## CRIME ANALYSIS AND PREDICTION USING OPTIMIZED KNN ALGORITHM

*Submitted by*

**AISHWARYA T M 211418104009**

**ASHMIN ANGEL A 211418104028**

**SRINIDHI ISWARYA B.K 211418104263**

***in partial fulfillment for the award of the degree***

*of*

**BACHELOR OF ENGINEERING**

IN

**COMPUTER SCIENCE AND ENGINEERING**



**PANIMALAR ENGINEERING COLLEGE**

**(An Autonomous Institution, Affiliated to Anna University, Chennai)**

**MAY 2022**

# PANIMALAR ENGINEERING COLLEGE

# (An Autonomous Institution, Affiliated to Anna University, Chennai)

**BONAFIDE CERTIFICATE**

Certified that this project report **“CRIME PREDICTION USING OPTIMIZED KNN ALGORITHM”** is the bonafide work of **“AISHWARYA T M [REG NO: 211418104009], ASHMIN ANGEL A [REG NO: 211418104028] and SRINIDHI ISWARYA B.K [REG NO: 211418104263]”** who carried out the project work under my supervision.

### SIGNATURE SIGNATURE Dr.S.MURUGAVALLI,M.E.,Ph.D., Mrs. DEVI .R, M.E.,

**HEAD OF THE DEPARTMENT SUPERVISOR**

**ASSOCIATE PROFESSOR**

DEPARTMENT OF CSE, DEPARTMENT OF CSE, PANIMALAR ENGINEERING COLLEGE, PANIMALAR ENGINEERING

COLLEGE,

NAZARATHPETTAI, NAZARATHPETTAI,

POONAMALLEE, POONAMALLEE,

CHENNAI-600 123. CHENNAI-600 123.

Certified that the above mentioned students were examined in End Semester

project viva voice held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### INTERNAL EXAMINER EXTERNAL EXAMINER

**DECLARATION BY THE STUDENT**

### We AISHWARYA T M[211418104009], ASHMIN ANGEL A [211418104028] and SRINIDHI ISWARYA B.K [211418104263] here by declare that this project report titled “CRIME PREDICTION USING OPTIMIZED KNN ALGORITHM”, under the guidance of Mrs. DEVI R, M.E., is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

**AISHWARYA T M**

**ASHMIN ANGEL A**

**SRINIDHI ISWARYA B.K**

**ACKNOWLEDGEMENT**

We express our deep gratitude to our respected Secretary and Correspondent **Dr.P.CHINNADURAI, M.A., Ph.D.** for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

We would like to express our heartfelt and sincere thanks to our Directors

### Tmt.C.VIJAYARAJESWARI, Dr.C.SAKTHIKUMAR, M.E., Ph.D., and

**Tmt. SARANYASREE SAKTHIKUMAR B.E., M.B.A.,** for providing us with the necessary facilities for completion of this project.

We also express our gratitude to our Principal **Dr.K.Mani, M.E., Ph.D.** his timely concern and encouragement provided to us throughout the course.

We thank the HOD of CSE Department, **Dr. S.MURUGAVALLI, M.E.,Ph.D.,** for the support extended throughout the project.

We would like to thank my Project Guide, **Mrs. DEVI R, M.E.,** and all the faculty members of the Department of CSE for their advice and suggestions for the successful completion of the project.

### AISHWARYA T M

**ASHMIN ANGEL A**

**SRINIDHI ISWARYA B.K**

**ABSTRACT**

The objective of this project is to tackle a vital issue in the society - Crimes. Analyzing and examining of crimes happening in the world will give us a Broadview in understanding the crime regions and can be used to take necessary precautions to mitigate the crime rates. Identifying Crime patterns will allow us to tackle problems with unique approaches in specific crime category regions and improve more security measures in society. Current studies show the reason of increase in crime rates is more in areas that are economically backward. In few decades’ property crime will be a target. The following approach involves predicting crimes classifying, pattern detection and visualization with effective tools and technologies. Use of past crime data trends helps us to correlate factors which might help understanding the future scope of crimes. In this day and age security is a perspective which is given higher need by all political and government worldwide and intending to decrease wrongdoing frequency. As information mining is the proper field to apply on high volume wrongdoing dataset and information picked up from information mining approaches will be helpful and bolster police power. So In this paper wrongdoing investigation is finished by performing k-implies bunching on wrongdoing dataset utilizing quick digger apparatus.

from the active server when any updation takes place in the active server.

**TABLE OF CONTENTS**

**CHAPTER NO.** **TITLE PAGE NO.**

**ABSTRACT** iv

**LIST OF FIGURES**  vii

**LIST OF ABBREVATIONS** viii

**1**. **INTRODUCTION**

1.1 Overview 2

1.2 Problem Definition 3

**2.** **LITERATURE SURVEY** 4

**3. SYSTEM ANALYSIS**

3.1 Existing System 9

3.2 Proposed System 9

3.3 Feasibility Study 10

3.4 Hardware Environment 12

3.5 Software Environment 12

**4.** **SYSTEM DESIGN**

4.1 ER Diagram 14

4.2 Data Flow Diagram 16

4.3 UML Diagrams 18

**5.** **SYSTEM ARCHITECTURE** 21

5.1 Module Design Specification 22

5.2 Algorithms

5.2.1 KNN Algorithm 23

5.2.2 Decision Tree Algorithm 23

5.2.3 Naïve Bayes Algorithm 23

5.2.4 Random Forest Algorithm 24

5.2.5 Logistic Regression Algorithm 24

**6.** **SYSTEM IMPLEMENTATION** 25

6.1 Server Side Coding 26

6.2 Client Side Coding 44

**7.** **PERFORMANCE ANALYSIS** 53

7.1 Function Testing 54

7.2 Integration Testing 54

**8.** **CONCLUSION**  55

8.1 Results And Discussion 56

8.2 Future Enhancements 56

**APPENDICES**

A.1 Sample Screens 57

**REFERENCES** 59

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIG NO.** | **FIGURE DESCRIPTION** | **PAGE NO.** |
| 4.1 | ER Diagram | 14 |
| 4.2 | Data Flow Diagram | 16 |
| 4.3 | UML Diagrams | 18 |
| 5.1 | System Architecture | 21 |
| A.1.1 | Crime data Collection | 57 |
| A.1.2  A.1.3  A.1.4 | Random Forest Algorithm  Comparison Graph  Hypertuning | 57  58  58 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### LIST OF ABBREVATIONS

|  |  |  |
| --- | --- | --- |
| **S. NO.** | **ABBREVATION** | **EXPANSION** |
| 1 | KNN | K-Nearest Neighbour |
| 2 | UML | Unified Modelling Language |
| 3 | ER | Entity Relationship |
| 4 | OMT | Object Modelling Technique |
|  |  |  |
| 6 | OMG | Object Management Group |
|  |  |  |
| 7 | ML | Machine Learning |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION**

### 

### 1.INTRODUCTION

* 1. **OVERVIEW**

In present situation hoodlums are getting mechanically complex in carrying out wrongdoing and one test looked by knowledge and law requirement offices is trouble in breaking down enormous volume of information associated with wrongdoing and psychological militant exercises along these lines offices need to realize procedure to get criminal and stay ahead in the interminable race between the crooks and the law implementation. So suitable field need to picked to perform wrongdoing examination and as information mining alludes to extricating or mining information from a lot of information, information mining is utilized here on high volume wrongdoing dataset and information picked up from information mining approaches is helpful and bolster police powers. San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. From Sunset to SOMA, and Marina to Excelsior, this project analysis 12 years of crime reports from across all of San Francisco's neighbourhoods to create a model that predicts the category of crime that occurred, given time and location.

* 1. **PROBLEM DEFINITION**

To perform wrongdoing investigation proper information mining approach should be picked and as bunching is a methodology of information mining which bunches a lot of items so that object in a similar gathering are more comparative than those in different gatherings and included different calculations that vary essentially in their thought of what establishes a group and how to effectively discover them. In this paper k implies grouping strategy of information mining used to separate helpful data from the high-volume wrongdoing dataset and to decipher the information which help police in distinguish and break down wrongdoing examples to lessen further events of comparative rate and give data to decrease the wrongdoing. In this paper k mean grouping is executed utilizing open source information mining apparatus which are investigative devices utilized for breaking down information .Among the accessible open source information mining suite, for example, R, Tanagra ,WEKA ,KNIME ,ORANGE ,Rapid miner.k implies bunching is finished with the assistance of quick excavator device which is an open source factual and information mining bundle written in Java with adaptable information mining bolster alternatives. Likewise, for wrongdoing examination dataset utilized is Crime dataset an offense recorded by the police in England and Wales by offense and police power territory from 1990 to 2011-12. In this paper manslaughter which is wrongdoing carried out by human by executing another human is being broke down.

**CHAPTER 2**

**LITERATURE SURVEY**

### 2.LITERATURE SURVEY

[1] De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data mining approaches to criminal carrer analysis ,”in Proceedings of the Sixth International Conference on Data Mining (ICDM”06) ,Pp. 171-177

[2] Nazlena Mohamad Ali1, Masnizah Mohd2, Hyowon Lee3, Alan F. Smeaton3, Fabio Crestani4 and Shahrul Azman Mohd Noah2 ,2010 Visual Interactive Malaysia Crime News Retrieval System

[3] Malathi.A 1 ,Dr.S.Santhosh Baboo 2 and Anbarasi . A 31 Assistant professor ,Department of Computer Science ,Govt Arts College ,Coimbatore , India . 2 Readers , Department of Computer science , D.G. Vaishnav Collge ,Chennai , India , 2011 An intelligent Analysis of a city Crime Data Using Data Mining

[4] Kadhim B.Swadi al-Janabi . Department of Computer Science . Faculty of Mathematics and Computer Science .University of Kufa/Iraq , 2011 A Proposed Framework for Analyzing Crime DataSet using Decision Tree and Simple K-means Mining Algorithms.

[5] Aravindan Mahendiran, Michael Shuffett, Sathappan Muthiah, Rimy Malla, Gaoqiang Zhang,2011 Forecasting Crime Incidents using Cluster Analysis and Bayesian Belief Networks

[6] Sutapat Thiprungsri,2012 Cluster Analysis for Anomaly Detection in Accounting Data : An Audit Approach1, In the future, this research can also be further improved by analyzing the sensitivities of different load forecasting methods with the load data tampered with cyber attacks.

[7]Akanksha Upadhyaya, Dr. Vinod Shokeen, Dr. Garima Srivastava, 2018, Analysis of Counterfeit Currency Detection Techniques for Classification Model, Counterfeit currency is one of the threats which creates vice to nation’s economy and hence impacts the growth worldwide. Producing forge currency or fabricating fake features in the currency considered to be a crime. Over the past few years many researchers have proposed various techniques to identify and detect forged currency. They used techniques like fake currency, classification model, statitical techniques, machine learning, Logistic regression etc. The effort is also made to analyze and compare the prediction and classification statistical technique i.e. logistic regression and LDA.

[8]Shruti Ranjan, Prayati Garhwal, Anupama Bhan, Monika Arora, Anu Mehra, 2018, Framework For Image Forgery Detection And Classification Using Machine Learning. They used algorithms like Artificial Neural Networks; GLCM features; Graphical User Interface; Machine Learning ; Support Vector Machine. Future work will continue to refine the methodology so that there are lesser loop-holes in the analysis and will hopefully come up with a better method in future. The advancement in the field of science and technology, the introduction of various advance images editing tools are also surging up.

[9]Guiming Zhanga , Jihui Wanga, Aiqing Nib , Haixiao Hub , Anxin Dingb , Shuxin Lib, 2018, Process-induced deformation of L-shaped variable-stiffness composite structures during cure, They used techniques like Composite materials; L-shape; Variable stiffness; Process-induced deformation; Spring-in. It is the parameterized L-shaped parts were modeled to investigate the effects of the radius of curvature, the angle, the laminate thickness and the length of flange on spring-in angle of the parts. In this paper the parameterized investigation on the process-induced spring-in of L-shaped variable-stiffness composites was presented.

[10]Mingjian Cui, Jianhui Wang, Meng Yue, 2018, Machine Learning Based Anomaly Detection for Load Forecasting Under Cyberattacks. They used techniques like Anomaly detection, cyberattack, dynamic programming, load forecasting, machine learning. Accurate load forecasting can make both economic and reliability benefits for power system operators. However, the cyberattack on load forecasting may mislead operators to make unsuitable operational decisions for the electricity delivery. The predicted load data is first used to reconstruct the benchmark and scaling data by using the k-means clustering. The naive Bayes classification is then used to determine the specific attack template.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

Extraction of crime patterns by analysis of available crime and criminal data. Prediction of crime based on spatial distribution of existing data and anticipation of crime rate using different data mining techniques by actualizing bunching calculation on wrongdoing dataset utilizing quick excavator instrument and here we do wrongdoing investigation by thinking about wrongdoing manslaughter and plotting it regarding year and got into end that murder is diminishing from 1990 to 2011 .From the grouped outcomes it is anything but difficult to distinguish wrongdoing pattern over years and can be utilized to structure insurance techniques for future.

DISADVANTAGES:

• Visual and intuitive criminal and intelligence investigation techniques can

be developed for crime pattern

• Also, we can perform analysis on various dataset such as enterprise survey dataset, poverty dataset, aid effectiveness dataset, etc.

**3.2 PROPOSED SYSTEM**

In this we accept that wrongdoing information mining has a promising future for expanding the adequacy and proficiency of criminal and knowledge examination. Visual and instinctive lawbreaker and knowledge examination methods can be created for wrongdoing design. As we have applied bunching procedure of information digging for wrongdoing investigation, we can likewise perform different strategies of information mining, for example, arrangement. Additionally, we can perform investigation on different dataset, for example, undertaking review dataset, neediness dataset, help viability dataset, and so forth.

ADVANTAGES:

In this we will improve the efficiency, visualization and try to detect the wrong information taken while compared the previous one and collect more dataset.

**3.3 FEASIBILITY STUDY**

1. Helps websites or network services remain accessible in the event of outage.
2. Obtains the IP address or other required details and resolves the query, by returning the DNS record to the client.
3. Replication offers disaster recovery and preparedness capabilities in Windows.
4. Storage Replica may allow you to decommission existing file replication systems such as DFS Replication that were pressed into duty as low-end disaster recovery solutions.

**AREAS OF FEASIBILITY**

**3.3.1 SOCIAL FEASIBILITY:**

We hope to raise people's awareness of dangerous places in different eras by providing an information mining method for determining the most criminal hotspots and discovering the type, place, and time of committed wrongdoings. As a result, our proposed arrangement could allow people to avoid certain areas at particular times while also saving lives.

**3.3.2 ECONOMIC FEASIBILITY:**

Crime is a common societal problem that has an effect on people's personal happiness and economic growth. It is regarded as a vital factor in deciding whether or not people should migrate to another city and what places should be avoided while travelling. With the rise of violations, law enforcement agencies are requesting more advanced regional data systems and modern information mining methods to help them better investigate corruption and protect their organisations.

**3.3.3 OPERATIONAL FEASIBILITY:**

* It will be highly useful for the people so the operational system is feasible and it is also way more better than the existing system

**3.3.4 SCHEDULE FEASIBILITY:**

* Based on the designed timeline chart the proposed system only requires 2-3 months for developing it without any delay.

**3.3.5 TECHNICAL FEASIBILITY:**

**PYTHON:**

* Since this is an artificial intelligence project using python gives us the advantage for various inbuilt libraries like TensorFlow, Scikit-Learn, and NumPy.
* Python is an object-oriented programming language that makes coding easier and less prone to syntax errors.
* Python being platform-independent will help in future up-gradation to deploy as software.

**KNN:**

* K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

## 

## 3.4 HARDWARE ENVIRONMENT

• Processor - I3

• RAM - 4 GB (min)

• Hard Disk - 500 GB

• Key Board - Standard Windows Keyboard

• Mouse - Two or Three Button Mouse

• Monitor - LCD

**3.5 SOFTWARE ENVIRONMENT:**

Operating System - Unix/Linux

Coding Language - Python >= 3.8.0

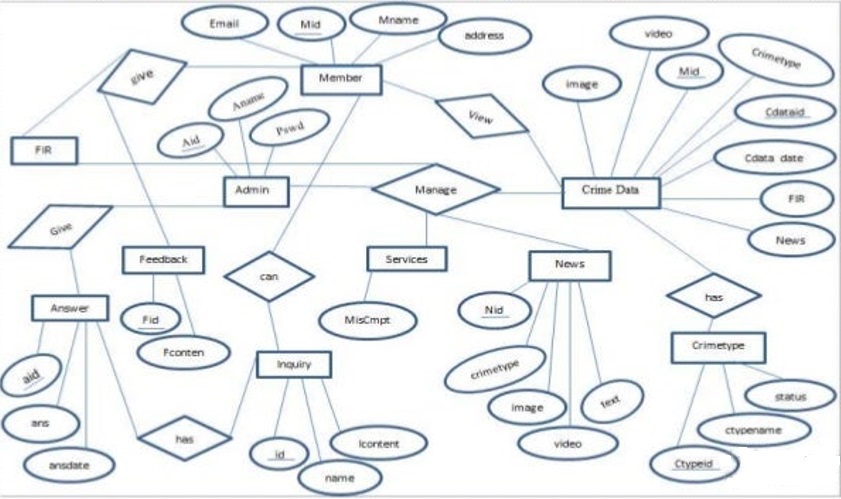
* Operating system : Windows 7/8/10(64bit) / Unix/ Linux
* Additional : Anaconda Navigator, Jupyter Notebook

**CHAPTER 4**

**SYSTEM DESIGN**

**4.SYSTEM DESIGN**

**4.1 ER DIAGRAM**



**Fig 4.1 ER DIAGRAM**

An Entity Relationship (ER) Diagram is a type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research. Also known as ERDs or ER Models, they use a defined set of symbols such as rectangles, diamonds, ovals and connecting lines to depict the interconnectedness of entities, relationships and their attributes. Entity : A definable thing—such as a person, object, concept or event—that can have data stored about it. Think of entities as nouns. Examples: a customer, student, car or product. Typically shown as a rectangle.

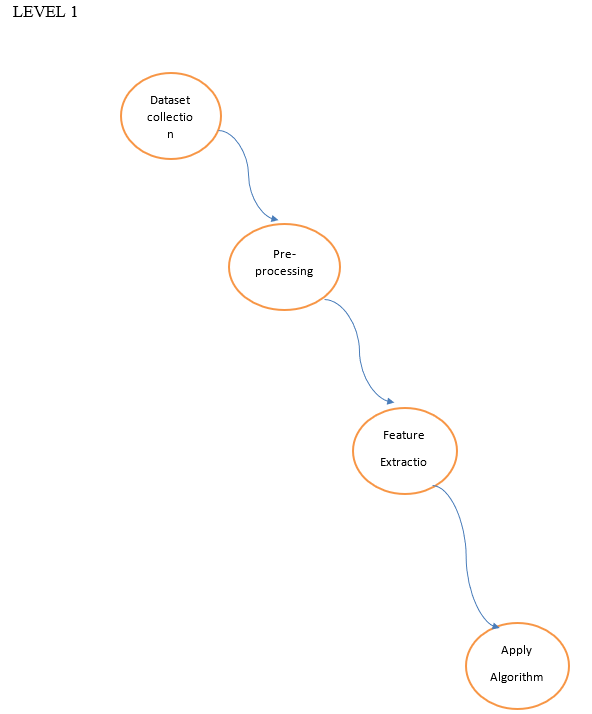
Entity type: A group of definable things, such as students or athletes, whereas the entity would be the specific student or athlete. Other examples: customers, cars or products. Entity set: Same as an entity type, but defined at a particular point in time, such as students enrolled in a class on the first day. Other examples: Customers who purchased last month, cars currently registered in Florida. A related term is instance, in which the specific person or car would be an instance of the entity set.

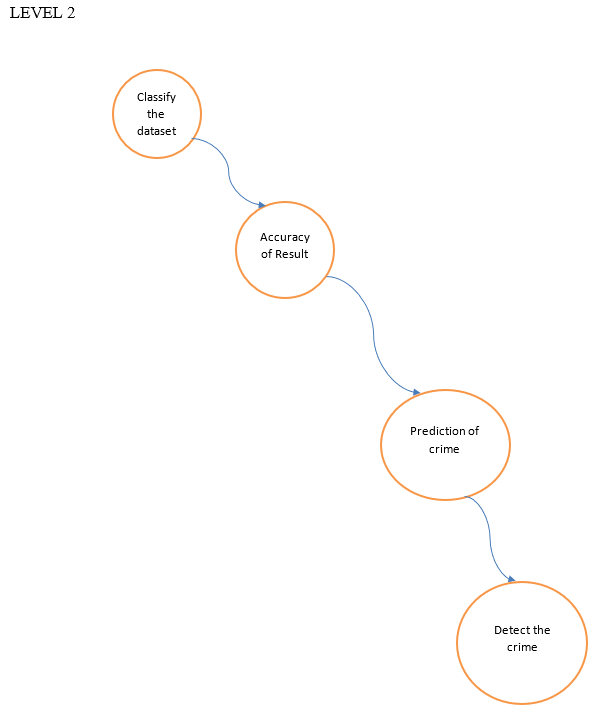
Entity categories: Entities are categorized as strong, weak or associative. A strong entity can be defined solely by its own attributes, while a weak entity cannot. An associative entity associates entities (or elements) within an entity set.

Entity keys: Refers to an attribute that uniquely defines an entity in an entity set. Entity keys can be super, candidate or primary. Super key: A set of attributes (one or more) that together define an entity in an entity set. Candidate key: A minimal super key, meaning it has the least possible number of attributes to still be a super key. An entity set may have more than one candidate key. Primary key: A candidate key chosen by the database designer to uniquely identify the entity set. Foreign key: Identifies the relationship between entities. Relationship : How entities act upon each other or are associated with each other. Think of relationships as verbs. For example, the named student might register for a course. The two entities would be the student and the course, and the relationship depicted is the act of enrolling, connecting the two entities in that way. Relationships are typically shown as diamonds or labels directly on the connecting lines. Recursive relationship: The same entity participates more than once in the relationship. Attribute : A property or characteristic of an entity. Often shown as an oval or circle. Descriptive attribute: A property or characteristic of a relationship (versus of an entity.) Attribute categories: Attributes are categorized as simple, composite, derived, as well as single-value or multi-value. Simple: Means the attribute value is atomic and can’t be further divided, such as a phone number. Composite: Sub-attributes spring from an attribute. Derived: Attributed is calculated or otherwise derived from another attribute, such as age from a birthdate.

Multi-value: More than one attribute value is denoted, such as multiple phone numbers for a person. Single-value: Just one attribute value. The types can be combined, such as: simple single-value attributes or composite multi-value attributes.

**4.2 DATA FLOW DIAGRAM**



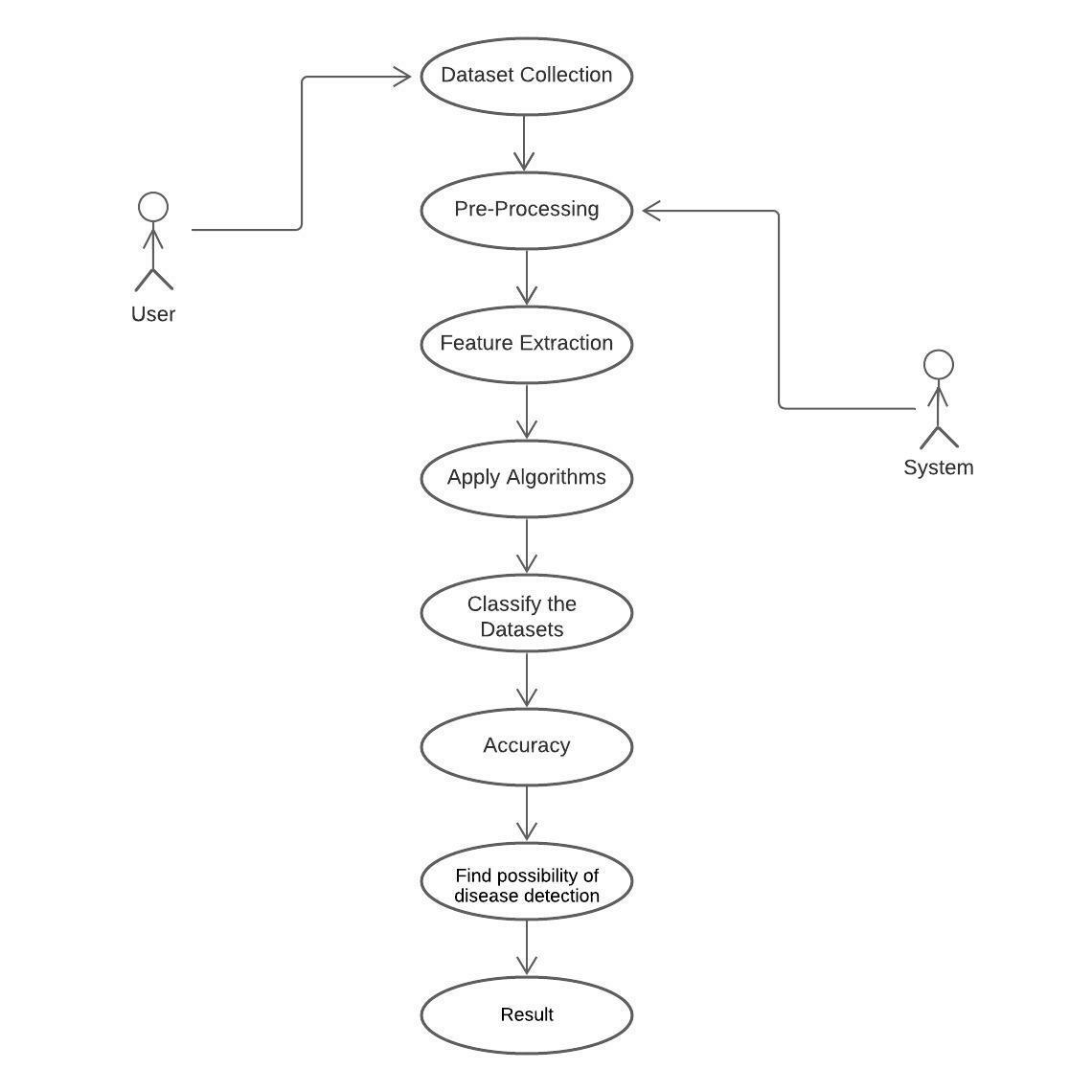


**Fig 4.2 DATA FLOW DIAGRAM**

A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination. Data flowcharts can range from simple, even hand-drawn process overviews, to in-depth, multi-level DFDs that dig progressively deeper into how the data is handled.

**4.3 UML DIAGRAMS**

**4.3.1 USECASE DIAGRAM**

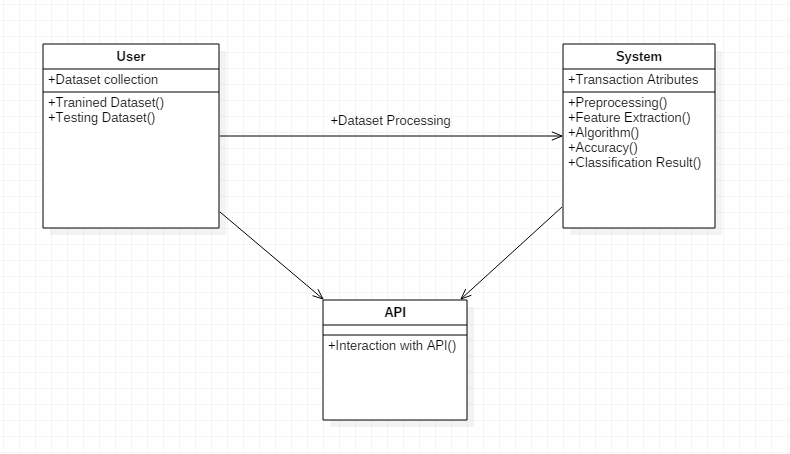
****

**Fig 4.3.1 USECASE DIAGRAM**

Use case diagrams are typically developed in the early stage of development and people often apply use case modeling for the following purposes:

Specify the context of a system, Capture the requirements of a system, Validate a systems architecture, Drive implementation and generate test cases, Developed by analysts together with domain experts.

**4.3.2 CLASS DIAGRAM**

****

**Fig 4.3.2 CLASS DIAGRAM**

The class diagram is the main building block of object-oriented modeling. It is used for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main elements, interactions in the application, and the classes to be programmed. In the diagram, classes are represented with boxes that contain three compartments: The top compartment contains the name of the class. It is printed in bold and centered, and the first letter is capitalized. The middle compartment contains the attributes of the class. They are left-aligned and the first letter is lowercase. The bottom compartment contains the operations the class can execute. They are also left-aligned and the first letter is lowercase.

**CHAPTER 5**

**SYSTEM ARCHITECTURE**

**5.SYSTEM ARCHITECTURE**

****

**Fig 5 SYSTEM ARCHITECTURE**

First, we take wrongdoing dataset, Channel dataset as indicated by prerequisite and make new dataset which has ascribe as per investigation to be done. Open quick digger device and read exceed expectations record of wrongdoing dataset and apply "Supplant Missing worth administrator" on it and execute activity. Perform "Standardize administrator" on resultant dataset and execute activity. Perform k implies grouping on resultant dataset framed after standardization and execute activity. From plot perspective on result plot information among wrongdoings and get required group. Examination should be possible on bunch framed.

# 5.1 MODULE DESIGN SPECIFICATION

There are four modules in our project . The list of modules are as follows:

* DATA COLLECTION
* DATA PRE-PROCESSING
* FEATURE EXTRATION
* EVALUATION MODEL

**5.1.1 DATA COLLECTION**

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand. Data for which you already know the target answer is called labelled data.

**5.1.2 DATA PRE-PROCESSING**

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

● Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

● Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

● Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

**5.1.3. FEATURE EXTRACTION:**

Features selection is done which can be used to build the model. The attributes used for feature selection are Block, Location, District, Community area, like dates, crime description, day of week, X, Y, Location.

**5.1.4. EVALUATION MODEL:**

In this module we implement the algorithm which one give best accuracy value. Based on that applying the algorithm it should calculate and give the accuracy. Here we are applying the algorithm and get the result on the form of graphical format.

# 5.2 ALGORITHMS:

**5.2.1 K-NEAREST NEIGHBOR(KNN):**

K-Nearest Neighbor(KNN) is a supervised algorithm in machine learning that is used for classification and regression analysis. This algorithm assigns the new data based on how close or how similar the data is to the points in training data.

Here, ‘K’ represents the number of neighbors that are considered to classify the new data point. KNN is called a lazy learning algorithm because it uses all the data during training for the classification of a new point.

In other words, it doesn’t learn from training data rather it stores data and when new data is introduced it classifies that new point in the course of training.

**5.2.2DECISION TREE ALGORITHM**

A Decision Tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question, and the leaves represent the actual output or class label. They are used in non-linear decision making with a simple linear decision surface.

The Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can solve regression and classification problems.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

**5.2.3 NAÏVE BAYES ALGORITHM**

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis. These classifiers are widely used for machine learning because they are simple to implement.

**5.2.4 RANDOM FOREST**

“Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**5.2.5 LOGISTIC REGRESSION**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

**CHAPTER 6**

**SYSTEM IMPLEMENTATION**

**6.SYSTEM IMPLEMENTATION**

**SERVER SIDE CODING**

{

"cells": [

{

"cell\_type": "code",

"execution\_count": 2,

"metadata": {

"id": "Vyz62OTzP7tS"

},

"outputs": [],

"source": [

"import numpy as np\n",

"import pandas as pd\n",

"import matplotlib.pyplot as plt\n",

"import seaborn as sns\n",

"from sklearn.model\_selection import train\_test\_split\n",

"from sklearn.preprocessing import StandardScaler\n",

"from sklearn.naive\_bayes import GaussianNB\n",

"from sklearn import tree\n",

"from sklearn.ensemble import RandomForestClassifier\n",

"from sklearn.linear\_model import LogisticRegression\n",

"from sklearn import metrics\n",

"from sklearn.metrics import classification\_report\n",

"from sklearn import preprocessing\n",

"from sklearn.metrics import plot\_confusion\_matrix\n",

"from mlxtend.plotting import plot\_decision\_regions\n",

"%matplotlib inline\n",

"sns.set\_style(\"darkgrid\")\n",

"import warnings\n",

"warnings.filterwarnings('ignore')\n",

"warnings.simplefilter('ignore')"

]

},

{

"cell\_type": "code",

"execution\_count": 3,

"metadata": {

"id": "ZZ6\_F2dFP7tS"

},

"outputs": [],

"source": [

"crimes\_data = pd.read\_csv('crimedata.csv')"

]

},

{

"cell\_type": "code",

"execution\_count": 4,

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 482

},

"id": "z\_g-SyLXloGc",

"outputId": "5db275f2-d3cc-4556-c916-2beeb3b1132d"

},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>ID</th>\n",

" <th>Case Number</th>\n",

" <th>Date</th>\n",

" <th>Block</th>\n",

" <th>IUCR</th>\n",

" <th>Primary Type</th>\n",

" <th>Description</th>\n",

" <th>Location Description</th>\n",

" <th>Arrest</th>\n",

" <th>Domestic</th>\n",

" <th>...</th>\n",

" <th>Ward</th>\n",

" <th>Community Area</th>\n",

" <th>FBI Code</th>\n",

" <th>X Coordinate</th>\n",

" <th>Y Coordinate</th>\n",

" <th>Year</th>\n",

" <th>Updated On</th>\n",

" <th>Latitude</th>\n",

" <th>Longitude</th>\n",

" <th>Location</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>0</th>\n",

" <td>12583965</td>\n",

" <td>JF102303</td>\n",

" <td>01/03/2022 01:30:00 PM</td>\n",

" <td>003XX E 76TH ST</td>\n",

" <td>1320</td>\n",

" <td>CRIMINAL DAMAGE</td>\n",

" <td>TO VEHICLE</td>\n",

" <td>STREET</td>\n",

" <td>False</td>\n",

" <td>True</td>\n",

" <td>...</td>\n",

" <td>6.0</td>\n",

" <td>69</td>\n",

" <td>14</td>\n",

" <td>1179942.0</td>\n",

" <td>1854699.0</td>\n",

" <td>2022</td>\n",

" <td>01/10/2022 03:51:23 PM</td>\n",

" <td>41.756565</td>\n",

" <td>-87.616111</td>\n",

" <td>(41.756565176, -87.616111074)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>1</th>\n",

" <td>12583949</td>\n",

" <td>JF102446</td>\n",

" <td>01/03/2022 05:15:00 PM</td>\n",

" <td>021XX W 72ND PL</td>\n",

" <td>1365</td>\n",

" <td>CRIMINAL TRESPASS</td>\n",

" <td>TO RESIDENCE</td>\n",

" <td>APARTMENT</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>17.0</td>\n",

" <td>67</td>\n",

" <td>26</td>\n",

" <td>1163445.0</td>\n",

" <td>1856564.0</td>\n",

" <td>2022</td>\n",

" <td>01/10/2022 03:51:23 PM</td>\n",

" <td>41.762045</td>\n",

" <td>-87.676517</td>\n",

" <td>(41.762044507, -87.676517323)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2</th>\n",

" <td>12583314</td>\n",

" <td>JF101736</td>\n",

" <td>01/03/2022 01:49:00 AM</td>\n",

" <td>024XX S ARCHER AVE</td>\n",

" <td>031A</td>\n",

" <td>ROBBERY</td>\n",

" <td>ARMED - HANDGUN</td>\n",

" <td>PARKING LOT / GARAGE (NON RESIDENTIAL)</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>11.0</td>\n",

" <td>60</td>\n",

" <td>03</td>\n",

" <td>1171749.0</td>\n",

" <td>1887900.0</td>\n",

" <td>2022</td>\n",

" <td>01/10/2022 03:51:23 PM</td>\n",

" <td>41.847856</td>\n",

" <td>-87.645163</td>\n",

" <td>(41.84785572, -87.645162578)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>3</th>\n",

" <td>12584353</td>\n",

" <td>JF103119</td>\n",

" <td>01/03/2022 03:00:00 AM</td>\n",

" <td>075XX S ESSEX AVE</td>\n",

" <td>0810</td>\n",

" <td>THEFT</td>\n",

" <td>OVER $500</td>\n",

" <td>STREET</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>7.0</td>\n",

" <td>43</td>\n",

" <td>06</td>\n",

" <td>1194171.0</td>\n",

" <td>1855517.0</td>\n",

" <td>2022</td>\n",

" <td>01/10/2022 03:51:23 PM</td>\n",

" <td>41.758472</td>\n",

" <td>-87.563939</td>\n",

" <td>(41.758472475, -87.563938551)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>4</th>\n",

" <td>12584669</td>\n",

" <td>JF102858</td>\n",

" <td>01/03/2022 12:53:00 PM</td>\n",

" <td>044XX S TRUMBULL AVE</td>\n",

" <td>0890</td>\n",

" <td>THEFT</td>\n",

" <td>FROM BUILDING</td>\n",

" <td>RESIDENCE</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>14.0</td>\n",

" <td>58</td>\n",

" <td>06</td>\n",

" <td>1154121.0</td>\n",

" <td>1875045.0</td>\n",

" <td>2022</td>\n",

" <td>01/10/2022 03:51:23 PM</td>\n",

" <td>41.812949</td>\n",

" <td>-87.710200</td>\n",

" <td>(41.812949384, -87.710200316)</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"<p>5 rows × 22 columns</p>\n",

"</div>"

],

"text/plain": [

" ID Case Number Date Block IUCR \\\n",

"0 12583965 JF102303 01/03/2022 01:30:00 PM 003XX E 76TH ST 1320 \n",

"1 12583949 JF102446 01/03/2022 05:15:00 PM 021XX W 72ND PL 1365 \n",

"2 12583314 JF101736 01/03/2022 01:49:00 AM 024XX S ARCHER AVE 031A \n",

"3 12584353 JF103119 01/03/2022 03:00:00 AM 075XX S ESSEX AVE 0810 \n",

"4 12584669 JF102858 01/03/2022 12:53:00 PM 044XX S TRUMBULL AVE 0890 \n",

"\n",

" Primary Type Description Location Description \\\n",

"0 CRIMINAL DAMAGE TO VEHICLE STREET \n",

"1 CRIMINAL TRESPASS TO RESIDENCE APARTMENT \n",

"2 ROBBERY ARMED - HANDGUN PARKING LOT / GARAGE (NON RESIDENTIAL) \n",

"3 THEFT OVER $500 STREET \n",

"4 THEFT FROM BUILDING RESIDENCE \n",

"\n",

" Arrest Domestic ... Ward Community Area FBI Code X Coordinate \\\n",

"0 False True ... 6.0 69 14 1179942.0 \n",

"1 False False ... 17.0 67 26 1163445.0 \n",

"2 False False ... 11.0 60 03 1171749.0 \n",

"3 False False ... 7.0 43 06 1194171.0 \n",

"4 False False ... 14.0 58 06 1154121.0 \n",

"\n",

" Y Coordinate Year Updated On Latitude Longitude \\\n",

"0 1854699.0 2022 01/10/2022 03:51:23 PM 41.756565 -87.616111 \n",

"1 1856564.0 2022 01/10/2022 03:51:23 PM 41.762045 -87.676517 \n",

"2 1887900.0 2022 01/10/2022 03:51:23 PM 41.847856 -87.645163 \n",

"3 1855517.0 2022 01/10/2022 03:51:23 PM 41.758472 -87.563939 \n",

"4 1875045.0 2022 01/10/2022 03:51:23 PM 41.812949 -87.710200 \n",

"\n",

" Location \n",

"0 (41.756565176, -87.616111074) \n",

"1 (41.762044507, -87.676517323) \n",

"2 (41.84785572, -87.645162578) \n",

"3 (41.758472475, -87.563938551) \n",

"4 (41.812949384, -87.710200316) \n",

"\n",

"[5 rows x 22 columns]"

]

},

"execution\_count": 4,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"crimes\_data.head()"

]

},

{

"cell\_type": "code",

"execution\_count": 5,

"metadata": {

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 499

},

"id": "vcGeMbTgnBMG",

"outputId": "2310d81e-e0cd-4a57-b07a-50ddf5ae4fe1"

},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>ID</th>\n",

" <th>Case Number</th>\n",

" <th>Date</th>\n",

" <th>Block</th>\n",

" <th>IUCR</th>\n",

" <th>Primary Type</th>\n",

" <th>Description</th>\n",

" <th>Location Description</th>\n",

" <th>Arrest</th>\n",

" <th>Domestic</th>\n",

" <th>...</th>\n",

" <th>Ward</th>\n",

" <th>Community Area</th>\n",

" <th>FBI Code</th>\n",

" <th>X Coordinate</th>\n",

" <th>Y Coordinate</th>\n",

" <th>Year</th>\n",

" <th>Updated On</th>\n",

" <th>Latitude</th>\n",

" <th>Longitude</th>\n",

" <th>Location</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>29981</th>\n",

" <td>12634579</td>\n",

" <td>JF160345</td>\n",

" <td>03/01/2022 12:00:00 PM</td>\n",

" <td>012XX W BYRON ST</td>\n",

" <td>1330</td>\n",

" <td>CRIMINAL TRESPASS</td>\n",

" <td>TO LAND</td>\n",

" <td>APARTMENT</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>44.0</td>\n",

" <td>6</td>\n",

" <td>26</td>\n",

" <td>1167111.0</td>\n",

" <td>1926026.0</td>\n",

" <td>2022</td>\n",

" <td>03/08/2022 03:49:42 PM</td>\n",

" <td>41.952577</td>\n",

" <td>-87.661088</td>\n",

" <td>(41.952576753, -87.661087678)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>29982</th>\n",

" <td>12632233</td>\n",

" <td>JF161640</td>\n",

" <td>03/01/2022 03:00:00 PM</td>\n",

" <td>086XX S COLFAX AVE</td>\n",

" <td>0820</td>\n",

" <td>THEFT</td>\n",

" <td>$500 AND UNDER</td>\n",

" <td>RESIDENCE</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>7.0</td>\n",

" <td>46</td>\n",

" <td>06</td>\n",

" <td>1194993.0</td>\n",

" <td>1848202.0</td>\n",

" <td>2022</td>\n",

" <td>03/08/2022 03:49:42 PM</td>\n",

" <td>41.738379</td>\n",

" <td>-87.561167</td>\n",

" <td>(41.738379322, -87.561166634)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>29983</th>\n",

" <td>12631112</td>\n",

" <td>JF160508</td>\n",

" <td>03/01/2022 08:00:00 PM</td>\n",

" <td>004XX N WOOD ST</td>\n",

" <td>0910</td>\n",

" <td>MOTOR VEHICLE THEFT</td>\n",

" <td>AUTOMOBILE</td>\n",

" <td>STREET</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>27.0</td>\n",

" <td>24</td>\n",

" <td>07</td>\n",

" <td>1164313.0</td>\n",

" <td>1903426.0</td>\n",

" <td>2022</td>\n",

" <td>03/08/2022 03:49:42 PM</td>\n",

" <td>41.890621</td>\n",

" <td>-87.672014</td>\n",

" <td>(41.890620703, -87.672013889)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>29984</th>\n",

" <td>12630796</td>\n",

" <td>JF160104</td>\n",

" <td>03/01/2022 06:20:00 PM</td>\n",

" <td>001XX E 51ST ST</td>\n",

" <td>0560</td>\n",

" <td>ASSAULT</td>\n",

" <td>SIMPLE</td>\n",

" <td>GAS STATION</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>3.0</td>\n",

" <td>40</td>\n",

" <td>08A</td>\n",

" <td>1178181.0</td>\n",

" <td>1871245.0</td>\n",

" <td>2022</td>\n",

" <td>03/08/2022 03:49:42 PM</td>\n",

" <td>41.802009</td>\n",

" <td>-87.622063</td>\n",

" <td>(41.8020092, -87.622063199)</td>\n",

" </tr>\n",

" <tr>\n",

" <th>29985</th>\n",

" <td>12630458</td>\n",

" <td>JF159699</td>\n",

" <td>03/01/2022 12:13:00 PM</td>\n",

" <td>067XX S ELIZABETH ST</td>\n",

" <td>0560</td>\n",

" <td>ASSAULT</td>\n",

" <td>SIMPLE</td>\n",

" <td>CHA APARTMENT</td>\n",

" <td>False</td>\n",

" <td>False</td>\n",

" <td>...</td>\n",

" <td>17.0</td>\n",

" <td>67</td>\n",

" <td>08A</td>\n",

" <td>1169148.0</td>\n",

" <td>1860172.0</td>\n",

" <td>2022</td>\n",

" <td>03/08/2022 03:49:42 PM</td>\n",

" <td>41.771824</td>\n",

" <td>-87.655511</td>\n",

" <td>(41.771823924, -87.655510895)</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"<p>5 rows × 22 columns</p>\n",

"</div>"

],

**CLIENT SIDE CODING: (PYTHON SCRIPT TO VALIDATE)**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.naive\_bayes import GaussianNB

from sklearn import tree

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear\_model import LogisticRegression

from sklearn import metrics

from sklearn.metrics import classification\_report

from sklearn import preprocessing

from sklearn.metrics import plot\_confusion\_matrix

from mlxtend.plotting import plot\_decision\_regions

%matplotlib inline

sns.set\_style("darkgrid")

import warnings

warnings.filterwarnings('ignore')

warnings.simplefilter('ignore')

crimes\_data = pd.read\_csv('crimedata.csv')

crimes\_data.head()

crimes\_data.tail()

crimes\_data.info()

crimes\_data = crimes\_data.dropna()

crimes\_data = crimes\_data.reset\_index(drop = True)

crimes\_data

crimes\_data.columns = crimes\_data.columns.str.strip()

crimes\_data.columns = crimes\_data.columns.str.replace(',', '')

crimes\_data.columns = crimes\_data.columns.str.replace(' ', '\_')

crimes\_data.columns = crimes\_data.columns.str.lower()

crimes\_data[crimes\_data.duplicated(keep=False)]

crimes\_data.drop(['id','case\_number','location'],axis=1,inplace=True)

crimes\_data

crimes\_data.isnull().sum()

crimes\_data.info()

crimes\_data.date = pd.to\_datetime(crimes\_data.date)

crimes\_data['day\_of\_week'] = crimes\_data.date.dt.day\_name()

crimes\_data['month'] = crimes\_data.date.dt.month\_name()

crimes\_data['time'] = crimes\_data.date.dt.hour

primary\_type\_map = {

('BURGLARY','MOTOR VEHICLE THEFT','THEFT','ROBBERY') : 'THEFT',

('BATTERY','ASSAULT','NON-CRIMINAL','NON-CRIMINAL (SUBJECT SPECIFIED)') : 'NON-CRIMINAL\_ASSAULT',

('CRIM SEXUAL ASSAULT','SEX OFFENSE','STALKING','PROSTITUTION') : 'SEXUAL\_OFFENSE',

('WEAPONS VIOLATION','CONCEALED CARRY LICENSE VIOLATION') : 'WEAPONS\_OFFENSE',

('HOMICIDE','CRIMINAL DAMAGE','DECEPTIVE PRACTICE','CRIMINAL TRESPASS') : 'CRIMINAL\_OFFENSE',

('KIDNAPPING','HUMAN TRAFFICKING','OFFENSE INVOLVING CHILDREN') : 'HUMAN\_TRAFFICKING\_OFFENSE',

('NARCOTICS','OTHER NARCOTIC VIOLATION') : 'NARCOTIC\_OFFENSE',

('OTHER OFFENSE','ARSON','GAMBLING','PUBLIC PEACE VIOLATION','INTIMIDATION','INTERFERENCE WITH PUBLIC OFFICER','LIQUOR LAW VIOLATION','OBSCENITY','PUBLIC INDECENCY') : 'OTHER\_OFFENSE'

}

primary\_type\_mapping = {}

for keys, values in primary\_type\_map.items():

for key in keys:

primary\_type\_mapping[key] = values

crimes\_data['primary\_type\_grouped'] = crimes\_data.primary\_type.map(primary\_type\_mapping)

zone\_mapping = {

'N' : 'North',

'S' : 'South',

'E' : 'East',

'W' : 'West'

}

crimes\_data['zone'] = crimes\_data.block.str.split(" ", n = 2, expand = True)[1].map(zone\_mapping)

season\_map = {

('March','April','May') : 'Spring',

('June','July','August') : 'Summer',

('September','October','November') : 'Fall',

('December','January','February') : 'Winter'

}

season\_mapping = {}

for keys, values in season\_map.items():

for key in keys:

season\_mapping[key] = values

crimes\_data['season'] = crimes\_data.month.map(season\_mapping)

crimes\_data.arrest = crimes\_data.arrest.astype(int)

crimes\_data.domestic = crimes\_data.domestic.astype(int)

crimes\_data = crimes\_data.dropna()

crimes\_data = crimes\_data.reset\_index(drop = True)

crimes\_data\_2018 = crimes\_data[crimes\_data.year == 2018]

crimes\_data\_2019 = crimes\_data[crimes\_data.year == 2019]

crimes\_data\_2020 = crimes\_data[crimes\_data.year == 2020]

plt.figure(figsize=(15,5))

zone\_plot = sns.countplot(data=crimes\_data,x='day\_of\_week',hue='year',order=crimes\_data.day\_of\_week.value\_counts().index,palette='Set2')

plt.figure(figsize=(20,5))

zone\_plot = sns.countplot(data=crimes\_data,x='month',hue='year',order=crimes\_data.month.value\_counts().index,palette='Set2')

zone\_plot = sns.countplot(data=crimes\_data,x='zone',hue='year',order=crimes\_data.zone.value\_counts().index,palette='Set2')

zone\_plot = sns.countplot(data=crimes\_data,x='season',hue='year',palette='Set2')

arrest\_plot = sns.countplot(data=crimes\_data,x='year',hue='arrest',palette='Set2')

new\_crimes\_data = crimes\_data.loc[(crimes\_data['x\_coordinate']!=0)]

sns.lmplot('x\_coordinate',

'y\_coordinate',

data=new\_crimes\_data[:],

fit\_reg=False,

hue="primary\_type\_grouped",

palette='Dark2',

height=12,

ci=2,

scatter\_kws={"marker": "o",

"s": 10})

ax = plt.gca()

ax.set\_title("Crimes by Type of crime")

crimes\_data.year = pd.Categorical(crimes\_data.year)

crimes\_data.time = pd.Categorical(crimes\_data.time)

crimes\_data.domestic = pd.Categorical(crimes\_data.domestic)

crimes\_data.arrest = pd.Categorical(crimes\_data.arrest)

crimes\_data.beat = pd.Categorical(crimes\_data.beat)

crimes\_data.district = pd.Categorical(crimes\_data.district)

crimes\_data.ward = pd.Categorical(crimes\_data.ward)

crimes\_data.community\_area = pd.Categorical(crimes\_data.community\_area)

crimes\_data\_prediction = crimes\_data.drop(['date','block','iucr','primary\_type','description','location\_description','fbi\_code','updated\_on','x\_coordinate','y\_coordinate'],axis=1)

crimes\_data\_prediction.head()

crimes\_data\_prediction = pd.get\_dummies(crimes\_data\_prediction,drop\_first=True)

crimes\_data\_prediction.head()

X\_train, X\_test, y\_train, y\_test = train\_test\_split(crimes\_data\_prediction.drop(['arrest\_1'],axis=1),crimes\_data\_prediction['arrest\_1'], test\_size=0.25, random\_state=42)

classifier = GaussianNB()

classifier.fit(X\_train, y\_train)

y\_pred = classifier.predict(X\_test)

conf\_matrix = metrics.confusion\_matrix(y\_test, y\_pred)

print(conf\_matrix)

sns.heatmap(conf\_matrix, annot = True, fmt = ".3f", square = True, cmap = plt.cm.Blues)

plt.ylabel('Actual')

plt.xlabel('Predicted')

plt.title('Confusion matrix')

plt.tight\_layout()

print('Error = ',1 - metrics.accuracy\_score(y\_test, y\_pred))

print('Precision = ',metrics.precision\_score(y\_test, y\_pred,))

print('Recall = ',metrics.recall\_score(y\_test, y\_pred))

print('F-1 Score = ',metrics.f1\_score(y\_test, y\_pred))

print('Classification Report\n',metrics.classification\_report(y\_test, y\_pred))

model4 = metrics.accuracy\_score(y\_test, y\_pred)

model = tree.DecisionTreeClassifier(criterion = "entropy", random\_state = 42)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

conf\_matrix = metrics.confusion\_matrix(y\_test, y\_pred)

print(conf\_matrix)

sns.heatmap(conf\_matrix, annot = True, fmt = ".3f", square = True, cmap = plt.cm.Blues)

plt.ylabel('Actual')

plt.xlabel('Predicted')

plt.title('Confusion matrix')

plt.tight\_layout()

print('Error = ',1 - metrics.accuracy\_score(y\_test, y\_pred))

print('Precision = ',metrics.precision\_score(y\_test, y\_pred,))

print('Recall = ',metrics.recall\_score(y\_test, y\_pred))

print('F-1 Score = ',metrics.f1\_score(y\_test, y\_pred))

print('Classification Report\n',metrics.classification\_report(y\_test, y\_pred))

model3 = metrics.accuracy\_score(y\_test, y\_pred)

model = RandomForestClassifier(n\_estimators = 10,criterion='entropy',random\_state=42)

model.fit(X\_train,y\_train)

y\_pred = model.predict(X\_test)

conf\_matrix = metrics.confusion\_matrix(y\_test, y\_pred)

print(conf\_matrix)

sns.heatmap(conf\_matrix, annot = True, fmt = ".3f", square = True, cmap = plt.cm.Blues)

plt.ylabel('Actual')

**p**lt.xlabel('Predicted')

plt.title('Confusion matrix')

plt.tight\_layout()

print('Accuracy = ',metrics.accuracy\_score(y\_test, y\_pred))

print('Error = ',1 - metrics.accuracy\_score(y\_test, y\_pred))

print('Precision = ',metrics.precision\_score(y\_test, y\_pred,))

print('Recall = ',metrics.recall\_score(y\_test, y\_pred))

print('F-1 Score = ',metrics.f1\_score(y\_test, y\_pred))

print('Classification Report\n',metrics.classification\_report(y\_test, y\_pred))

classifier = LogisticRegression(random\_state=42)

classifier.fit(X\_train,y\_train)

y\_pred = classifier.predict(X\_test)

conf\_matrix = metrics.confusion\_matrix(y\_test, y\_pred)

print(conf\_matrix)

sns.heatmap(conf\_matrix, annot = True, fmt = ".3f", square = True, cmap = plt.cm.Blues)

plt.ylabel('Actual')

plt.xlabel('Predicted')

plt.title('Confusion matrix')

plt.tight\_layout()

print('Error = ',1 - metrics.accuracy\_score(y\_test, y\_pred))

print('Precision = ',metrics.precision\_score(y\_test, y\_pred,))

print('Recall = ',metrics.recall\_score(y\_test, y\_pred))

print('F-1 Score = ',metrics.f1\_score(y\_test, y\_pred))

print('Classification Report\n',metrics.classification\_report(y\_test, y\_pred))

model2 =metrics.accuracy\_score(y\_test, y\_pred)

feature\_scaler = StandardScaler()

X\_train = feature\_scaler.fit\_transform(X\_train)

X\_test = feature\_scaler.transform(X\_test)

classifier = RandomForestClassifier(n\_estimators=10, random\_state=42)

from sklearn.model\_selection import cross\_val\_score

all\_accuracies = cross\_val\_score(estimator=classifier, X=X\_train, y=y\_train, cv=5)

model1 =all\_accuracies.max()

import matplotlib.pyplot as plt; plt.rcdefaults()

import numpy as np

import matplotlib.pyplot as plt

objects = ('Random Forest','Logistic Regression','Decision Tree','Naive Bayes')

y\_pos = np.arange(len(objects))

performance = [model1,model2,model3,model4]

plt.bar(y\_pos, performance, align='center', alpha=0.5)

plt.xticks(y\_pos, objects)

plt.ylabel('Accuracy Score')

plt.title("Performance Comparision")

plt.show()

**CHAPTER 7**

**PERFORMANCE ANALYSIS**

**7.PERFORMANCE ANALYSIS**

**Functional Testing**

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

● Functions: Identified functions must be exercised.

● Output: Identified classes of software outputs must be exercised.

● Systems/Procedures: system should work properly

**Integration Testing**

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

# CHAPTER 8

**CONCLUSION**

**8.CONCLUSION**

**8.1 Results and discussion:**

First, we take wrongdoing dataset, Channel dataset as indicated by prerequisite and make new dataset which has ascribe as per investigation to be done. Open quick digger device and read exceed expectations record of wrongdoing dataset and apply "Supplant Missing worth administrator" on it and execute activity. Perform "Standardize administrator" on resultant dataset and execute activity. Perform k implies grouping on resultant dataset framed after standardization and execute activity. From plot perspective on result plot information among wrongdoings and get required group. Examination should be possible on bunch framed.

**8.2 Future Enhancements:**

In the present study, ML models with machine learning algorithms (ensemble and simile), SVm, SVM-Random forest, SVM-stacking and Naive Bias, were designed and were implemented. Each predetermined factor was feed into a violent crime training dataset (murder, rape, robbery, etc.). We discovered a major conclusion after successfully training and validating simodels.

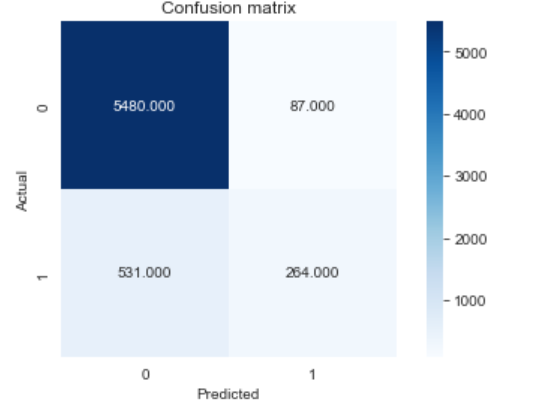
In the future, the whole model will be converted into an opensource library and connected to the crime site, allowing it to function at the highest level of expertise Function on a framework where the threshold for class crime rates can be set. Instead of a limited crime, the highest crime rate can be measured. A small grouping is used to assess the success of all of the criminal figures examined in the proposal. A Calculation can be measured on a sufficiently broad scale of local or cloud-based crime with heavy datasets will pay forecast of multi-label charging, expand more possibilities, and realistically increase our research

**APPENDICES**

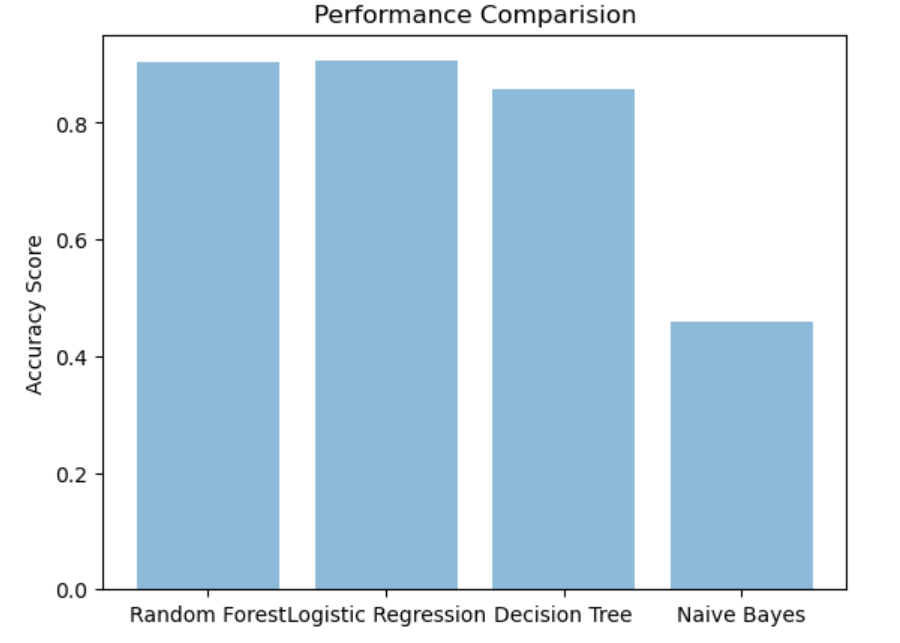
**A.1 SAMPLE SCREENS**

****

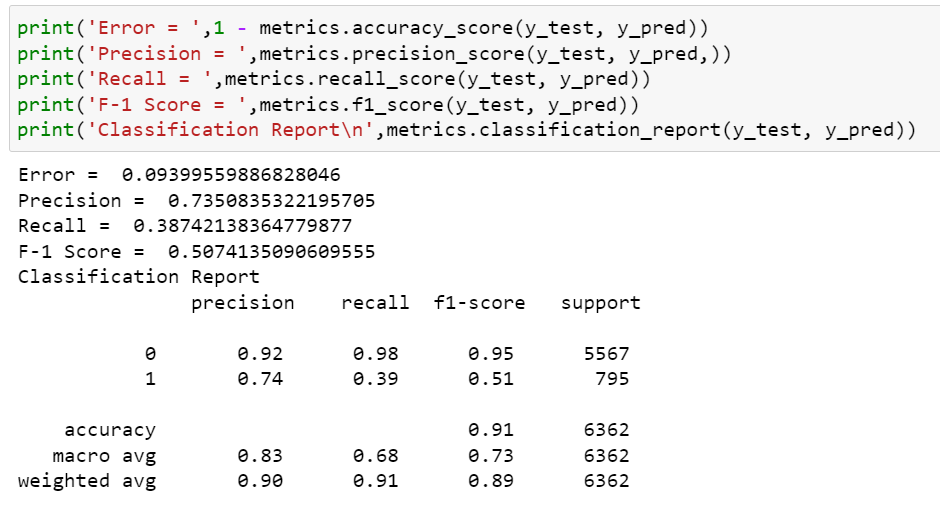
**Fig A.1.1 CRIME DATA COLLECTION**



**Fig A.1.2 RANDOM FOREST CONFUSION MATRIX**



**Fig A.1.3 COMPARISION GRAPH**



**Fig A.1.4 HYPERTUNING**

# REFERENCES

[1] M. Cahill and G. Mulligan, ‘‘Using geographically weighted regression to explore local crime patterns,’’ Social Sci. Comput. Rev., vol. 25, no. 2, pp. 174–193, May 2007, doi: 10.1177/0894439307298925.

[2] J. M. Caplan, L. W. Kennedy, and J. Miller, ‘‘Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting,’’ Justice Quart., vol. 28, no. 2, pp. 360–381, Apr. 2011, doi: 10.1080/07418825.2010.486037.

[3] A. Almehmadi, Z. Joudaki, and R. Jalali, ‘‘Language usage on Twitter predicts crime rates,’’ in Proc. 10th Int. Conf. Secur. Inf. Netw., Oct. 2017, pp. 307–310, doi: 10.1145/3136825.3136854.

[4] V. K. Borooah and N. Ireland, ‘‘Deprivation, violence, and conflict: An analysis of Naxalite activity in the districts of India,’’ Int. J. Conf. Violence, vol. 2, no. 2, pp. 317–333, 2008, doi: 10.4119/UNIBI/ijcv.42.

[5] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, ‘‘Improved method of classification algorithms for crime prediction,’’ in Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST), Aug. 2014, pp. 250–255, doi: 10.1109/ISBAST.2014.7013130.

[6] P. Pławiak, M. Abdar, and U. R. Acharya, ‘‘Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring,’’ Appl. Soft Comput., vol. 84, Nov. 2019, Art. no. 105740, doi: 10.1016/j.asoc.2019.105740.

[7] Z. Li, T. Zhang, Z. Yuan, Z. Wu, and Z. Du, ‘‘Spatio-temporal pattern analysis and prediction for urban crime,’’ in Proc. 6th Int. Conf. Adv. Cloud Big Data (CBD), Aug. 2018, pp. 177–182, doi: 10.1109/CBD.2018.00040.

[8] A. Almaw and K. Kadam, ‘‘Survey paper on crime prediction using ensemble approach,’’ Int. J. Pure Appl. Math., vol. 118, no. 8, pp. 133–139, 2018. [Online]. Available: https://internal-pdf://107.93.182.66/18. pdf%0Ahttp://www.ijpam.eu

[9] T. B. Hyde, H. Dentz, S. A. Wang, H. E. Burchett, S. Mounier-Jack, and C. F. Mantel, ‘‘The impact of new vaccine introduction on immunization and health systems: A review of the published literature,’’ Vaccine, vol. 30, no. 45, pp. 6347–6358, 2015, doi: 10.1016/j.vaccine.2012.08.029

[10] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, ‘‘Crime pattern detection, analysis & prediction,’’ in Proc. Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Apr. 2017, pp. 225–230, doi: 10.1109/ICECA.2017.8203676.