

Crime Analysis Framework for Predicting Criminal Behavior Patterns with Machine Learning

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Technology degree in Information Technology
by

ARJUN A (Reg. No 39120013)
SHAM VENKAT S (Reg. No 39120107)



**DEPARTMENT OF INFORMATION TECHNOLOGY
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, TAMILNADU**

MAY - 2023



DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **ARJUN A (Reg. No:39120013)** and **SHAM VENKAT S (39120107)** who underwent the professional training in "Crime Analysis Framework for Predicting Criminal Behaviour Patterns with Machine Learning" under my supervision from January 2023 to May 2023.

Internal Guide

Dr. R. Jeberson Retna Raj M.Tech., Ph.D.,

Head of the Department

Dr. R. Subhashini M.E., Ph.D.,

Submitted for Viva voce Examination held on 10/05/2023

Internal Examiner

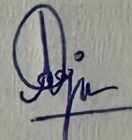
External Examiner

DECLARATION

I, ARJUN A (RegNo:39120013) here by declare that the Project Phase-2 Report entitled "Crime Analysis Framework for Predicting Criminal Behaviour Patterns with Machine Learning" done by me under the guidance of Dr.R.Jeberson Retna Raj M.Tech., Ph.D is submitted in partial fulfilment of the requirements for the award of Bachelor of Technology degree in **INFORMATION TECHNOLOGY**.

DATE: 10/05/2023

PLACE: CHENNAI



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** 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. R.Subhashini M.E., Ph. D.**, Head of the Department of Information Technology 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 **Dr.R.Jeberson Retna Raj M.Tech.,Ph.D.**, for him 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 Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

In this era of recent times, crime has become an evident way of making people and society under trouble. An increasing crime factor leads to an imbalance in the constituency of a country. In order to analyse and have a response ahead this type of criminal activities, it is necessary to understand the crime patterns. This project imposes one such crime pattern analysis by using crime data obtained from Kaggle open source which in turn used for the prediction of most recently occurring crimes. The major aspect of this project is to estimate which crime rate contributes the most along with time period and location where it has happened. Some machine learning algorithms such as Random Forest, XG boost, Decision tree, etc is implied in this work in order to classify among various crime patterns and the accuracy achieved was comparatively high when compared to precomposed works. The input features used in the algorithms include time, state, district, year, murder, rape, theft, and dowry deaths. The proposed framework uses different visualization techniques to show the trends of crimes and various ways that can predicts the crime using machine learning algorithm. This project presents a detailed analysis of the various relevant crime patterns and statistical analysis of crime data. This project will be helpful to law enforcing agencies in making strategies and tactics to address crime prevention and control.

TITLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	viii
1	INTRODUCTION	9
	1.1 Aim of the Project	10
	1.2 Objective of the Project	10
2	LITERATURE SURVEY	11
	2.1 Inference of Literature survey	11
3	REQUIREMENT ANALYSIS SYSTEM	17
	3.1 Software Requirement	17
	3.2 Hardware requirement	19
	3.3 Use Case Diagram	21
4	DESCRIPTION OF PROPOSED SYSTEM	23
	4.1 Architecture of Proposed System	23
	4.2 Proposed System	23
	4.3 Existing System	24
	4.4 Input System	24
	4.5 Data Flow Diagram	25
5	IMPLEMENTATION DETAILS	26
	5.1 Data Collection	26
	5.2 Data Preparation	26
	5.3 Model Selection	27
	5.4 Analyse and Prediction	29
6	RESULT AND DISCUSSION	30

7	CONCLUSSION & FUTURE WORK	31
7.1	Conclusion	31
7.2	Future Work	31
	REFERENCE	33
	APPENDIX	
	A. SOURCE CODE	36
	B. SCREENSHOT	42
	C.RESEARCH PAPER	46

LIST OF FIGURES

Figure No.	Title	Page No.
3.1	Use Case Diagram for Crime Rate	22
4.1	System Architecture	23
4.2	Data Flow Diagram	25
6.1	Comparison of Different Algorithm	30

CHAPTER 1

INTRODUCTION

Prevention is better than Cure. Preventing a crime from occurring is better than investigating what or how the crime had occurred. Just like vaccination is given to a child to prevent disease, in today's world with such higher crime rate and brutal crime happenings, it has become necessary to have a vaccination system that prevents from crimes happening. By vaccinating society against crime, it refers to various methods such as educating peoples, creating awareness, increasing efficiency and proactive policing methods and other deterrent techniques. Crime incident prediction has depended mainly on the historical crime record and various geospatial and demographic information. Data mining is a process of extracting information and discovering patterns from large amount of data. Crime Analysis can be a lucrative application of data mining. Data mining can play an important role in analyzing and predicting crimes using the data stored in repositories. Crime rate in India is increasing day by day, becoming a topic of major concern, hindering good governance in India. Due to awful growth in crime rate, it has become impossible to analyze those crime related data and detect crime patterns or predict future crimes by intelligence agencies or local law enforcement agencies. This project presents a detailed analysis of the various relevant crime patterns and statistical analysis of crime data. This study will be helpful to law enforcing agencies in making strategies and tactics to address crime and disorder. The output is the class of crime that is likely to have occurred. We try out with multiple classification algorithms, such as k-means, Decision Trees, Random Forests, etc.

1.1 AIM OF THE PROJECT

Much of the current work is focused in two major directions:

- Predicting the crime rate based on location, and
- Understanding patterns of criminal behaviour that could help in solving criminal investigations.

1.2 OBJECTIVE OF THE PROJECT

The objective of our work is to:

- Predicting crime before it takes place.
- To develop an algorithm which includes Machine Learning algorithms to check the accuracy.
- To compare the performance of the algorithm with existing system.
- Understanding crime pattern and Classify crime based on location.
- Analysis of crime in Tamilnadu.

CHAPTER 2

LITERATURE SURVEY

2.1 INFERENCE FROM LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system, the above consideration is taken into account for developing the proposed system. The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

Shiju Sathyadevan et. al presented a crime analysis and prevention system this approach is utilized for detecting and scrutinizing criminal trends and patterns. It has the ability to forecast areas that are at a greater risk of crime occurrence and can also provide visual representation of such crime-prone locations. The process involves implementing data mining techniques to extract significant insights from unstructured data that were previously unknown [1].

Dnyaneshwari S et.al conveys a classification model used to classify crimes that are clustered based on their occurrence frequency over several years. Data mining algorithms are employed to analyse, investigate, and detect patterns related

to various crimes. The crime dataset used in this study is obtained from the police records of England and Wales from 1990 to 2011. Additionally, the study utilizes the RapidMiner tool to optimize the Outlier Detection operator parameters using the Genetic Algorithm (GA) [2].

Benjamin et.al introduced a detailed survey on crime analysis based on supervised and unsupervised learning techniques. This model used to identify the crime and forecasting the crime. In The system employed with FCA and other machine learning techniques used for crime analysis. [3].

Tushar Sonawanev et.al Measures are being implemented to prevent the escalation of crimes against women. Data classification and correlation facilitate the identification of similarities and differences among data objects. Clustering techniques are employed to group data objects based on predetermined criteria. By utilizing data mining methods, it becomes possible to anticipate and visually represent criminal activity. patterns in various formats, improving our understanding of the crime pattern. [4]

RajKumar. Set. al presented a Various data mining techniques are utilized in criminal analysis and crime prediction, including the prediction of crime hotspots based on past data. Using pre-processed datasets, a predictive model is created by applying the Naïve Bayesian algorithm. This model analyses the data and aids in predicting crime trends for a particular region in the future [5].

S. Shankaraiah et.al presented the objective is to utilize K-means, Naive Bayes, and linear regression will be used in machine learning to predict the state that displays the highest crime rate. The algorithm that demonstrates the most effective combination of efficiency and accuracy will be selected as the optimal approach for crime prediction and analysis [6].

Ayisheshim Almaw et. al presented a method it focuses on application of the application of different data mining algorithms and ensemble learning methods is critical in analysing crime data. Data mining techniques play a vital role in uncovering obscure criminal patterns and addressing crime-related concerns. The study analyses and discusses different methods that are utilized for crime prediction and analysis, including the application of ensemble classification and

data mining approaches methods to predict and discover potential future crimes [7].

Harha Vardhan Reddy et.al conveys his system can forecast regions with a high likelihood of crime occurrences and visualize areas prone to crime. Data mining allows us to extract valuable, previously unknown information from unstructured data. This includes analysing and detecting patterns using decision trees, linear regression, k-means clustering, and various other data mining techniques for criminal analysis and prediction. [8]

K.S. N. Murthy et. al presented a Crime pattern analysis can be performed by utilizing various Frequent Pattern Mining and Rule Mining algorithms. These algorithms facilitate comprehensive comprehension of frequent pattern mining in diverse fields. Frequent pattern mining is essential for rule mining, and the commonly employed algorithms for frequent pattern mining include Apriorism and FP-Growth [9].

Smitha Vinod et.al conveys the process method for detecting crime in India, the implementation of data mining methods comprises a set of procedures including pre-processing data, clustering, categorization, and visualisation. Data is analysed for categorization utilising techniques such as Random Forest and Neural Networks. Crime hotspots are located and visualised using the Google marker clustering method [10].

Bhavana Saini et. al presented a Different the Machine learning algorithms and visualisation techniques are used to forecast how crime will be distributed in a given area. The use of machine learning methods such as Naive Bayes follows. to extract insights from extensive datasets and uncover hidden correlations among the data to analyse criminal networks [11].

Suhong Kim et.al In This method it discusses the use of machine learning algorithms to analyse various types of crime data. Two distinct data processing techniques were employed to analyse data on crime in Vancouver during the fifteen years before. When forecasting crime in Vancouver, the accuracy range boosted decision tree and K-nearest neighbour machine learning models was 39% to 44% [12].

L. -C. Yu et. al presented data-driven approach that combines the Using machine learning methods with the goal of predicting developing crime hotspots requiring added police attention. Among the machine learning techniques used, the Deep Learning algorithm demonstrated superior performance when compared to others like Random Forest and Naïve Bayes in anticipating potential crime hotspots.[13].

Mary Shermila et.al In This a system is designed to detect crime patterns by analysing the information Gathered from the site of the crime, then predict potential the system aims to provide a description of the possible perpetrator responsible for the crime by utilizing machine learning algorithms, including Multilinear Regression, K-Neighbour Classifier, and Neural Networks. [14].

M. Aasha et. al presented a method investigates the study focuses on utilizing various data mining clustering methods to analyse crime data in Tamilnadu, India, It has been compiled using data obtained by the NCRB. Some of the methods used in this process are K-Means clustering, Agglomerative clustering, and Density Based Spatial Clustering with Noise (DBSCAN) algorithms. [15].

Mahmud et.al conveys order to forecast crime rates, the research uses data mining and machine learning approaches. Various methods including regression analysis, Random forests and decision trees are categorised and evaluated for their effectiveness in predicting crime rates [16].

Priyanka et. al presented the utilization of online newspaper reports to analyse crimes against women. It classifies the challenges of data collection and analysis from online sources and recommends using machine learning techniques like text mining and sentiment analysis to extract relevant information. The main objective is to gain a better understanding and analysis of crimes against women [17].

G. Rani et.al In Several machines learning algorithms, the study uses a variety of techniques, including Fuzzy-C, SVM, K-means A priori Algorithm, FP trees, CART algorithm to find patterns and compare them to actual data. This procedure normalises the data, gets rid of any irregularities, and yields more

precise results. By implementing this approach, it is possible to identify regions and criminals in a timely and appropriate manner [18].

Nair et. al presented an approach that multiple techniques are employed to identify crime hotspots using deep learning technique. The process includes using Long-Term, Short-Term Memory then CNN can provide significant insights into the potential of deep learning in crime analysis and identifying crime hotspots [19].

Gorkem et.al conveys this approach presents an overview of how Crime prediction involves numerous machines learning methods, including k-means clustering and support vector machines (SVM) are utilized to analyse diverse forms of crime data. These algorithms are implemented to predict crime patterns, detect crime hotspots, and allocate resources for preventing further crime [20].

T. S. Sudha et. al presented a model utilizes flexible computational tools and approaches for machine learning to analyse the frequency of cybercrimes. various states of a country, helping to classify these crimes. By using security analytics in combination with data analytic approaches, it becomes possible to analyse and classify cyber offenses based on integrated data, whether structured or unstructured [21].

N. Kanimozhi et.al implemented the idea it compares six different algorithms to determine which one gives the best results, ultimately finding that Random Forest performs the best. The primary goal is to determine which types of crime are most prevalent, along with when and where they occur. High accuracy in classifying different crime patterns is achieved through the use of algorithms for machine learning such as Naïve Bayes [22].

Verma et. al presented the use of the K-Nearest neighbour's algorithm for predicting the type and location of crimes. The algorithm's performance is evaluated, and it is found to be effective, although with limitations such as a small sample size and not considering the temporal aspects of crime [23].

Sukumar & Hassan et.al presented the publication provides a thorough analysis of the many methods used discipline of crime hotspot identification. It explores the advantages and disadvantages of various techniques as well as potential issues with data quality and privacy. The study also highlights the significance of deep learning and time series analysis for predicting crime patterns

as well as the value of machine learning and data mining methodologies, particularly clustering algorithms, for identifying offence hotspots [24].

Fadratul Hafinaz et. al presented a Various data mining techniques are utilized in criminal analysis and crime prediction, including the prediction of crime hotspots based on past data. Using pre-processed datasets, a predictive model is created by applying the Naïve Bayesian algorithm. This model analyses the data and aids in predicting crime trends for a particular region in the future [25].

CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 SOFTWARE REQUIREMENTS

- Windows 10 OS 64-bit
- Python
- Html
- CSS
- Anaconda with Jupiter Notebook

3.1.1 WINDOWS 10

Windows 10 is a Microsoft operating system for personal computers, tablets, embedded devices and internet of things devices. Microsoft released Windows 10 in July 2015 as a follow-up to Windows 8. The company has said it will update Windows 10 in perpetuity rather than release a new, full-fledged operating system as a successor

3.1.2 PYTHON

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library. Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

3.1.3 HTML

HTML (Hypertext Markup Language) is a markup language used to create and structure content on the web. HTML uses tags and attributes to define elements such as headings, paragraphs, links, images, tables, and forms. The content of a webpage is created using HTML and is then styled using CSS (Cascading Style Sheets) and often enhanced with interactive functionality using JavaScript. HTML is an essential part of building websites and web applications, and is supported by all modern web browsers.

3.1.4 CSS

CSS (Cascading Style Sheets) is a style sheet language used to describe the presentation of a document written in HTML or XML. CSS provides a way to separate the presentation from the content, allowing designers to control the look and feel of a webpage without having to modify the underlying HTML code. CSS works by applying styles to HTML elements using selectors. Selectors can target specific elements based on their tag name, class name, ID, or other attributes. CSS styles can be defined inline within the HTML code, in an external stylesheet file, or in the head section of an HTML document. CSS offers a wide range of styling options, including font styles, colours, backgrounds, borders, layouts, and animations. CSS can also be used to make web pages responsive, adjusting the layout and style of a page based on the screen size and device type of the user. CSS is an essential tool for web designers and developers, and is supported by all modern web browsers.

3.1.5 ANACONDA WITH JUPITER NOTEBOOK

Anaconda is a distribution of the Python programming language that includes a comprehensive set of packages for scientific computing and data analysis. It includes popular packages like NumPy, Pandas, Matplotlib, and SciPy, as well as development environments like Jupiter Notebook and Spyder. Jupiter Notebook is a web-based interactive computing environment that allows users to create and share documents that contain live code, equations, visualizations, and narrative text. It supports several programming languages, including Python, R and others. Anaconda comes with Jupiter Notebook pre-installed, making it easy for users to create and run Jupiter notebooks without having to install any additional

software. Users can launch Jupiter Notebook from the Anaconda Navigator or from the command line, and then create new notebooks or open existing ones. Jupiter Notebook provides a flexible and powerful environment for data analysis, machine learning, and scientific computing. It allows users to combine code, data, and narrative text in a single document, making it easy to document and share their work with others.

3.2 HARDWARE REQUIREMENT

- Hard Disk
- RAM – 1 GB
- Processor – Dual Core or Above
- Mouse
- Keyboard
- Monitor

3.2.1 *HARD DISK*

A hard disk or fixed disk, is an electro-mechanical data storage device that uses magnetic storage to store and retrieve digital information using one or more rigid rapidly rotating disks (platters) coated with magnetic material. During the mid-1990s the typical hard disk drive for a PC had a capacity of about 1 gigabyte. As of January 2019, desktop hard disk drives typically had a capacity of 1 to 6 terabytes, with the largest-capacity drives reaching 15 terabytes

3.2.2 *RAM*

Random access memory (RAM) is a type of data storage used in computers that is generally located on the motherboard. This type of memory is volatile and all information that was stored in RAM is lost when the computer is turned off. The RAM chip may be individually mounted on the motherboard or in sets of several chips on a small board connected to the motherboard. Older memory types were in the form of chips called dual in-line package (DIP). Although DIP chips are still used today, the majority of memory is in the form of a module, a narrow printed circuit board attached to a connector on the motherboard. The three main memory circuit boards 21 types containing chips are: RIMMs (Rambus in-line memory

modules), DIMMs (dual inline memory modules) and SIMMs (single in-line memory modules). Most motherboards today use DIMMs. Sometimes, people get confused about the difference between memory and storage, in part because both can be measured in megabytes (MB), gigabytes (GB) and terabytes (TB). By contrast, storage is slower, but it can retain data when the device is powered down. So, for example, if a document has been saved to a hard drive prior to a power outage or system crash, the user will still be able to retrieve it when the system is back up and running.

3.2.3 PROCESSOR

A processor is an integrated electronic circuit that performs the calculations that run a computer. A processor performs arithmetical, logical, input/output (I/O) and other basic instructions that are passed from an operating system (OS). Most other processes are dependent on the operations of a processor. Dual core is a CPU that has two distinct processors that work simultaneously in the same integrated circuit. This type of processor can function as efficiently as a single processor but can perform operations up to twice as quickly.

3.2.4 MOUSE

A mouse is a small handheld input device that controls a computer screen's cursor or pointer in conjunction with the way it is moved on a flat surface. The mouse term name originates from its likeness to a small, corded and elliptical shaped device that looks like a mouse tail. Some mouse devices have integrated features, such as extra buttons that may be programmed and assigned with different commands. Because the mouse reduces the use of a keyboard, its invention and continuous innovation is considered one of the most important breakthroughs in computer ergonomics.

3.2.5 KEYBOARD

A computer keyboard is an input device used to enter characters and functions into the computer system by pressing buttons, or keys. It is the primary device used to enter text. A keyboard typically contains keys for individual letters, numbers and special characters, as well as keys for specific functions. A keyboard is connected to a computer system using a cable or a wireless connection. A

computer keyboard is an input device used to enter characters and functions into the computer system by pressing buttons, or keys. It is the primary device used to enter text. A keyboard typically contains keys for individual letters, numbers and special characters, as well as keys for specific functions.

3.2.6 MONITOR

A computer monitor is an electronic device that shows pictures for computer, Monitor often look similar to televisions. The main difference between a monitor and a television is that a monitor does not have a television tuner to change channels. Monitors often have higher display resolution than televisions. The advent of display technology has paved the way for the continuous evolution of the monitor, whether it's for computers, television, mobile devices or any device that has a display. The current contenders for top-tier technology being used for display devices includes Super LCD 3 (SLCD3) and Super AMOLED. It should be noted that LED displays are actually just a kind of LCD display that use LED lights as backlight illumination.

3.3 USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

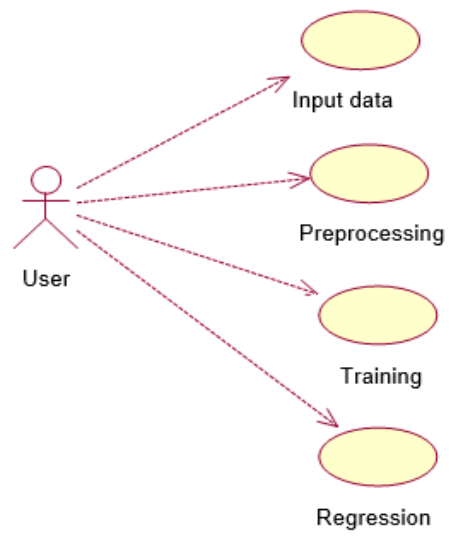


Figure 3.1 – Use Case Diagram for Crime Prediction

CHAPTER 4

DESCRIPTION OF THE PROPOSED SYSTEM

4.1. ARCHITECTURE OF PROPOSED SYSTEM

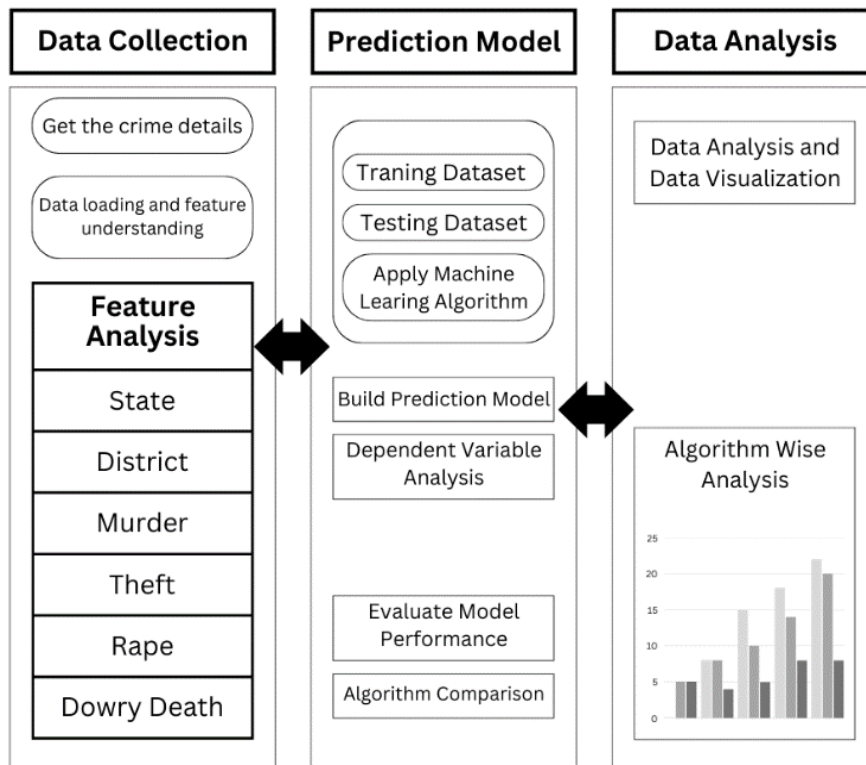


Figure 4.1 – System Architecture

4.2 PROPOSED SYSTEM

We introduced data mining algorithm to predict crime. K-means algorithm plays an important role in analysing and predicting crimes. K-means algorithm will cluster co-offenders, collaboration and dissolution of organized crime groups, identifying various relevant crime patterns, hidden links, link prediction and statistical analysis of crime data. In this paper, k-means algorithm is used for crime prediction. The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the

accuracy. The k-means algorithm will be used for crime prediction. Visualization of dataset is done to analyse the crimes which may have occurred in the country.

4.2.1 ADVANTAGES OF PROPOSED SYSTEM

- The purpose of this work is to improve our previously proposed prediction framework through alternative crime mapping and feature engineering approaches, and provide an open-source implementation that police analysts can use to deploy more effective predictive policing.
- This work helps the law enforcement agencies to predict and detect crimes in India with improved accuracy and thus reduces the crime rate.

4.3 EXISTING SYSTEM

Liu et al. Compared the random forest and spatiotemporal KDE method, found that the random forest algorithm is more efficient than the traditional spatiotemporal KDE method in the smaller time scale and grid space unit. Gabriel et al. used the Gated Localized Diffusion Network for crime prediction at the street segment level. Compared with the traditional Network-time KDE method, the diffusion network approach significantly increased the prediction accuracy. The ability of machine learning algorithm in processing non-linear relational data has been confirmed in many fields, including crime prediction. It has a faster training speed, can handle very high-dimensional data, and can also extract the characteristics of the data.

4.4 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

4.5 DATA FLOW DIAGRAM

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

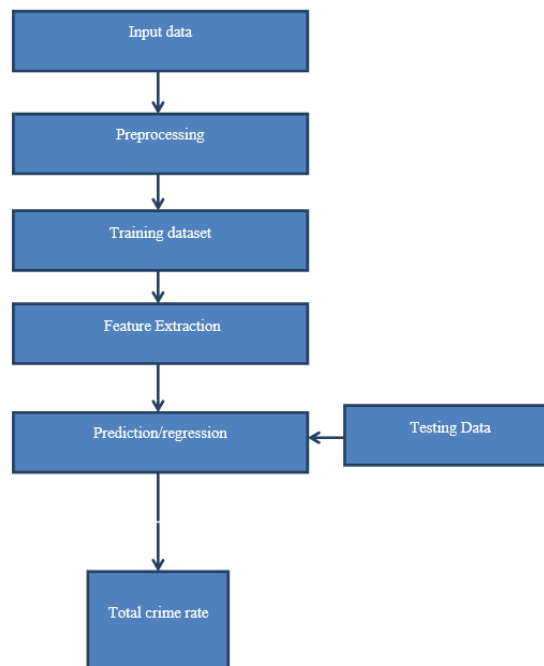


Figure 4.2 – Data Flow Diagram

CHAPTER 5

IMPLEMENTATION DETAILS

5.1 MODULES

The system has four main modules

- Data Collection
- Data Preparation
- Model Selection
- Analyse and Prediction

5.1 DATA COLLECTION

This is the first real step towards the real development of a machine learning model, collection data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions and etc. Comparison of Machine Learning Algorithms for Predicting Crime Hotspots taken from Kaggle and some other source.

5.1.1 DATASET:

The dataset consists of 821 individual data. There are 27 columns in the dataset, which are described below.

- STATE : State in India
- DISTRICT : District in the state of India.
- YEAR : 2001-2018
- MURDER : Total number of murder rate
- RAPE : Total number of rape rate
- THEFT : Total number of theft rate
- TOTAL CRIME : Total number of total crime rate

5.2 DATA PREPARATION

We will transform the data. By getting rid of missing data and removing some columns. First, we will create a list of column names that we want to keep or retain.

Next, we drop or remove all columns except for the columns that we want to retain. Finally, we drop or remove the rows that have missing values from the data set.

5.3 MODEL SELECTION

While creating a machine learning model, we need two datasets, one for training and other for testing. But now we have only one. So, let's split this in two with a ratio of 80:20. We will also divide the data frame into feature column and label column. Here we imported `train_test_split` function of `sk learn`. Then use it to split the dataset. Also, `test_size = 0.2`, it makes the split with 80% as train dataset and 20% as test dataset. Once the model is trained, we need to Test the model. For that we will pass `test_x` to the `predict` method.

5.3.1 THE RANDOM FOREST ALGORITHM:

A machine learning algorithm called random forest generates an output by creating a group of decision trees and utilizing their predictions. To avoid overfitting, each tree is trained on a subset of the data and considers a random subset of features at each node. Random forest employs techniques like bagging and boosting to enhance its performance. It can handle classification and regression tasks using numerical and categorical data. With minimal hyperparameter tuning required, it has been employed in various applications and shown to outperform other Decision trees and Two examples of machine learning techniques that are frequently utilized as algorithms include support vector machines.

5.3.2 DECISION TREE:

Decision trees and support vector machines are two examples of machine learning approaches to make predictions, which is presented in a flowchart-like structure. It includes nodes that represent decisions or outcomes, and edges that depict the potential paths to reach these outcomes. Decision trees are typically utilized in classification and regression problems due to their ability to interpret and visualize data easily. Nevertheless, they are susceptible to overfitting, which can be managed by applying techniques such as pruning or ensemble methods.

5.3.3 XG BOOST:

XGBoost is a machine learning library that employs gradient boosting to create predictive models. This open-source platform starts with a single decision tree and progressively adds more trees to correct prediction errors. To prevent overfitting, XGBoost provides L1 and L2 regularization, and it has an "early stopping" feature that halts training when the model's performance does not improve further. XGBoost is a fast and scalable library that outperforms other machine learning algorithms in many domains, such as click-through rate prediction, customer churn prediction, and image classification. Furthermore, it has been the champion model in various machine learning competitions, including the prestigious Kaggle competition.

5.3.4 SUPPORT VECTOR MACHINE

SVM are subset machine learning algorithms that are capable of classification and regression analysis. SVMs separate using a hyperplane data into two classes, maximizing the gap between them. These algorithms can process high-dimensional data, are resistant to overfitting, and are effective in classifying intricate data sets. Kernel functions can be used to transform the input data. However, SVMs are highly sensitive to the selection of hyperparameters and kernel functions. SVMs are widely used in image classification, text classification, and bioinformatics.

5.3.5 MULTILAYER PERCEPTRON

The multilayer perceptron (MLP) is an artificial neural network utilized for machine learning in classification and regression. It involves multiple layers of nodes that apply nonlinear activation functions to generate outputs. Backpropagation and regularization techniques can be used to train MLPs. MLPs can model nonlinear relationships and work with diverse data types, but may be expensive and require tuning of hyperparameters. MLPs have found applications in speech recognition, image classification, and financial forecasting, and have been combined with other neural networks to enhance Capability

5.4 ANALYSE AND PREDICTION:

In the actual dataset, we chose only 3 features:

- STATE: State in India
- DISTRICT: District in the state of India.
- Year: 2001-2018

For Prediction,

- Total number of murder rate
- Total number of rape rate
- Total number of theft rate
- Total number of total crime rate

5.4.1 ACCURACY ON TEST SET:

We got a accuracy of 95.1%,97.1%, 98.1%, 96.5%, on test set.

CHAPTER 6

RESULTS AND DISCUSSION

This project uses Python software to analyse a dataset containing both categorical and numerical values. The main goal of the analysis is to predict and estimate crime rates based on location and time. This are done by using machine algorithms to achieve this goal. The use of algorithms in the analysis is important because they can help identify patterns in the dataset that may not be immediately obvious. Which are tools that can assist in the analysis process by providing a fast and convenient way to explore data. These algorithms can help identify key features of the dataset that may be important predictors of crime rates. To evaluate the accuracy of the models developed using the algorithms, cross-validation is used. Cross-validation is a technique that involves dividing the dataset into a training set and a test set, and then repeating this process multiple times (in this case, 10 times) for each fold. This helps to ensure that the models developed are accurate and can be used to predict crime rates with a reasonable degree of confidence. Overall, the project suggests that Python and its associated tools can be useful in analysing complex datasets and identifying patterns that can be used to predict real-world phenomena such as crime rates. The use of algorithms and cross-validation can help ensure that the models developed are accurate and reliable, which is essential for making informed decisions based on the analysis.

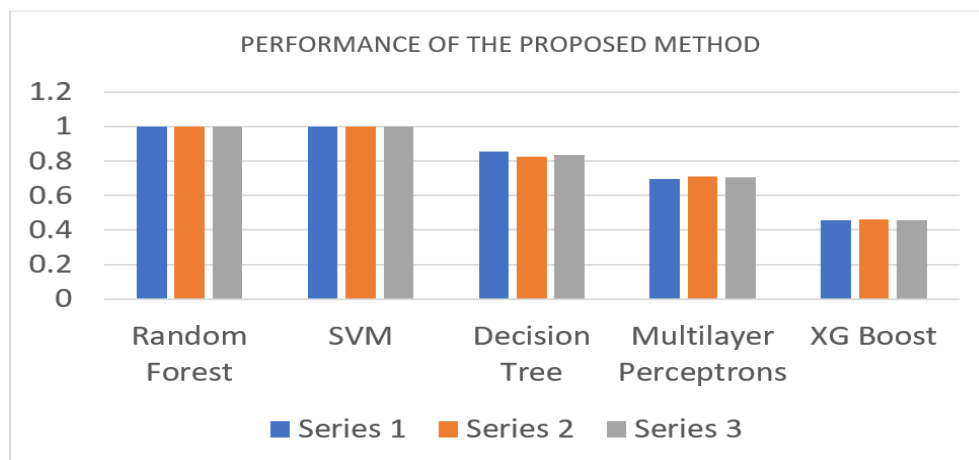


Figure 6. 1 – Comparison of different algorithms

CHAPTER 7

CONCLUSION

7.1 CONCLUSION

The project focuses on using machine learning technology to predict the type of crime that may occur based on the location of its occurrence. The first step involved cleaning and transforming the training data set. Once the data set was prepared, a model was built using machine learning techniques to predict the type of crime with good accuracy. This process has made it easier to identify patterns and relationships among different data sets, helping to understand the factors that contribute to crime in society. The use of machine learning technology has proven to be an effective tool in predicting and preventing crime. To aid in the analysis of the data set, data visualization techniques were used to generate various graphs such as bar, pie, line, and scatter graphs. Each type of graph had its own unique characteristics that helped to extract interesting statistics from the Indian crime data set. The data visualization techniques used were instrumental in providing insights into the nature of the crimes committed in India. The resulting statistics and insights could prove invaluable in developing strategies to make society safer by understanding the factors that contribute to crime. The project demonstrates how machine learning and data visualization technologies can be used to develop solutions that can make a difference in society.

7.2 FUTURE WORK

The proposed project has limitations that need to be addressed. One such limitation is the absence of class labels, which may result in zero evaluation opportunities. To overcome this limitation, future extensions of this work may involve using more machine learning classification models to improve the accuracy of crime rate prediction and overall performance. These improvements could pave the way for better research in the future. In addition, the project could incorporate information on neighbourhood incomes to identify any potential relationships between an individual's income levels in certain neighbourhoods and the crime

rate. This could provide valuable insights for future research and development. By addressing these limitations, the project has the potential to enhance the accuracy of crime prediction and contribute to a safer society through the use of machine learning and data visualization techniques.

REFERENCES

- [1]. Perna G. Kapoor, Prem Singh, and Aswani Kumar Cherukuri (2020). Study of a Crime Data Collection Using Formal Concept Analysis (FCA)
- [2]. Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.
- [3]. Khushabu A. Bokde, Tisksha P. Kakade, Dnyaneshwari S. Tumasare, and Chetan G. Wadhai, B.E. students, "Crime Detection Methods Using Data Mining and K-Means," International Journal of Engineering Research & Technology (IJERT), 2018
- [4]. H. Benjamin Fredrick David and A. Suruliandi published a study in 2017 in the ICTACT Journal on Soft Computing titled "Study on Crime Analysis and Prediction Using Data Mining Techniques."
- [5]. Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization And prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.
- [6]. S. Rajkumar, M. Sakkarai Pandi, and The International Journal of Recent Trends in Engineering & Research published a 2019 article titled "Crime Analysis and Prediction Using Data Mining Techniques."
- [7]. D. M S and S. Shankaraiah, "Crime Analysis and Prediction using Machine Learning Algorithms," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022
- [8]. Ayisheshim Almaw and Kalyani Kadam, "Survey Paper on Crime Prediction Using Ensemble Method," International Journal of Pure and Applied Mathematics, 2018.
- [9]. Dr. M. Sreedevi, A. Harha Vardhan Reddy, and Ch. Venkata Sai Krishna Reddy published a review titled "Review on Crime Analysis and Prediction Using Data Mining Methods" in the International Journal of Innovative Research in Science, Engineering, and Technology in 2018.
- [10]. K.S. N. Murthy, A.V.S. Pavan Kumar, Gangu Dharmaraju, international journal of engineering, Science and mathematics, 2017.

- [11]. Deepiika k.K, Smitha Vinod, Crime analysis in india using data mining techniques, international journal of Engineering and technology, 2018.
- [12]. Hitesh Kumar Reddy ToppyiReddy, Bhavana Saini, Ginika mahajan, Crime Prediction &Monitoring Framework Based on Spatial Analysis, International Conference on Computational Intelligence Data Science (ICCIDS 2018).
- [13]. Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri Fraser International College, Simon Fraser Universit.
- [14]. Y. -L. Lin, T. -Y. Chen and L. -C. Yu, "Using Machine Learning to Assist Crime Prevention," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, Japan, 2017
- [15]. Mary Shermila, A. B. Bellarmine and N. Santiago, "Crime Data Analysis and Prediction of Perpetrator Identity Using Machine Learning Approach," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018
- [16]. S. Sivaranjani, S. Sivakumari and M. Aasha, "Crime prediction and forecasting in Tamilnadu using clustering approaches," 2016 International Conference on Emerging Technological Trends (ICETT), Kollam, India, 2016
- [17]. Mahmud, Sakib & Nuha, Musfika & Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining.
- [18]. Das, Priyanka & Das, Asit. (2017). Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment.
- [19]. P. Kaur, G. Rani, T. Sharma and A. Sharma, "A Comparative Study to analyze crime threats using data mining and machine learning approach," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021
- [20]. Nair, Sankar & E.S., Gopi. (2020). Deep Learning Techniques for Crime Hotspot Detection.
- [21]. Jenga, Karabo & Catal, Cagatay & Kar, Gorkem. (2023). Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing.

- [22]. T. S. Sudha and C. Rupa, "Analysis and Evaluation of Integrated Cyber Crime Offences," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019
- [23]. N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021
- [24]. Kumar, Akash & Verma, Aniket & Shinde, Gandhali & Sukhdeve, Yash & Lal, Nidhi. (2020). Crime Prediction Using K-Nearest Neighboring Algorithm.
- [25]. Butt, Umair & Letchmunan, Sukumar & Hassan, Fadratul Hafinaz & Ali, Mubashir & Baqir, Anees & Sherazi, Husnain. (2020). Spatio-Temporal Crime Hotspot Detection and Prediction: A Systematic Literature Review. IEEE Access.

APPENDIX

A. SOURCE CODE

```
from flask import Flask, render_template, request, url_for, Markup, jsonify
import pickle
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from werkzeug.utils import secure_filename
import pickle
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

app = Flask(__name__) #Initialize the flask App
rap = pickle.load(open('kmean.pkl','rb'))
regresso = pickle.load(open('murder.pkl','rb'))
sandy = pickle.load(open('theft.pkl','rb'))
raps = pickle.load(open('rap.pkl','rb'))
last = pickle.load(open('total.pkl','rb'))
scaler = MinMaxScaler()
#scaler = MinMaxScaler()
#scaler = MinMaxScaler()
@app.route('/')
@app.route('/first')
def first():
    return render_template('first.html')
@app.route('/login')
def login():
    return render_template('login.html')
@app.route('/chart')
def chart():
    return render_template('chart.html')
```

```

@app.route('/murder')
def murder():
    return render_template('murder.html')
@app.route('/theft')
def theft():
    return render_template('theft.html')
@app.route('/rape')
def rape():
    return render_template('rape.html')
@app.route('/upload')
def upload():
    return render_template('upload.html')
@app.route('/preview',methods=["POST"])
def preview():
    if request.method == 'POST':
        dataset = request.files['datasetfile']
        df = pd.read_csv(dataset,encoding = 'unicode_escape')
        df.set_index('Id', inplace=True)
        return render_template("preview.html",df_view = df)
@app.route('/prediction')
def prediction():
    return render_template("prediction.html")
@app.route('/predict',methods=['POST'])
def predict():
    features = [float(x) for x in request.form.values()]
    final_features = [np.array(features)]
    #final_features = pca.transform(final_features)
    y_pred = rap.predict(final_features)
    if y_pred[0] == 1:
        label="High rate crime area"
    elif y_pred[0] == 0:
        label="Low rate crime area"
    return render_template('prediction.html', prediction_text= label)

```

```

@app.route('/crime')
def crime():
    return render_template("crime.html")
@app.route('/predicts',methods=['POST'])
def predicts():
    int_feature = [x for x in request.form.values()]
    final_features = [np.array(int_feature)]
    prediction=regresso.predict(final_features)
    pred=format(int(prediction[0]))
    return render_template('crime.html', prediction_text= pred)
@app.route('/total')
def total():
    return render_template("total.html")
@app.route('/prediction',methods=['POST'])
def totals():
    int_feature = [x for x in request.form.values()]
    final_features = [np.array(int_feature)]
    prediction=last.predict(final_features)
    pred=format(int(prediction[0]))
    return render_template('total.html', prediction_text= pred)
@app.route('/crimes')
def crimes():
    return render_template("crimes.html")
@app.route('/predictions',methods=['POST'])
def predictions():
    int_feature = [x for x in request.form.values()]
    final_features = [np.array(int_feature)]
    prediction=raps.predict(final_features)
    pred=format(int(prediction[0]))
    return render_template('crimes.html', prediction_text= pred)
@app.route('/last')
def last():
    return render_template("last.html")

```

```

@app.route('/data',methods=['POST'])
def data():
    int_feature = [x for x in request.form.values()]
    final_features = [np.array(int_feature)]
    prediction=sandy.predict(final_features)
    pred=format(int(prediction[0]))
    return render_template('last.html', prediction_text= pred)

#@app.route('/chart')
#def chart():
    #abc = request.args.get('news')
    #input_data = [abc.rstrip()]
    # transforming input
    #tfidf_test = tfidf_vectorizer.transform(input_data)
    # predicting the input
    #y_pred = pac.predict(tfidf_test)
    #output=y_pred[0]
    #return render_template('chart.html', prediction_text='Review is
    {}.format(y_pred[0]))
if __name__ == '__main__':
    app.run(debug=True)

```

```

<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="utf-8">
    <meta content="width=device-width, initial-scale=1.0" name="viewport">
    <title>crime</title>
    <meta content="" name="description">
    <meta content="" name="keywords">
    <link href="../static/img/favicon.png" rel="icon">
    <link href="../static/img/apple-touch-icon.png" rel="apple-touch-icon">

```

```

<link
href="https://fonts.googleapis.com/css?family=Open+Sans:300,300i,400,400i,600
,600i,700,700i|Playfair+Display:ital,wght@0,400;0,500;0,600;0,700;1,400;1,500;1
,600;1,700|Poppins:300,300i,400,400i,500,500i,600,600i,700,700i"
rel="stylesheet">
<link href="../static/vendor/animate.css/animate.min.css" rel="stylesheet">
<link href="../static/vendor/aos/aos.css" rel="stylesheet">
<link href="../static/vendor/bootstrap/css/bootstrap.min.css" rel="stylesheet">
<link href="../static/vendor/bootstrap-icons/bootstrap-icons.css" rel="stylesheet">
<link href="../static/vendor/boxicons/css/boxicons.min.css" rel="stylesheet">
<link href="../static/vendor/glightbox/css/glightbox.min.css" rel="stylesheet">
<link href="../static/vendor/swiper/swiper-bundle.min.css" rel="stylesheet">
<link href="../static/css/style.css" rel="stylesheet">
<body>
<div id="topbar" class="d-flex align-items-center fixed-top">
  <div class="container d-flex justify-content-center justify-content-md-between">
    </div>
  </div>
<header id="header" class="fixed-top d-flex align-items-cente">
  <div class="container-fluid container-xl d-flex align-items-center justify-content-
lg-between">
    <h1 class="col-lg-8"><a href="index.html"> <span>Crime Prediction and
Analysis</span></a></h1>
    <nav id="navbar" class="navbar order-last order-lg-0">
      <ul>
        <li>
<a class="nav-link scrollto active" href="{{ url_for('first')}}">Home</a></li>
        <li><a class="nav-link scrollto" href="{{ url_for('login')}}">Login</a></li>
      </ul>
      <i class="bi bi-list mobile-nav-toggle"></i>
    </div>
  </header>
<section id="hero" class="d-flex align-items-center">

```

```

<div class="container position-relative text-center text-lg-start" data-
aos="zoom-in" data-aos-delay="100">
  <div class="row">
    <div class="col-lg-8">
      <h1>crime rate prediction </h1>
      <h1><span>using ML algorithm</span></h1>
    </div>
  </div>
</div>
</section>
<main id="main">
<script src="../static/vendor/aos/aos.js"></script>
<script src="../static/vendor/bootstrap/js/bootstrap.bundle.min.js"></script>
<script src="../static/vendor/glightbox/js/glightbox.min.js"></script>
<script src="../static/vendor/isotope-layout/isotope.pkgd.min.js"></script>
<script src="../static/vendor/php-email-form/validate.js"></script>
<script src="../static/vendor/swiper/swiper-bundle.min.js"></script>
<script src="../static/js/main.js"></script>
</body>
</html>

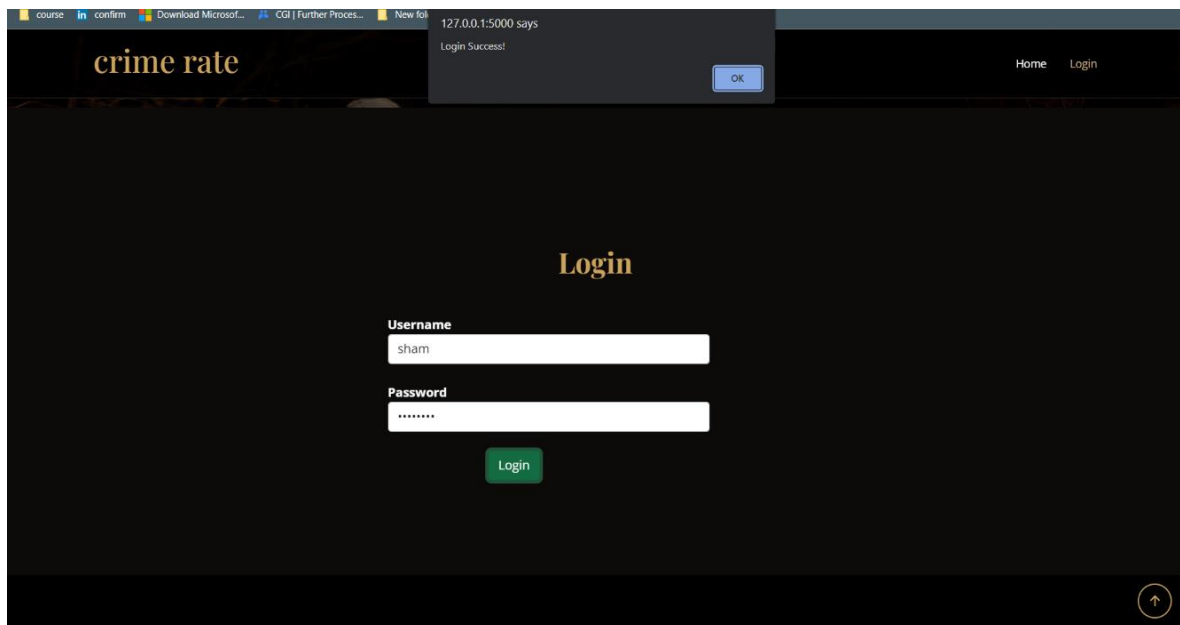
```

B. SCREENSHOTS

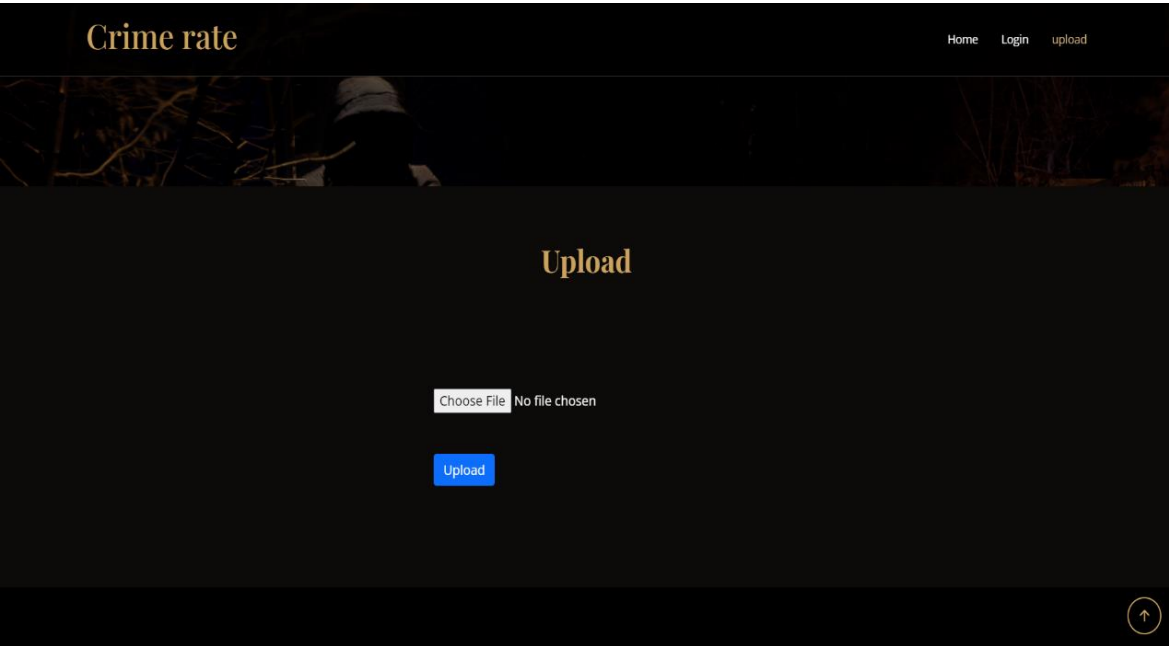
Home Page



Login Page



File Upload



Data sets

Crime rate

HomeLoginupload

Preview

PREVIEW

	DISTRICT	STATE- todel	ASSAULT ON WOMEN	DOWRY DEATHS	KIDNAPPING AND ABDUCTION	MAJOR_CRIMES_TOTAL	MURDER	RAPE	RIOTS	ROBBI	
Id											
1	Ariyalur	Tamil nadu	64.0	0.0	1.0	26.0	233.0	27	20	19.0	9.0
2	Chennai	Tamil nadu	45.0	2.0	16.0	47.0	1858.0	103	47	60.0	74.0
3	Chennai rly	Tamil nadu	2.0	1.0	2.0	0.0	250.0	0	0	14.0	58.0
4	Chennai city	Tamil nadu	61.0	1.0	18.0	30.0	1059.0	86	29	85.0	102.0
5	Coimbatore	Tamil nadu	39.0	11.0	1.0	62.0	664.0	41	13	83.0	78.0

[course](#)
[confirm](#)
[Download Microsof...](#)
[CGI | Further Proces...](#)
[New fol](#)

Crime rate

809 NORTH Delhi 171.0 -

Delhi 333.0 -

811 NORTH WEST Delhi 371.0 -

812 OUTER Delhi 530.0 - - - -

813 RAILWAY Delhi 4.0 - - - -

814 SOUTH Delhi 708.0 - - - -

815 SOUTH WEST Delhi 510.0 - - - -

816 SOUTH-WEST Delhi 389.0 - - - -

817 spl CELL Delhi 1.0 - - - -

818 SPUWAC Delhi 0.0 - - - -

819 VIGILANCE Delhi 0.0 - - - -

820 WEST Delhi 450.0 - - - -

localhost:5000 says

training finished!

OK

26 65 3.0 233.0 7280 3725

51 223 15.0 8 Home Login upload

46 190 17.0 597.0 12135 3458

101 372 5.0 852.0 15033 2796

6 7 0.0 60.0 3837 697

43 246 7.0 325.0 13965 1504

58 275 5.0 265.0 10812 2711

58 171 5.0 225.0 6872 6936

0 0 0.0 0.0 0 1756

0 0 0.0 0.0 0 3676

0 0 0.0 0.0 0 5243

64 256 5.0 442.0 16944 1635

Click to Train | Test

Prediction of Crime Rate

Crime rate

Home Login upload cluster_prediction prediction >

Crime Rate prediction

State: Tamil nadu District: Ariyalur MURDER: 456

RAPE: 43 THEFT: 65 DOWRY DEATH: 43 Year: 2024

Predict

Total crime rate is:

High rate crime area

Crime Rate Analysis



C.RESEARCH PAPER:

Crime Analysis Framework for Predicting Criminal Behaviour Patterns with Machine Learning

Sham Venkat S, R.Jeberson Retna Raj, Arjun A
Department of Information Technology
Sathyabama Institute of Science and Technology
Chennai, India
jebersonretnarajr@gmail.com

Abstract— The objective of the project is to create a system that uses machine learning techniques to predict and look at criminal behavior patterns. The framework consists of three phases: data preprocessing, model training, and prediction and analysis. By analyzing crime data using Kaggle's open-source crime records, this project aims to understand crime patterns, determine the level of crime, and the location and time period in which it occurred. The project employs various machine learning algorithms, including the random forest classifier and k-means clustering, to accurately categorize diverse criminal patterns. The proposed framework employs various visualization techniques to reveal crime trends and predict crime using machine learning algorithms. The input features used in the algorithms include time, state, district, year, murder, rape, theft, and dowry deaths.

Keywords— Machine learning, Crime prediction, Data analysis, Security and privacy, Crime analysis, Crime pattern.

I.INTRODUCTION

Prevention is preferable to treatment. It is preferable to prevent crimes from happening rather than look into how or why they were committed. With today's rising crime rate and violent crime occurring, it has become essential to establish a vaccine system that stops crimes from happening, just as vaccinations are provided to children to prevent disease. By immunizing society against crime, it refers to a variety of strategies including enlightening citizens, raising awareness, enhancing the effectiveness of proactive policing techniques, and other deterrent strategies. Predicting crime incidents has mostly relied on past crime statistics as well as various geospatial and demographic data. Data mining is a process of taking information out of a lot of data and finding patterns in it. Crime analysis can be a successful use of mining data. Using the data kept in repositories, data mining can be very helpful in evaluating and forecasting crimes. The crime rate in India is rising daily, posing a serious threat to the country's ability to govern itself well. It is now hard for intelligence agencies or local law enforcement organizations to examine crime-related data, discover crime patterns, or predict future crimes as a result of the terrible increase in crime rate. In this study, numerous significant crime trends are thoroughly examined, and crime data are statistically analyzed. Law enforcement agencies will find this study useful in developing their plans and techniques to combat crime and disturbance.

A. Scope of the Project

The current efforts are primarily concentrated on two main aspects: forecasting crime rates based on geographical location and comprehending criminal conduct patterns that can assist in resolving criminal inquiries.

B. Objective of the Project

Our work aims to anticipate criminal activity before it occurs by forecasting crime rates, comprehending crime patterns, categorizing crimes according to geographical location, and examining crime trends in Tamil Nadu.

II.LITERATURE SURVEY

In [1] The system employed with FCA and other machine learning techniques used for crime analysis.

Shiju Sathyadevan et. al presented a crime analysis and prevention system this approach is utilized for detecting and scrutinizing criminal trends and patterns. It has the ability to forecast areas that are at a greater risk of crime occurrence and can also provide visual representation of such crime-prone locations. The process involves implementing data mining techniques to extract significant insights from unstructured data that were previously unknown [2].

In [3] a classification model used to classify crimes that are clustered based on their occurrence frequency over several years. Data mining algorithms are employed to analyze, investigate, and detect patterns related to various crimes. The crime dataset used in this study is obtained from the police records of England and Wales from 1990 to 2011. Additionally, the study utilizes the RapidMiner tool to optimize the Outlier Detection operator parameters using the Genetic Algorithm (GA).

Benjamin et al introduced a detailed survey on crime analysis based on supervised and unsupervised learning techniques. This model used to identify the crime and forecasting the crime [4].

In [5] Measures are being implemented to prevent the escalation of crimes against women. Data classification and correlation facilitate the identification of similarities and differences among data objects. Clustering techniques are employed to group data objects based on predetermined criteria. By utilizing data mining methods, it becomes possible to anticipate and visually represent criminal activity, patterns in

various formats, improving our understanding of the crime pattern.

RajKumar. Set. al presented a Various data mining techniques are utilized in criminal analysis and crime prediction, including the prediction of crime hotspots based on past data. Using pre-processed datasets, a predictive model is created by applying the Naïve Bayesian algorithm. This model analyses the data and aids in predicting crime trends for a particular region in the future [6].

In [7] This the objective is to utilize linear regression, K-means, Naïve Bayes will be used in machine learning to predict the state that displays the highest crime rate. The algorithm that demonstrates the most effective combination of efficiency and accuracy will be selected as the optimal approach for crime prediction and analysis.

Ayisheshim Almaraw et. al presented a method it focuses on application of the application of different data mining algorithms and ensemble learning methods is critical in analyzing crime data. Data mining techniques play a vital role in uncovering obscure criminal patterns and addressing crime-related concerns. The study analyses and discusses different methods that are utilized for crime prediction and analysis, including the application of ensemble classification and data mining approaches methods to predict and discover potential future crimes [8].

In [9] This system can forecast regions with a high likelihood of crime occurrences and visualize areas prone to crime. Data mining allows us to extract valuable, previously unknown information from unstructured data. This includes analyzing and detecting patterns using decision trees, linear regression, k-means clustering, and various other data mining techniques for criminal analysis and prediction.

K.S. N. Murthy et. al presented a Crime pattern analysis can be performed by utilizing various Frequent Pattern Mining and Rule Mining algorithms. These algorithms facilitate comprehensive comprehension of frequent pattern mining in diverse fields. Frequent pattern mining is essential for rule mining, and the commonly employed algorithms for frequent pattern mining include Apriorism and FP-Growth [10].

In [11] This process method for detecting crime in In India, the implementation of data mining methods comprises a set of procedures including pre-processing data, clustering, categorization, and visualization. For the purpose of identifying crime, k-means clustering method is used, groups created based on similarity of crime attributes. Data is analyzed for categorization utilizing techniques such as Random Forest and Neural Networks. Crime hotspots are located and visualized using the Google marker clustering method.

Bhavana Saini et. al presented a Different the Machine learning algorithms and visualization techniques are used to forecast how crime will be distributed in a given area. Data pre-processing and the creation of the necessary visuals are part of the initial stage. The use of machine learning methods such as Naive Bayes follows. to extract insights from extensive datasets and uncover hidden correlations among the data to analyze criminal networks [12].

In [13] This method it discusses the use of machine learning algorithms to analyze various types of crime data. Two distinct data processing techniques were

employed to analyze data on crime in Vancouver during the fifteen years before. When forecasting crime in Vancouver, the accuracy range boosted decision tree and K-nearest neighbor machine learning models was 39% to 44%.

L. -C. Yu et. al presented data-driven approach that combines the Using machine learning methods with the goal of predicting developing crime hotspots requiring added police attention. Among the machine learning techniques used, the Deep Learning algorithm demonstrated superior performance when compared to others like Random Forest and Naïve Bayes in anticipating potential crime hotspots.[14].

In [15] This a system is designed to detect crime patterns by analyzing the information Gathered from the site of the crime, then predict potential the system aims to provide a description of the possible perpetrator responsible for the crime by utilizing machine learning algorithms, including Multilinear Regression, K-Neighbor Classifier, and Neural Networks.

M. Aasha et. al presented a method investigates the study focuses on utilizing various data mining clustering methods to analyze crime data in Tamilnadu, India, It has been compiled using data obtained by the NCRB. Some of the methods used in this process are K-Means clustering, Agglomerative clustering, and Density Based Spatial Clustering with Noise (DBSCAN) algorithms. [16].

In [17] In order to forecast crime rates, the research uses data mining and machine learning approaches. Various methods including regression analysis, Random forests and decision trees are categorized and evaluated for their effectiveness in predicting crime rates.

Priyanka et. al presented the utilization of online newspaper reports to analyze crimes against women. It classifies the challenges of data collection and analysis from online sources and recommends using machine learning techniques like text mining and sentiment analysis to extract relevant information. The main objective is to gain a better understanding and analysis of crimes against women [18].

In [19] Several machines learning algorithms, the study uses a variety of techniques, including Fuzzy-C, SVM, K-means A priori Algorithm, FP trees, CART algorithm to find patterns and compare them to actual data. This procedure normalizes the data, gets rid of any irregularities, and yields more precise results. By implementing this approach, it is possible to identify regions and criminals in a timely and appropriate manner.

Nair et. al presented a approach that multiple techniques are employed to identify crime hotspots using deep learning technique. The process includes using Long-Term, Short-Term Memory then CNN can provide significant insights into the potential of deep learning in crime analysis and identifying crime hotspots [20].

In [21] This approach presents an overview of how Crime prediction involves numerous machines learning methods, including k-means clustering and support vector machines (SVM) are utilized to analyze diverse forms of crime data. These algorithms are implemented to predict crime patterns, detect crime hotspots, and allocate resources for preventing further crime.

T. S. Sudha et. al presented a model utilizes flexible computational tools and approaches for machine learning to analyze the frequency of cybercrimes. various states of a country, helping to classify these crimes. By using security analytics in combination with data analytic approaches, it becomes possible to analyze and classify cyber offenses based on integrated data, whether structured or unstructured [22].

In [23] This it compares six different algorithms to determine which one gives the best results, ultimately finding that Random Forest performs the best. The primary goal is to determine which types of crime are most prevalent, along with when and where they occur. High accuracy in classifying different crime patterns is achieved through the use of algorithms for machine learning such as Naïve Bayes.

Verma et. al presented the use of the K-Nearest neighbour's algorithm for predicting the type and location of crimes. The algorithm's performance is evaluated, and it is found to be effective, although with limitations such as a small sample size and not considering the temporal aspects of crime [24].

In [25] The publication provides a thorough analysis of the many methods used discipline of crime hotspot identification. It explores the advantages and disadvantages of various techniques as well as potential issues with data quality and privacy. The study also highlights the significance of deep learning and time series analysis for predicting crime patterns as well as the value of machine learning and data mining methodologies, particularly clustering algorithms, for identifying offence hotspots.

III. PROPOSED ARCHITECTURE

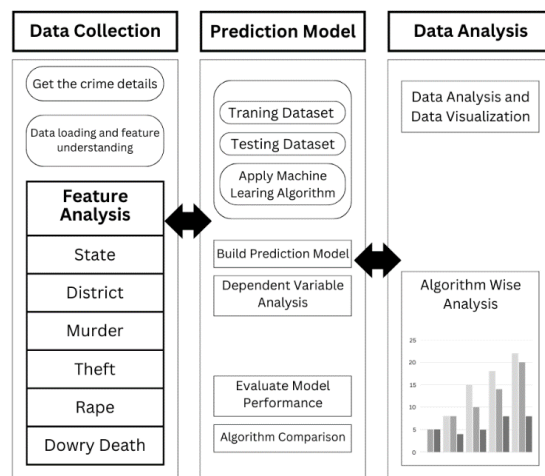


Fig. 1. System Architecture

Modules

In this project. The modules contain 4 types. Which are described below.

- Gathering Data
- Preparing Data
- Choosing a Model
- Analyzing and Predicting

A) Gathering Data:

The actual process of creating a machine learning model and accumulating data starts now. This stage is critical since the amount and quality of data we can gather will determine how effectively the model performs. Among the techniques used to acquire data include manual interventions, web scraping, and others. Comparison of Crime Hotspot Prediction Machine Learning Algorithms from Kaggle.

a) Data set:

There are 821 distinct data points in the dataset. The dataset contains 27 columns, each of which is detailed below. State, district, rape, theft, and murder.

B) Preparing Data:

Process the facts and prepare it for training. Clean up what is probably wanted (do away with duplicates, restoration mistakes, cope missing values, normalize, convert information kinds to eliminate the impact of the order in which records were collected or organized, randomizing facts is a useful technique. Data visualization can assist in identifying significant relationships among variables or class imbalances, but one should be cautious of potential biases. Exploratory analysis can be performed in various ways. Finally, dividing the data into smaller units can facilitate education and evaluation purposes

C) Choosing a Model:

When creating a machine learning model, it is necessary to have two datasets - one for training and another for testing. However, in this scenario, only one dataset is available, so it is suggested to split the 80:20 split the dataset into two pieces. There should be a feature column and a label column in the dataset. To do this, the scikit-learn library's train test split function is utilized. The dataset is divided into a 20% test dataset and an 80% train dataset by selecting test size=0.2. after exercise the model, the next step is to test the model using the predict method and passing test as input.

a) Random Forest Algorithm

A machine learning algorithm called random forest generates an output by creating a group of decision trees and utilizing their predictions. To avoid overfitting, each tree considers a random selection of features at each node and is trained on a subset of the data. Random forest employs techniques like bagging and boosting to enhance its performance. It can handle classification and regression tasks using numerical and categorical data. With minimal hyperparameter tuning required, it has been employed in various applications and shown to outperform other Decision trees and Two examples of machine learning techniques that are frequently utilized as algorithms include support vector machines.

b) Decision Tree Algorithm

Decision trees and support vector machines are two examples of machine learning approaches to make predictions, which is presented in a flowchart-like structure. It includes nodes that represent decisions or outcomes, and edges that depict the potential paths to reach these outcomes. Decision trees are typically utilized in classification and regression problems due to their ability to interpret and visualize data easily. Nevertheless, they are susceptible to overfitting, which

can be managed by applying techniques such as pruning or ensemble methods.

c) Support Vector Machine Algorithm

SVM are subset machine learning algorithms that are capable of classification and regression analysis. SVMs separate using a hyperplane data into two classes, maximizing the gap between them. These algorithms can process high-dimensional data, are resistant to overfitting, and are effective in classifying intricate data sets. Kernel functions can be used to transform the input data. However, SVMs are highly sensitive to the selection of hyperparameters and kernel functions. SVMs are widely used in image classification, text classification, and bioinformatics.

d) XGBoost Algorithm

XGBoost is a machine learning library that employs gradient boosting to create predictive models. This open-source platform starts with a single decision tree and progressively adds more trees to correct prediction errors. To prevent overfitting, XGBoost provides L1 and L2 regularization, and it has an "early stopping" feature that halts training when the model's performance does not improve further. XGBoost is a fast and scalable library that outperforms other machine learning algorithms in many domains, such as click-through rate prediction, customer churn prediction, and image classification. Furthermore, it has been the champion model in various machine learning competitions, including the prestigious Kaggle competition.

e) Multilayer Perceptron Algorithm

The multilayer perceptron (MLP) is an artificial neural network utilized for machine learning in classification and regression. It involves multiple layers of nodes that apply nonlinear activation functions to generate outputs. Backpropagation and regularization techniques can be used to train MLPs. MLPs can model nonlinear relationships and work with diverse data types, but may be expensive and require tuning of hyperparameters. MLPs have found applications in speech recognition, image classification, and financial forecasting, and have been combined with other neural networks to enhance Capability

D. Analysis and prediction:

In the given dataset, only eight features were selected for analysis. The dataset contains a total of 821 data points and is composed of 27 columns, which are elaborated on below.

- **STATE:** Refers to the state in India where the crime incident occurred
- **DISTRICT:** Refers to the district in the respective state where the crime incident occurred
- **Year:** Specifies the time frame of the crime incident between 2001 to 2018
- **MURDER:** Represents the overall death toll incidents
- **RAPE:** Represents the number of rape incidents overall
- **THEFT:** Represents the overall quantity of thefts

- **Total crime:** Represents the total number of incidences of all sorts crime The longitude the place where crime occurrence took place is indicated. Where exactly is slightly altered for anonymity purposes but remains in the same block.

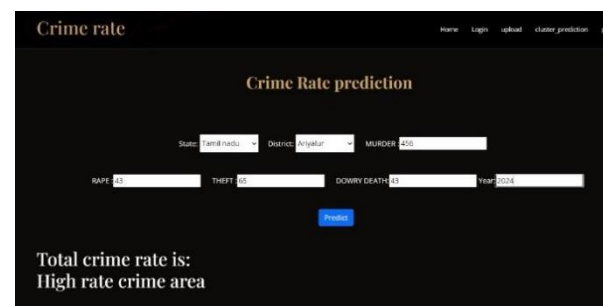
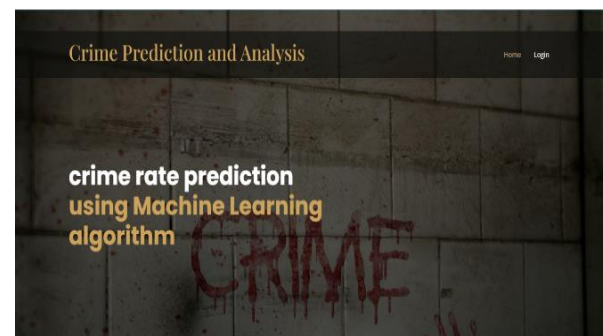
E. Accuracy on check set:

Our model achieved test set accuracies of 95.1%, 97.1%, 98.1%, and 96.5%.

IV. EXPERIMENTAL RESULT

To conduct this project, I used Python software-based approach and familiarity with its use. It contains equipment for records preprocessing, category, regression, and visualization. It can be used to find out various hidden patterns in our dataset and find the maximum determinants of the records. The file uses records set consisting of every of them and an aggregate of specific and numerical values. Thus, the thing discusses buildings for predicting and estimating the extent of crime, which are based on the usage of laptops to perceive algorithms. The objective is to utilize algorithms on datasets to analyze crime rates, specifically in terms of country, district, and year. Helper algorithms are used in the area because they provide a convenient and fast manner to find out about a characteristic. Experiments are run using cross-validation with the default putting folds = 10. Cross-validation is a method for evaluating and comparing predictive models by dividing the original dividing a dataset into a test set and a training set for a model for evaluation. This process is repeated 10 times for each fold in the dataset.

a) Output Screenshot's



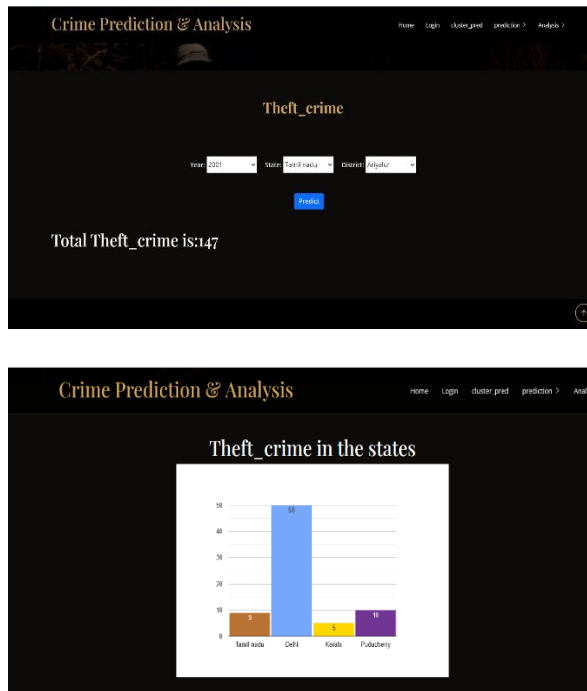


Fig. 2. Execution of web application

b)Table

Algorithm	Precision	Recall	Accuracy
Random Forest	1.000	1.000	1.000
SVM	1.000	1.000	1.000
Decision Tree	0.858	0.820	0.838
Multilayer Perceptron	0.697	0.710	0.704
XGBoost	0.455	0.462	0.458

Table.1. The effectiveness of the suggested method

c) Graph

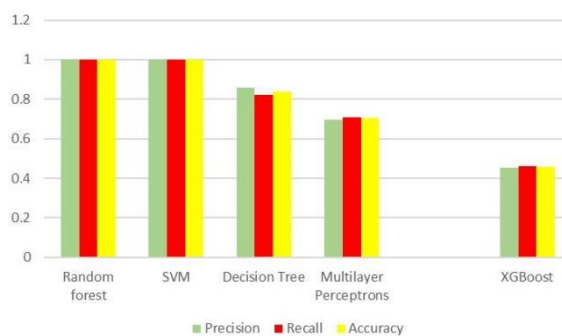


Fig. 3. Graph For Performance of the Proposed Method

V. CONCLUSION

It has become simple to recognize links and patterns thanks to machine learning technologies. within diverse sets of data. The paintings in this mission especially revolves round predicting the extent of crime which can occur if we understand the vicinity wherein it came about. Using the idea of gadget mastering, we built a version the usage of a fixed of training facts that underwent statistics cleansing and records transformation. The crime rate is predicted with

considerable accuracy by the model. Understanding the dataset is greatly aided by data visualization. Histograms, pie, line, and scatter charts are different types of graphs, each with unique characteristics. By creating several graphs and examining the results, we have discovered intriguing facts about crime datasets in India, which could aid in identifying factors that promote community safety.

REFERENCES

- [1] Perna G. Kapoor, Prem Singh, and Aswani Kumar Cherukuri (2020). Study of a Crime Data Collection Using Formal Concept Analysis (FCA)
- [2] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.
- [3] Khushabu A. Bokde, Tisksha P. Kakade, Dnyaneshwari S. Tumasare, and Chetan G. Wadhai, B.E. students, "Crime Detection Methods Using Data Mining and K-Means," International Journal of Engineering Research & Technology (IJERT), 2018[
- [4] H. Benjamin Fredrick David and A. Suruliandi published a study in 2017 in the ICTACT Journal on Soft Computing titled "Study on Crime Analysis and Prediction Using Data Mining Techniques."
- [5] Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization And prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.
- [6] S. Rajkumar, M. Sakkarai Pandi, and The International Journal of Recent Trends in Engineering & Research published a 2019 article titled "Crime Analysis and Prediction Using Data Mining Techniques."
- [7] D. M S and S. Shankaraiah, "Crime Analysis and Prediction using Machine Learning Algorithms," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022
- [8] Ayisheshim Almwaw and Kalyani Kadam, "Survey Paper on Crime Prediction Using Ensemble Method," International Journal of Pure and Applied Mathematics, 2018.
- [9] Dr. M. Sreedevi, A. Harha Vardhan Reddy, and Ch. Venkata Sai Krishna Reddy published a review titled "Review on Crime Analysis and Prediction Using Data Mining Methods" in the International Journal of Innovative Research in Science, Engineering, and Technology in 2018.
- [10] K.S. N. Murthy, A.V.S. Pavan Kumar, Gangu Dharmaraju, international journal of engineering, Science and mathematics, 2017.
- [11] Deepiika k.K, Smitha Vinod, Crime analysis in india using data mining techniques, international journal of Engineering and technology, 2018.
- [12] Hitesh Kumar Reddy ToppyReddy, Bhavana Saini, Ginika mahajan, Crime Prediction &Monitoring Framework Based on Spatial Analysis, International Conference on Computational Intelligence Data Science (ICCIDS 2018).
- [13] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri Fraser International College, Simon Fraser Universit.
- [14] Y. -L. Lin, T. -Y. Chen and L. -C. Yu, "Using Machine Learning to Assist Crime Prevention," 2017 6th IIAI International Congress on Advanced

- Applied Informatics (IIAI-AAI), Hamamatsu, Japan, 2017
- [15] Mary Shermila, A. B. Bellarmine and N. Santiago, "Crime Data Analysis and Prediction of Perpetrator Identity Using Machine Learning Approach," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018
 - [16] S. Sivaranjani, S. Sivakumari and M. Aasha, "Crime prediction and forecasting in Tamilnadu using clustering approaches," 2016 International Conference on Emerging Technological Trends (ICETT), Kollam, India, 2016
 - [17] Mahmud, Sakib & Nuha, Musfika & Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining.
 - [18] Das, Priyanka & Das, Asit. (2017). Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment.
 - [19] P. Kaur, G. Rani, T. Sharma and A. Sharma, "A Comparative Study to analyze crime threats using data mining and machine learning approach," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021
 - [20] Nair, Sankar & E.S., Gopi. (2020). Deep Learning Techniques for Crime Hotspot Detection.
 - [21] Jenga, Karabo & Catal, Cagatay & Kar, Gorkem. (2023). Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing.
 - [22] T. S. Sudha and C. Rupa, "Analysis and Evaluation of Integrated Cyber Crime Offences," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019
 - [23] N. Kanimozhi, N. V. Keerthana, G. S. Pavithra, G. Ranjitha and S. Yuvarani, "CRIME Type and Occurrence Prediction Using Machine Learning Algorithm," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021
 - [24] Kumar, Akash & Verma, Aniket & Shinde, Gandhali & Sukhdeve, Yash & Lal, Nidhi. (2020). Crime Prediction Using K-Nearest Neighboring Algorithm.
 - [25] Butt, Umair & Letchmunan, Sukumar & Hassan, Fadratul Hafinaz & Ali, Mubashir & Baqir, Anees & Sherazi, Husnain. (2020). Spatio-Temporal Crime Hotspot Detection and Prediction: A Systematic Literature Review. IEEE Access.

