Identification of Crime Prone Areas

Project Synopsis

Project Work Phase-1(ML) (EAI 753)

BACHELOR OF TECHNOLOGY (AI+ML+DL)

PROJECT GUIDE:

Mr. Ashish Bishnoi

Dr. Saurabh Pathak

SUBMITTED BY:

Divyanshu Jain (TCA1959012)
Aashvi Jain (TCA1959004)

October, 2022

FACULTY OF ENGINEERING & COMPUTING SCIENCES

TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD

Table of Contents

1 **Project Title-** Identification of Crime Prone Areas

2 **Domain –** Machine Learning and Web Development

3 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.

4 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.

4.1 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 avialable 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.

4.2 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, you need to specify the following:

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(UI) 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.

5 Implementation Methodology

1. **Literature Survey & Planning: -** In This phase problem statement is being read carefully, so as to understand the requirements of users and the solution being planned considering its all the outcomes. We have gone through some similar problems and as well as their solution to find out the current solutions available and the modifications required.
2. **Data Collection:-** After analysing the problem we have gather the data from 112 Helpline and some other sources which contains the crime data of Lucknow district. Data Collection is the process of collecting data required for model training. Before collecting data we find out know that what kind of problem we are solving, we check for the sources of data available, then we check for is data available publically and at the end we check for format of data.

   Then after all these assumptions we collect the crime data of **Lucknow District** in csv format.
3. **Data Preprocessing:-** The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

   3.1 Data formatting- The importance of data formatting grows when data is acquired from various sources by different people. The first task for a data scientist is to standardize record formats. A specialist checks whether variables representing each attribute are recorded in the same way. Titles of products and services, prices, date formats, and addresses are examples of variables. The principle of data consistency also applies to attributes represented by numeric ranges.

   3.2 Data cleaning.- This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible. This stage also includes removing incomplete and useless data objects.
4. **Feature Selection:-** selecting the most valuable features of your dataset to model. Potentially reducing overfitting and training time(less overall data and less redundant data to train on) and improving accuracy.

   4.1 Dimensionality reduction: A common dimensionality reduction method, PCA or principal component analysis taken a large number of dimensions (features) and uses

linear algebra to reduce them to fewer dimensions. For example, say you have 10 numerical features, you could run PCA to reduce it down to 3.

4.2 Feature importance (post modelling): Fit a model to a set of data, then inspect which features were most important to the results, remove the least important ones.

4.3 Wrapper methods- such as genetic algorithms and recursive feature elimination involve creating large subsets of feature options and then removing the ones which don't matter.

5. **Dataset Splitting:-** A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

   Training set.- A *data scientist* uses a training set to train a model and define its optimal parameters — parameters it has to learn from data.

   Test set- A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

   Validation set- The purpose of a validation set is to tweak a model's hyperparameters — higher-level structural settings that can't be directly learned from data. These settings can express, for instance, how complex a model is and how fast it finds patterns in data.

6. **Training Model:-** Model Training is done using Clustering algorithm K-Means and Decision Tree ID3 algorithm for classify the crime areas. Model Training is done on training data whoch we get split in splitting process. After we preprocessed the collected data and split it into three subsets, we proceed with a model training. This process entails "feeding" the algorithm with training data. our algorithm will process data and output a model that is able to cluster and categorized the crime prone areas on the basis of threshold value.The purpose of model training is to develop a model. We have used unsupervised clustering algorithm for this purpose.
7. **Testing The Model:-** The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to achieve an algorithm's best performance.

   One of the more efficient methods for model evaluation and tuning is cross-validation.

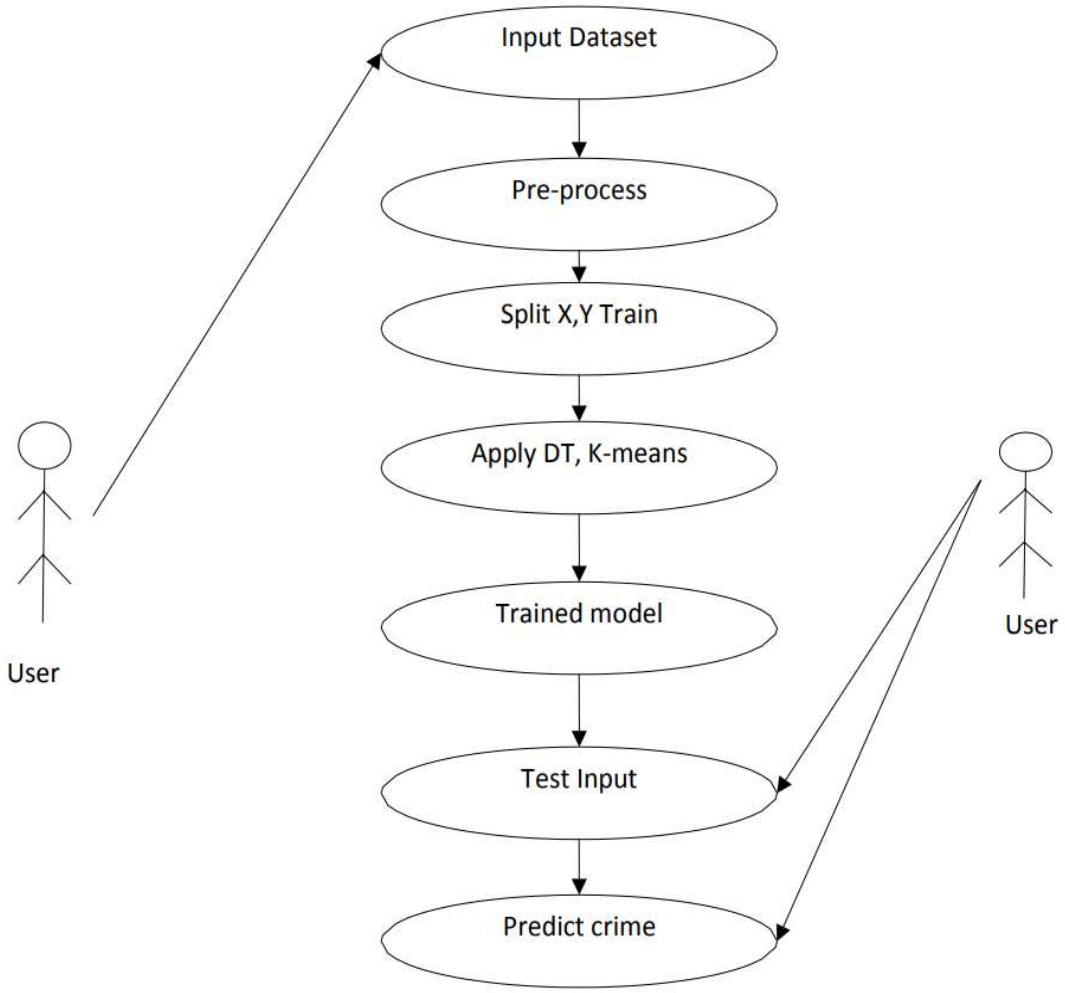
   7.1 Confusion Matrix- A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

A good model is one that has high TP and TN rates, while low FP and FN rates. It's always better to use a confusion matrix as your evaluation criteria for the machine learning model so we have used this in our model.

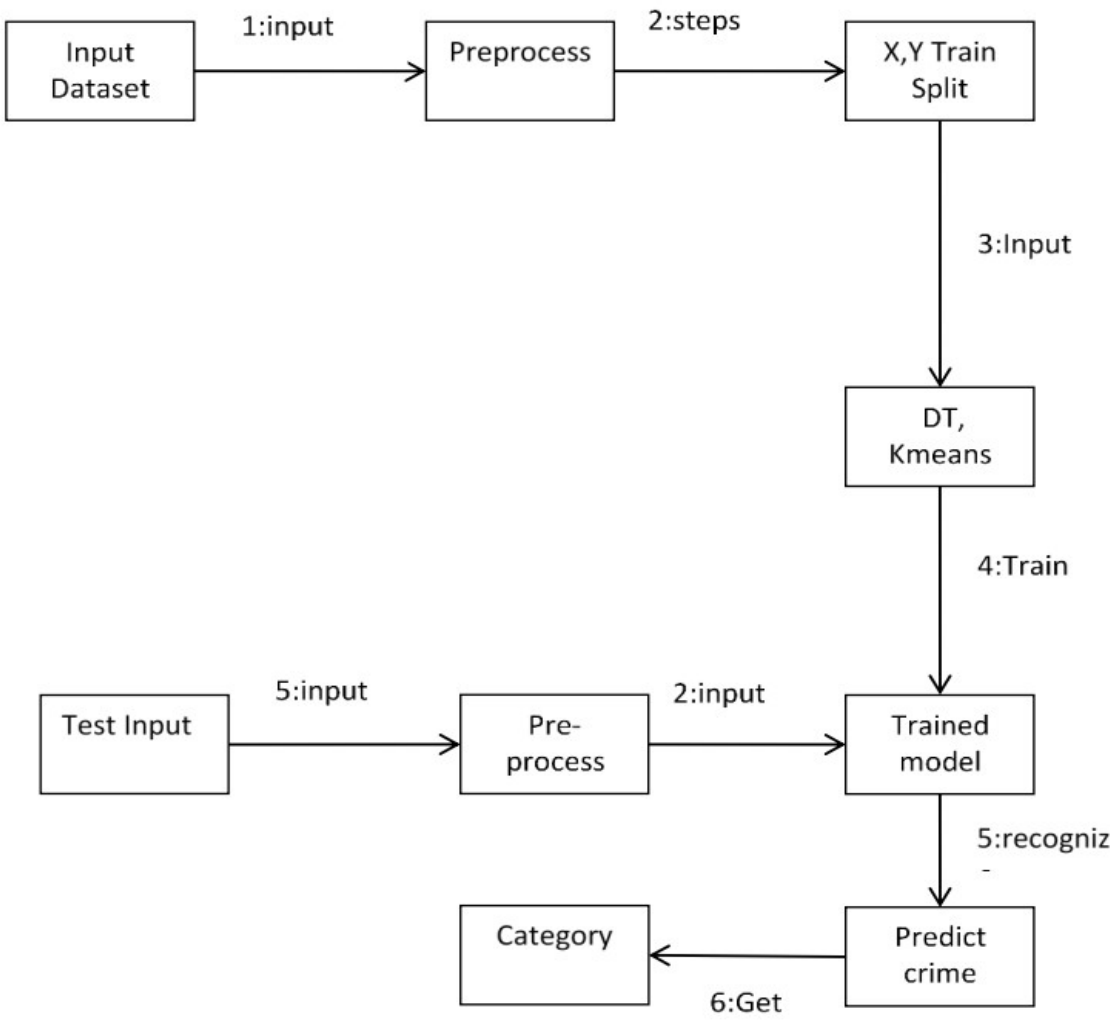
7.2 Cross-validation- Cross-validation is the most commonly used tuning method. It entails splitting a training dataset into ten equal parts (folds). A given model is trained on only nine folds and then tested on the tenth one (the one previously left out). Training continues until every fold is left aside and used for testing. As a result of model performance measure, a specialist calculates a cross-validated score for each set of hyperparameters. A data scientist trains models with different sets of hyperparameters to define which model has the highest prediction accuracy. The cross-validated score indicates average model performance across ten hold-out folds.

8. **Retrain the Model:-** After testing the model on test data then evaluate it using various model evaluation techniques. If error is more then we will retrain the model until the error got minimized.
9. **Deployment:-** After successful training of the model followed by testing it on test data then the calculation of error and if retraining needed we retrain the model and then deploy the model using flask library to generate API so that we can use it to develop our Webpage.

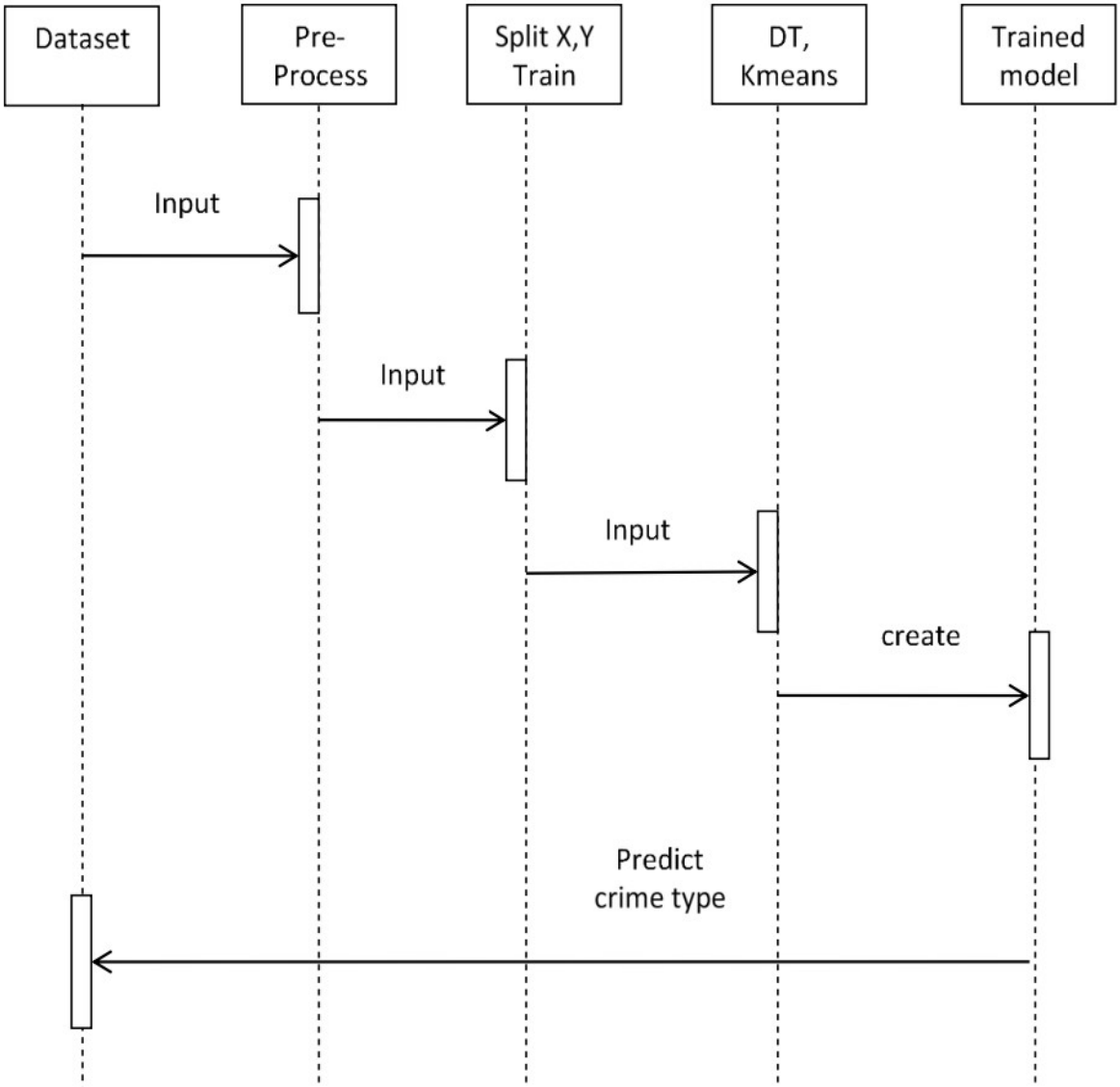
USE CASE DIAGRAM (ML Model):



COLLABORATION DIAGRAM:



SEQUENCE DIAGRAM:



6 Technologies to be used

6.1 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 unable to handle noisy data and outliers and not suitable to discover 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.

6.2 Hardware Platform

OPERATING SYSTEM	Windows 7 or higher
RAM	Minimum 4 GB
PROCESSOR	Above 500 MHz
BROWSER	Chrome, Edge
HARD DISK	1 GB Minimum

6.3 Tools

1. **Jupyter Notebook:-** JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.
2. **Visual Studio Code: -** Visual Studio Code is a code editor redefined and optimized for building and debugging modern web and cloud applications. Visual Studio Code is free and available on your favorite platform - Linux, macOS, and Windows

7 Advantages of this Project

1. Identify the Crime Prone Areas in a efficient way.
2. Reduce the Chances of Crimes.
3. Help Police and administrative authorities for planning better preventive measures.
4. Ensure Strong Public Safety.
5. Help new people to examine the area which is new for them.
6. Can be easily access by the end user through web page.

8 Future Scope and further enhancement of the Project

- For Further enhancement we will train the model on different algorithms so as to check the accuracy and results and try to build an ensemble model.
- For now our model is based on dataset of Lucknow District but in future we want to work on other datasets also.
- As we have applied clustering technique of data mining for crime analysis we can also perform other techniques such as classification.
- We will also build the better web interface in future along with the Facility of search by city name.
- Crime prone areas will also be able to search through Keyword names of that Geographical area.

9 Team Details

Project Name & ID	Course Name	Student ID	Student Name	Role	Signature
Identification of Crime Prone Areas FPA-1	Project Work Phase-1(ML) (EAI 753)	TCA1959004	Aashvi Jain	Developer, Testing	
		TCA1959012	Divyanshu Jain	Data Analayzer, Developer	

10 Conclusion

Our system takes elements attributes of an area and preprocessing offers the frequent patterns of that place. The pattern is used for constructing a model for decision tree. Corresponding to each place we build a model by training on these frequent patterns. Crime patterns cannot be static since patterns change over time. By training means we are teaching the system based on some particular inputs.

So the machine automatically learns the converting patterns in crime through examining the crime patterns. Also the crime elements trade over time. By sifting through the crime data we have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes of places instead of fixing certain attributes. Till now we trained our system using certain attributes but we are planning to include more factors to improve accuracy. Our software predicts crime prone regions in Lucknow on a particular day. It will be more accurate if we consider a particular state/region. Also another problem is that we are not predicting the time in which the crime is happening. Since time is an important factor in crime we have to predict not only the crime prone regions but also the proper time.

11 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