Identification of Crime Prone Areas

System Requirement Specification (SRS)

Project work Phase 1 (EAI753)

BACHELOR OF TECHNOLOGY (AI+ML+DL)

PROJECT GUIDE:

Dr. Saurabh Pathak

SUBMITTED BY:

Divyanshu Jain (TCA1959012)
Aashvi Jain (TCA1959004)

Dec, 2022

FACULTY OF ENGINEERING & COMPUTING SCIENCES
TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD

Table of Contents

1 Introduction

Our Project title is **Identification of Crime Prone Areas** as the name suggest in this project we are identifying and predicting such areas which are more crime prone mean the areas where probability of occurring of crimes is high. We have taken this problem after considering the 112-helpline data which consist of all the past crimes of Lucknow District in a given range of timeframe. We have taken this idea after exploring the present manual work of police to classify crime prone areas based on F.I.R reported.

Now to automate this our idea is to gather the data from emergency services (112) and after analyzing and setting the threshold value for the crime rate of crime-prone areas, we can categorize the crime-prone areas on the basis of the crime rate. Police can now get information about predicted crime type to happen in particular areas and month at some particular time, through a model generated which gradually decrease the manual work.

We are using the modern Machine Learning techniques to find out and predict crime prone areas and also we are visualizing the results in graphical form on a Web Interface.

1.1 Problem Statement

Crimes are increasing at a rapid rate, thus safety & security is becoming a major concern for us. While people should know whether a particular area is safe or not. People who are new to a place, have no idea about the safe areas of that particular region. Still now the police is using the traditional ways of filter out the Crime Prone areas (the areas where crime rate is high). Crime cannot be predicted since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds. According to Crime Records Bureau crimes like burglary, arson etc. have been decreased while crimes like murder have been increased. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence. For building such a powerful crime analytics tool we have to collect crime records and evaluate it.

The task is to find an effective solution in terms of Machine Learning Prediction and analyses which can Identify the Crime prone areas on the basis of Locations and also can predict crime in such areas.

Points that we recognized in this Problem:

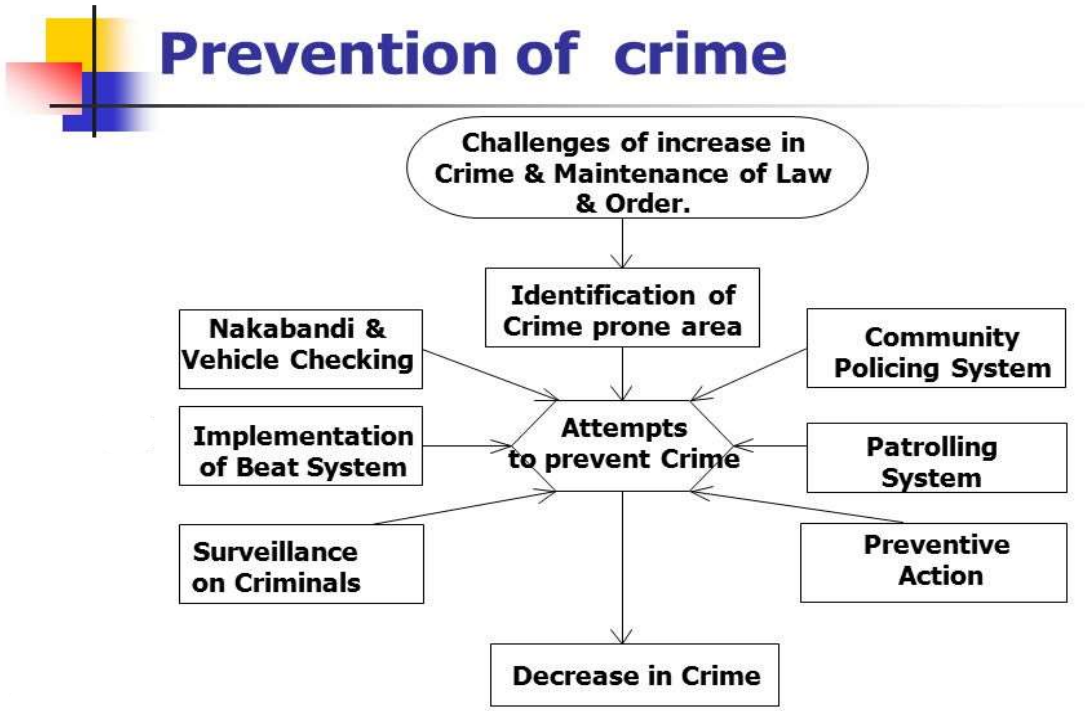
- Increase in the Crime Rate in last few years.
- Not availability of correct preventive measures of crime.
- Need of automation in current traditional methods used by Police.

1.2 Scope of the project

In the proposed system, we have done crime data analysis of with many parameters and factors including Event Id, Circle name, Police Station, Caller source, Event Type, Event-sub-type, Data of crime, Latitude, and Longitude of the location of the crime. Using Decision Tree algorithm and K-means clustering algorithm, we are predicting the type of crime for the given latitude and longitude. As a Outcome of our solution we can detect the crime prone areas on the basis of available factors which will facilitate in taking preventive actions against crime in such areas and hence crime will gradually decrease. We have plan to develop a webpage for the end user and to integrate our model with that webpage so that we can visualize the results on frontend.

The SRS is developed consistent with and in conjunction with the full set of software development activities identified in this project software Project Management Plan(SPMP). This SRS is the originating requirements source for Identification of Crime Prone Areas.

Benefits:



1.3 Definitions, Acronyms, and Abbreviations

Definitions:-

- **Machine Learning:** Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.
- **API:** API stands for Application Programming Interface. In the context of APIs, the word Application refers to any software with a distinct function. Interface can be thought of as a contract of service between two applications. This contract defines how the two communicate with each other using requests and responses.

Acronyms:

- **SPMP:** Software Project Management Plan
- **IP:** Internet Protocol
- **GUI:** Graphical User Interface
- **FIR:** First Information Report
- **SRS:** Software Requirements Specification
- **API:** Application Programming Interface

1.4 References

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. Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), 2013.

3. https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/

4. https://www.analyticsvidhya.com/blog/2021/04/steps-to-complete-a-machine-learning-project/

5. https://towardsdatascience.com/5-unique-python-modules-for-creating-machine-learning-and-data-science-projects-that-stand-out-a890519de3ae

6. https://manthan.mic.gov.in/sampledata/PS7%20Predictive%20Policing/PS 7%20predictive%20 policing%20sample %20data.xlsx

7. https://www.researchgate.net/figure/Map-showing-crime-prone-areas_fig7_280722606

2 Project Description

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict the type of crime activity which have high probability for given location in terms of latitude and longitude and date and also we can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc. we are focusing mainly on crime factors of each day.

We have taken this idea after exploring the present manual work of police to classify crime prone areas on the basis of F.I.R reported. Now to automate this our idea is to gather the data from emergency services (112) and after analyzing and setting the threshold value for the crime rate of crime-prone areas, we can categorize the crime-prone areas on the basis of the crime rate. Police can now get information about predicted crime type to happen in particular areas and month at some particular time, through a model generated which gradually decrease the manual work.

2.1 Scope of the work

In the proposed system, we have done crime data analysis of with many parameters and factors including Event Id, Circle name, Police Station, Caller source, Event Type, Event-sub-type, Data of crime, Latitude, and Longitude of the location of the crime. Using Decision Tree algorithm and K-means clustering algorithm, we are predicting the type of crime for the given latitude and longitude. As a Outcome of our solution we can detect the crime prone areas on the basis of available factors which will facilitate in taking preventive actions against crime in such areas and hence crime will gradually decrease. We have plan to develop a webpage for the end user and to integrate our model with that webpage so that we can visualize the results on frontend.

Future Scope:

1. In future scope we will try to generalize the model using different dataset
2. We will also implement notification system so that user can get notification automatically.

2.2 Project Modules

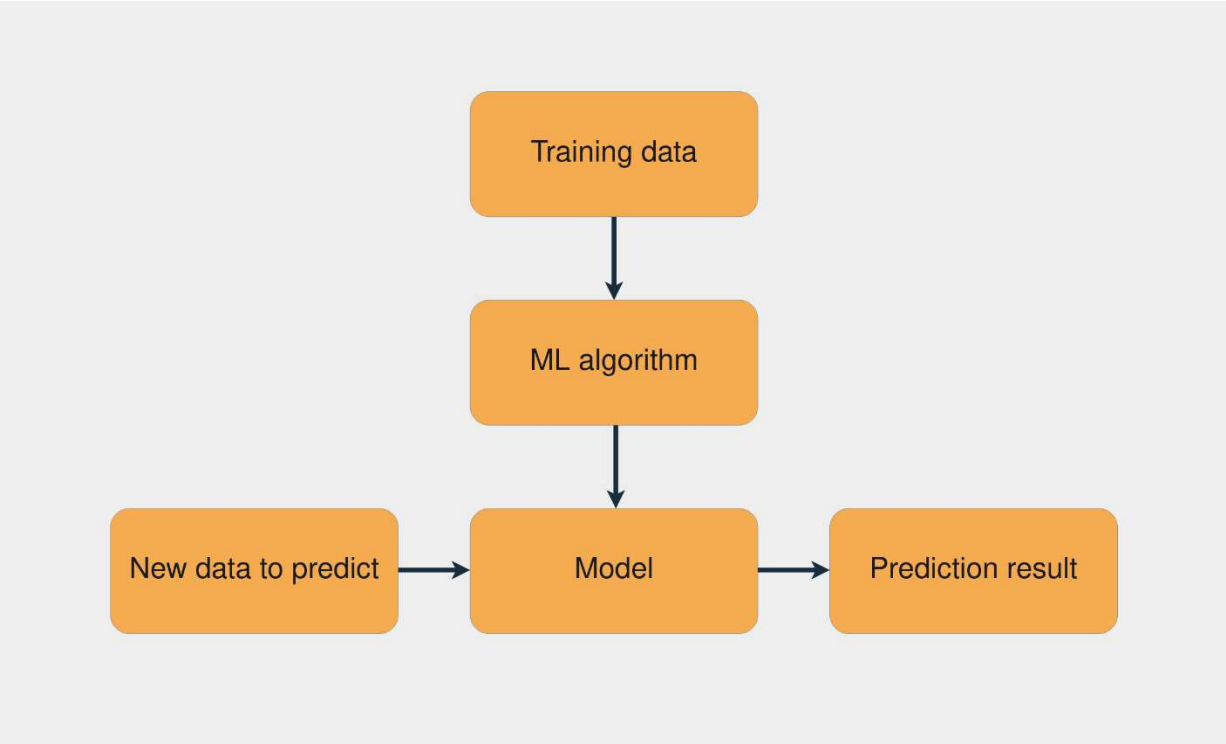
For Machine Learning Model:

1. **Data input Module:** In this we are importing the datasets in the form of csv file using pandas library of python. It will read the csv file and return dataframe object of the dataset. Pandas is an open source library in Python. It provides ready to use high-performance data structures and data analysis tools. Pandas module runs on top of NumPy and it is popularly used for data science and data analytics.
2. **Data Preprocessing Module:** In this we analyse the data and remove the null values and unnecessary data and split the dataset in to train and test data. .By preprocessing data, we make it easier to interpret and use. This process eliminates inconsistencies or duplicates in data, which can otherwise negatively affect a model's accuracy. Data preprocessing also ensures that there aren't any incorrect or missing values due to human error or bugs.
3. **Feature Extraction Module:** Feature Extraction is done for finding out the most relevant features from the given datasets. The feature in our datasets is Event Id, Circle name, Police Station, Caller source, Event Type, Event-sub-type, Data of crime, Latitude, and Longitude of the location of the crime.
4. **Training Module:** This is the next phase of our model development in this we will start the model training on the preprocessed dataset using the K-Means and Decision Tree Algorithms. Model Training is done on training dataset which is approx 70-80 percent of total data.

   The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process.

To train an ML model, these things need to be specify :

- **I.** Input training dataset.
- **II.** Name of the data attribute that contains the target to be classified.
- **III.** Required data transformation instructions
- **IV.** Training parameters to control the learning algorithm



5. **Web (GUI) Interface Module:**

1. **Home Module:** This module contains the frontend part and this is the first page of website which contains a form for enter the location, longitude and latitude for finding out weather this area is a crime prone area or not.

2. **Search Module:** Search button send the form data to the flask Machine leaning model which will process the input from user and sent back the result which can be shown on webpage.

3. **Hosting Server:** We will deployed and host our Machine Learning Integrated Web Page on Heroku. It will provide a hosted link which can be access remotely anywhere.

2.3 User Characteristics

There are 2 types of users for this Project:

- **Administrators:** this is the group of users that has the highest level of permission.They can:
    - View/Retrain/rebuild the machine learning model.
    - Can Change the User Interface settings and working of Model.
- **Users:** This is the end user who will interact with the web page.
    - Can search for crime prone areas according to area name.
    - Can search for crime prone areas by latitude and longitude.

2.4 Constraints of project

- **Availability of Data:** As we know, data is absolutely essential to train machine learning algorithms, but you have to obtain this data from somewhere and it is not cheap. Creating a data collection mechanism that adheres to all of the rules and standards imposed by governments is a difficult and time-consuming task. You need to plan out in advance how you will be classifying the data, ranking, cluster regression and many other factors. Even when the data is obtained, not all of it will be useable. In order to refine the raw data, you will have to perform attribute and record sampling, in addition to data decomposition and rescaling. Even if you have a lot of room to store the data, this is a very complicated, time-consuming and expensive process.
- **Poor Quality of Data:** Data plays a significant role in machine learning, and it must be of good quality as well. Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.
- **Irrelevant Features:** Although machine learning models are intended to give the best possible outcome, if we feed garbage data as input, then the result will also be garbage. Hence, we should use relevant features in our training sample. A machine learning model is said to be good if training data has a good set of features or less to no irrelevant features.

- **Ethical Concern:** There are, of course, many advantages to trusting algorithms. Humanity has benefited from relying on computer algorithms to automate processes, analyze large amounts of data, and make complex decisions. However, trusting algorithms has its drawbacks. Algorithms can be subject to bias at any level of development. And since algorithms are developed and trained by humans, it's nearly impossible to eliminate bias.
- **Internet Failure:** As the web interface is hosted on cloud and is accessible on internet so if somehow internet goes down then the project stops working.

2.5 Assumptions and Dependencies

- **Trustfulness of Dataset:** We are assuming that the data we collect from 112 helpline number is correct and relevant. It does not contain any fake entry.
- **Data is Clean and Correlated:** Another assumption is that our Crime Dataset is clean and the features it contains have some relationship between them.
- **Hardware and Software Requirements:** We are assuming that all the hardware and software resources that are needed in this model and web page are available on user's machine.
- **Internet Dependency:** The web page is accessible on internet and is totally dependent on the availability of an internet connection.
- **Reliability of Hosting Server:** Our Project is depend on hosting server as long as hosting server is available the web page of our project will be accessible.

3 Specific Requirements

3.1 External Interfaces

User Interfaces

This project will contain following user interfaces,

- Home Page for displaying the results on the basis of past data.
- Search form for taking user input for predicting whether it is a crime prone area or not.

Software Interface

This Project will integrate with the following software interfaces,

- Python and Machine Learning Algorithms for training the model.
- Model API using flask for integrating model with web page.
- Jupyter Notebook for the backend Model generation.
- Visual Studio Code for the Frontend Web Interface.

Communication Interface

- API build using flask will be work as a communication mechanism between model and the web interface.
- When the user enter details in the form using web page it will call the API for getting the predicted output from the backend model and show the output on the page.

3.2 Functions

Model 1.0 Import Data

Input Selection of the Link to the path of Crime Dataset in the form of csv file.

Action Working of pandas module and function will run and search for the given path.
Output Data will be imported in the model.

Notes Imported Data should be analyzed and converted in to a dataframe.
Priority High

Model 1.1 Data Cleaning

Input Selection of the imported data from the previous function in the form of data frame.
Action Data cleaning approach will run and remove noisy data.
Output Cleaned, Noise free data without null values.

Notes Null values should be check again in data.
Priority High

Model 1.1 Data Splitting

Input Selection of the imported data from the previous function in the form of data frame.
Action Train-test split will run and work on given splitting ratio.
Output Data is split in to train and test data.

Notes N/A
Priority High

Model 1.1 Model Training

Input Selection of Train data from previous function.
Action Import the required algorithms for training and start the training of the model on given train dataset.
Output Model will be trained.
Notes N/A.
Priority High.

Model 1.1 Model Testing

Input Selection of Test data from data splitting.

Action After model training the test data is feed in to the trained model and output will be checked on test data.

Output It will give report about model whether it is trained well or it may cause overfitting and underfitting.

Notes N/A

Priority High

Model 1.1 Generate Model API

Input Trained Machine Learning Model.

Action Generation of Model API started using Flask.

Output We will get the required API of Model for further use.

Notes Test the API before using.

Priority High

Web Page 1.2 Search for Crime Prone Areas

Input Filled the necessary form details like city, area, latitude and longitude and press submit button.

Action Page will call Model API in the backend and collect data.

Output The Web Page will display whether it's a crime prone area or not in text and graphical format.

Notes Put form validation to avoid information missing.

Priority High.

3.3 Performance Requirements

- **Fast Loading:** Web Page should load within seconds on internet throughout the world.
- **Fast Response:** Search result should be display within seconds.
- **Correct Prediction:** Model should predict the correct result on the basis of input data.

3.4 Dataset Requirements

Dataset to be used includes the following columns:

Event Id, Circle name, Police Station, Caller source, Event Type, Event-sub-type, Data of crime, Latitude, and Longitude of the location of the crime.

Reliability

Dataset should provide reliable and relevant information 100% of times. It should be according to the need of the Model to be trained.

Accuracy:

The dataset we are using should be accurate, contain the correct information, and be free from wrong or fake information/entries.

Scalable:

The dataset should be scalable so that if new data is available, we can add to it.

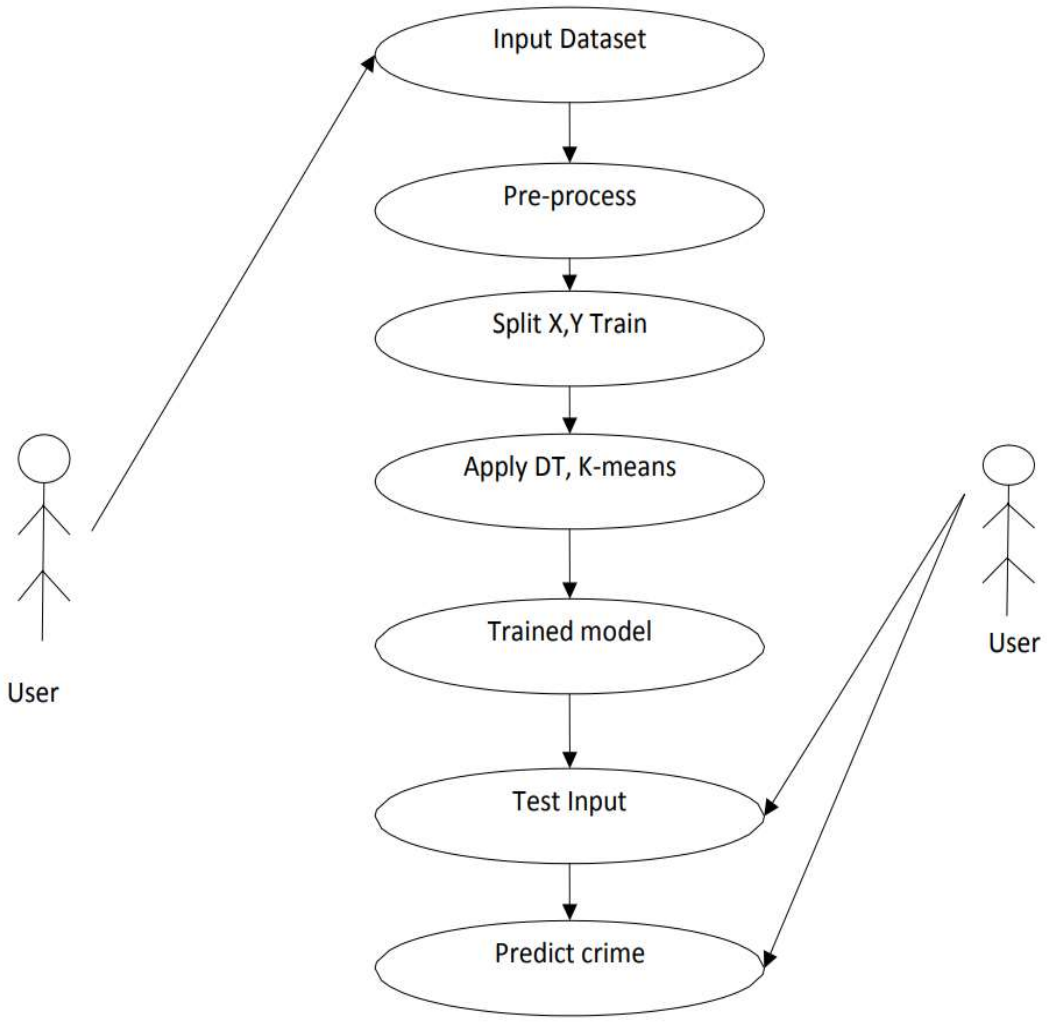
Diversify:

The Dataset should contain various diversified features so that we can train the model in an efficient approach.

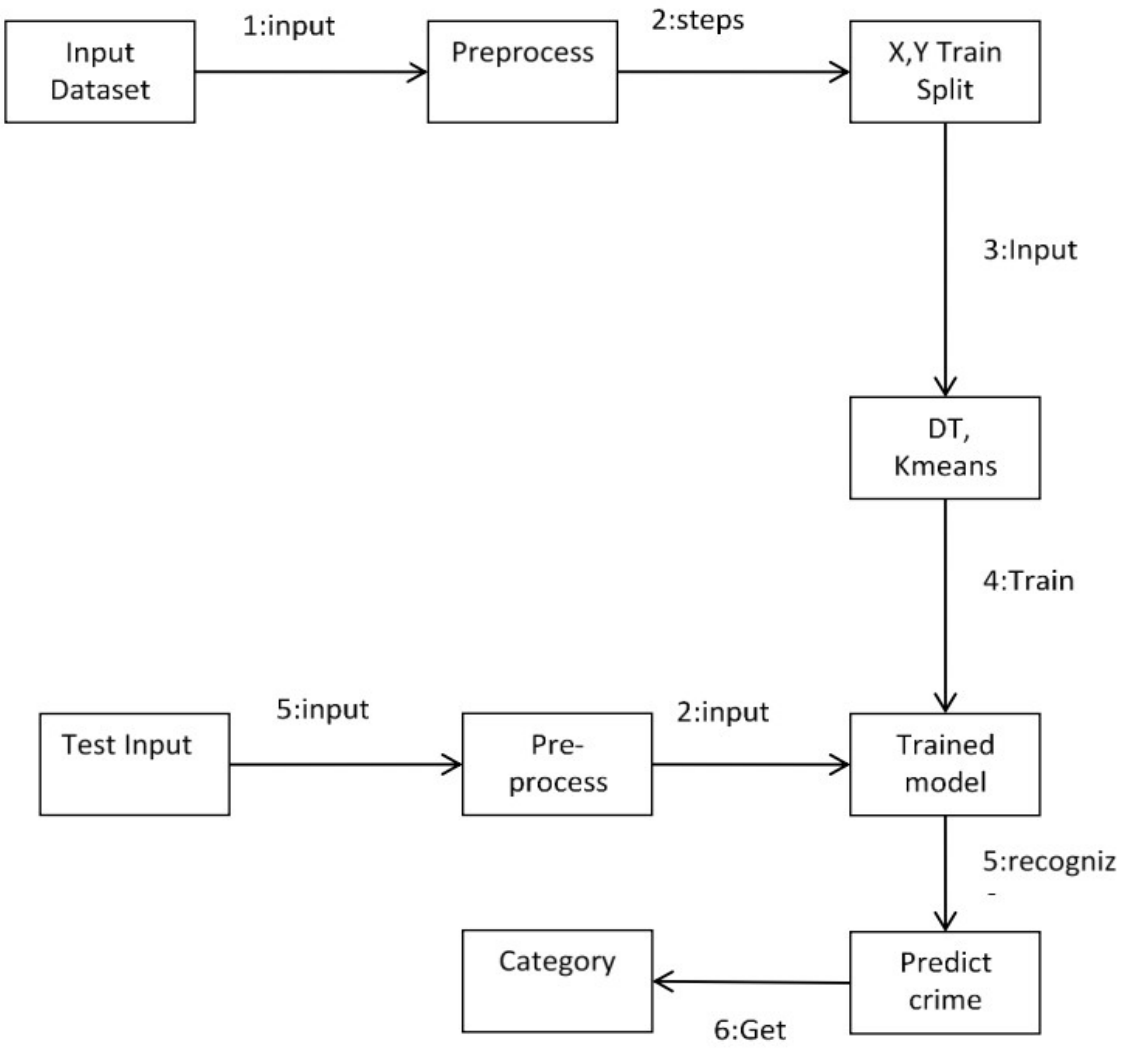
3.5 Design Constraints

- **Use Standard Version:** Web Page should be design using html 5 standards.
- **Stable Python Version:** Machine Learning model should be build using stable reliable python version.

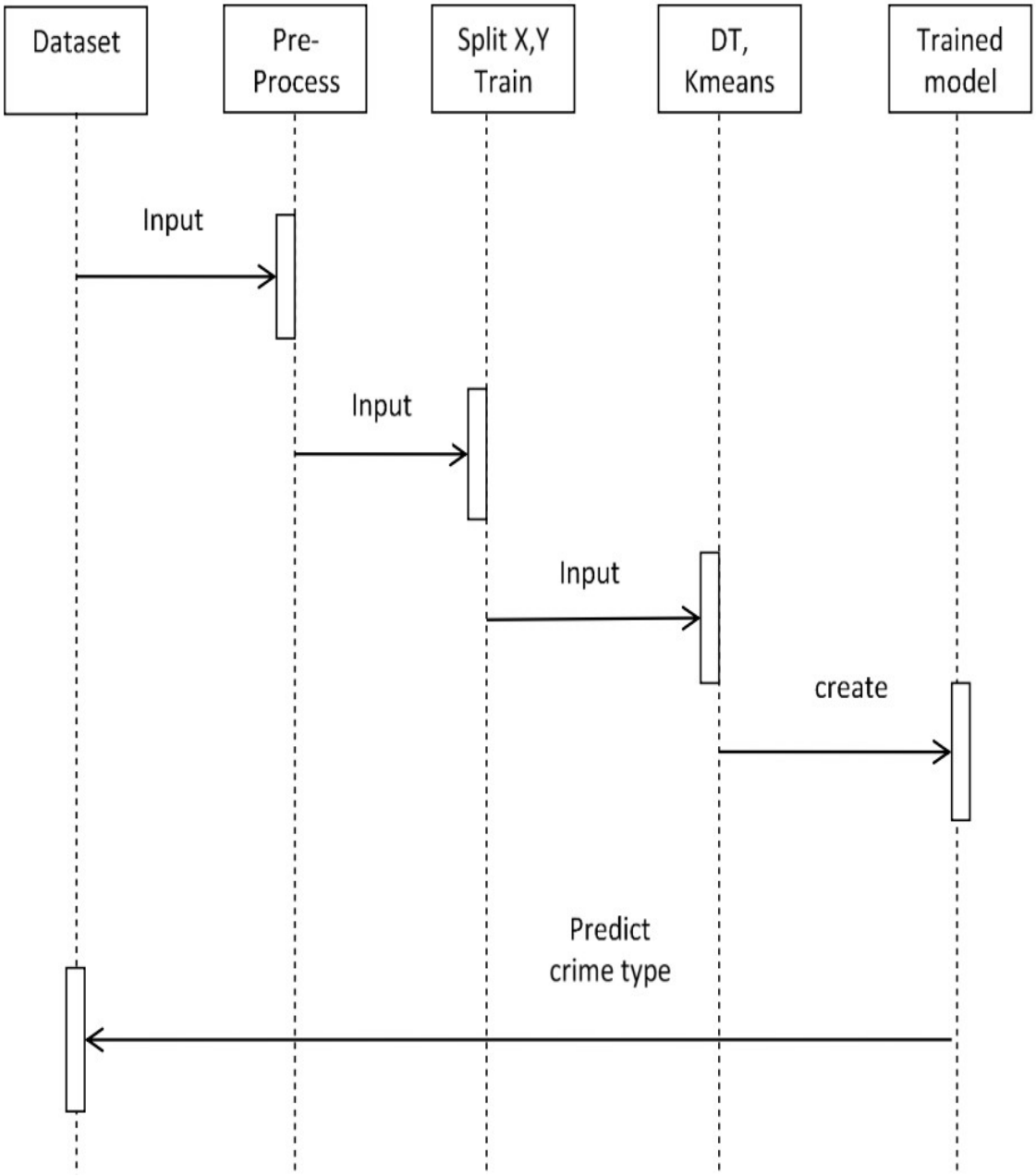
USE CASE DIAGRAM (ML Model):



COLLABORATION DIAGRAM:



SEQUENCE DIAGRAM:



3.6 Software and Hardware Requirements

Software Platform

a) Front-end

a.1.HTML5: - HTML stands for Hyper Text Markup Language. It is used to design web pages using markup language. HTML is the combination of Hypertext and Markup language. Hypertext defines the link between the web pages.

a.2.CSS 3: - Cascading Style Sheets Level 3 (CSS3) is the iteration of the CSS standard used in the styling and formatting of Web pages. CSS3 incorporates the CSS2 standard with some changes and improvements. A key change is the division of standard into separate modules, which makes it easier to learn and understand.

a.3.BOOTSTRAP: - Bootstrap is the most popular HTML, CSS and JavaScript framework for developing a responsive and mobile friendly website. It is absolutely free to download and use. It is a front-end framework used for easier and faster web development. It includes HTML and CSS based design templates for typography, forms, buttons, tables, navigation, modals, image carousels and many others. It can also use JavaScript plugins. It facilitates you to create responsive designs.

b) Back-end

b.1.Python- (Version Used 3.10)- Python is a very popular general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is a dynamically-typed and garbage-collected programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).

b.2. Flask:- Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on the Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

b.3.Libraries Used:

- Numpy
- Pandas
- SeaBorn
- Matplotlib
- Plotly
- Sklearn
- Flask

b.4. Algorithms Used:

1.K-Means Clustering: K-means clustering is one of the method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K means algorithm complexity is O(tcn), where n is instances, c is clusters, and t is iterations and relatively efficient .

It often terminates at a local optimum. Its disadvantage is applicable only when mean is defined and need to specify c, the number of clusters, in advance. It is unable to handle noisy data and outliers and not suitable for discovering clusters with non-convex shapes.K-Means clustering investigation plans to partition n perceptions into k bunch during which each perception includes a place with the bunch with the nearest centroid.

Algorithm Illustration Process:

1. Initially, the number of clusters must be known let it be k
2. The initial step is to choose a set of K instances as centres of the clusters.
3. Next, the algorithm considers each instance and assigns it to the cluster which is closest.
4. The cluster centroids are recalculated either after whole cycle of re-assignment or each instance assignment.
5. This process is iterated.

2.Decision Tree ID3 Algorithm: ID3 algorithm, stands for Iterative Dichotomiser 3, is a classification algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H).

The steps in ID3 algorithm are as follows:

1.Calculate entropy for dataset.

2.For each attribute/feature.

2.1. Calculate entropy for all its categorical values.

2.2. Calculate information gain for the feature.

3.Find the feature with maximum information gain.

4.Repeat it until we get the desired tree.

Hardware Platform

OPERATING SYSTEM	Windows 8 or higher
RAM	Minimum 4 GB
PROCESSOR	Above 500 MHz
BROWSER	Chrome, Edge
HARD DISK	1 GB Minimum