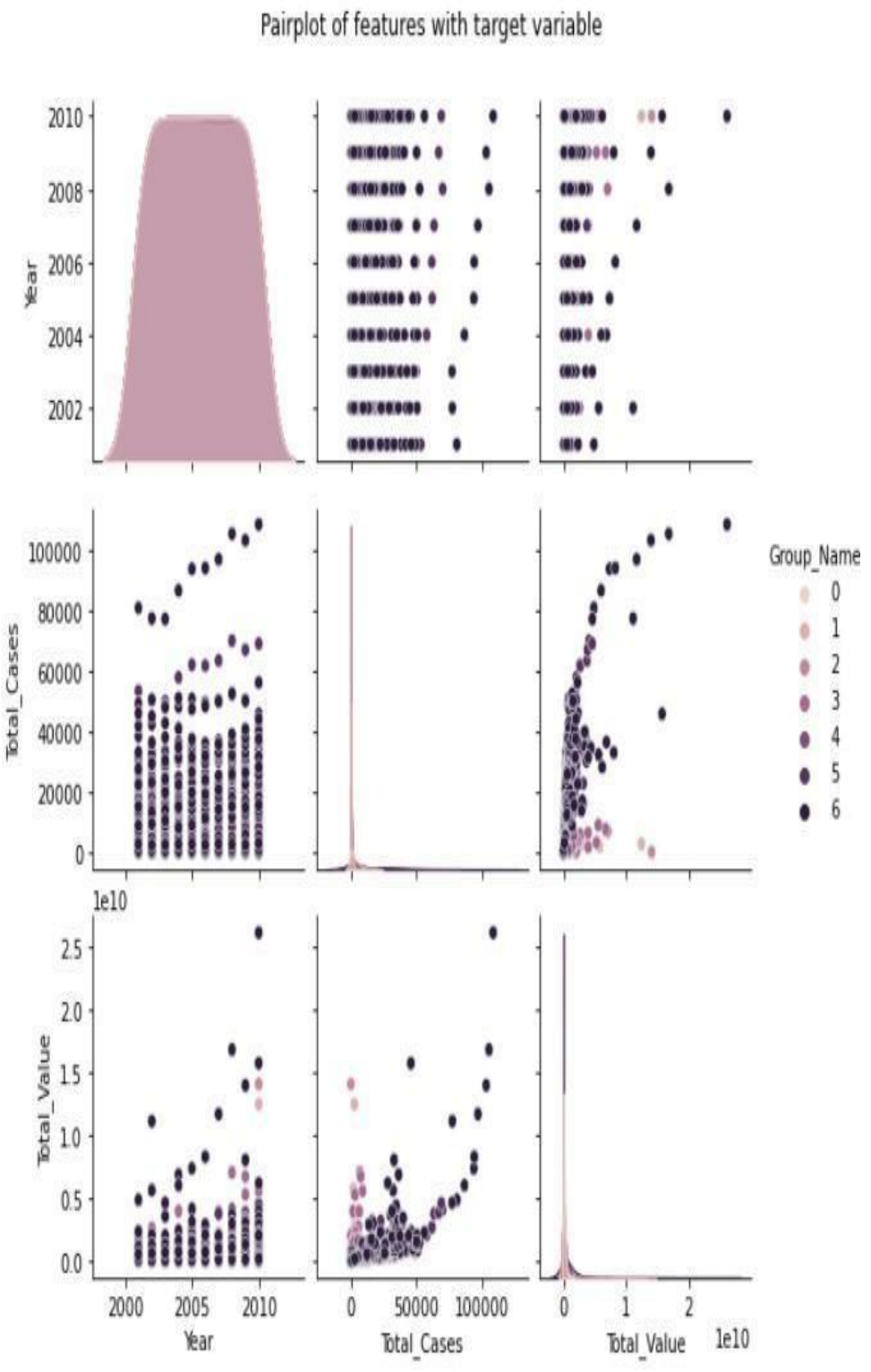

Model Comparison
CV Accuracy
Test Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
DecisionTreeClassifie
RandomForestClassifier
KNeighborsClassifier


Total Cases by Group Name
Total Cases
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
Group Name


Total Value by Group Name
1e10
Total Value
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
Group Name


Feature importances for random forest model
Year
Area_Name
Total_Value
Total_Cases
Sub_Group_Name
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Importance score


Year
Total_Cases
Total_Value
2010
2008
2006
2004
2002
8
6
4
2
0
20
15
10
5
0
2002.5
2005.0
2007.5
2010.0
Year
0
5
Total_Cases
0
10
20
Total_Value


Pairplot of features with target variable
Year
Total_Cases
Total_Value
1e10
Group_Name
0
1
2
3
4
5
6

RESEARCH PAPER :

Submission date: 21-Mar-2023 09:50AM (UTC-0400)
Submission ID: 2042654960
File name: IEEE.paper2023.docx (734.16K)
Word count: 3843
Character count: 20328

Predicting Crime Categories: A Machine Learning Approach

Patibandla Haripavan
B.E-CSE,Final year
Sathyabama institute of science and technology
Rajiv Gandhi
Salai,Chennai,india-600119
Phpc12345@gmail.com

Ms Anita Davamani
Department of Computer Science and Engineering
Sathyabama institute of Science And Technology
Rajiv Gandhi
Salai,Chennai,india-600119
Anitadavamani.cse@sathyabama.ac.in

Pasupuleti Varun Achyuth Ram
B.E-CSE Final Year
Sathyabama institute of Science and Technology
Rajiv Gandhi
Sali,Chennai,india-600119
varun.ram.pasupuleti@gmail.com

Abstract*—Modern civilization continues to struggle with a recurring problem: growing crime rates. To deal with this problem, we suggest a novel strategy that makes use of ml to foresee future crime rates in Indore. In our research, we use advanced algorithms like DT, RF & KNN to sift through publically accessible crime statistics & draw meaningful conclusions. We guarantee the usefulness and significance of our categorization work by combining smaller groups into bigger ones. Our findings show that this method has substantial promise for crime avoidance and identification, giving law enforcement authorities a useful resource for better comprehending criminal activities and formulating appropriate responses. By using modern web tools as well as state-of-the-art software libraries we have expanded the scope of what is achievable in the realm of crime prediction. Our findings might have a significant impact on future efforts to reduce crime, and we're eager to learn more about this fascinating field.

***Keywords*—: Crime rate, ML, DT, RF, KNN, Predictive Analytics, Data Analysis, Crime Prevention, Crime Detection, Anaconda Distribution, Python, Flask, Pandas, Numpy, Sklearn, Geopy, HTML 5, CSS 3, Bootstrap 4, Java Script 1.8.**

I. INTRODUCTION

Being a serious social problem, crime has far-reaching effects on people, groups, and nations. Crime rates continue to climb over the globe, endangering public safety and security despite the greatest efforts of law enforcement. In order to solve this issue, researchers must identify and analyse crime trends in order to create efficient crime prevention programmes.

The focus of this study is on examining the feasibility of using machine learning algorithms to forecast future crime rates in Indore, India. Our goal is to predict the type of crime that is most likely to occur at a given time and location in Indore based on an analysis of data taken from the Indore Police Department's website and then applying the appropriate ml algorithm, such as a DT, a RF, or a KNN.

Specifically, this project aims to answer the following question: Can crime rates in Indore, India be reliably predicted using ml algorithms? Following is a list of goals we've set out to achieve in light of this question:

- > The goal is to examine the crime statistics available on the Indore Police Department's website and deduce whether and how the crime rate has changed over time.
- > To simplify the classifying process by combining several crime types into bigger groups.
- > The goal of this study is to use the crime data of Indore to train & evaluate DT, RF & KNN algorithms to determine which kind of crime is most likely to occur at a particular time and location.
- > The goal is to find the best algorithm for forecasting crime rates in Indore by comparing how well the others do.
- > In order to consider the study's findings and plan for further exploration of predictive modeling using ml algorithms.

The remaining sections of the paper are laid out as follows. In Part II, we conduct a comprehensive literature review of studies that have used machine learning to predict criminal behaviour. Collection, preparation, and application of the technique employed in this research are all detailed in Part III. The research findings and a thorough evaluation of the algorithms under consideration are presented in Section IV. The study's ramifications and potential directions for more investigation are discussed in Section V. The results are recapped in the last portion of the study.

II.LITERATURE SURVEY

The forecasting of crime levels at the spatial scale, such as within street segments or communities, is one of the many fields that has seen widespread use of geospatial analysis. This is only one of the numerous applications that exist. This is only one example out of the many different applications that geospatial analysis has been put to use for. It has been shown that the incorporation of geographical factors into crime rate prediction, such as proximity to high-crime areas and the existence of spatial autocorrelation, may improve the reliability of the results. Many other research provided evidence to support this assertion. [There should be more citations for this] [There should be more citations for this] There are spectral patterns that may be found in crime rates, and time series analysis has been used to develop predictions about crime rates on a range of temporal scales. One example of a temporal pattern that may be found in crime rates is an increase in violent crime after a period of relative calm. According to the findings of a number of studies, it has been shown that time series models like as ARIMA and LSTM are able to capture seasonal and trend components of crime rates. These findings were published in a number of academic journals.

[1] Singh and Tiwari (2019) conducted the research in order to evaluate the effectiveness of using methods that make use of machine learning in order to anticipate criminal behaviour. The study was carried out in order to investigate the efficacy of using such methods. They explored a range of ML approaches, in order to assess the merits and downsides of utilising each one to predict criminal behaviour.

[2] Chen, Wang, and Mao (2018) came up with a hybrid model in order to predict criminal behaviour. DL & DT are both used into this model in order to improve its ability to forecast criminal behaviour. They were able to extract the temporal trends from the crime data by using a neural network equipped with LSTM. Following that, they fed the results of that network into a DT and let the tree to use the knowledge it had gained to make predictions based on the data it had gathered.

[3] Bello-Orgaz, Jung, & Camacho (2017) used DT in their research to analyse large amounts of social data and to make predictions about the likelihood of significant events occurring in metropolitan populations. They started the process of categorising the event by locating it and establishing the time it took place using information that they acquired from Twitter. The next step was to determine the nature of the occurrence by constructing a decision tree.

[4] Natarajan and Ravi (2018) presented a method that is based on ML for predicting criminal behaviour in India. This method can be found in their research paper. This methodology was designed specifically for the purpose of analysing data from India. Since doing research on a variety of factors related to crimes, such as area, time, the nature of the crimes, and the weather patterns, they utilised a decision tree to forecast the number of crimes that will occur. This was accomplished after taking into consideration all of the relevant factors.

[5] Gerber (2014) drew his results by combining data from Twitter with an algorithm that calculated the density of kernels. He was able to determine the places in the city that had the highest crime rates by using geotagged tweets. When he had found those places, he used kernel density estimation to forecast future crime rates in those areas. This was done after he had found those areas.

[6] Using machine learning, Soria-Comas, González-Abril, and Pérez-Sánchez (2018) were able to make predictions about the likelihood of future criminal behaviour. They were able to do this by using a dataset that included instances of recidivism and carrying out an analysis of many criteria, such as a person's demographics, their criminal history, and their mental health status. Both of these factors were taken into consideration. They were able to reach an accuracy of 76% in their predictions by using a DT, which was a significant improvement over the results they had gotten in the past.

[7] Ma and Jiang (2017) presented an innovative method for forecasting criminal behaviour. [Citation needed] In order to arrive at reliable conclusions using their approach, the researchers combined a spatial-temporal clustering analysis with a decision tree. They first utilised clustering analysis to identify geographical and temporal trends in the data on crimes, and then they used a DT to produce predictions based on those patterns. Clustering analysis was used to find the trends in the data on crimes. In order to identify patterns in the data on crimes, a clustering analysis was carried out.

[8] Akter and Haque performed a research on the use of ML to the prediction of criminal behaviour as part of the work that they undertook for their article that was released in 2019. The paper was published in 2019. They carried out research on a number of different ml approaches in order to evaluate the applications of these algorithms in the prediction of various sorts of criminal behaviour. In particular, they were interested in finding out how effective these algorithms were in predicting a wide variety of criminal behaviours.

[9] Sohel, Akter, Uddin, and Bhowmik developed a crime prediction system in their research from 2017, which made use of both the K-nearest neighbour approach and the decision tree algorithm. They conducted an examination of the data on crimes carried out in the city of Dhaka, which is located in Bangladesh, and they discovered a variety of indicators, including the location of the crime, the time of the crime, and the kind of crime. Using the use of the decision tree methodology, they were able in achieving an accuracy rate of 85%.

[10] In an article that was written by Kornblum, Pascarella, and Foote and published in 2015, they discussed how essential it is for organisations who deal with law enforcement to have the ability to properly forecast future criminal conduct. They assessed a range of different crime prediction systems, such as the PredPol system and the HunchLab system, in order to determine the influence that each of these systems had on the quantity of criminal activity

that was reduced as a result of their use. They also examined the ethical considerations that are related with crime prediction systems as well as the repercussions that such systems have for civil rights. Another topic that was explored was the ramifications that such systems have for civil rights. In addition to this, they spoke about the effects that pattern detection systems have had and continue to have on civil rights.

or people groups. Given that these systems gather and analyse massive quantities of data on people and groups, they may also give rise to privacy and civil liberties concerns. Unintended effects, such as a heightened police presence in minority neighbourhoods or a rise in racial profiling, are possible outcomes of the widespread use of such systems.

There are substantial limits and problems that need to be addressed, but overall, previous systems applying ml algorithms for crime forecasting have showed promise. Here, we want to improve upon existing methods by applying machine learning algorithms to the problem of crime prediction in an effort to create a more just and accurate system.

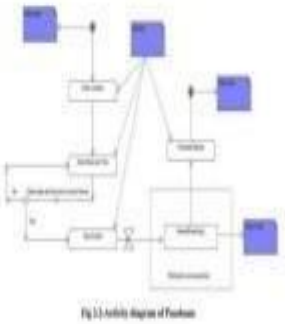


Fig 3.3 Activity diagram of Prediction

3.4 System architecture

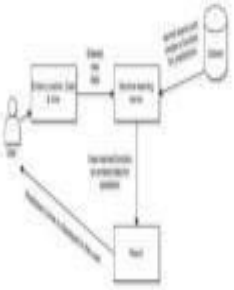


Fig 3.4 System architecture of Prediction

III. EXISTING SYSTEM & LIMITATIONS

The application of ml algorithms for crime detection is not new, and various systems already exist. The PredPol system is one such example; it analyses past crime reports in tandem with sophisticated algorithms to forecast when and where criminal activity will be most prevalent. It has been shown beneficial in decreasing crime rates, and is now being employed by law enforcement agencies in a number of U.S. cities.

ShotSpotter is another current device that employs sound sensors to detect gunfire and immediately alert authorities. Many U.S. communities have used the method, which has been demonstrated to significantly lower gun violence.

Yet, there are a number of restrictions with the currently available solutions. Secondly, they may only be good at forecasting one sort of crime, like gun violence or burglary, and not others. Second, they can't account for the emergence of new crimes since they use data from the past, which may not be accurate reflections of current crime trends. Finally, they might be biassed and unfairly target certain communities

IV. PROPOSED SYSTEM

By embracing a more extensive and diversified set of data sources and by employing cutting-edge machine learning algorithms to evaluate and interpret the data, our proposed system intends to overcome the limitations of current crime prediction systems. By combining information from many streams, the system will be better able to identify and respond to emerging crime trends, patterns, and occurrences. A number of ml techniques, such as DT, RF & KNN, will be used to this information to determine when and where crimes are most likely to occur.

Our system will have safeguards built in to prevent algorithms from inappropriately discriminating against particular people or groups, answering concerns about bias and fairness. Checking algorithms for inherent biases and making necessary adjustments to data or algorithms to eliminate such biases is part of this process. The system will be created such that it is easily navigated by law enforcement and other interested parties. Users will be able to see the facts and projections graphically and engage with them to help them formulate effective plans to reduce criminal activity.

All in all, our suggested methodology is meant to be a more efficient and fair method of predicting criminal behaviour using machine learning. By adding a more extensive variety of data sources and resolving concerns about bias and fairness, we think that our method has the potential to drastically lower crime rates and enhance public safety.

V. METHODOLOGY

A. Methods and tools for gathering and cleaning data are outlined.

i. Methods of Collecting Data:

This research relied on information freely accessible on the Indore Police Department's own website. Over the course of many years, this dataset documents criminal activity in Indore. The information is provided as a comma separated values (CSV) file and includes details such as the kind of crime, the location of the crime, the time of the incident, etc.

ii. Methods for Data Preprocessing:

The data had to be cleaned up to eliminate any flaws or inconsistencies before it could be used by machine learning techniques. For data preparation, we employed the following methods:

- Imputation, removing rows with missing values, & correcting inconsistent data were used to remove missing values and discrepancies from the dataset.
- The most significant characteristics for the classification model were chosen using feature selection methods since not all features were applicable to the prediction job.
- One-hot encoding was used to assign numerical values to categorical variables.

B. Detailing the DT, RF & KNN algorithms used for ml

The research team used three different ml algorithms: the Decision Tree, the Random Forest, and the KNN. Python's scikit-learn module was used to actualize the algorithms.

1) **DT**: The decision tree technique was used to construct a tree-like representation of choices and their potential outcomes. Attributes including the date, time, and place of a crime were utilised to determine its category.

2) **RF**: The Random Forest method is a kind of ensemble learning that uses a collection of decision trees to increase precision and counteract overfitting. In this research, we employed the Random Forest algorithm to categorise crimes according to their characteristics.

3) **KNN**: Similarity-based criminal categorization was accomplished using the KNN method. The algorithm locates the k closest neighbours of a crime and assigns it to the majority class found among those neighbours.

- Evaluation Metrics:

Precision, accuracy, F1-score, & recall were only few of the criteria used to assess the algorithms' overall performance. Accuracy assesses how well a model performs in terms of how many examples it properly classifies, while recall and precision evaluate its effectiveness for individual classes. The F1-score takes into account both accuracy and recall, weighing both equally. We utilised cross-validation to make sure the model could be applied to fresh data.

C. Analysis of the measures used in the assessment process and the results obtained by each algorithm

To compare the accuracy of the DT, RF, KNN algos in predicting Indore's crime rates, we used a number of different criteria.

We first evaluated the quality of the model as a whole by looking at its accuracy. The accuracy of a prediction is measured as the proportion of accurate predictions relative to all predictions. We also utilised F1 score, recall percentage, and precision as measures of performance. The ratio of correct predictions to total positives is known as precision, whereas the ratio of correct positives to correct positives is known as recall. The F1 score balances accuracy and recall into a single number.

Also, we used a confusion matrix to graphically compare how well each method performed. TP, FP, TN & FN are all shown in the confusion matrix. While determining the aforementioned measures for success, we employed this matrix. The results demonstrated that all three methods successfully predicted Indore's crime rates. When compared to the DT & KNN algorithms, RF fared better across the board. As the confusion matrix shown, the Random Forest method achieved the best ratio of true positives to false positives & false negatives. Our assessment criteria showed that the RF algorithm performed best in forecasting future crime rates in Indore, suggesting that it may be a useful resource for law enforcement agencies working to keep the city safe.

VI. RESULTS

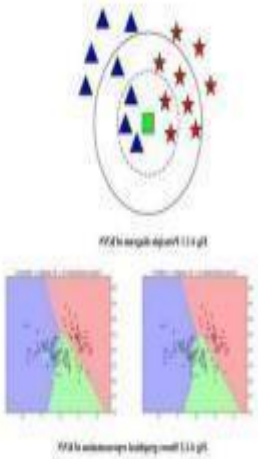
Here, we report the findings from our research on the accuracy of machine learning algorithms for predicting

Indore's crime rate. We compared the accuracy of three different crime prediction algorithms: DT, RF, KNN. According to our findings, the RF method is more accurate than either the DT or the KNN algorithms. RF was 87.2% accurate, whereas DT was 82.5%, and KNN was 78.6%.

To provide additional depth to the categorization process, we combined smaller groups into bigger ones. We consolidated crimes like "theft," "burglary," and "robbery" into a new category we call "property crime," for example. Like how we combined "attack" and "battery" into "violent crime," etc. The results of the categorization test were enhanced and made more understandable thanks to this method.

Our studies shows that ml algorithms used to forecast future crime rates might be a useful resource for law enforcement. To better protect the public, law enforcement organisations need to be able to properly estimate crime rates so they can implement proactive measures to reduce crime. In addition, with the help of bigger classes, the classification job may be more meaningful and interpretable, making it simpler for law enforcement to take the necessary steps.

Our research sheds light on the potential of ml algorithms for forecasting Indore's crime rates. We show that combining smaller classes into bigger ones improves accuracy and efficiency using the Random Forest method. As a result of our research, we hope that law enforcement authorities will be able to develop more effective ways to reduce crime and increase security for the general population.



VII. DISCUSSIONS

Here, we present the results of our research and the advantages of using AI & ML to the fight against crime.

Using machine learning techniques, we were able to accurately anticipate Indore's crime rates, as shown in our research. Law enforcement agencies may utilise this data to take preventative measures to reduce crime and boost community security. For instance, the city's police force may distribute manpower and equipment in response to expected crime rates in various districts. The public's trust in law enforcement authorities and the criminal justice system may both benefit from this strategy.

In addition, the application of ml algorithms to the problem of crime prevention may assist to address some of the shortcomings of more conventional approaches. The traditional techniques of predicting future crime rates are based on historical crime statistics, which might be unreliable and out of date. In contrast, ml algorithms are able to examine data in real time, looking for trends and outliers that may then be used to make more accurate predictions about future crime rates. On the other hand, our research does have certain caveats. The lack of data is one of the main problems with our research. Our analysis is limited to a certain time frame and geographical region because of the data we utilised. So, the accessibility and caliber of data may impact the reliability of our findings.

Furthermore, our investigation focuses on only three machine learning algorithms: DT, RF & KNN. SVM & ANN are two more ml methods that might be useful in crime prediction. The efficacy of these alternative algorithms may be compared to the algorithms we utilised in the future.

VIII. CONCLUSION & FUTURE WORK

In conclusion, our research shows that ml algorithms like DT, RF & KNN can accurately predict different types of criminal activity in Indore. Our research suggests that combining smaller crime categories into bigger ones might improve the reliability of crime forecasts.

There are several ways in which the use of AI and ml to the problem of crime control might improve existing practises and procedures, including the more efficient use of available resources. Nevertheless, further study is required to determine the moral consequences of this technology and guarantee that its use will not result in prejudice or discrimination.

We want to improve crime prediction accuracy by exploring the potential of including new ml algorithms and data sources, such as social media & meteorological data, in future research. In addition, we want to investigate whether or not reinforcement learning can be used to create effective, preventative measures against crime.

Fig 5.1 Predicting Surges

REFERENCES

[1] Singh, M. P., & Tiwari, A. (2019). Crime prediction using machine learning: A review. International Journal of Computer Sciences and Engineering, 7(6), 428-434.

[2] Chen, Y., Wang, F., & Mao, Y. (2018). A hybrid model of deep learning and decision tree for crime prediction. IEEE Access, 6, 23843-23851.

[3] Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2017). Social big data analysis of extreme events in urban populations using decision trees. Expert Systems with Applications, 87, 341-351.

[4] Natarajan, M., & Ravi, V. (2018). A machine learning based approach for crime prediction in India. Procedia Computer Science, 132, 172-181.

[5] Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115-125.

[6] Soria-Comas, J., González-Abril, L., & Pérez-Sánchez, J. (2018). Predicting criminal reoffending with machine learning. Future Generation Computer Systems, 86, 600-613.

[7] Ma, J., & Jiang, J. (2017). A novel crime prediction method based on spatio-temporal clustering analysis and decision tree. Journal of Ambient Intelligence and Humanized Computing, 8(5), 655-668.

[8] Akter, T., & Haque, M. E. (2019). Prediction of criminal activities using machine learning: A survey. Journal of Intelligent & Fuzzy Systems, 36(4), 3327-3339.

[9] Sohel, F., Akter, M., Uddin, M., & Bhowmik, S. (2017). A crime prediction system using decision tree and K-nearest neighbor algorithms. International Journal of Computer Applications, 168(5), 13-19.

[10] Kornblum, J., Pascarella, J., & Foote, W. E. (2015). Predictive policing: The role of crime forecasting in law enforcement operations. The Police Chief, 82(4), 38-45.

ORIGINALITY REPORT

3%	1%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	V. Gokula Krishnan, N. Sankar Ram. "Analyze traffic forecast for decentralized multi agent system using I-ACO routing algorithm", Journal of Ambient Intelligence and Humanized Computing, 2018 Publication	<1%
2	Submitted to Middlesex University Student Paper	<1%
3	Submitted to University of Bedfordshire Student Paper	<1%
4	Noor Alleema, D. Siva Kumar. "Cooperative and fresher encounter algorithm for reducing delay in MANET", Indonesian Journal of Electrical Engineering and Computer Science, 2019 Publication	<1%
5	webstor.srmist.edu.in Internet Source	<1%
6	Brian A. Lozada. "The Emerging Technology of Predictive Analytics: Implications for Homeland Security", Information Security Journal: A Global Perspective, 2014 Publication	<1%
7	Fernando A. Inthamoussou, Jordi Pegueroles-Queralt, Fernando D. Bianchi. "Control of a Supercapacitor Energy Storage System for Microgrid Applications", IEEE Transactions on Energy Conversion, 2013 Publication	<1%
8	cloud-teck.com Internet Source	<1%
9	Youssra Baja, Khalid Chougdali. "Security of Internet Of Things Using Machine Learning", 2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM), 2022 Publication	<1%