PREDICTING A CRIME IN A PARTICULAR REGION USING DECISION TREES

Project Report

Submitted in partial fulfillment for the award of degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Submitted By

KOSIREDDI SAI DIVYA (Regd.No.18L31A1288)

PANDRANKI SHANMUKHA RAO (Regd.No.18l31A1246)

CHINTAGUNTI DILIP KUMAR (Regd.No.18L31A1255)

NALI RAVI TEJA (Regd.No.18L31A1281)

Under the esteemed guidance of

Dr. G. Rajendra Kumar

Professor

VIGNAN's INSTITUTE OF INFORMATION TECHNOLOGY

(AUTONOMOUS)

(Approved by AICTE-New Delhi & Affiliated to JNTUK, Kakinada)

Beside VSEZ, Duvvada, Vadlapudi Post, Gajuwaka, Visakhapatnam - 530 049.

VIGNAN's INSTITUTE OF INFORMATION TECHNOLOGY
(AUTONOMOUS)
(Approved by AICTE-New Delhi & Affiliated to JNTUK, Kakinada)
Beside VSEZ, Duvvada, Vadlapudi Post, Gajuwaka, Visakhapatnam - 530 049.

Department of Information Technology

CERTIFICATE

This is to certify that the Project report entitled **"Predicting a crime in a particular region using Decision Trees"** is a bonafide record of project work carried out under my supervision by **K. Sai Divya** bearing Regd.No.**18L31A1288**, **P. Shanmukha Rao** bearing Regd.No.**18L31A1246, Ch. Dilip Kumar** bearing Regd.No.**18L31A1255, N. Ravi Teja** bearing Regd.No.**18L31A1281** in partial fulfillment of the degree of **Bachelor of Technology in Information Technology of** Vignan's Institute of Information Technology(A) affiliated to Jawaharlal Nehru Technology University Kakinada, during the academic year 2018-2022

Signature
Name of the Guide

Signature
Dr. G. Rajendra Kumar (HOD)

EXTERNAL EXAMINER

DECLARATION

We here by declare that this project report entitled **“Predicting a crime in a particular region using Decision Trees”** has undertaken by us for the fulfillment of Bachelor of Technology in Information Technology. We declare that this project report has not been submitted anywhere in the part of fulfillment for any degree of any other University.

PLACE: Visakhapatnam
DATE:

K.SAI DIVYA (18L31A1288)
P. SHANMUKHA RAO (18L31A1246)
CH. DILIP KUMAR (18L31A1255)
N. RAVITEJA (18L31A1281)

ACKNOWLEDGEMENT

An endeavour over a long period can be successfully with the advice and support of many well-wishers. I take this opportunity to express our gratitude and appreciation to all of them.

I express my sincere gratitude to my internal guide, **DR.G. RAJENDRA KUMAR**

for his/her encouragement and cooperation in completion of my project. I am very fortunate in getting the generous help and constant encouragement from him/her.

I would be very grateful to our project coordinator, **A. SIRISHA** for the continuous

monitoring of my project work. I truly appreciate for her time and effort spend helping me

I would like to thank our Head of the Department, **DR.G. RAJENDRA KUMAR** and all other teaching and non-teaching staff of the department for their cooperation and guidance during my project.

I sincerely thank to **Dr. B. Arundhati,** Principal of VIGNAN'S INSTITUTE OF INFORMATION TECHNOLOGY (A) for her inspiration to undergo this project.

I wanted to convey my sincere gratitude to **Dr. V. Madhusudhan Rao**, Rector of

VIGNAN'S INSTITUTE OF INFORMATION TECHNOLOGY (A) for allocating the required resources and for the knowledge sharing during my project work.

I extended my grateful thanks to our honorable Chairman **Dr. L. Rathaiah** for giving me an opportunity to study in his esteemed institution.

K.SAI DIVYA (18L31A1288)

P. SHANMUKHA RAO (18L31A1246)

CH. DILIP KUMAR (18L31A1255)

N. RAVI TEJA (18L31A1281)

VIGNAN's INSTITUTE OF INFORMATION TECHNOLOGY
(AUTONOMOUS)
(Approved by AICTE-New Delhi & Affiliated to JNTUK, Kakinada)
Beside VSEZ, Duvvada, Vadlapudi Post, Gajuwaka, Visakhapatnam - 530 049.

INFORMATION TECHNOLOGY

VISION:

To be a center of excellence in high quality education and research producing globally competent IT professionals with ethical /human values to meet the needs of Information Technology sector and related services

MISSION:

- To impart high quality of education through innovative teaching –learning practices resulting in strong software and hardware knowledge and skills to enable students to meet the challenges of IT profession

- To facilitate faculty and students to carry out research work by providing necessary latest facilities and a conductive environment

- To mould the students into effective professionals with necessary communication skills, team spirit, leadership qualities, managerial skills, integrity, social and environmental responsibility and lifelong learning ability with professional ethics and human values

Program Educational Objectives (PEOs):

PEO1: To work in core IT companies/allied industries, educational institutions, research organizations and/or be entrepreneurs

PEO2: To pursue higher education/ research in the field of Information Technology

PEO3: To demonstrate communication skills, team spirit, leadership qualities, managerial skills, integrity, social & environmental responsibility and lifelong learning ability, professional ethics and human values in profession/career

PSOs: Program Specific outcomes:

PSO1: Analyze and design the solutions for data storage & computing systems.

PSO2: Implement the solutions for network and communication problems of Information Technology.

PROGRAM OUTCOMES	
PO1	**Engineering Knowledge:** Apply the knowledge of mathematics science engineering fundamentals and mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems engineering problems.
PO2	**Problem analysis:** Identify, formulate, review research Literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, and natural sciences, and engineering sciences.
PO3	**Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural societal, and environmental considerations
PO4	**Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5	**Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
PO6	**The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO7	**Environment and sustainability:**
	Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development and need for sustainable development.
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO9	**Individual and team work:** Function effectively as an individual and as a member or leader in diverse teams and individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11	**Project management and finance:** Demonstrate knowledge and understanding of the engineering and knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

ABSTRACT

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. In general, while we are going out to any place it was quite normal to check for weather and traffic through network. Similarly, we can also check for the crime rate before going. Using the concept of Machine Learning, we can extract previously unknown, useful information from a structured data i.e., crime data. Predicting a crime in a particular region can be done based on the data that was collected. Using this concept, we can predict the high probability of crime rate using Decision Trees. It also provides visualization techniques to understand easily. During prediction we will undergo various techniques like Data collection, Data cleaning, Data handling, Predictive modelling, Model selection, Prediction, Visualization using Random Forest classification.

With the results, we can provide identification of crime by predicting the crime rate with an improved accuracy than with other methods which will directly help police to practice and for better strategies.

INDEX

S. No	Content	Page No
1	**Introduction**	01
	1.1 About Machine Learning	02
	1.2 Algorithms used	04
	1.3 About project	05
	1.4 Workflow of the project	07
2	Literature Survey (list of papers referred)	11
3.	**System Analysis**	19
	3.1 Existing System	20
	3.2 Proposed System	20
	3.3 Feasibility study	20
4.	**System specifications**	22
	4.1 Functional Requirements	23
	4.2 Non-functional requirements	23
	4.3 Hardware requirements	23
	4.4 Software requirements (SRS)	23
5.	**System Design**	25
	5.1 System Architecture	26
	5.2 UML Diagrams	27
	5.3 Data Flow Diagram	33
6.	**System Implementation**	34
	6.1 Project Modules	35
	6.2 Methodology (Algorithms)	36
	6.3 Source Code	37
7.	**System Testing**	44
	7.1 Testing Methods	45
	7.2 Testing Approaches	46
	7.3 Test cases	48
8.	Experimental Results	50
9.	Conclusion & Future scope	53
10.	Bibliography	55

INTRODUCTION

1.1 About Machine Learning:

The term 'machine learning' is often, incorrectly, interchanged with Artificial Intelligence [JB1], but machine learning is actually a sub-field/type of AI. Machine learning is also often referred to as predictive analytics, or predictive modelling. Coined by American computer by scientist Arthur Samuel in 1959. At its most basic, machine learning uses programmed algorithms that receive and analyze input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimize their operations to improve performance, developing 'intelligence' over time. There are four types of machine learning algorithms: Supervised, Semi-supervised, Unsupervised, and Reinforcement.

1.1.1 Supervised Learning:

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance. Under the umbrella of supervised learning: Classification, Regression and Forecasting.

Classification: In classification tasks, the machine learning program must conclude observed values and determine to what category new observations belong. For example, when filtering emails as 'spam' or 'not spam', the program must look at existing observational data and filter the emails accordingly. **Regression:** In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables –making by Predicting and forecasting.

Forecasting: Forecasting is the process of making predictions about the future based on the past and present data and is commonly used to analyze trends.

1.1.2 Semi Supervised Learning:

Semi-supervised learning is similar to supervised learning but instead uses both labeled and unlabeled data. Labeled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabeled data lacks that information. By using this combination, machine learning algorithms can learn to label unlabeled data.

1.1.3 Unsupervised Learning:

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analyzing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly.

Clustering: Clustering involves grouping sets of similar data (based on defined criteria). It’s useful for segmenting data into several groups and performing analysis on each data set to find the problems.

Dimension reduction: Dimension reduction reduces the number of variables being considered to find the exact information required.

1.1.4 Reinforcement Learning:

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

1.2 Algorithms Used:

1.2.1 Decision Trees (Supervised Learning–classification/Regression):

A decision tree is a flow-chart-like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific variable – and each branch is the outcome of that test.

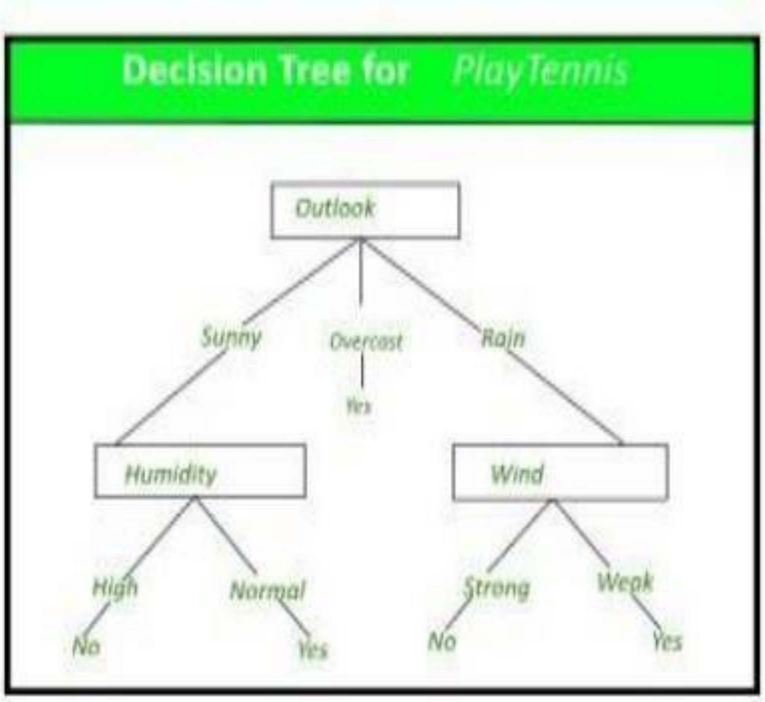


Fig 1.2.1: Representation of decision Tree

1.2.2 Random Forest (Supervised Learning – Classification/Regression):

Random forests or 'random decision forests' is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a 'decision tree' (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.

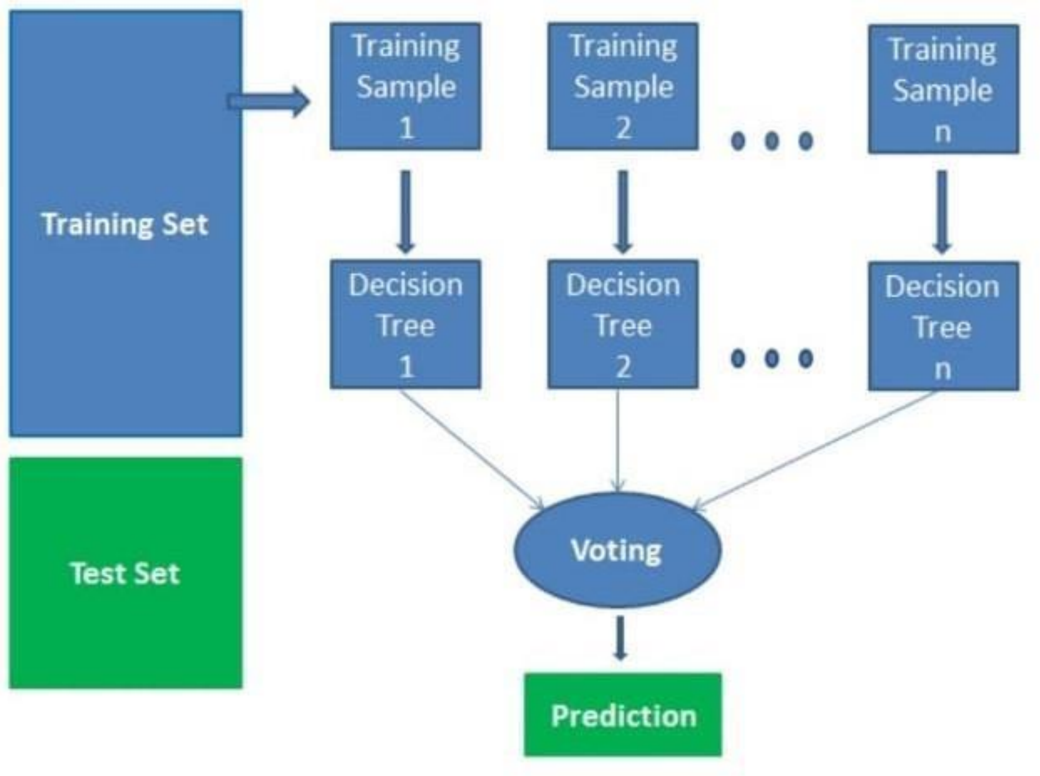


Fig 1.2.2: Prediction process in Random Forest

1.3 About the Project:

Crimes now-a-days are increasing day by day and with different level of intensity and versatility. The result is great loss to society in terms of monitory loss, social loss and further it enhances the level of threat against the smooth livelihood in the society. To overcome this problem the computing era can help to reduce the crime or even may be helpful in predicting the crime so that sufficient measures can be taken to minimize the loss to property and life. The crime rate prediction strategies can be applied on historical data available in the police records by examining the data at various angles like reason of crime, frequency of similar kind of crimes at specific location with other parameters to prepare the machine learning model for crime prediction. It is the major challenge to understand the versatile data available with us then model it to predict the crime incidence with acceptable accuracy and further to reduce the crime rate.

The main scope of the project is to predict the crime rate in a particular region based on the historic data and visualize it graphically using statistical tools so that, it is easy to look and understand the data to support public safety, financial success and better outcomes. Public safety and protection related 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 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. This project presents the visualization techniques and classification algorithms that can be used for predicting the crimes and helps the law agencies. In future, there is a plan for applying other classification algorithms on the crime data and improving the accuracy in prediction.

In recent time, many researchers have conducted various experiments to predict crimes using various machine learning methods and inputs. For crime prediction, KNN (K- Nearest Neighbor), Decision trees and some other algorithms are used. Data collected from various websites and newsletter were used for prediction and classification of crime using Naïve Bayes, Decision trees, Support Vector Machine (SVM) and Artificial neural networks (ANN) but there does not exist any particular method that can solve different crime datasets problems.

To predict the crime of a particular region some of the standards should be followed. In this model the standards we followed are mentioned as below:

- Collecting the dataset from Kaggle and exploring the dataset as per requirement. In this dataset 1994 areas are present. Size of the dataset is 1994 rows and 128 columns. The dataset features are: the occurrence month, the occurrence day of the week, the occurrence time and the crime location.
- Importing the required modules into the jupyter notebook and implementing data cleaning, analyzing correlation (relationship between variables using heat map) on the data set are done.

- After that, we have used machine learning algorithms like regression and decision tree classifier to train the dataset and build machine learning model to get good accuracy.
- The accuracy rate is 85% with Decision tree classifier and 0.65 R2 score with Linear Regression. To increase the previous accuracy rate, Decision tree classifier algorithm is optimized.

After following these standards, the accuracy increased to 90%.

1.4 Work flow of the project:

Workflow can mean different things to different people, but in the case of ML it is the series of various steps through which a ML project goes on. It means that passing each and every stage under the workflow to complete the project successfully and in time.

We will follow the general Machine Learning workflow steps:

1. Gathering data
2. Data per-processing
3. Splitting the data for training and testing the model
4. Researching the model that will be best for the type of data
5. Training and testing the model
6. Evaluation

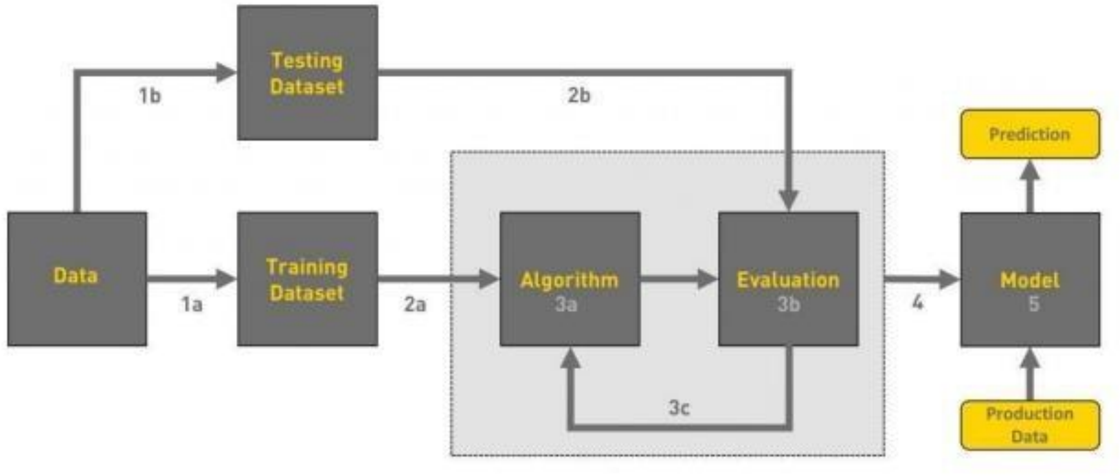


Figure 1.4: Workflow of machine learning Project.

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e., the trained model will provide false or wrong predictions.

1. **Gathering Data:**

   The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

2. **Data pre-processing:**

Data pre-processing is a process of cleaning the raw data i.e., the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects. Most of the real-world data is messy, some of these types of data are:

Missing Data: Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).

Noisy Data: This type of data is also called outliners; this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data. Inconsistent Data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

3. **Splitting the data for training and testing the model:**

For training a model we initially split the model into 3 three sections which are 'Training data' and 'Testing data'.

Training Data: The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data used for learning, that is to fit the parameters of the classifier.

Test Data: A set of unseen data used only to assess the performance of a fully-specified classifier.

4. **Researching the model that will be best for the type of data:**

Moving on forward to the next step, we have to choose the best suited model. Models are compared on the basis of the accuracy score that they generate. One way to choose the best model is to train each and every model and take the results of that model that is showing the best results out of them (obviously, a time taking process, but quite interesting if we get familiar). This step also includes training the data set and fitting our data in the model and then testing it to predict and get the accuracy score.

5. **Training and testing the model on data:**

You train the classifier using 'training data set', tune the parameters using 'validation set' and then test the performance of your classifier on unseen 'test data set'. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier. In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

6. **Evaluation Model:**

Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

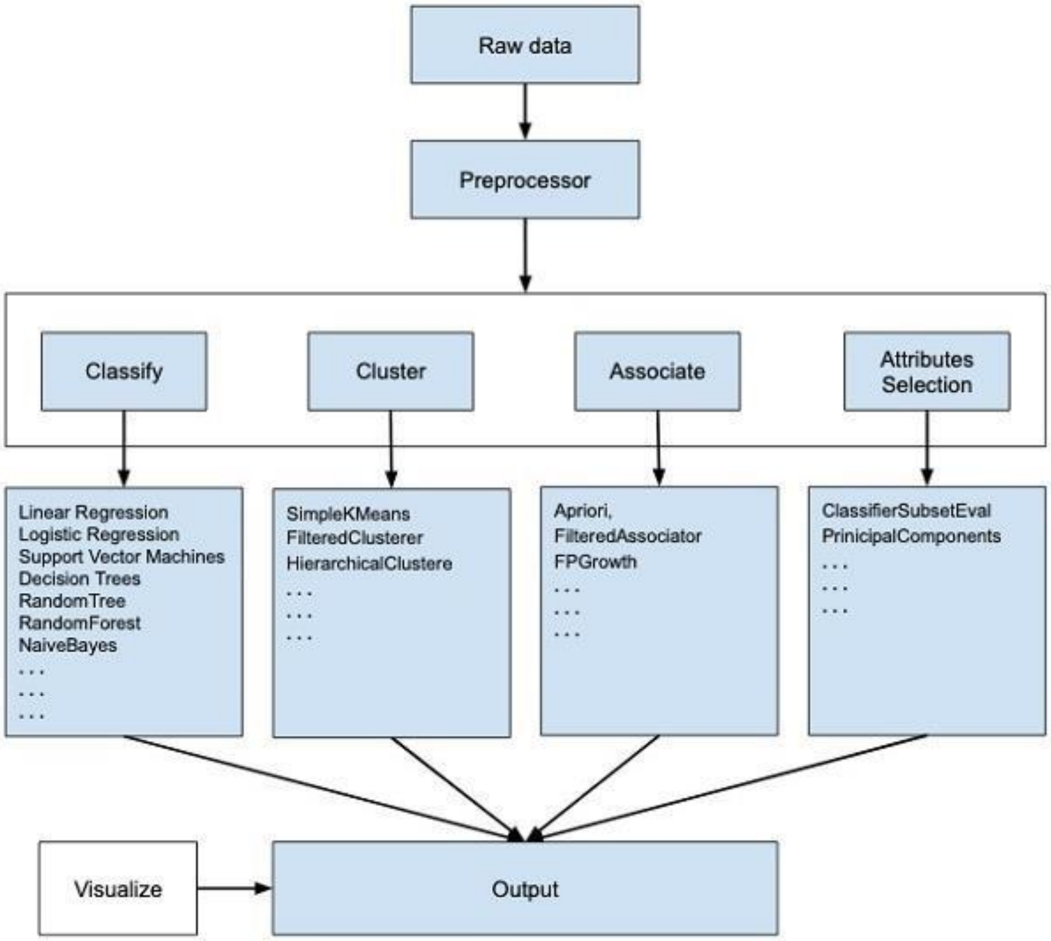
LITERATURE SURVEY

Predicting a Crime in a particular region using Decision Trees

[1] prediction is done using ID3(Iterative Dichotomiser 3) algorithm with the help of WEKA mining software as tool at an accuracy of 72.7%.

WEKA:

WEKA - an open-source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram –



WEKA is a powerful tool for developing machine learning models. It provides implementation of several most widely used ML algorithms. Before these algorithms are applied to your dataset, it also allows you to pre-process the data. The types of algorithms that are supported are classified under Classify, Cluster, Associate, and Select attributes. The result at various stages of processing can be visualized with a beautiful and powerful visual representation. This makes it easier for a Data Scientist to quickly apply the various machine learning techniques on his dataset, compare the results and create the best model for the final use.

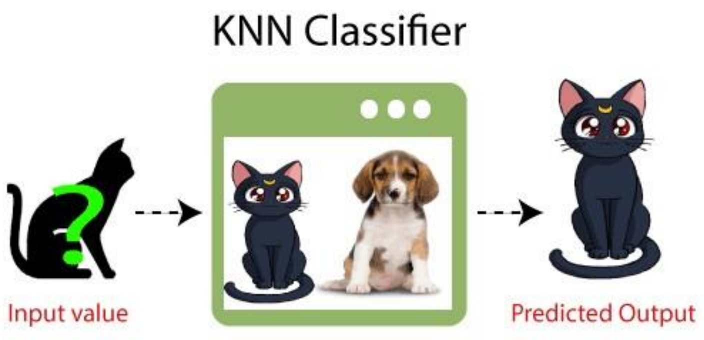
ID3 Algorithm:

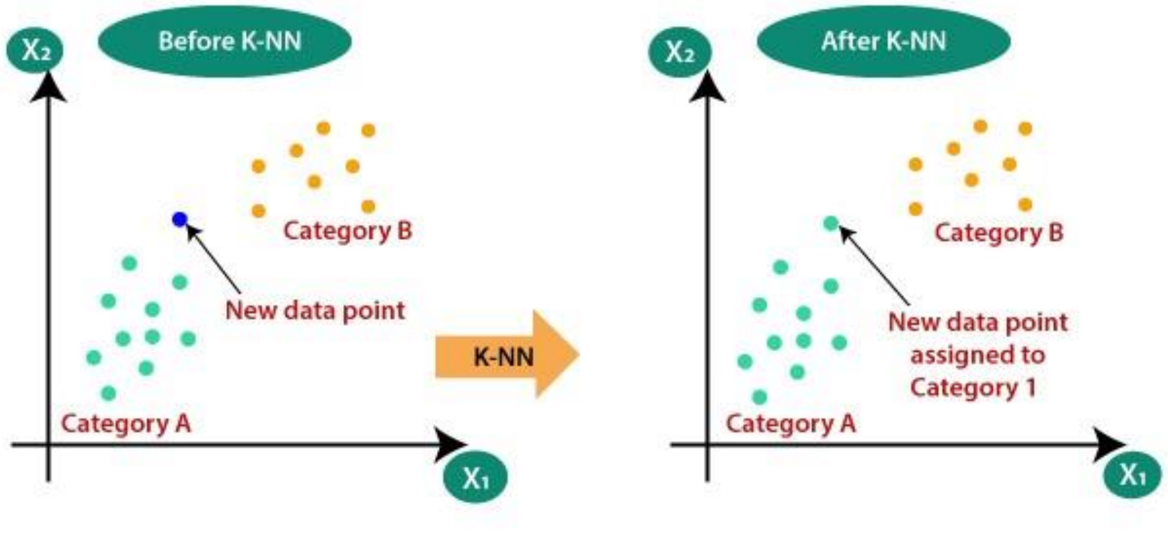
Iterative Dichotomiser 3 algorithm builds a decision on two metrics: Entropy and Information Gain of each featured data column. Entropy is the measure of uncertainty of a label in a particular column, whereas Information Gain is the measure of feature information of a column in the dataset. Using this method, decision trees are built iteratively by finding out the maximum Information Gain among all the featured data columns to be represented as the node of the tree.

[2] prediction and classification of crime is done based on the only feature that is location that is Indore City and it is done by using KNN, decision trees and random forests but it doesn't achieve that much of accuracy.

KNN (K- Nearest Neighbour):

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. ○ K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. ○ KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



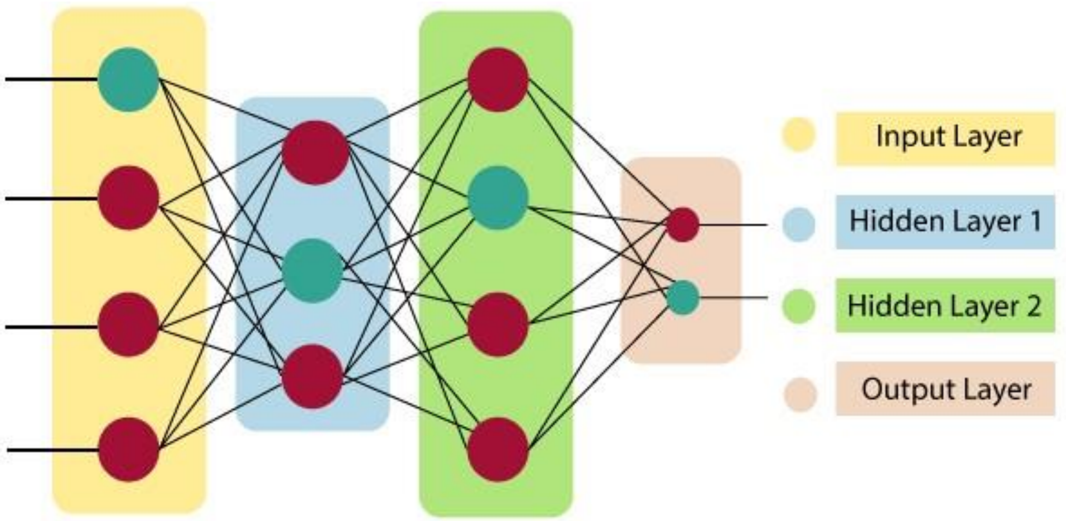


[3] The author predicts the rate based on the features like sex, age and relationship and the algorithms used are KNN and Artificial Neural Networks (ANN) and got an accuracy around 85%.

ANN (Artificial Neural Networks):

The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modelled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.

Artificial Neural Network primarily consists of three layers:



Input Layer:

As the name suggests, it accepts inputs in several different formats provided by the programmer.

Hidden Layer:

The hidden layer presents in-between input and output layers. It performs all the calculations to find hidden features and patterns.

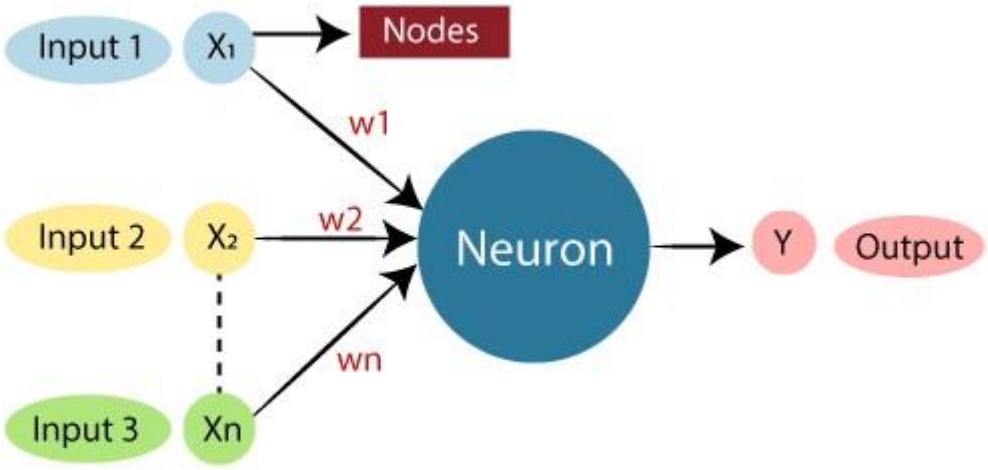
Output Layer:

The input goes through a series of transformations using the hidden layer, which finally results in output that is conveyed using this layer.

The artificial neural network takes input and computes the weighted sum of the inputs and includes a bias. This computation is represented in the form of a transfer function.

$$\sum_{i=1}^{n} Wi * Xi + b$$

It determines weighted total is passed as an input to an activation function to produce the output. Activation functions choose whether a node should fire or not. Only those who are fired make it to the output layer. There are distinctive activation functions available that can be applied upon the sort of task we are performing.



[4] prediction is done using the techniques like KNN, Decision Tree an Extra Tree Classifier with an accuracy of 88%.

Extra Tree Classifier:

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple decorrelated decision trees collected in a “forest” to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple decorrelated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.

[5] For prediction the author used Deep learning techniques along with some Machine learning techniques like XGBoost and KNN whereas deep learning techniques like LSTM (for analysis), RME and MAE are used for visualization. The project is main purpose is to predict and forecast the crime in Chicago and Los Angeles.

XGBoost:

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. XGBoost is a popular implementation of gradient boosting.

LSTM:

Long Short-Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

MAE:

Mean Absolute Error, MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group. MAE can also be referred as L1 loss function. MAE helps users to formulate learning problems into optimization problems. It also serves as an easy-to-understand quantifiable measurement of errors for regression problems.

[6] Based on the past crime data, the prediction is done using Linear regression, Naïve Bayes and KNN around with an accuracy of 73.6%,69.5% and 76.9% respectively.

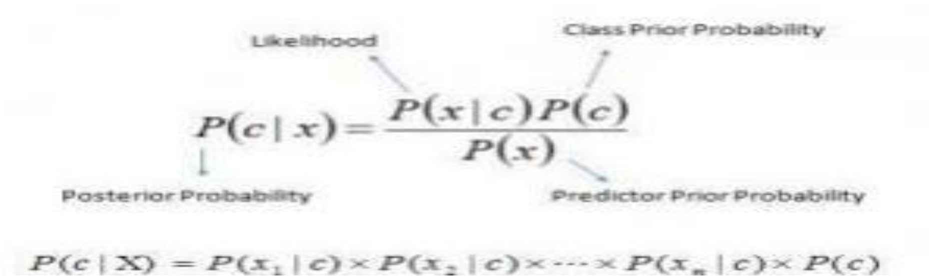
Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Naïve Bayes:

The Naive Bayes classifier is based on Bayes' theorem and classifies every value as independent of any other value. It allows us to predict a category, based on a given set of features, using probability. Despite its simplicity, the classifier does surprisingly well and is often used due to the fact it outperforms more sophisticated classification methods.



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), Above, P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

SYSTEM ANALYSIS

3.1 Existing System:

In recent times, many researches have conducted various experiments to predict crimes using various machine learning methods and inputs. For crime prediction, KNN (K-Nearest Neighbour), Decision trees, Neural Networks and some other algorithms are used. Data collected is from various websites and newsletter were used for prediction and classification of crime using Naïve Bayes, Decision trees, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) but there does not exist any particular method that can solve different crime datasets problems and predict crime accurately.

3.2 Proposed System:

In the proposed system, the Machine Learning Model is built in such a way that it can predict the crime based on the data collected from various features like the type of the crime. The model classifies the data based on the location given as input using decision trees and random forests and then it performs prediction using prediction functions. This system provides the accurate result than the existing systems.

3.3 Feasibility Study:

A feasibility study is used to determine the feasibility of an idea, for example to ensure that a project is legally and technically feasible as well as economically justifiable. The feasibility study is generally conducted before undertaking any initiative concerning a project, including planning. It is one of the fundamental factors, if not the most important, which determine whether the project should be carried out or not.

Three Key Considerations involved in the feasibility analysis are:

Economic feasibility
Technical feasibility
Operational feasibility

3.3.1 Economic feasibility:

This assessment typically involves a cost/ benefits analysis of the project, helping organizations determine the viability, cost, and benefits associated with a project before financial resources are allocated. It also serves as an independent project assessment and enhances project credibility—helping decision-makers determine the positive economic benefits to the organization that the proposed project will provide.

For the project, there is no investment and economically this is cost effective.

3.3.2 Technical feasibility:

A large part of determining resources has to do with assessing technical feasibility. It considers the technical requirements of the proposed project. The technical requirements are then compared with technical capability of the organization. The systems project is considered technically feasible if the internal technical capability is sufficient to support the project requirements.

The system can be technically feasible on the following grounds.

1. All the necessary technology exists to develop the system.
2. The existing resources are capable and can hold all necessary data.
3. The system is too flexible and it can be expanded further.
4. The system can give guarantees of accuracy, ease of use, reliability and the data security.

The technologies used for the project are feasible and capable for holding the data. It also guarantees the accuracy and reliability.

3.3.3 Operational feasibility:

Operational feasibility is a measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development.

The project provides the better accuracy by predicting the crime rate of a particular area with user friendly environment and feasibility.

System Specifications

4.1 Functional Requirements:

Functional requirements are product features or functions those developers must implement to enable users to accomplish their tasks. So, it's important to make them clear both for the development team and the stakeholders. Generally, functional requirements describe system behaviour under specific conditions. Functional requirements define what a software system will deliver based on certain inputs. A system delivering its functional requirements does what its users expect.

4.2 Non-Functional Requirements:

Non-functional requirements describe user-visible aspects of the system that are not directly related to functionally of the system. These requirements define what qualities is exhibited by the system.

The following are the non-function requirements of the system:

1. Performance
2. Usability
3. Scalability
4.Maintainability
5. Interoperability

4.3 Hardware requirements:

1. Operating Systems	: Windows, Linux and Mac 2.
2. Browsers	: Chrome, Edge and Firefox
3. RAM	: 4 GB or above.

4.4 Software requirements:

1.Python-version 3.6 or above
2. IDE: Jupyter Notebook, Visual Studio
3.Framework: Flask
4. Dataset: source-Kaggle.

SRS (Software Requirement Specification):

The software, Site Explorer is designed for management of web sites from a remote location.

INTRODUCTION

Purpose: The main purpose for preparing this document is to give a general insight into the analysis and requirements of the existing system or situation and for determining the operating characteristics of the system.

Scope: This Document plays a vital role in the development life cycle (SDLC) and it describes the complete requirement of the system. It is meant for use by the developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.

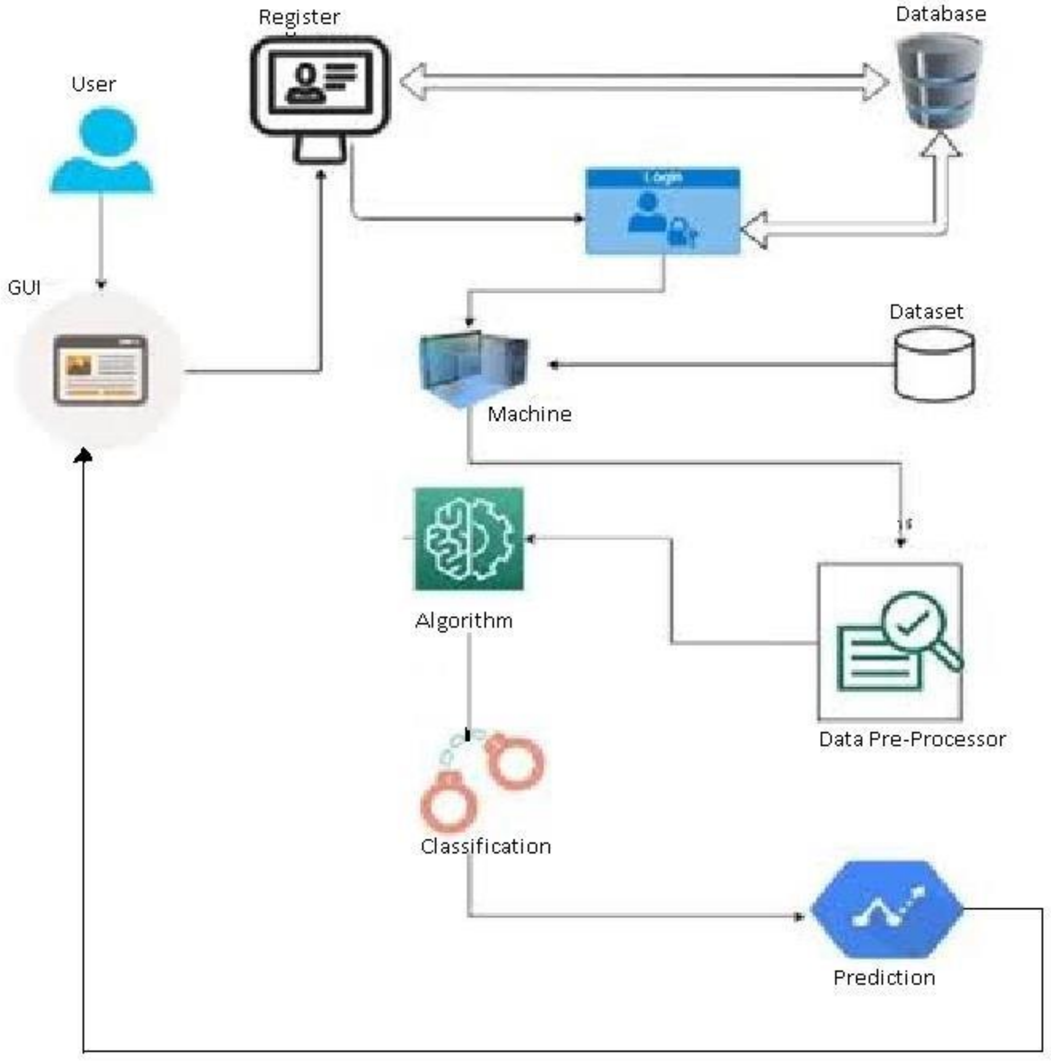
DEVELOPERS RESPONSIBILITIES OVERVIEW:

The developer is responsible for:

- Developing the system, which meets the SRS and solving all the requirements of the system?
- Demonstrating the system and installing the system at client's location after the acceptance testing is successful.
- Submitting the required user manual describing the system interfaces to work on it and also the documents of the system.
- Conducting any user training that might be needed for using the system.

SYSTEM DESIGN

5.1 System Architecture:

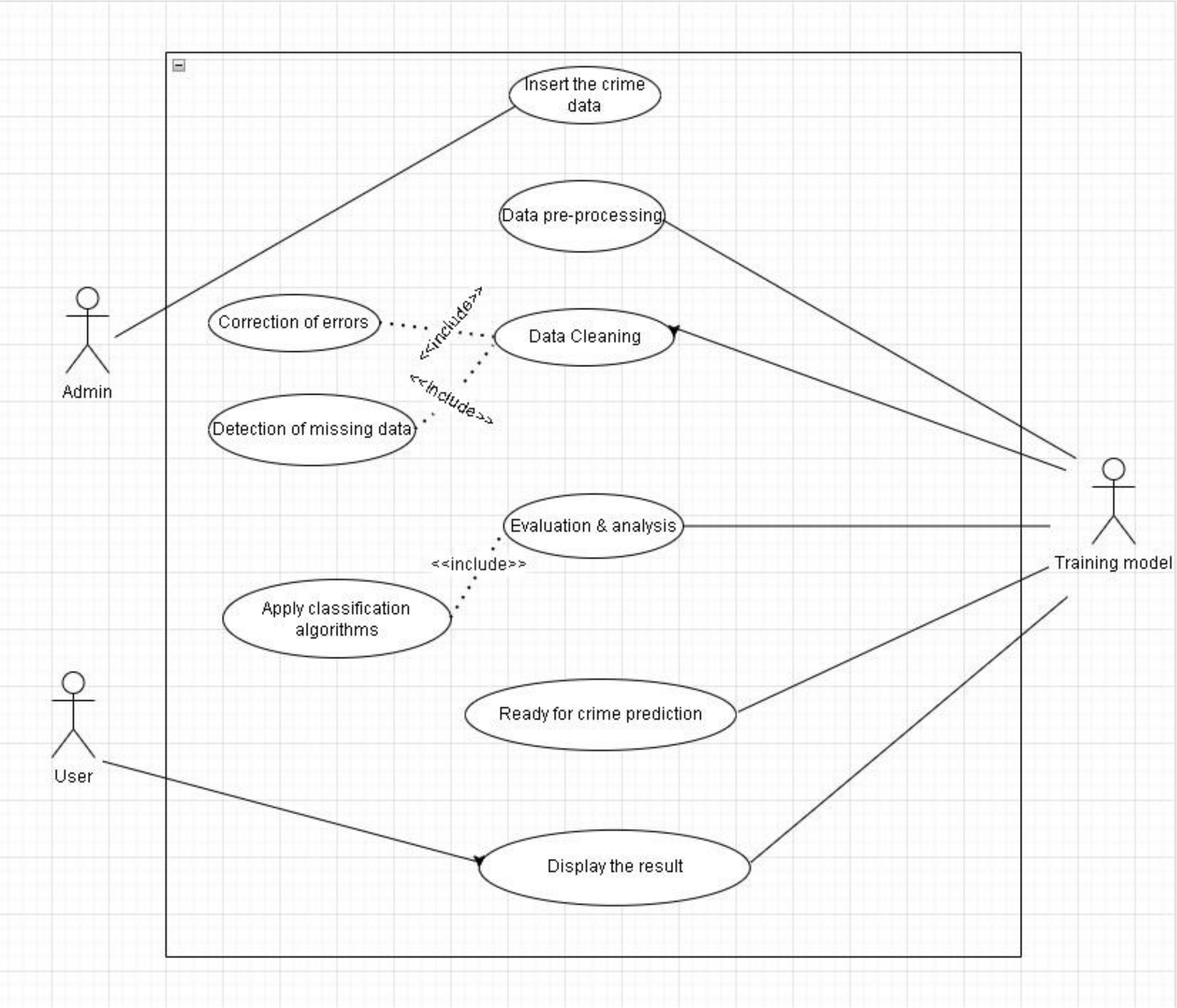


The above figure depicts the architecture of the proposed system. It can be accessed by interacting the user through web application. Through that the user can enter the location of the place that he wants to know the crime rate. After the location is entered the respective area's data is automatically inserted to the machine/model from the dataset and the machine will undergo all the steps of the algorithm, thus undergo for classification. At last, with the help of some predict functions the machine will predict the output and displays the result to the user's application.

5.2 UML Diagrams:

5.2.1 Use case diagram:

A use case diagram is a dynamic or behaviour diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform. In this context, a "system" is something being developed or operated, such as a web site. The "actors" are people or entities operating under defined roles within the system.

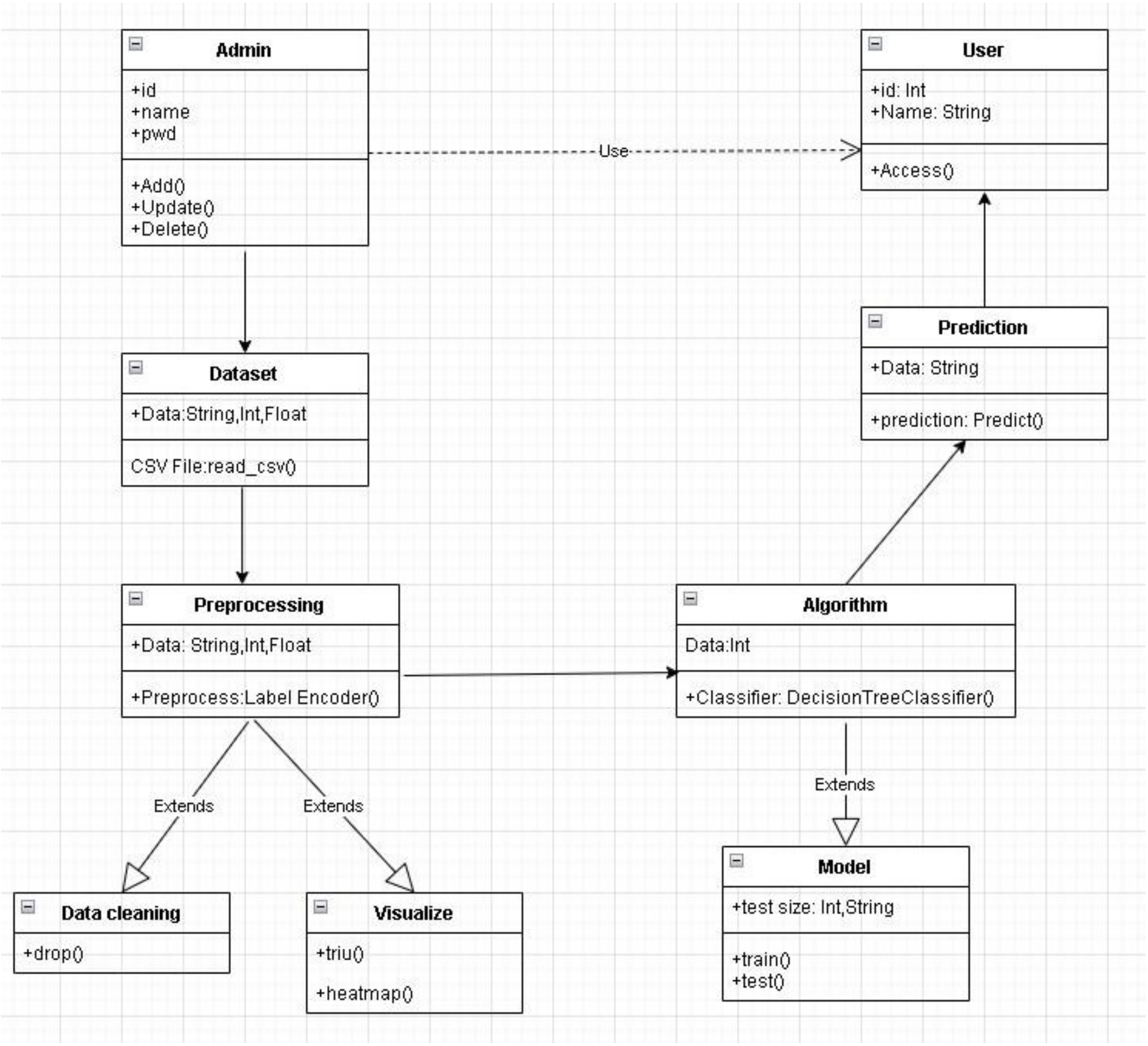


Use case diagrams are valuable for visualizing the functional requirements of a system that will translate into design choices and development priorities. They also help identify any internal or external factors that may influence the system and should be taken into consideration.

They provide a good high-level analysis from outside the system. Use case diagrams specify how the system interacts with actors without worrying about the details of how that functionality is implemented.

5.2.2 Class Diagram:

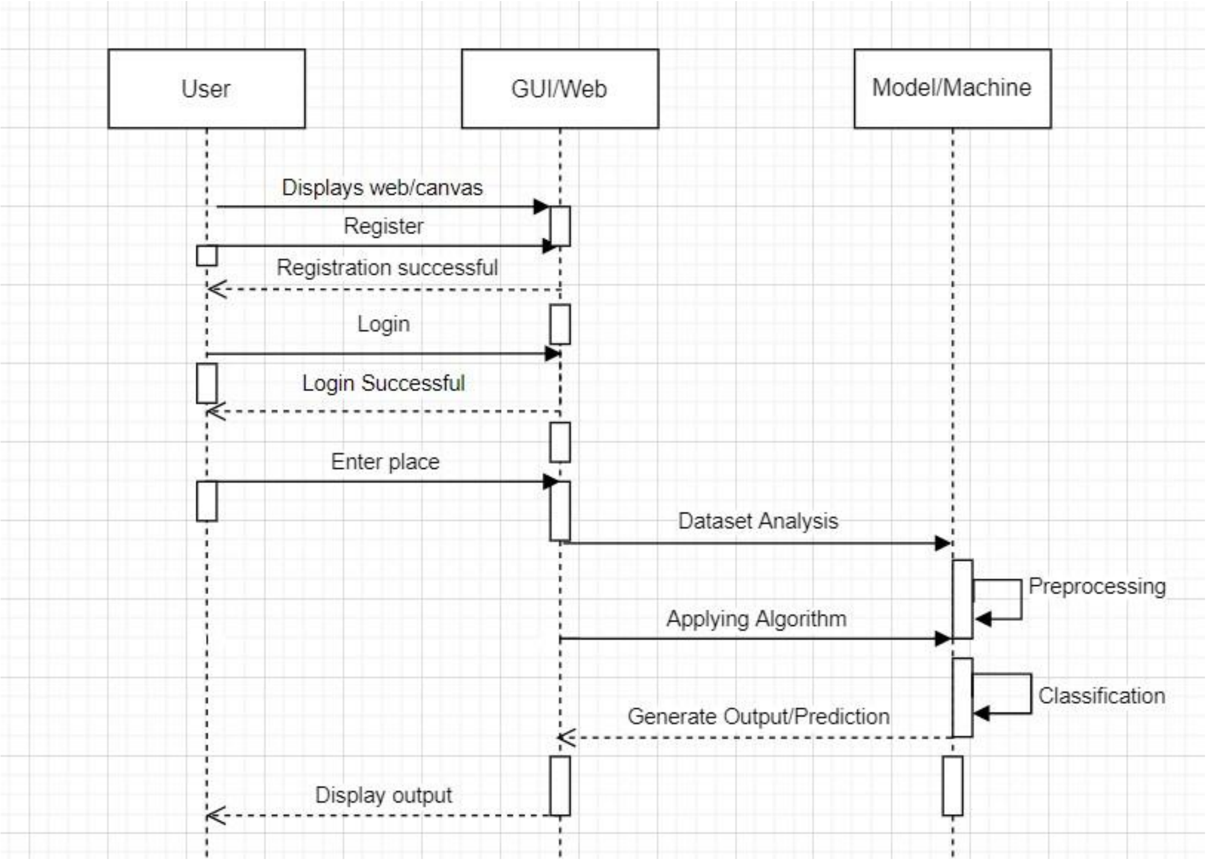
The class diagram depicts a static view of an application. It represents the types of objects residing in the system and the relationships between them.



A class consists of its objects, and also it may inherit from other classes. A class diagram is used to visualize, describe, document various different aspects of the system, and also construct executable software code. It shows the attributes, classes, functions, and relationships to give an overview of the software system. It constitutes class names, attributes, and functions in a separate compartment that helps in software development. Since it is a collection of classes, interfaces, associations, collaborations, and constraints, it is termed as a structural diagram.

5.2.3 Sequence Diagram:

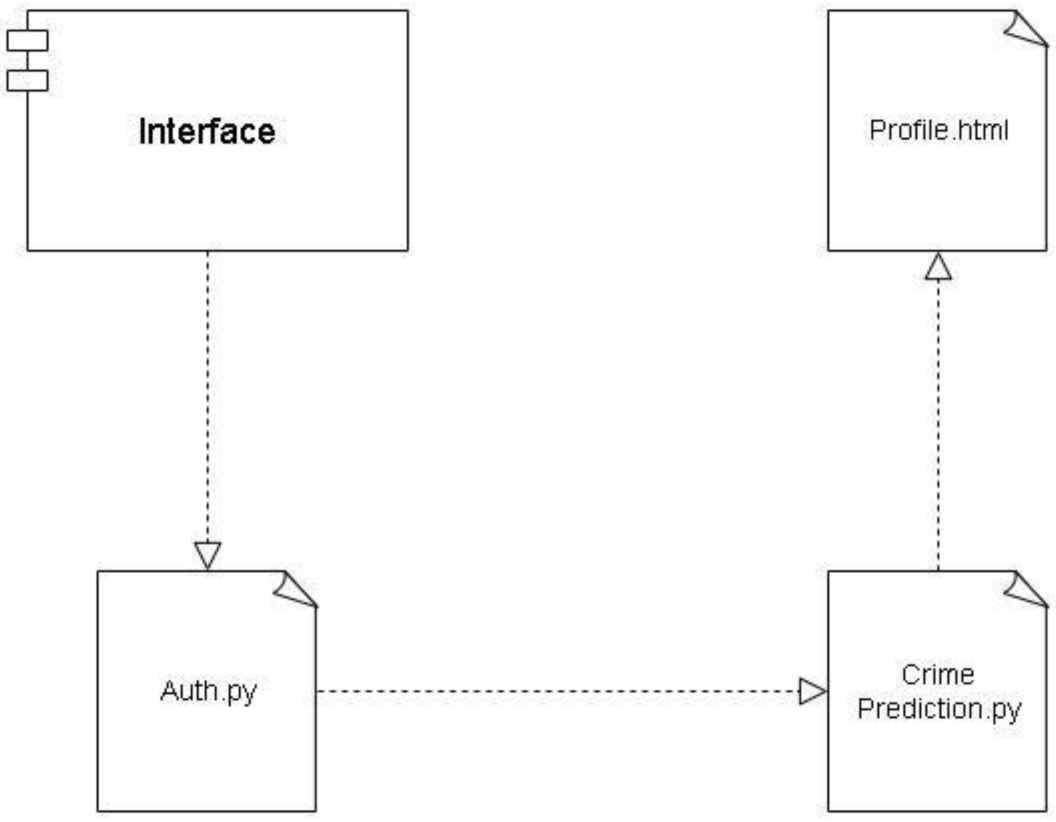
The sequence diagram represents the flow of messages in the system and is also termed as an event diagram. It helps in envisioning several dynamic scenarios. It portrays the communication between any two lifelines as a time-ordered sequence of events, such that these lifelines took part at the run time. In UML, the lifeline is represented by a vertical bar, whereas the message flow is represented by a vertical dotted line that extends across the bottom of the page. It incorporates the iterations as well as branching.



5.2.4 Component Diagram:

A component diagram is used to break down a large object-oriented system into the smaller components, so as to make them more manageable. It models the physical view of a system such as executables, files, libraries, etc. that resides within the node.

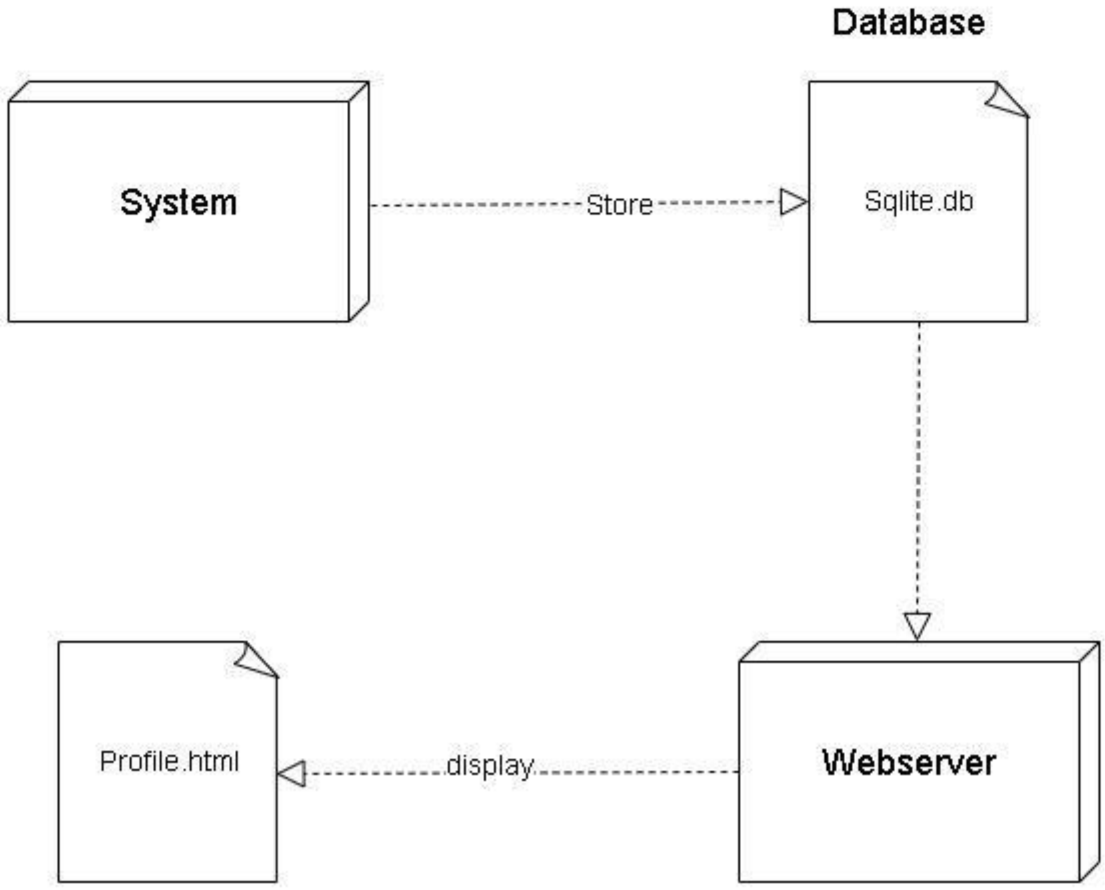
It visualizes the relationships as well as the organization between the components present in the system. It helps in forming an executable system. A component is a single unit of the system, which is replaceable and executable. The implementation details of a component are hidden, and it necessitates an interface to execute a function. It is like a black box whose behaviour is explained by the provided and required interfaces.



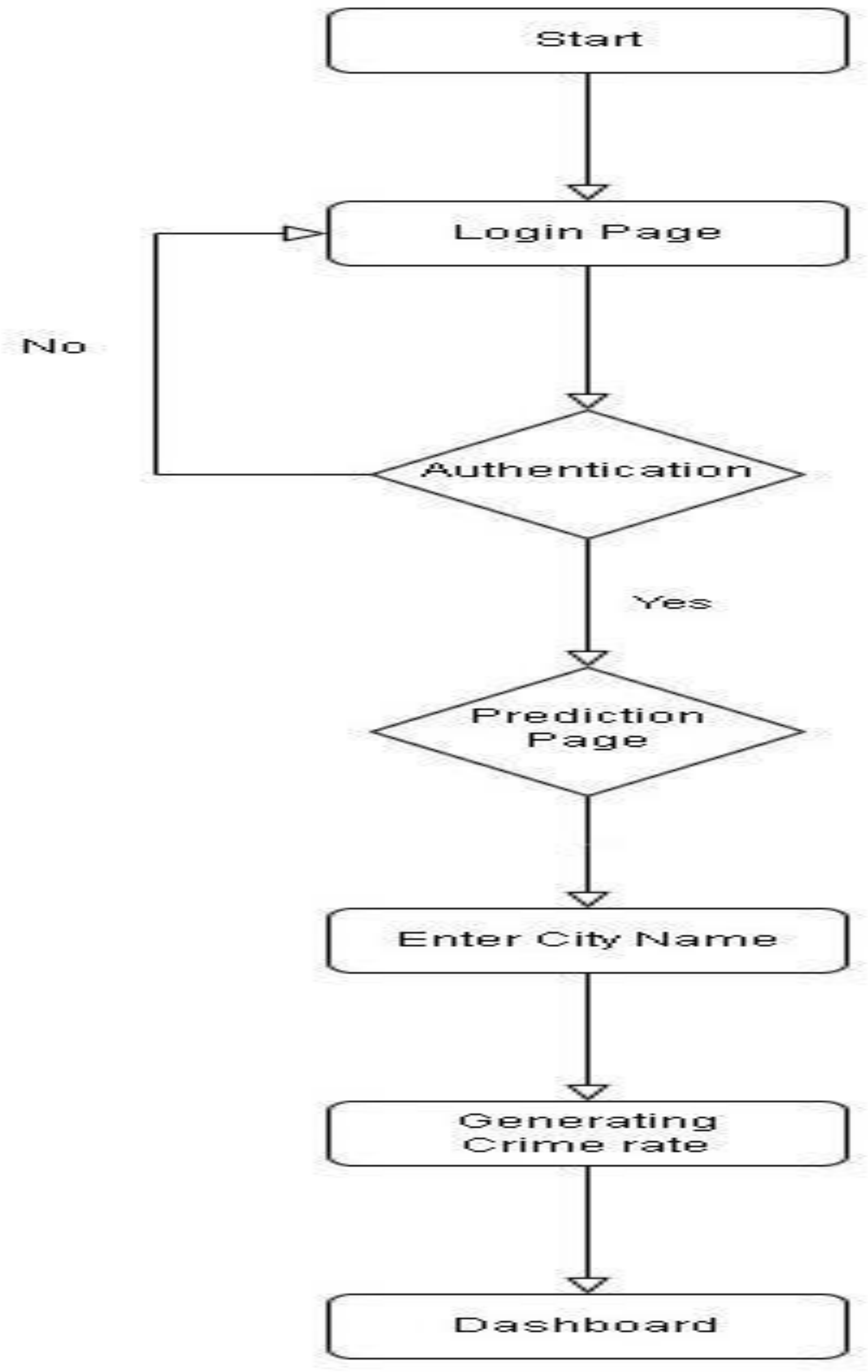
5.2.5 Deployment Diagram:

The deployment diagram visualizes the physical hardware on which the software will be deployed. It portrays the static deployment view of a system. It involves the nodes and their relationships.

It ascertains how software is deployed on the hardware. It maps the software architecture created in design to the physical system architecture, where the software will be executed as a node. Since it involves many nodes, the relationship is shown by utilizing communication paths. The main purpose of the deployment diagram is to represent how software is installed on the hardware component. It depicts in what manner a software interacts with hardware to perform its execution.

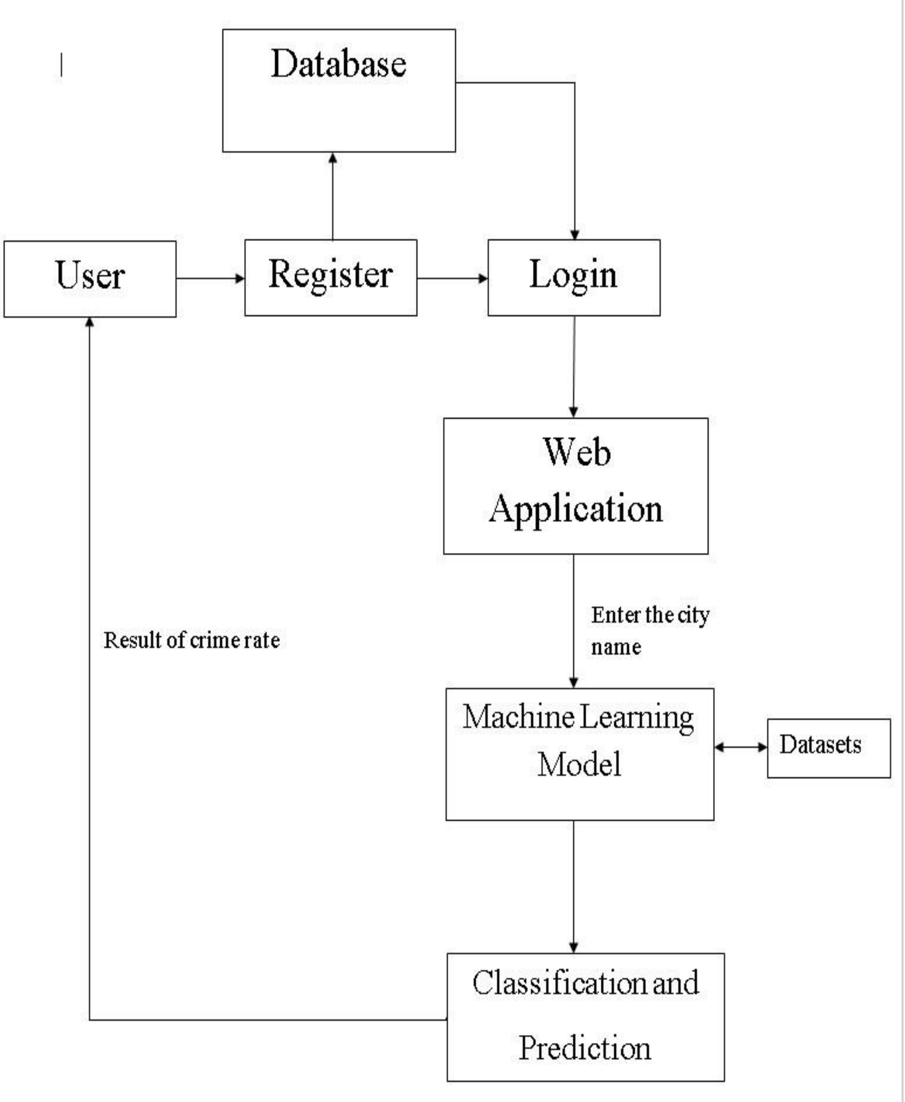


5.2.6 Activity Diagram:



In UML, the activity diagram is used to demonstrate the flow of control within the system rather than the implementation. It models the concurrent and sequential activities. The activity diagram helps in envisioning the workflow from one activity to another. It put emphasis on the condition of flow and the order in which it occurs. The flow can be sequential, branched, or concurrent, and to deal with such kinds of flows, the activity diagram has come up with a fork, join, etc. It is also termed as an object-oriented flowchart. It encompasses activities composed of a set of actions or operations that are applied to model the behavioural diagram.

5.3 Data Flow Diagram:



SYSTEM IMPLEMENTATION

6.1 Project Modules:

6.1.1 Front End:

- Register module
- Login module
- Data Access

Register module:

Registration module contains name, password and email as input fields and it stores these values into the database(sqlite).

Login module:

Login module performs the query operation and authenticates the user and provides access to the application.

Data Access Module:

Data Access module helps the user to enter a particular place into the application for analysis.

6.1.2 Back End:

- Analysis module
- Evaluation module
- Output module

Analysis Module:

Analysis module take place at backend which performs data cleaning and data pre-processing so that it takes the data and removes the missing values and standardize the data.

Evaluation Module:

In Evaluation module it takes data from the analysis and takes the classification algorithm as input and perform classification to predict the output.

Output Module:

The output module performs the prediction operation and provides the result to the user.

6.2 Methodology:

The methodology includes Machine Learning approach.

Data Collection:

Data is collected from digitalized information of crime reports which is of 1994 areas and the size of the dataset is 1994 rows and 128 columns.

Data Pre-processing:

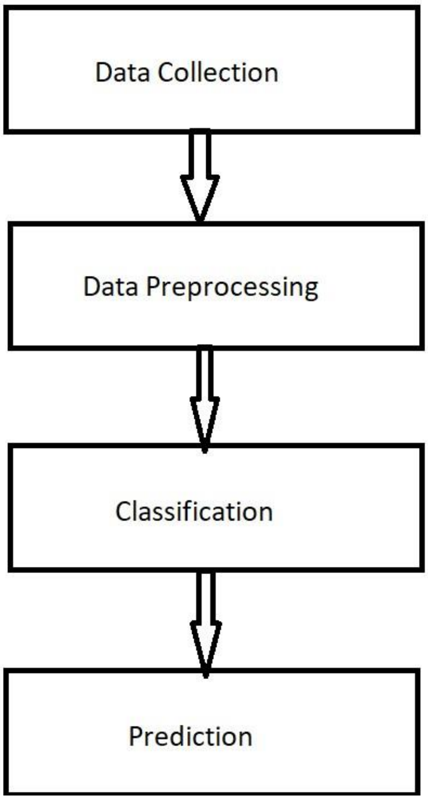
Data pre-processing includes data cleaning which removes missing values and noisy data and resolve inconsistencies. Further data integration and reduction also takes place.

Analysis:

The analysis includes graphical representation of data which provides the keen view towards the prediction.

Training and Testing:

The splitting of dataset for training and testing. For training, 80% of the dataset is used and for testing, 20% of the data is used.



Classification:

Classification is the process of separating or analysing the data based on the features. Decision Tree classifier is used for classification along with that Random tree classifier is used for bagging concept.

Prediction:

Prediction is performed based on the trained data. Whenever the data is inserted by the user it will undergo prediction functions and predicts the crime.

Algorithm:

1.Select the given samples from the dataset.

2.Construct a decision tree for each sample and get a prediction result from each decision tree.

3.Perform a vote for each predicted result.

4.Select the prediction result with the most votes as the final prediction

To train and test split:

```
x_train, x_test, y_train, y_test=train_test_split(x,y,test_size,random_state,stratify)
```

To train the model:

```
Clf_tree=DecisionTreeClassifier(criterion,max_depth,random_stat)
Clf_tree.fit(x_train, y_train)
```

To plot the tree:

```
fig.ax=plt.subplots(figsize)
tree.plot_tree(clf_tree,fontsize) plt.show()
```

6.3 Source Code:

For Sign up and Login of user to the application and to authenticate the user:

```
//Importing the modules that are required
from flask import Blueprint, render_template, redirect, url_for, request, flash
from werkzeug.security import generate_password_hash,check_password_hash
from flask_login import login_user, login_required, logout_user
from. models import User
from. import db
auth = Blueprint('auth', __name__)

@auth.route('/login')
def login():
            return  render_template('login.html')

@auth.route('/login',methods=['POST'])
def login_post():     # login code goes here

 email = request.form.get('email’)
 password = request.form.get('password')
 remember = True
 if request.form.get('remember') else False

 user = User.query.filter_by(email=email).first()

   # check if the user actually exists
   # take the user-supplied password, hash it, and compare it to the hashed password in the database
if not user or not check_password_hash(user.password, password):
       flash('Please check your login details and try again.')
       return redirect(url_for('auth.login')) # if the user doesn't exist or password is wrong, reload the page
   # if the above check passes, then we know the user has the right credentials
login_user(user, remember=remember)
   # if the above check passes, then we know the user has the right credentials
return redirect(url_for('main.profile'))

@auth.route('/signup')
def signup():
   return  render_template('signup.html')


@auth.route('/signup', methods=['POST'])
def signup_post():
   # code to validate and add user to database goes here
email = request.form.get('email')
name = request.form.get('name')
password = request.form.get('password')
```

```
user = User.query.filter_by(email=email).first() # if this returns a user, then the email already
exists in database

     if user: # if a user is found, we want to redirect back to signup page so user can try again
  flash('Email address already exists')
        return redirect(url_for('auth.signup'))

    # create a new user with enecccbftcerigulndiulnvvcvelkujfhvcffbldgftb
# the form data. Hash the password so the plaintext version isn't saved.
    new_user = User(email=email, name=name, password=generate_password_hash(password,
method='sha256'))

    # add the new user to the database
db.session.add(new_user)
    db.session.commit()

    return redirect(url_for('auth.login'))

@auth.route('/logout)
@login_required
def logout():
                logout_user()
    return redirect(url_for('main.index'))
```

Machine Learning code for prediction:

//Importing Modules

```
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
from sklearn.model_selection import KFold
from scipy.spatial import distance
from sklearn.decomposition import PCA
from numpy import linalg as LA
from sklearn.metrics import roc_curve, auc
import math from scipy.stats
import pearsonr import copy
from random import randrange
import random
```

```
from sklearn import preprocessing
from numpy import int64
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import itertools
import csv
from sklearn import metrics
import tensorflow as tf
import numpy as np
import pandas as pd
from sklearn import linear_model
from sklearn.ensemble import RandomForestClassifier
import seaborn as sb
from sklearn import utils
from sklearn.linear_model import LogisticRegression
import warnings warnings.filterwarnings('ignore')
plt.switch_backend('Agg')

#uploading the dataset
df = pd.read_csv("project/crime_prep.csv")
df.head()

# Handling the NULL values
df = df.fillna(df.mean())
df.head()

# Encoding the data to make it fit for any algorithm
lab_enc = preprocessing.LabelEncoder()
df['v_cat_0'] = lab_enc.fit_transform(df['v_cat_0'].values.reshape(-1, 1))
df['v_cat_1'] = lab_enc.fit_transform(df['v_cat_1'].values.reshape(-1, 1))
df['v_cont_0'] = lab_enc.fit_transform(df['v_cont_0'].values.reshape(-1, 1))
df.head()
```

```
# Removing the leading and trailing spaces in the resultant dataframe.
df.columns = df.columns.str.strip()

# Descriptive statistics
df.describe()

# Check if there are Null values
df.isnull().sum()

#Create correlation matrix for all variables in the dataframe
sb.set(rc={'figure.figsize':(20,25)})
dff = df[:10][df.columns[:10]]
matrix = np.triu(dff.corr())
sb.heatmap(dff.corr(), annot = True,cmap= 'coolwarm',linewidths=3,linecolor='black',
square=True, mask=matrix)

# Selecting the DEPENDENT and INDEPENDENT variables.
y = df['target']
x = df.drop(['target','v_cat_2','v_cont_7'], axis = 1)

# Splitting the dataset into TRAIN and TEST data.
from sklearn.preprocessing import MinMaxScaler
min_max_scaler = MinMaxScaler()
x = min_max_scaler.fit_transform(x)

#Split data into training and testing datasets 80% for training and 20% for testing
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)



import pandas as pd
```

Predicting a Crime in a particular region using Decision Trees

```
from sklearn.tree import DecisionTreeRegressor # Import Decision Tree Regressor
from sklearn.model_selection import train_test_split # Import train_test_split function
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier


# Handling the continous values in dataset(converting to categorical values)
df = df.drop('v_cat_2',axis=1) df = df.round(0).astype(int) df.head()

# Check if there are Null values
df.isnull().sum()

# Handling the NULL values
df = df.fillna(df.mean())
df.head()

# Encoding the data to make it fit for any algorithm
df['v_cat_0'] = lab_enc.fit_transform(df['v_cat_0'].values.reshape(-1, 1))
df['v_cat_1'] = lab_enc.fit_transform(df['v_cat_1'].values.reshape(-1, 1))
df['v_cont_0'] = lab_enc.fit_transform(df['v_cont_0'].values.reshape(-1, 1))
df.head()

df.columns = df.columns.str.strip()
df.columns

# Selecting the DEPENDENT and INDEPENDENT variables.
y = df['target']
x = df.drop(['target','v_cont_7'], axis = 1)

# Splitting the dataset into TRAIN and TEST.
min_max_scaler = MinMaxScaler()
x = min_max_scaler.fit_transform(x)

#Split data into training and testing datasets 80% for training and 20% for testing
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

# Create Decision Tree classifer object

clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
```

```
clf = clf.fit(x_train,y_train)
#Predict the response for test dataset
y_pred = clf.predict(x_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

print("Accuracy % = :",metrics.accuracy_score(y_test, y_pred)*100)

clf = DecisionTreeClassifier(criterion="entropy", max_depth=3)

# Train Decision Tree Classifer
clf = clf.fit(x_train,y_train)
#Predict the response for test dataset
y_pred = clf.predict(x_test)
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

print("Accuracy % :",metrics.accuracy_score(y_test, y_pred)*100)

import joblib

def result(city):

    filename='finalized_model.sav'
    joblib.dump(clf, filename)
       # load the model from disk
    loaded_model = joblib.load(filename)
    result = loaded_model.score(x_test, y_test)
    result = "{:.2f}".format(result*100)+"%"
    df = pd.read_csv("project/crime_prep.csv")
    val = 'City is out of range.'
    try:
            if len(city) < 4:
                    return result, val


     data = df[df["v_cat_2"].str.contains(city)]
     data_ = data.filter(regex='target').values[0]
     val = ' '.join(map(str,data_))
     val = int(float(val)*100)
     if val < 60:
           val=random.uniform(50.1, 90.5)
     val = str("{:.2f}”. format(float(val))) +"%"
       return result, val

     except IndexError as e:
            return result,val
```

SYSTEM TESTING

Software testing is a process, to evaluate the functionality of a software application with an intent to find whether the developed software met the specified requirements or not and to identify the defects to ensure that the product is defect free in order to produce the quality product. In simple terms, Software Testing means Verification of Application Under Test (AUT).

There are two types of software testing:

1. Manual Testing
2. Automation Testing

Manual Testing:

Manual testing is the process of testing software by hand to learn more about it, to find what is and isn't working. This usually includes verifying all the features specified in requirements documents, but often also includes the testers trying the software with the perspective of their end users in mind. Manual test plans vary from fully scripted test cases, giving testers detailed steps and expected results, through to high-level guides that steer exploratory testing sessions. There are lots of sophisticated tools on the market to help with manual testing, but if you want a simple and flexible place to start, take a look at testpad.com.

Automation Testing:

Automation testing is the process of testing the software using an automation tool to find the defects. In this process, testers execute the test scripts and generate the test results automatically by using automation tools. Some of the famous automation testing tools for functional testing are QTP/UFT and selenium.

7.1 Testing Methods:

There are two types of testing methods:

1. Static Testing
2. Dynamic Testing

7.1.1 Static Testing:

Static Testing is also known as Verification in Software Testing. Verification is a static method of checking documents and files. Verification is the process, to ensure that whether we are building the product right i.e., to verify the requirements which we have and to verify whether we are developing the product accordingly or not.

7.1.2 Dynamic Testing:

Dynamic Testing is also known as Validation in Software Testing. Validation is a dynamic process of testing the real product. Validation is the process, whether we are building the right product i.e., to validate the product which we have developed is right or not.

7.2 Testing Approaches:

There are three types of testing approaches:

1.Black Box Testing

2.White Box Testing

3.Grey Box Testing

7.2.1 Black Box Testing:

It is also called as Behavioural/Specification-Based/Input-Output Testing. BLACK BOX TESTING is defined as a testing technique in which functionality of the Application Under Test (AUT) is tested without looking at the internal code structure, implementation details and knowledge of internal paths of the software. This type of testing is based entirely on software requirements and specifications. In Black Box Testing we just focus on inputs and output of the software system without bothering about internal knowledge of the software program.

Input
Black Box
Output

Fig 7.2.1: Black box testing

7.2.2 White Box Testing:

WHITE BOX TESTING is testing of a software solution's internal structure, design, and coding. In this type of testing, the code is visible to the tester. It focuses primarily on verifying the flow of inputs and outputs through the application, improving design and usability, strengthening security. White box testing is also known as Clear Box testing, Open Box testing, Structural testing, Transparent Box testing, Code-Based testing, and Glass Box testing. It is usually performed by developers.

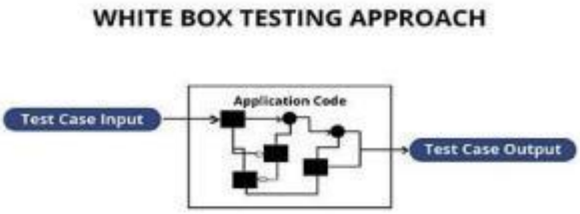


Fig 7.2.2: White Box Testing

7.2.3 Grey box Testing:

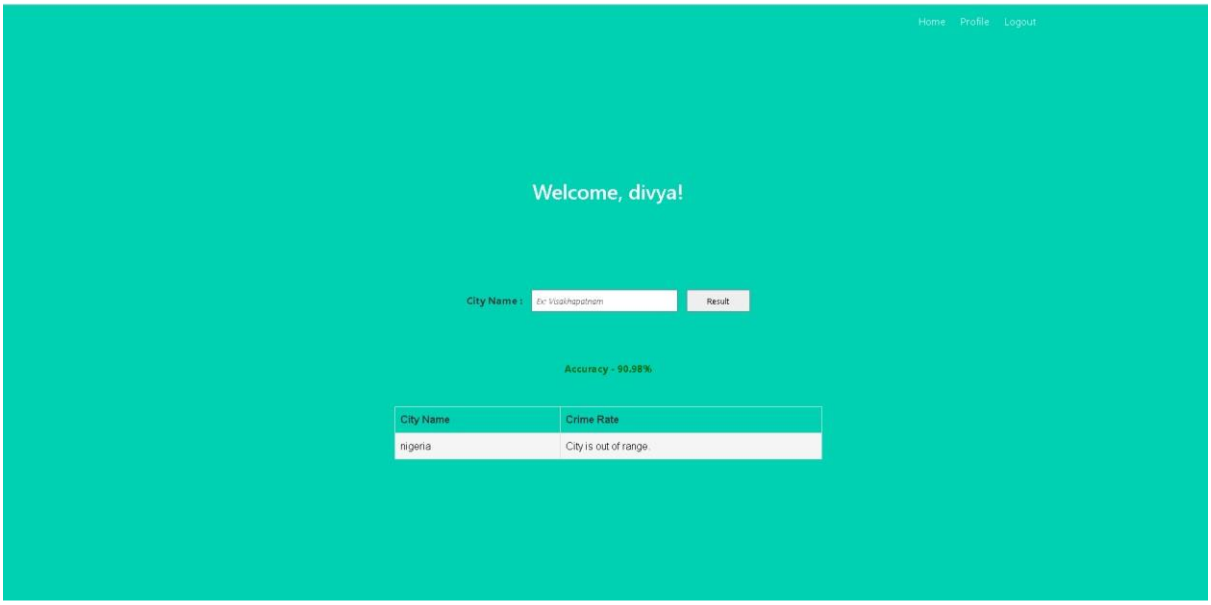
Grey box is the combination of both White Box and Black Box Testing. The tester who works on this type of testing needs to have access to design documents. This helps to create better test cases in this process.

Grey Box Testing
Black Box Testing
+
White Box Testing
=
Grey Box Testing
(Tester having no knowledge of internal functionality)
(Tester having the knowledge of internal functionality)
www.educba.com

Fig 7.2.3: Grey Box Testing

7.3 Test cases:

The following are the test cases depicts about the pass or fail of the application.



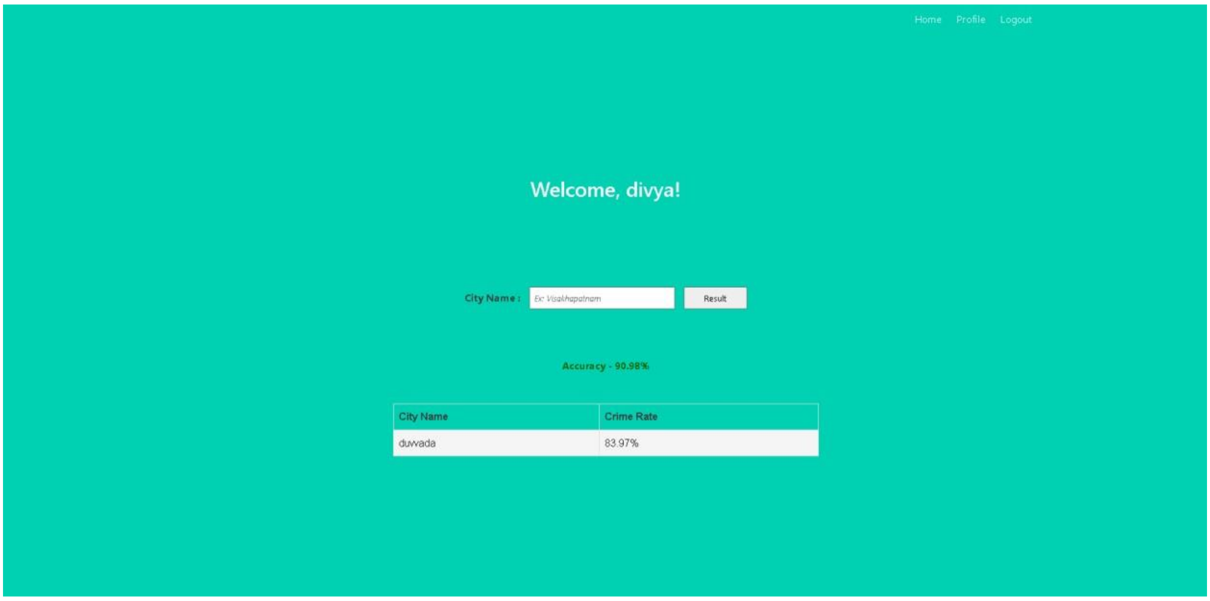
Welcome, divya!

City Name : Ex: Visakhapatnam

Accuracy - 90.98%

City Name	Crime Rate
nigeria	City is out of range.

The above figure explains the failure case of the application i.e., whenever a city entered which is not in range then it occurs the above case.



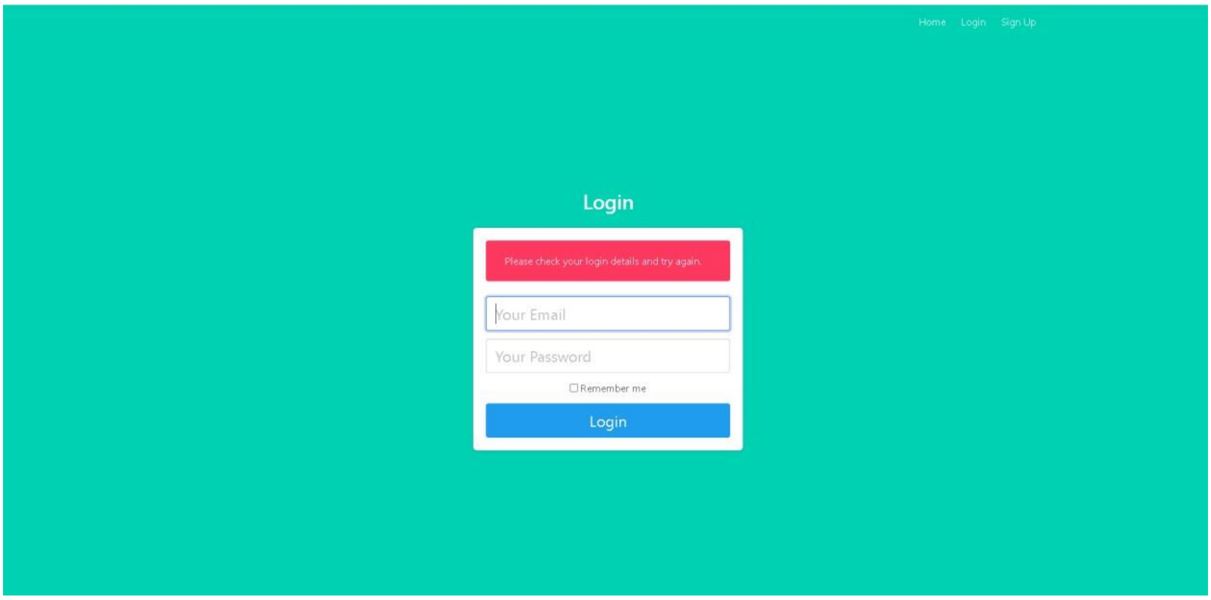
Welcome, divya!

City Name : Ex: Visakhapatnam

Accuracy - 90.98%

City Name	Crime Rate
duvvada	83.97%

The above figure explains the success case of the application i.e., whenever a city entered which is in the range then it provides the respective area's crime rate.



Please check your login details and try again.

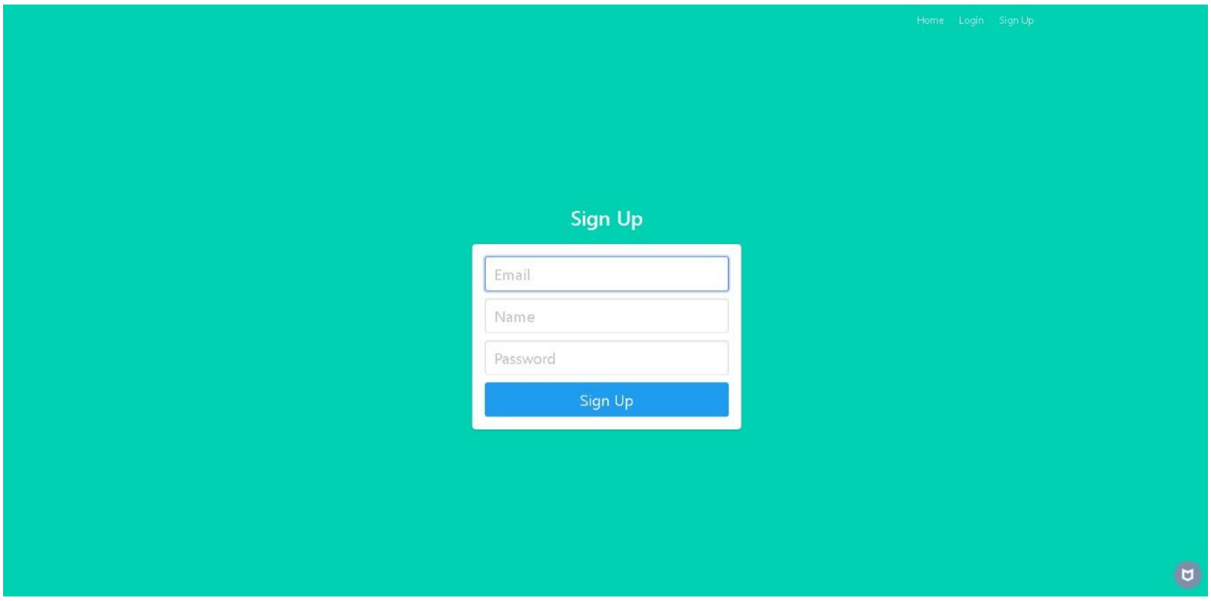
The above figure depicts the test case regarding the failed case of login with wrong credentials.

EXPERIMENTAL RESULTS

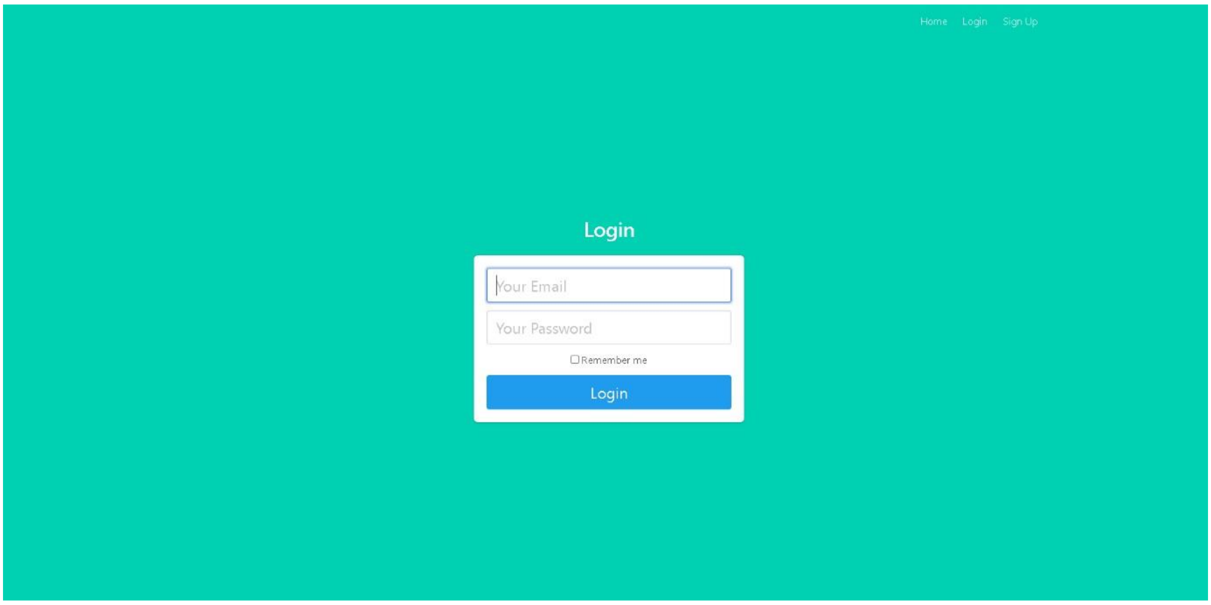
Predicting a Crime in a particular region using Decision Trees

Our model gives an accuracy of 90% which was greater than the previous models to predict the crime in a particular area.

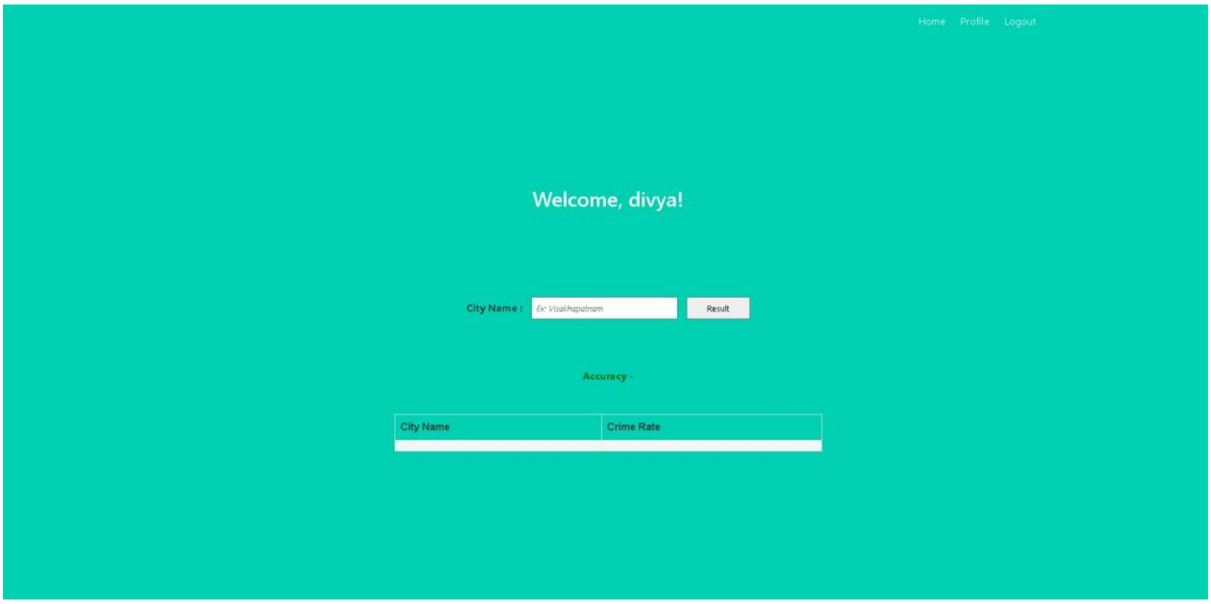
Signup page:



Login Page:



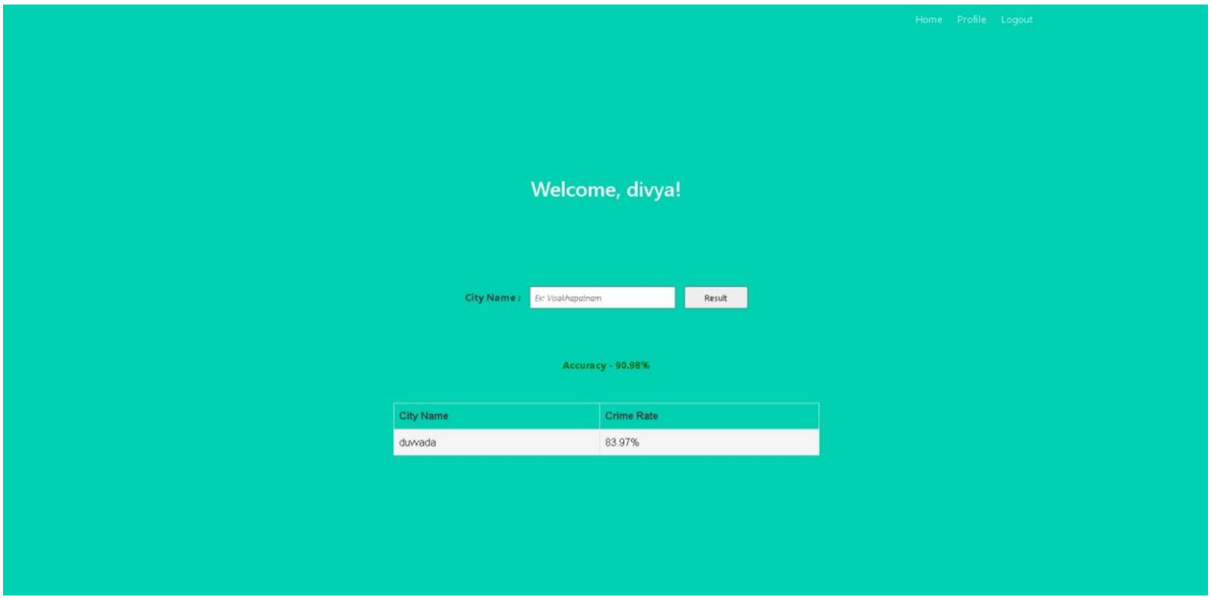
Crime prediction / Result Page:



Welcome, divya!

Accuracy -

City Name	Crime Rate



Welcome, divya!

Accuracy - 90.98%

City Name	Crime Rate
duvvada	83.97%

Conclusion and Future Scope

Conclusion:

With the help of machine learning technology, it has become easy to find out relation and patterns among various data and provides awareness of the crime. The work in this project mainly revolves around predicting the crime of region where it has occurred. Using the concept of machine learning we have built a model using data sets that have undergone data cleaning. The model is classified using decision trees and prediction is done using random forest classifier.

Future work:

In future this model can be built for predicting crime in any area by inserting the data to the model based on the track of user's location and also, they can find the crime rate for other places rather than the areas in their location.

BIBLIOGRAPHY

[1] Adewale Opeoluwa Ogunde, Gabriel Opeyemi Ogunleye, Oluwaleke Oreoluwa. A Decision Tree Algorithm Based System for Predicting Crime in the University. Machine Learning Research. Vol. 2, No. 1, 2017, pp

[2] Kunal Diwan, Sahitya Nigan, Sourabh Tiwari, Vikram Aditya Singh Bhatti "PAASBAAN"-Crime Prediction and classification in Indore city

[3] Prof. ShivPrasadMore, Sakshi Mench, Saloni Kuge, Hafsa Bagwan "Crime Prediction using ML approach" by International Journal of Advanced Research in Computer and Communication Engineering Vol. 10, Issue 5, May 2021.

[4] Pratibha Kumari, Lokesh Chouhan, Akanksha Gotalot "Crime prediction and analysis" Conference Paper by Research Gate in February 2020

[5] Wajiha Safat, Sohail Asghar, Saira Andleeb GillAni "Empirical Analysis for crime prediction and forecasting using Machine Learning and Deep Learning techniques" by IEEE date of publication May 6, 2021

[6] Sakib Mahmud, Musfika Nuha, Abdur Sattar "Crime rate prediction using Machine Learning and Data Mining" by Research Gate in January 2021

[7] Prajakta Yerpude and Vaishnavi Gudur "Predictive Modelling of Crime Dataset using Data Mining" by International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017

[8] Sirivanth Paladugu, Tarun Sai Yakkala, Neeraj Boggarapu, Sri Krishna Kumar Modekurty "Crime rate Prediction using Machine Learning" by International Journal of Research in Engineering, Science and Management Volume 4, Issue 9, September 2021

[9] Abhishek Yadav, Rakshit adgaonkar, Rohit Ingale, Omkar Pathak "Crime Prediction Using Machine Learning Algorithms" by International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 4, Issue 2, April 2021

[10] Anish Krishnan, Aditya Sarguru, A.C. Shantha Sheela "Predictive Analysis of Crime Data using Deep Learning" by International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 4023-4031

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10537

Crime Prediction Using Machine Learning Approach

Prof. Shivaprasad More[1], Sakshi Mench[2], Saloni Kuge[3], Hafsa Bagwan[4]

Assistant Professor, Sanjay Ghodawat University, Kolhapur, Maharashtra, India[1]

Research Scholars, Sanjay Ghodawat Group of Institutions, Kolhapur, Maharashtra, India[2, 3, 4]

Abstract: Crime is one of the serious issues in our society. It is the most predominant aspect of our society. It is also predominant in society. So, the prevention of crime is one of the important tasks. The crime analysis should be in a systematic way. As the analysis makes it important in the detecting and prevention of crime. The analysis detects the investigating patterns and helps in the detection of trends in crime. The main of this paper is the analysis of the efficiency of the crime investigation. The model is designed for the detection of crime patterns from inferences. The inferences are collected from the crime scene and these inferences, the paper demonstrates the prediction of the perpetrator. The paper gives the research way for the prediction of perpetrator age and gender. This paper gives two major aspects of crime prediction. One is perpetrator gender and the other is perpetrator age. The parameters used are analysis of the various factors like the year, month, and weapon used in the unsolved crimes. The analysis part identifies the number of unsolved crimes. The prediction task involves the description of the perpetrator's age, sex, and relationship with the victim. The dataset used in this paper is taken from the Kaggle. The system predicts the output using multi-linear regression, K-Neighbor's classifier, and neural networks. It was trained and tested using a machine learning approach.

Keywords: Crime Prediction, KNN, Decision Tree. Multilinear Regression; K-Neighbors Classifier, Artificial Neural Networks.

I. Introduction

A crime is nothing but it's an action. It constitutes an offense. It's punishable by law. The identification and analysis of hidden crime is a very difficult task for the police department. Also, there is voluminous data of the crime is available. So, there should some methodologies that should help in the investigation. So, the methodology should help to solve the crime.

The machine learning approach can better help in the prediction and analysis of the crime. The machine learning approach provides regression algorithms. The classification techniques provide help to fulfill the purpose of investigation. Regression techniques such as multilinear regression are a statistical method. This method helps to find the relationship between two quantitative values or variables. This approach predicts the values of the dependent variables based on the independent variables. The classifier techniques such as K-Neighbor's classifier. These classifiers are used to classify the multiclass target variables. The neural networks are used to improve the accuracy. The neural network has an input layer dense and has an output layer. Based on the above algorithms the perpetrator description such as sex, age, and the relationship are predicted. The model is thus expected to help to remove the burden of the police investigation. Thus, it helps to solve homicide cases.

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

II. Literature review

Ling Chen, Xu Lai (2011) [1] has compared the experimental result that is obtained by the ANN (Artificial Neural Network). Jyoti Agarwal, Renuka Nagpal, et al., (2013) [2] has studied the crime analysis using K-means clustering on the crime dataset. They have developed this model using the rapid miner tool. The clustered results are obtained and analysed by plotting the values over the years. This model gives the result of the analysis that the number of homicides decreased from 1990 to 2011.

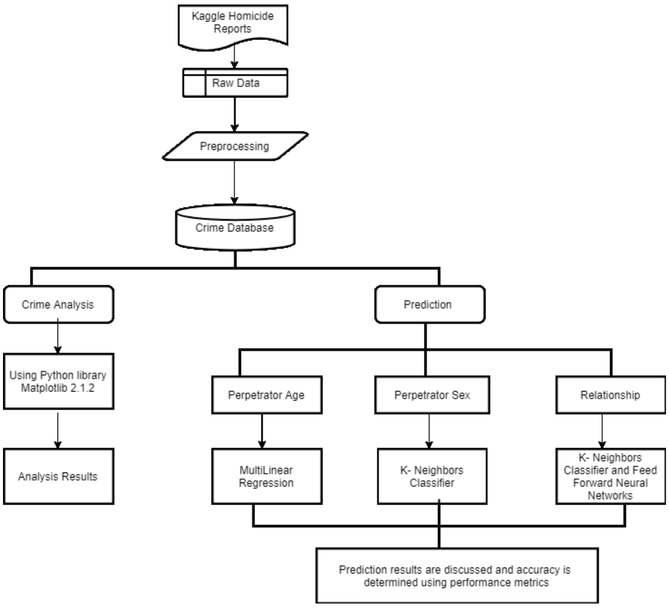
Shiju Sathyadevan, Devan M. S, et al., (2014) [3] have predicted the regions where there is a high probability of the crime occurred. They have visualized crime-prone areas also. They have classified the data using Naive Bayes classifiers. This algorithm is a supervised learning algorithm that also gives the statistical method for classification. This classification gives an accuracy of the 90%.

Lawrence McClendon and Natarajan Meghanathan (2015) [4] have used Linear Regression, Additive Regression, and Decision Stump algorithms using the same set of input (features), on the Communities and Crime Dataset. Overall, the linear regression algorithm gave the best results compared to the three selected algorithms.
Chirag Kansara, Rakhi Gupta, et al., (2016) [8] proposed a model which analyses the sentiments of the people on Twitter and predicts whether they can become a threat to a particular person or society. This model is implemented using the Naive Bayes Classifier which classifies the people by sentiment analysis.

III. Limitations of the Existing System

The existing system gives an accuracy of only 65 %. The model is used only using linear regression. The multiple approaches of machine learning are not implemented. Also, the model has used the dataset of the limited crimes.

IV. Proposed System Model



V. Implementation and Analysis

The dataset we have used contains almost 63000 values. The dataset is taken from the Kaggle website where the dataset is freely available. It has entries from 1980 to 2014.
The analysis includes the number of unsolved crimes, the weapons used in the crimes. The month when the maximum crime took place. The places and occurrence of the crime. The state where the crime rate is high.

VI. Methodology

The dataset is obtained from the Kaggle repository. This is the domain for the various research-oriented dataset. The dataset contains homicide entries collected from the FBI's supplementary Homicide Report. The dataset consists of 638454 rows and 17 columns and the column metadata. From the dataset, the significant features like State, Year, Month, Crime Type, Crime Solved, Victim Gender, Victim Age, Victim Race, Victim Count and Weapon are chosen as the input features for the system. The features Perpetrator Age, Perpetrator Sex and Relationship of the perpetrator with the victim are chosen as the target variable to be predicted by the system. We have used two algorithms for the prediction one is multilinear regression and the other is K-neighbors classifier.

a) MultiLinear Regression

This algorithm gives the mathematical approach to find the relationship between the dependent variable with the given set of independent variables. In our research, the perpetrator's age is a dependent variable, and the independent

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10537

variables are pieces of evidence collected from the crime scene. This algorithm predicts the perpetrator's age based on input features such as state, year, month, place, and crime solved, etc.
The equation for the Multilinear Regression line is given as:
$Y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp$
Where,
Y is the dependent variable,
x is the independent variable,
βi are coefficients of the regression equations.

b) K-Neighbors Classification

This classification algorithm is used when the target variable has more than two classes to classify [11]. In our dataset, the target variable is nothing, but its perpetrator sex and it has classified namely as male, female and unknown. Also, the target variable relationship has 27 unique values such as friend, wife, nephew, etc. so the K-Neighbors classifier is used to classify these target variables. The target variables are perpetrator sex and relationship.

Pseudo Code:

```
K_Nearest_Classifier (input variables);
Assign K -> the number of clusters
A set of K instances are chosen to be centres for the
clusters
For each data point in the input:
Calculate the Euclidian distance
Assign the cluster which is near to the data point
Recalculate the centroids and reassign the
variables in the clusters.
```

VII. IMPLEMENTATION DETAILS

The implementation details include the machine learning approach.

Data-collection:
The data collection for the implementation is from the Kaggle. The dataset is freely available. The record collected is almost 63000.

Pre-processing:
Once the dataset is collected, it must be pre-processed to get the clean dataset. The pandas and NumPy libraries are available in python for the pre-processing. it is removing of empty values from the dataset or repeated records should be removed.

Analysis:
The analysis includes the graphical representation of different values to analyse the dataset property. The different graphs are plotted by Matplotlib libraries. The graphical analysis gives a direction towards the prediction.

Training and Testing:
The dataset is divided into training and testing. Generally, 70 % dataset is kept for training and 30% for testing. The dataset ratio can be 70: 30 or 80:20.

Validation:
Once the model is created, it should be validated with the real-time data values. This is called validation. The validation is nothing, but its predicted value and it's also called the output value.

VIII. RESULTS AND COMPARATIVE STUDY

Our model gives an accuracy of 85 %. The previous model gives an accuracy of 65%. The below graph gives the comparison of the model with the previous results.

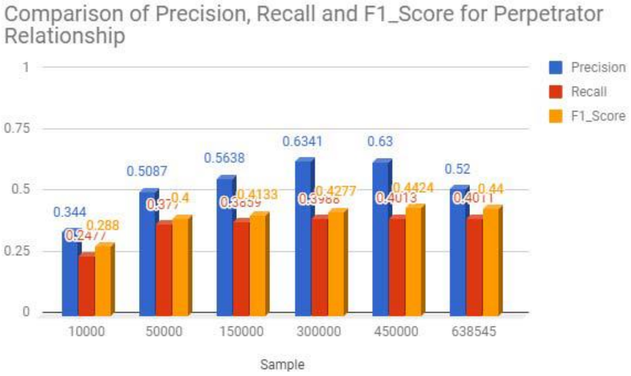


Fig. 1Comparison of Precision, Recall and F1_Score for Perpetrator Relationship.

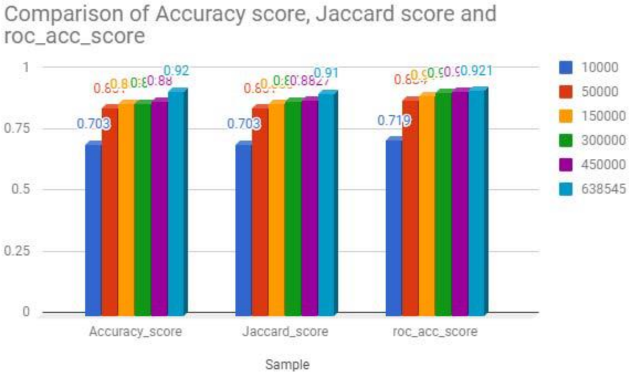


Fig. 2Comparison of Accuracy score, Jaccard score and roc_acc_score.

The ratio estimated through calculation of recall in is found to outscore those of precision and F1 score. However, in the case of a set of 10,000 samples, the values of precision and F1 score are observed to be greater than the recall score. This can be inferred as an indication of a larger number of false negatives present in the sample set as opposed to the number of false positives predicted by the model.

IX. CONCLUSION

This model helps to predict crime. The perpetrator's age, perpetrator sex, and relationship can be predicted using a machine learning approach. The regression and classifier are used here give almost 80 % accuracy. The dataset can be enhanced and can be used in other countries if the scenario is almost same. The model gives the overall prediction of any crime. This model can be enhanced by using deep learning techniques.

X. FUTURE WORK

This model gives an accuracy of almost 80 % for the perpetrator age, 82 % for the perpetrator sex, and 85 % for the relationship. The accuracy can be improved by using a complex neural network such as the recurrent neural network. Also, the deep learning approach can be used to enhance the accuracy of the model.

REFERENCES

[1]. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia- Pacific. IEEE, 2011.

[2]. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." International Journal of Computer Applications 83.4 (2013).

[3]. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014

[4]. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyse crime data." Machine Learning and Applications: An International Journal (MLAIJ)
2.1 (2015).

[5]. Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. "Analysis and prediction of crimes by clustering and classification." Analysis 4.8 (2015).

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.10537

[6]. Heartfield, Ryan, George Loukas, and Diane Gan. "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks." IEEE Access 4 (2016): 6910-6928.
[7]. Sivaranjani, S., S. Sivakumari, and M. Aasha. "Crime prediction and forecasting in TamilNadu using clustering approaches." Emerging Technological Trends (ICETT), International Conference on. IEEE, 2016.
[8]. Kansara, Chirag, et al. "Crime mitigation at Twitter using Big Data analytics and risk modelling." Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on. IEEE, 2016.
[9]. Tsunoda, Masateru, Sousuke Amasaki, and Akito Monden. "Handling categorical variables in effort estimation." Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement. ACM, 2012.
[10]. Su, Ya, et al. "Multivariate multilinear regression." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42.6 (2012): 1560-1573.
[11]. Viswanath, P., and T. Hitendra Sarma. "An improvement to K nearest neighbour classifier." Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE. IEEE, 2011.
[12]. Palocsay, Susan W., Ping Wang, and Robert G. Brookshire. "Predicting criminal recidivism using neural networks." Socio-Economic Planning Sciences 34.4 (2000): 271-284

International Journal of Research in Engineering, Science and Management
Volume 4, Issue 9, September 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Crime Rate Prediction Using Machine Learning

Sirivanth Paladugu[1*], Tarun Sai Yakkala[2], Neeraj Boggarapu[3], Sri Krishna Kumar Modekurty[4]
[1,2,3,4]Student, Department of Computer Science and Engineering, SRM University, Guntur, India

***Abstract*: The main objective of our project is to predict the Crime Rate in different regions using specific parameters like density, country, crime rate, central etc. In this project we considered a dataset of a particular Country. The region we selected is the United States of America. We used the Linear Regression model to predict the crime rate. Finally, a graph is plotted after implementing Linear Regression. A graph is plotted between density and crime rate to enhance distribution of crime rate in a particular region.**

***Keywords*: Data mining, data preprocessing, linear regression, predictive analysis.**

1. Introduction

Analysis of crime is a methodological approach to the identification and assessment of criminal patterns and trends. Before starting this project we have gone through many datasets all over the world, but in the USA a state named North Carolina has given us sufficient data to start off with. The variables with highest correlation in regards with crime rate are urban and density. The reason for highest correlation for urban and density variables is that the urban areas of North Carolina are densely populated. As a result, there is a high probability of multi-collinearity between the density and urban features. The combination of density and location features will help us in predicting the crime rate in North Carolina. The columns of wage, wfed and wtrd are positively correlated with the density feature. From this we can instinctively understand that the weekly wages would be higher in urban localities. The features wtrd and wfir are positively correlated with wfed and wloc respectively. Both the wfir and wtrd are moderately correlated. The features crime rate, density, urban, wfed, wtrd and taxpc are very highly correlated with the Crime Rate. The prediction of future crime location can be determined by geographical data mining approaches

2. Problem Survey

In our problem we will be evaluating and examining the large pre-existing databases in order to generate new information which would help us to find the solution. The prediction is based on the extraction of the new information using the existing datasets. The main aim of this problem is to perform the survey on certain algorithms which helps us to analyze the crime rate.

3. Dataset Description

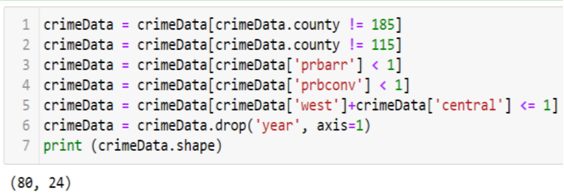
In our dataset we have considered several parameters like country, year, crime rate, density etc. After once we get the raw dataset we then clean the dataset. After we clean the dataset we then preprocess it. We then use linear regression to obtain our results, the results are obtained in the form of graphs. We then use a statistical approach to find values like count, mean, standard deviation, minimum and maximum values.

For each and every parameter we have given a model summary. A combination of density and location (west/ central/ urban) can help aid crime rate prediction. The variables with highest correlation in regards with crime rate are urban and density.

4. Data Preprocessing

The important step in the entire process is data preprocessing. Data preprocessing allows us to remove the unwanted data with the help of data cleaning, this allows the user to have a dataset which contains more valuable information after the preprocessing stage in the mining process.

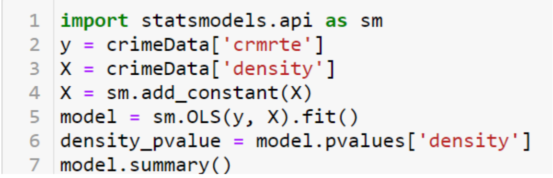
Data cleaning:



```
crimeData = crimeData[crimeData.county != 185]
crimeData = crimeData[crimeData.county != 115]
crimeData = crimeData[crimeData['prbarr'] < 1]
crimeData = crimeData[crimeData['prbconv'] < 1]
crimeData = crimeData[crimeData['west']+crimeData['central'] <= 1]
crimeData = crimeData.drop('year', axis=1)
print (crimeData.shape)
```

```
(80, 24)
```

Preprocessing the cleaned data:



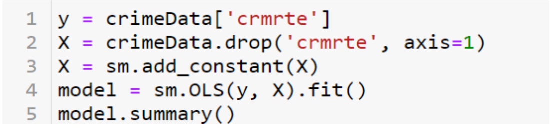
```
import statsmodels.api as sm
y = crimeData['crmrte']
X = crimeData['density']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
density_pvalue = model.pvalues['density']
model.summary()
```

Implementation:

We have implemented the problem using an algorithm called Linear Regression. We have developed a python code based on the algorithm. For this problem we also calculated R Square, P - Value Difference, Number of Observations, Covariance type, etc.

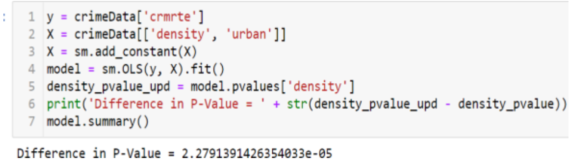
[*] Corresponding author: sirivanth2000@gmail.com

Parameters crime rate:



```
y = crimeData['crmrte']
X = crimeData.drop('crmrte', axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

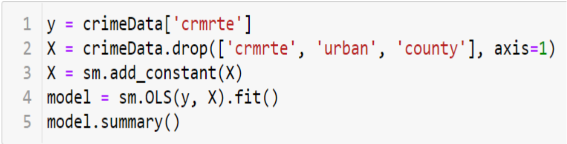
Parameters density urban:



```
y = crimeData['crmrte']
X = crimeData[['density', 'urban']]
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
density_pvalue_upd = model.pvalues['density']
print('Difference in P-Value = ' + str(density_pvalue_upd - density_pvalue))
model.summary()
```

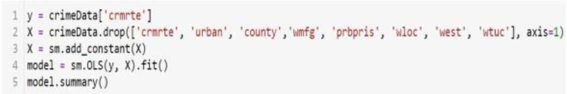
```
Difference in P-Value = 2.2791391426354033e-05
```

Parameters crime rate urban country:



```
y = crimeData['crmrte']
X = crimeData.drop(['crmrte', 'urban', 'county'], axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

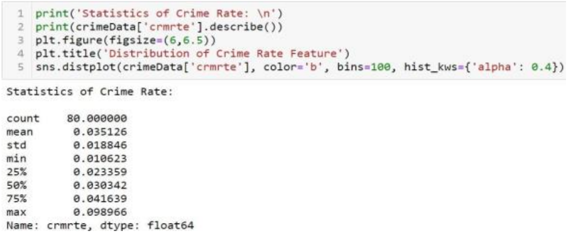
Parameters crime rate urban country:



```
y = crimeData['crmrte']
X = crimeData.drop(['crmrte', 'urban', 'county','wmfg', 'prbpris', 'wloc', 'west', 'wtuc'], axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Resultant Graphs:

Statistical approach towards crime rate prediction:



```
print('Statistics of Crime Rate: \n')
print(crimeData['crmrte'].describe())
plt.figure(figsize=(6,6.5))
plt.title('Distribution of Crime Rate Feature')
sns.distplot(crimeData['crmrte'], color='b', bins=100, hist_kws={'alpha': 0.4})
```

```
Statistics of Crime Rate:

count    80.000000
mean      0.035126
std       0.018846
min       0.010623
25%       0.023359
50%       0.030342
75%       0.041639
max       0.098966
Name: crmrte, dtype: float64
```

Graph Output:

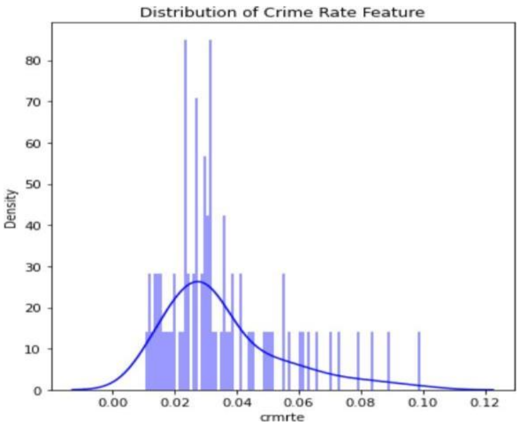
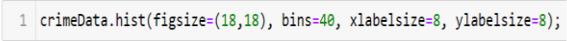


Fig. 1. Distribution of crime rate feature

Crime data mined in different perspectives:



```
crimeData.hist(figsize=(18,18), bins=40, xlabelsize=8, ylabelsize=8);
```

Graph Outputs:

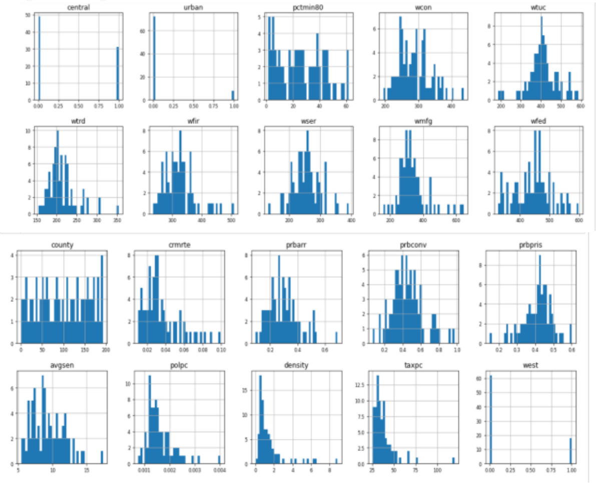


Fig. 2. Output

5. Conclusion

The map of Actual vs. Predicted is linear. This indicates that the forecast is right. The amount of data available for input is small. The plot could be more linear if there was more evidence.

We may also incorporate the Boolean features west, center, and urban into a single feature with categorical values 1,2, and 3 to create a single feature. A single function like this may be more useful for prediction. On features, functional transformations (such as log, function) can be useful. If there is a possibility to incorporate functionality, using 'unemployment rate' as a proxy for crime rate might be useful.

References

[1] Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, “The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan,” Cities, no. October 2017, pp. 0-1, 2018.

[2] Shah, N., Bhagat, N. & Shah, M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Vis. Comput. Ind. Biomed. Art* 4, 9 (2021).

[3] Prithi S, Aravindan S, Anusuya E, Kumar AM (2020) GUI based prediction of crime rate using machine learning approach. Int J Comput Sci Mob Comput 9(3):221–229

[4] Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, "Crime Prediction and Analysis," *2nd International Conference on Data, Engineering and Applications (IDEA)*, 2020, pp. 1-6.

[5] B. Sivanagaleela and S. Rajesh, "Crime Analysis and Prediction Using Fuzzy C-Means Algorithm," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 595-599.

[6] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve and N. Lal, "Crime Prediction Using K-Nearest Neighboring Algorithm," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1-4.

[7] S. Agarwal, L. Yadav and M. K. Thakur, "Crime Prediction Based on Statistical Models," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, 2018, pp. 1-3.

[8] A. Almaw and K. Kadam, "Crime Data Analysis and Prediction Using Ensemble Learning," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 1918-1923.

[9] D. M. Raza and D. B. Victor, "Data mining and Region Prediction Based on Crime Using Random Forest," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 980-987.

IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 4.819

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 4, Issue 2, April 2021

Crime Prediction Using Machine Learning Algorithms

Abhishek Yadav, Rakshit Adgaonkar, Rohit Ingale, Omkar Pathak
Students, Department of Computer Engineering
NBN Sinhgad School of Engineering, Pune

Abstract: *The most serious security challenges we face in these turbulent times are terrorist attacks and the transmission of disease. length and breadth are measured in hundredths of a centimetre. On a daily basis, we see the most minor offences committed by ordinary citizens. Details of breaches and recurring cases of items should be applied to files to ensure that they are up to date. When it is known that a crime has been committed, people believe that disciplinary action will be taken, even if there is no means of knowing which one. The study of criminology helps to broaden our understanding of who is likely to become a suspect. In the midst of his attempts to identify and deter alleged criminals from reoffending the legal system, he is incorporating both computer science and deep learning. Anyone interested in learning more about the workings of the Chicago Police Force should visit "The Chicago Police Department Site." The Crime Timeline will keep track of all criminal activity as well as the time and date of any incident that occurs. The data collection and modelling have been completed; all that remains is on-line modelling and compilation. To address this question, we must first determine if the case history of K-grooming and other related methods will help with criminal prediction. The invention is typically used as a testing tool, but it can also be used in conjunction with other technologies. Based on internal or external metrics, an algorithm can estimate how easily law enforcement authorities may be able to track, anticipate, and cope with, or preempt, risks, such as the ratio of those sentenced to those arrested, with a life sentence to those awaiting the risk of life imprisonment.*

Keywords: Crime Prediction, Data Analysis, Natural Language Processing, Machine Learning

I. Introduction

To understand this, it should be noted that criminals can be harmful to our wellbeing Nobody is actually imprisoned for just for ordinary crimes but a wide variety of misdemeanours occur on a daily basis. While nearly all galaxies are rotating at over the course of a million times a second, this galaxy can advance approximately 70 million miles a year. In small towns, crime is not just as common as in rural counties as it is in large cities. However numerous their names may be, all these actions may be, attempted murder, armed robbery, larceny, false imprisonment, sexual assault, false conviction, and battery are all identical. Criminal activity has increased, it is the responsibility of the police to act quickly and stop the problem from spreading. since all past criminal offences contain specific information and some indications of future criminal behaviour, the task of policing the city is difficult to predict, the cops' abilities to forecast criminal activity are highly limited To minimise the number of incidents, there are solutions that need to be employed to increase the speed of determination. Using an existing research approach, such as identifying places where crimes are most likely to occur, is a good way to better predict where potential ones will occur The definition of crime in this research uses different terminology throughout; terms related to the places it is incorporated into various geographic divisions are mentioned as well. the likelihood of committing a certain offence can be predicted based on knowing the circumstances in which it is more specific crimes have occurred before lots of crimes have happened in a certain locations also aids in the identification of hotspots, as crime prevention is sought in locations where crimes have been committed most frequently According to recent criminal rate figures, that must have happened with the same methodology as before. The use of a strong and highly functional cyber forensics platform is imperative for allowing us to detect trends in the database in a timely manner that can be viewed immediately so bugs can be addressed quickly.

Noncomprehending this phrase: Extortion, robbery, financial terrorism, and vandalism may also be examples of other kinds of criminal activity that are considered to be new for the industry, which may be described as unnatural. One important function of new technologies is the consumers get out of using it is to open doors to all the types of crooks, and customers getting a return on their investment out of the addition of new technology. Risks to financial, which include the possibility of financial losses due to cybercrime, an inability to transact business, and data leaks are covered under three distinct categories: 1) financial loss due to cybercrime; 2) incapacity to manage business due to technological factors; 3) possible data leakage.

II. Problem Statement

The models for the crime prediction methods will be used to search through data found in the police archives for data on specific types of crimes as well as the records on different details of their occurrences, with variables that influence the probability of occurrence studied to get better results.

III. Literature Survey

Alkesh Bharati1, Dr Sarvanaguru RA, “Crime Analysis and Prediction Using Fuzzy C-Means Algorithm”[1], The data was/is presented as Crime research is a tool used to define criminal activities and study them. If the research conducted so far can be seen to be more specifically useful, it is mostly because it indicates which criminal types are useful in controlling crime, then, mostly they would it be places where violent crimes are reduced. It is an excellent method for measuring the crime rate because of each region can be broken down by procedure and the data is collected for any of each process to be examined. Through the rapid increase in information technology, crime analysts will be able to continue to enhance the investigations and help them interpret the evidence. on the sample clustering and preprocessing to get unstructured evidence, and then look for crimes inside it Thus, persons formerly investigated and then arrested or identified as having committed the same criminal behaviour may often be looked at at for patterns such as criminal history, or incident reports, rather than only offences themselves. This is simply intended to direct law enforcement resources to where crimes can occur, without attention to identifying who is responsible. Bayesian classifiers were used as the current scheme was in use in place In the current methodology, the fuzzy C-Means algorithm will be used to group the crime data for all items that are apprehensible, apprehension of, physical assault, larceny-theft, and crime of women, as well as all criminal offences such as kidnapping, in the dataset.

Shubham Agarwal, Lavish Yadav , “Crime Prediction based on Statistical Models”[2], The data was/is presented as In societies that have less criminal activity, the increase in the number of various offences is still a matter of concern. [Through] developments in technology, freely accessible records, and services,] these people manage to go undetected by society and continue their illegal deeds, even in the act of involving far more people. As a result, crime is rising in countries with a steep increase in incidence in [either/many] developed and [or/in] and [under] developing Because of the preceding year's criminal events in Indian states, we provide two models —Working Average Geometric Progression and WAGP and Seasonal Augmented EP WAG with known past criminal occurrences in order to anticipate the crime activity that is likely to occur in following years. These recent examples are used to help researchers understand how crime has changed in Indian states between the years 2001 and 2013. It was discovered that information from police reports covering the years 2001 to 2011 was a valuable in forecasting data that followed. The calculated expected crime values were compared to crime data for the same years, as well as for 2014 and 2015. To estimate our prediction to be within 85% of the correct value obtained from real data, the difference between the real and expected values of each needs to be doubled.

Chitra Lekha, “Data Mining Techniques in Detecting and Predicting Cyber Crimes in Banking Sector”[3appearing as,also represented as Data mining is used in many industries such as client segmentation and efficiency, credit scoring, predicting payment failures, preventing fraud, and predicting which customers will default, but are also in the fields of advertisement and marketing Data Mining in banking applications is presented in this paper as a basic concept that serves as a foundation for understanding a range of diverse cyber criminals. It offers an inclusive look at the most powerful and current approaches used data mining practises for the purposes of criminal data analysis. the aim of

DOI: 10.48175/IJARSCT-1065

stealing personal data is to be able to identify behavioural habits in order to save you from doing what's likely to happen, and avoid the illegal things you have already done This paper employs novel data mining techniques like K-Means, which uses current events as a context signal to guide the challenge, and influences, and a novel data mining algorithm known as J48 Prediction Tree, to investigate sets of data about cyber crime and cyber threats. Cluster expansion is being used in unsupervised learning even though it is affected by the association (influenced) Using the J48 algorithm, the classifier finds the initial centres of things that may be indicative of cyber criminals and populates the list of candidate results with the predicted instances of those centres. In light of the above, a general intelligence outcome over all criminal justice and a prediction tree will definitely have expanded, and improved, and a significant general learning result for banking sector is a K-Means and J48 Incorporation. As a result, we need our law enforcement agencies to be properly equipped to face and avoid and deter online criminal activities.

Nafiz Mahmud, Khalid Ibn Zinna, "CRIMECAST: A Crime Prediction and Strategy Direction Service"[4], The data was/is presented as The wide variety of studies on criminology is valuable in providing us with a new information on criminal psychology. Criminals don't live in uncertain territory; they wait before they have an easy target to commit offences, in which case there are clustered areas like hotspots of people or strangers. It is possible to simulate a crime forecasting model using evidence that can be checked in the fact of past crimes, as long as it has been publicly available, there is enough time enough to verify. This paper intends to show how the CCRIMBA's artificial Neural Network has been broadened to include the CRIMAST, a crime prediction and threat management service which assists law enforcement in training and testing criminologists to work with Neural models. the CriMA employs spatial techniques, which concentrate on legitimate crime patterns of crime and generate defence strategies, designates areas vulnerable to criminal action, and then broadcasts security warnings. Our simulation with a large dataset shows that CRIEC can be much more effective than other models in terms of predicting crime.

Mary Shermila A, "Crime Data Analysis and Prediction of Perpetrator Identity using Machine Learning Approach"[5], The data was/is presented as Prevention is one of the most prominent and important tasks we have in the realm of civilization. In addition to being a means of identifying and researching the usual patterns and developments of violence, it is also a systemic, scientific approach. the aim of this model is to make systems more effective at detecting and apprehending criminals This statistical model can be employed at the crime scenes to discover crime dynamics and to forecast the description of the criminal most probable perpetrator to be present based on inferences drawn from the site. This is a long, involved process of both physically expanding and philosophically advancing. Predictions on who is likely to commit crimes and how serious such crimes are The phase includes determining the number of open offences, which gauges the importance of different variables, such as the year, month, the weapons used, and the social class or demographics of the perpetrators. the prediction process is able to deduce how old, whether the suspect is male, female, and/how many years they've known the victim There are several theories from the investigation based on the information gathered on this area. The method uses multilinomial regression, k-neigh regression, and neural networks for classifiers like Multilearate Regression, kNeighbors, and KNeighbors for static entity definition. The machine learning algorithm was developed and thoroughly tested using the San Francisco homicide dataset (1981-2014) and then deployed using Python.

IV. System Architecture

4.1 Explanation of System Architecture Modules:

An improvement on image data unwilling to distort, but necessary for further processing is achieved by means of geometric transformations, including rotation, scaling, and translation.

Likewise, the method of feature selection is feature extraction, which breaks the raw data into usable classes and thereafter divides it into a series of features is a source of data reduction. When you want to work on it, you will find it more agreeable. Also, one of the primary characteristics of these massive data sets is that it has a large numbers of variables.

Though supervised learning machines (i.e.e. classification or regression) can both classify and predict the future, it by getting information from past training examples, they are more useful for regression and less useful for prediction.

Classification issues account for the vast majority of the instances, however. We depict each object of the data with n-dimensional space (which could be described as a set of features with each having a value for each of those features), and the features are identified with various coordinates. And once the hyper-plane has distinguished the two classes, we use the classifier to locate the other point(s) that sits(s) in between them

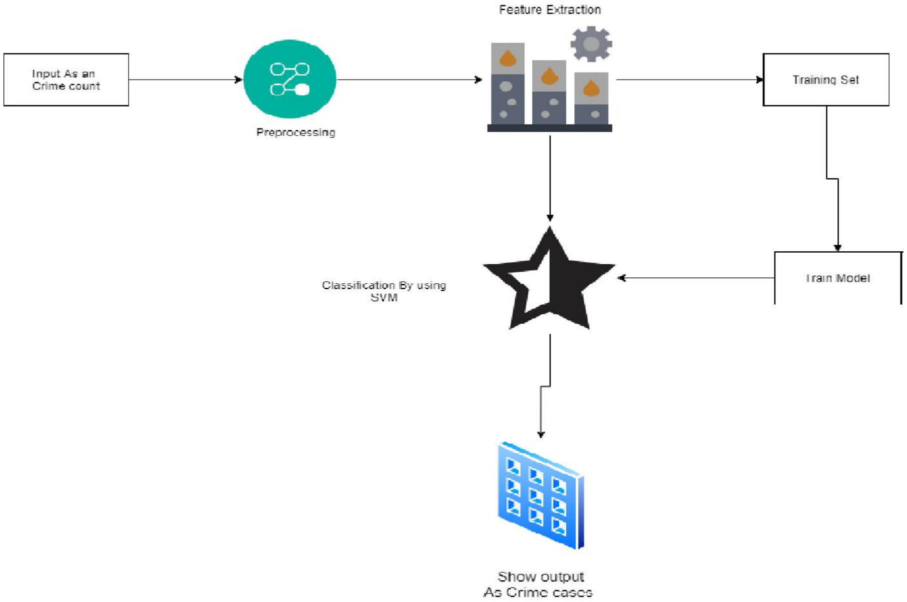


Figure: System Architecture

V. Conclusion

Since computers are now being used for both on- and off-site to help in data analysis, it is possible to recognise patterns using artificial intelligence. the bulk of the research in this project focuses on is focused on detecting crimes that have already occurred It was processed using the machine learning technique of data cleaning and normalisation. According to the theory, the prediction, this type of crime would have an accuracy of 7.88% Expanding on the original definition, it can be defined as "to assist in making better use of the dataset". When you look at the shapes of diagrams, you're often examining their properties, not when you determine what sort of information you're searching for. we designed multiple images to go along with this concept It was essential in obtaining datasets in Chicago to show patterns and trends in criminal behaviour in order to discover different factors that could further improve the likelihood of avoiding or combat crime. The new uses of AI and big data analysis to clarify and illustrate dynamic relationships and relationships that are based on big data have gained prominence in recent years the vast majority of the research done in this project has to this has been dedicated to estimating crimes that have already occurred It was through the use of the machine learning techniques that we developed a model using the data that has been through the whole data cleaning and data analysis.

References

[1]. McClendon, Lawrence, and Natarajan Meghanathan. ”Using machine learning algorithms to analyze crime data.” Machine Learning and Applications:An International Journal (MLAIJ) 2.1 (2015).

[2]. Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. ”Analysis and prediction of crimes by clustering and classification.” Analysis 4.8 (2015).

[3]. Heartfield, Ryan, George Loukas, and Diane Gan. ”You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks.” IEEE Access 4 (2016): 6910-6928.

DOI: 10.48175/IJARSCT-1065

[4]. Sivaranjani, S., S. Sivakumari, and M. Aasha. ”Crime prediction and forecasting in TamilNadu using clustering approaches.” Emerging Technological Trends (ICETT), International Conference on. IEEE, 2016.

[5]. Kansara, Chirag, et al. ”Crime mitigation at Twitter using Big Data analytics and risk modelling. ”Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on. IEEE, 2016.

[6]. Kim, Suhong, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. ”Crime Analysis Through Machine Learning.” In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 415-420. IEEE, 2018.

[7]. Shah, Riya Rahul. ”Crime Prediction Using Machine Learning.” (2003).

[8]. Lin, Ying-Lung, Tenge-Yang Chen, and Liang-Chih Yu. ”Using machine learning to assist crime prevention.”In 2017 6th II AI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 1029-1030. IEEE, 2017.

[9]. M. V. Barnadas, Machine learning applied to crime prediction, Thesis, Universitat Politecnica de Catalunya, Barcelona, Spain, Sep. 2016. ‘

[10]. Williams, Matthew L., Pete Burnap, and Luke Sloan. ”Crime sensing with big data: The a□ordances and limitations of using open-source communications to estimate crime patterns.” The British Journal of Criminology 57, no. 2 (2017): 320-340. 11 Agarwal, Shubham, Lavish Yadav, and Manish K. Thakur. ”Crime Prediction Based on Statistical Models.”In2018Eleventh International Conference on Contemporary Computing (IC3), pp. 1-3. IEEE, 2018.

DOI: 10.48175/IJARSCT-1065