Ain Shams University

Faculty of Computer and Information Sciences

SOFTWARE ENGINEERING

Crime Patterns Analysis and Prediction

By

Alaa Hassan Mahmoud

Bassant Saeed Abdo

Hana Ahmed Khater

Mennatallah Mohamed Hossam

Yousof Khaled

Under Supervision of

Sherine Rady

Associate in Information System Department,

Faculty of Computer and Information Sciences,

Ain Shams University

Esraa Karam

Assistant Lecturer in Information System Department

Faculty of Computer and Information Sciences,

Ain Shams University

Acknowledgment

First and foremost, all praise and due to Allah Subhanahu wa Ta'ala.

We have to thank our supervisors, Dr. Sherine Rady , T.A Esraa Karam for the help, guidance, and encouragement they provided throughout the year.

Thanks also to our families for their support and prayers during this year.

Abstract

Crime is an act strongly disapproved of by society. Crime exists in every society irrespective of its level of prosperity. Crime creates fear and indescribable suffering among people. It violates the law as no society can develop in an atmosphere where crime predominates. So this project is needed because the number and forms of criminal activities are increasing at an alarming rate, forcing authoritative agencies to develop innovative and efficient methods to take preventive measures to avoid crime and to highlight the areas with high severity to fight crime, hence citizens can feel safe and protected. Also, it helps tourists decide which place is safer for them to visit. To achieve this, we suggest including Data Mining algorithms and techniques. The sole purpose of this study project is to determine how Data Mining can be used by law agencies or authorities to detect, prevent, and solve crimes at a much more accurate and faster rate.

Table of Contents

List of Figures

List of Abbreviations

CSS	Cascading Style Sheets
DM	Data Mining
ODBC	Open Database Connectivity
HTML	Hyper Text Markup Language
XML	Extensible Markup Language

Chapter 1
Introduction

Chapter 1
Introduction

1.1 Motivation

Crime is a social dilemma that costs our society deeply in several ways. Crime does not just affect individuals. Communities that experience higher levels of crime are also adversely affected. Its costs and effects influence just about everyone to some degree.

Crime and violence experienced by individuals living in a community is an important public safety issue. People can be exposed to violence in many ways. They may be victimized directly, witness violence or property damage crimes in their community, or hear about crime and violence from other residents.

Crime knows no social or financial class as it affects people at all levels, so guiding people living in high crime hotspots can help prevent them from being victimized.

1.2 Problem Definition

Nowadays, the world economic crisis is increasing day by day, which leads to an increase in the crime rate all over the world regardless of any country's level of development.

High crime often stands as a barrier to the socio-economic growth of society, discourages investments, and fuels migration that creates disparities in economic development worldwide. The costs and effects of crime vary widely. The ultimate cost is loss of life. Other costs to victims can include injuries, property losses and mental health issues. Losses to both victims and their families can also come in the form of increased expenses.
The project helps highlight high-risk areas to find ways to reduce crime in these areas for the greater good.

1.3 Objective

Assist the government in developing effective and humane crime prevention strategies and in establishing and maintaining institutional frameworks for their implementation and review so as to achieve safety for individuals in society, and this is done by analyzing and anticipating crime hotspots.

Visualize data and generate statistical reports to monitor the progress of institutions responsible for crime prevention.

1.4 Document Organization

The rest of the document is organized as follows:

Chapter 2

This chapter presents background information related to the targeted research area, a description of the project field, survey of work done along with description of existing similar systems.

Chapter 3

This chapter includes a System overview accompanied by system architecture, analysis, and design.

Chapter 4

This chapter includes a detailed description of all the functions in the system, a detailed description of all the techniques and algorithms implemented, and a description of new technologies used in the implementation.

Chapter 5

This chapter describes in detail how the user can use the system & operate the project.

Chapter 6

This chapter includes a complete summary of the whole project along with a comparison of the results obtained using the different methods and algorithms and the conclusion of the best results reached.

What can be done in the future to improve the performance of the project & what additional functions could be added.

References

References to Papers and webpage links that were used to research and implement this project.

Chapter 2
Background

Chapter 2
Background

Our project is Data Mining based so we will discuss in this section what Data Mining is with a detailed description of it.

2.1 Data Mining

Data Mining (DM) is a type of artificial intelligence (AI) that allows software applications to predict statistical outcomes based on an input. Data Mining algorithms use historical data as input to predict new output values.

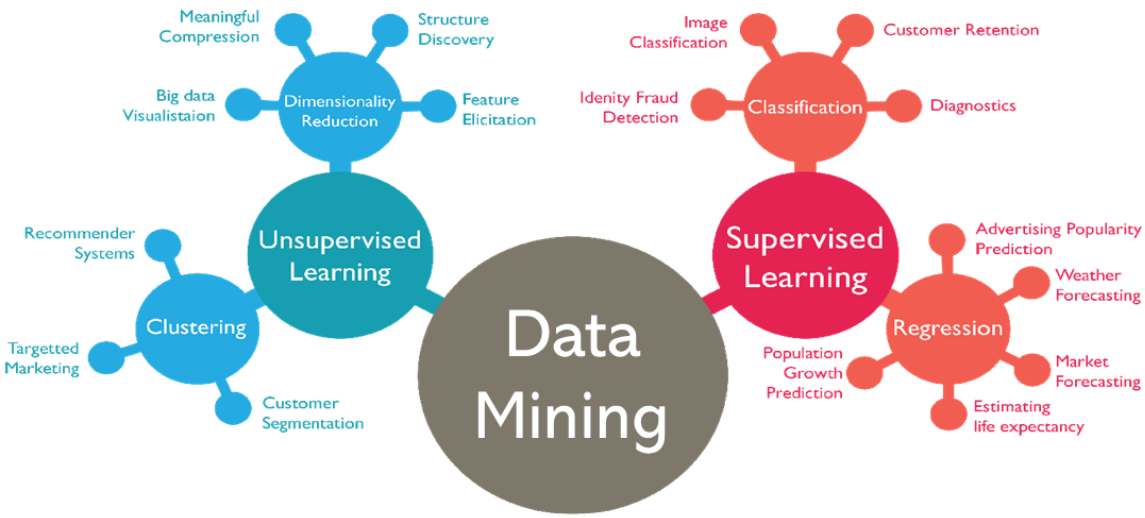

Figure 2.1 Machin Learning

2.1.1 Importance of Data Mining

Data Mining is important as it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google, and Uber, make Data Mining a central part of their operations. Data Mining has become a significant competitive differentiator for many companies.

2.1.2 Different Types of Data Mining

Classical Data Mining is often categorized by how an algorithm learns to become more accurate in its predictions. There are several basic approaches like supervised learning and unsupervised learning. but we focused on a supervised learning approach. The type of algorithms data scientists choose to use depends on what type of data they want to predict.

- **Supervised Learning:** In this type of Data Mining, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm are specified. It is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.
  Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

  Supervised learning can be separated into two types of problems when data mining—classification and regression:

  Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.

  Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as

for sales revenue for a given business. Linear regression, logistic regression, and polynomial regression are popular regression algorithms.

2.1.3 Advantages and Disadvantages of Data Mining

Data Mining has seen use cases ranging from predicting customer behavior to forming the operating system for self-driving cars.

- **Advantages**

    1. It helps companies gather reliable information.
    2. It's an efficient, cost-effective solution compared to other data applications.
    3. It helps businesses make profitable production and operational adjustments.
    4. Data mining uses both new and legacy systems.
    5. It helps data scientists easily analyze enormous amounts of data quickly.
    6. Data scientists can use the information to detect fraud, build risk models, and improve product safety.
    7. It helps data scientists quickly initiate automated predictions of behaviors and trends and discover hidden patterns.

- **Disadvantages**

    1. It can be expensive. Data Mining projects are typically driven by data scientists, who command high salaries. These projects also require software infrastructure that can be expensive.
    2. There is also the problem of Data Mining bias. Algorithms trained on data sets that exclude certain populations or contain errors can lead to inaccurate models of the world that, at best, fail and, at worst, are discriminatory. When an enterprise bases core business processes on biased models it can run into regulatory and reputational harm.

2.1.4 Choosing the Right Data Mining Model

The process of choosing the right Data Mining model to solve a problem can be time-consuming if not approached strategically.

Step 1: Define potential data inputs that should be considered for the solution. This step requires help from data scientists and experts who have a deep understanding of the problem.

Step 2: Collect data, format it and label the data if necessary. This step is typically led by data scientists.

Step 3: Choose which algorithm(s) to use and test to see how well they perform. This step is usually carried out by data scientists.

Step 4: Continue to fine-tune inputs until the model reaches an acceptable level of accuracy. This step is usually carried out by data scientists with feedback from experts who have a deep understanding of the problem.

2.2 Scientific background

The project tested multiple data mining algorithms

2.2.1 Random Forest Algorithm

2.2.1.1 What is Random Forest

The random forest classifier 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. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees.
Instead of relying on one decision tree, the random forest takes the prediction from each tree, and based on the majority votes of predictions, it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

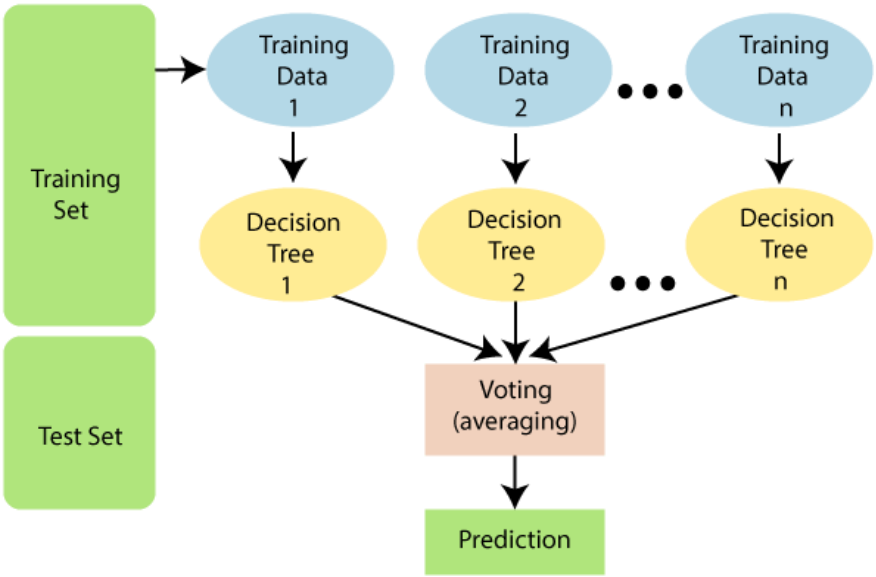

Figure 2.2 Working of Random Forest algorithm

2.2.1.2 Overview On Decision Trees

Let's quickly go over decision trees as they are the building blocks of the random forest model. Fortunately, they are pretty intuitive. They can be used to solve both regression and classification problems. It uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree.

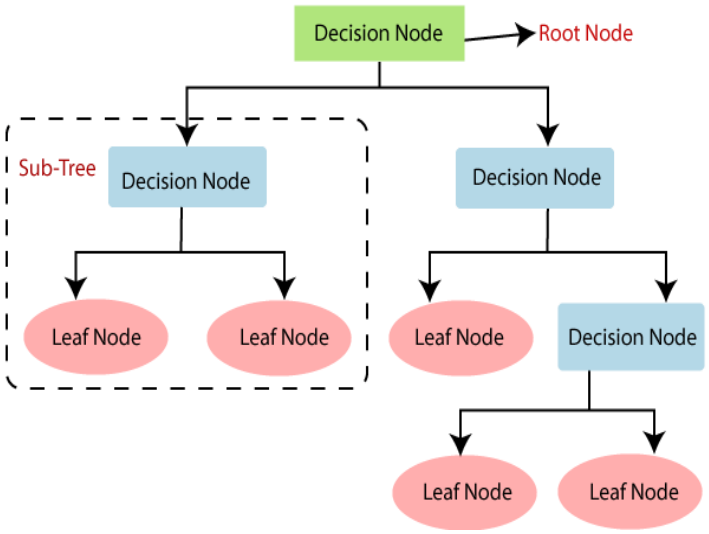

Figure 2.3 Decision Tree algorithm

In Decision Trees, the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures: Information Gain, and Gini Index.

- **Information Gain**

  It is a statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy.

  It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree.

$$Information\ Gain\ =\ Entropy(before) - \sum_{j=1}^{K} Entropy(j,\ after)$$

Figure 2.4 Information Gain equation

- **Gini Index**

  You can understand the Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and is easy to implement whereas information gain favors smaller partitions with distinct values.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Figure 2.5 Gini Index equation

2.2.1.3 Random Forest Prerequisites

In data science-speak, the reason that the random forest model works so well is:

A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for Random Forest to perform well are

- There needs to be some actual signal in our features so that models built using those features do better than random guessing.
- The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

2.2.1.4 Random Forest Feature Importance

Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

By looking at the feature importance you can decide which features to possibly drop because they don't contribute enough (or sometimes nothing at all) to the prediction process. This is important because a general rule in Data Mining is that the more features you have the more likely your model will suffer from overfitting and vice versa.

2.2.1.5 Difference Between Decision Trees and Random Forests

While Random Forest is a collection of Decision Trees, there are some differences.

If you input a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions.

For example, to predict whether a person will click on an online advertisement, you might collect the ads the person clicked on in the past and some features that describe their decision. If you put the features and labels into a decision tree, it will generate some rules that help predict whether the advertisement will be clicked or not. In comparison, the random forest algorithm randomly selects observations and features to build several decision trees and then averages the results.
Another difference is that "deep" decision trees might suffer from overfitting. Most of the time, random forest prevents this by creating random subsets of the features and building smaller trees using those subsets. Afterward, it combines the subtrees. It's important to note this doesn't work every time and it also makes the computation slower, depending on how many trees the random forest builds.

2.2.1.6 Advantages & Disadvantages of Random Forest

- Advantages
    a. It can be used for both regression and classification tasks.
    b. Random forest is also a very handy algorithm because the default hyperparameters it uses often produce a good prediction result.
    c. Deals with overfitting problems.

- Disadvantages
    a. A large number of trees can make the algorithm too slow and ineffective for real-time predictions
    b. predictive modeling tool and not a descriptive tool,

2.2.2 Logistic Regression Algorithm

Algorithm used for predicting the categorical dependent variable using a given set of independent variables, 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.

Logistic Regression is much similar to Linear Regression except that they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, whether a mouse is obese or not based on its weight, etc.
Can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Function (Sigmoid Function)

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

The S-form curve is called the Sigmoid function or the logistic function.
In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tend to 1, and a value below the threshold value tends to 0.

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Figure 2.6 Logistic Regression equation

2.2.3 Bayes Classifiers

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described next. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision trees and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.
Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naive."

Bayes Theorem

Bayes' theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. Let X be a data tuple. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis such as that the data tuple X belongs to a specified class C. For classification problems, we want to determine P(H|X), the

probability that hypothesis H holds given the "evidence" or observed data tuple X. In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

Naive Bayes

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X=(x1,x2,..,xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2.., An

2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if

$$P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X}) \quad \text{for } 1 \leq j \leq m, j \neq i.$$

Thus, we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem,

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i)P(C_i)}{P(\boldsymbol{X})}.$$

Figure 2.7 Bayes' Theorem

3. As P(X) is constant for all classes, only P(X|Ci) P(Ci) needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = .. = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci) P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = |Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). To reduce computation in evaluating P(X|Ci), the
naive assumption of class-conditional independence is made. This presumes that the attributes' values are conditionally independent of one another, given the

class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

Figure 2.8 Conditional Probability of X given Ci

5. To predict the class label of X, P(X|Ci) P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

$$P(\mathbf{X}|C_i)P(C_i) > P(\mathbf{X}|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i.$$

Figure 2.9 Condition of classifying X as Ci

In other words, the predicted class label is the class Ci for which P(X|Ci) P(Ci) is the maximum.

The results of testing all the previous algorithms on the Indian Dataset

Algorithm	Accuracy
Random Forest	90.4%
Decision Tree	86.5%
Logistic Regression	87.8%
Naïve Bayes	83.7%

Figure 2.10 Accuracy Results

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

Figure 2.11 Accuracy Equation

2.3 Work Done in the Field of Data Mining

A study was carried out on actual crime statistical data using a set of algorithms, namely Random Forest, Decision Tree, Logistic Regression, and Naive Bayes were tested on a set of demographic features found in India's districts along with their crime data.

The dataset that was used contains socio-economic data along with crime data of the 633 districts of India.

The dataset originally had 115 attributes rounded down using feature selection to 85 factors along with the class label which is the crime rate, the state name, and the district name.

The dataset was retrieved from censusindia.gov.in and data.gov.in, both websites having authorized datasets from the government of India.

The scope of the project was predicting crime severity and other applications, such as analyzing crime hotspots, visualizing crime hotspots on maps, and learning criminal trends.

Crime predictions were done based on DM. Crime data for 2011 in India was analyzed for prediction. This Data Mining-based crime analysis involves the collection of data, data classification, prediction, and visualization. Algorithms were also used to analyze the crime dataset to find which one gives the best accuracy among them. Crime prediction with an accuracy of between 83.7% and 90.4% was obtained.

Data Mining techniques were used to predict crime in the crime data set from India. The dataset consists of demographic information such as general age group, education level, number of males/females, and number of educated people in different stages. Various sets of models were tested, and the most accurate model which turned out to be Random Forest Classifier was used for prediction.

Interactive maps and statistical charts were developed which helped in visualizing the different attributes contributing to crime in India.

2.4 Similar Systems

Since crime is a topic of concern to the public, there should be some systems that try to spread awareness about the most dangerous places where crime is.

2.4.1 'Spot Crime'

'Spot Crime' provides nationwide crime information about arrests, arsons, assaults, burglaries, robberies, shootings, thefts, and vandalism. Spot Crime will map data for any police agency that supplies open and unrestricted access to crime data. The crime data is mainly from police departments and news reports. Anyone can access these maps and have the option to sign up to receive free crime alerts via email and text. The email alert includes a map and crime details of an incident that has occurred in the specified area. In 2012, Spot Crime launched its own crime tip service, CrimeTip.us, allowing users to anonymously report crimes in their area.

2.4.2 'CRIMEWATCH'

'CRIMEWATCH' is a platform built for Law Enforcement to drive transparency through the sharing of information and policy with local communities. A packaged solution to leverage your operational activity to meet emerging demands and build two-way communication within the community. It is a packaged solution to share policy, collect community feedback, and leverage your operational activity to create engagement and truth.

2.4.3 ‘Crime Mapping’

‘Crime Mapping’ has been developed by Central Square Technologies to help law enforcement agencies throughout North America provide the public with valuable information about recent crime activity in their neighborhood. Their goal is to assist police departments in reducing crime through a better-informed citizenry. Creating more self-reliance among community members is a great benefit to community-oriented policing efforts everywhere and has been proven effective in combating crime.

CrimeMapping.com utilizes an advanced mapping engine, which helps in providing a high level of functionality as well as flexibility to the agencies. Crime data is extracted on a regular basis from each department's records system so that the information being viewed through a Web browser is the most current available. This data is always verified for accuracy and all address information is generalized by block in order to help ensure privacy is protected.

Chapter 3
Analysis and Design

Chapter 3
Analysis and Design

3.1 System Overview

In this section we present a number of diagrams that provide important information in addition to the main function of the system.

3.1.1 System Architecture

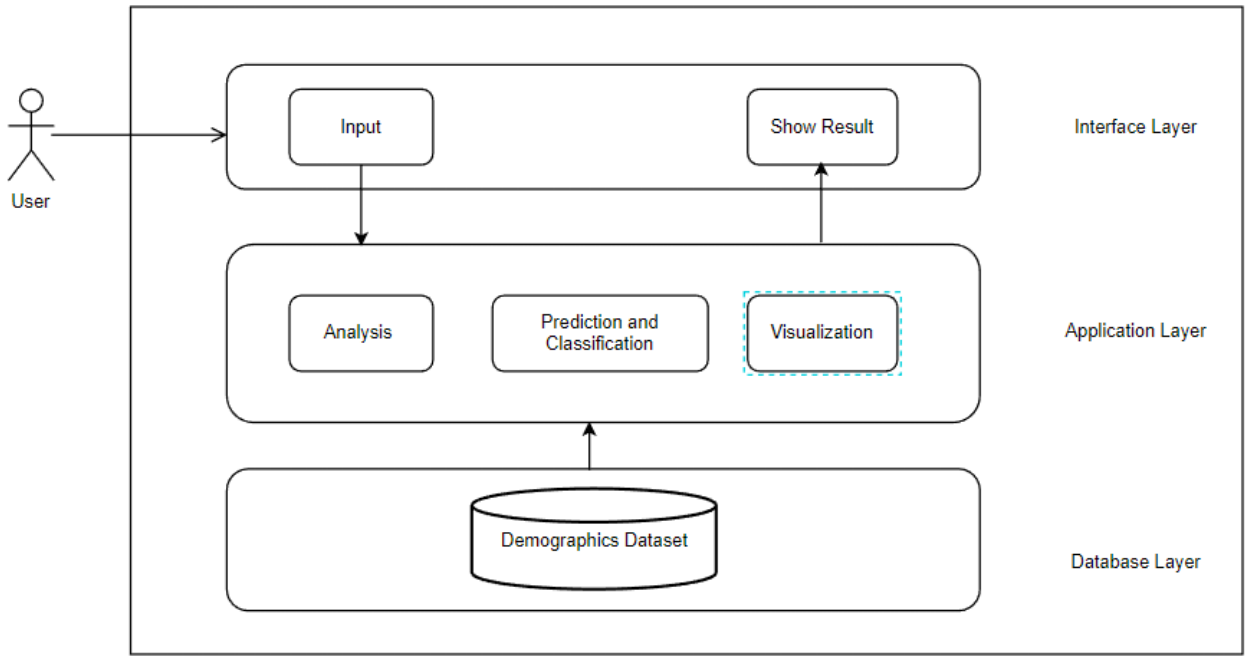

Figure 3.1 Architecture

- **Interface Layer**
    1. Input: User Input (Command)
    2. Show Result: System Output

- **Application Layer**

    1. Analysis: analysis of current crime hotspots.
    2. Visualization: visualization using an interactive map to present the crime rate for each district using different colors each color represents a range of crime severity, in addition to presenting top crime hotspots and a combination of more than one attribute to provide more information to the user
    3. Prediction and Classification: in this module, we used the Data Mining model to predict the future potential crime hotspots.

- **Database Layer**
    1. Demographics Dataset: The main dataset used by the system is the demographic dataset which is a mix between the socioeconomic factors and the crime rate.

3.1.2 Functional Requirements

- The user can register and log in.
- The user can text search for the current crime rate by area or any other attribute.
- The system can visualize crime hotspots for each district on an interactive map.
- The system can visualize different attributes individually on a map or on statistical charts.
- The system can visualize combinations of influential attributes on a map.
- The system can visualize the top X crimes on a map or on statistical charts.
- The system can predict a future crime hotspot by changing any attribute's value and then reclassifying it.
- The system can generate statistical reports for monitoring and decision-making.

3.1.3 Nonfunctional Requirements

1. **Accessibility**
   The system is meant to be accessed by anyone including police officers and law agencies. This prompted us to make sure that the system is easy to load, and that it only requires an internet connection and a browser.

2. **Usability**
   The application is so easy to use. Pages have names and titles on them, so the application is quite self-explanatory.

3. **Portability**
   The application is designed to work on any browser on a desktop. Thus, the team decided to create a responsive web application

4. **Fault Tolerance**
   The system is designed to deal with users' faults. Users can forget to enter the correct inputs, in this case, the application gives alert messages to users.

3.1.4 System Users

This system targeted special users. In this section, we will represent those users and their characteristics.

A. Intended Users

The target users of this system are represented first by authorities who use the system to know the future crime hotspots so that they take more caution and provide more protection for those areas, and secondly the moving citizens who need the system to know the safe and dangerous area around them which helps them in making decisions about where they can move, the last of which are tourists because the system helps them decide the safest place to visit and stay.

B. User Characteristics

- The user should have the basic knowledge of operating a browser and have access to it.
- The user should have an account.

3.2 System Analysis & Design

3.2.1 Use Case Diagram

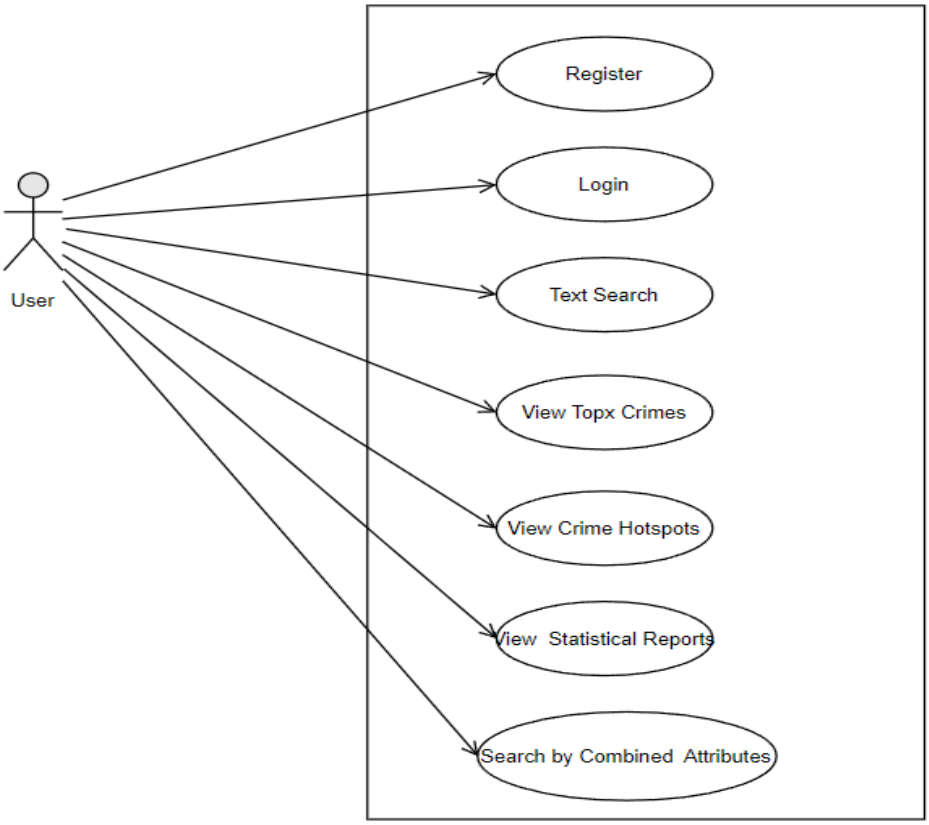

Figure 3.2 Use Case Diagram

Use Case Description

1. Register: This use case is used to enable users to create an account.
2. Login: This use case is used to enable users to login in the web application to use all the options that are provided.
3. Text search: Allow users to search for any district using any attribute's value.
4. View Top X Crimes: Users can view the top numbers of districts that have the highest rate of crimes on a map.
5. View Crime Hotspots: The system can display the crime rate of each district on a map.
6. View Statistical Reports: The system provides different kinds of statistical charts to analyze crime rates.
7. Search by Combined Attributes: The user can view districts that satisfy conditions chosen by him.

3.2.2 Class Diagram

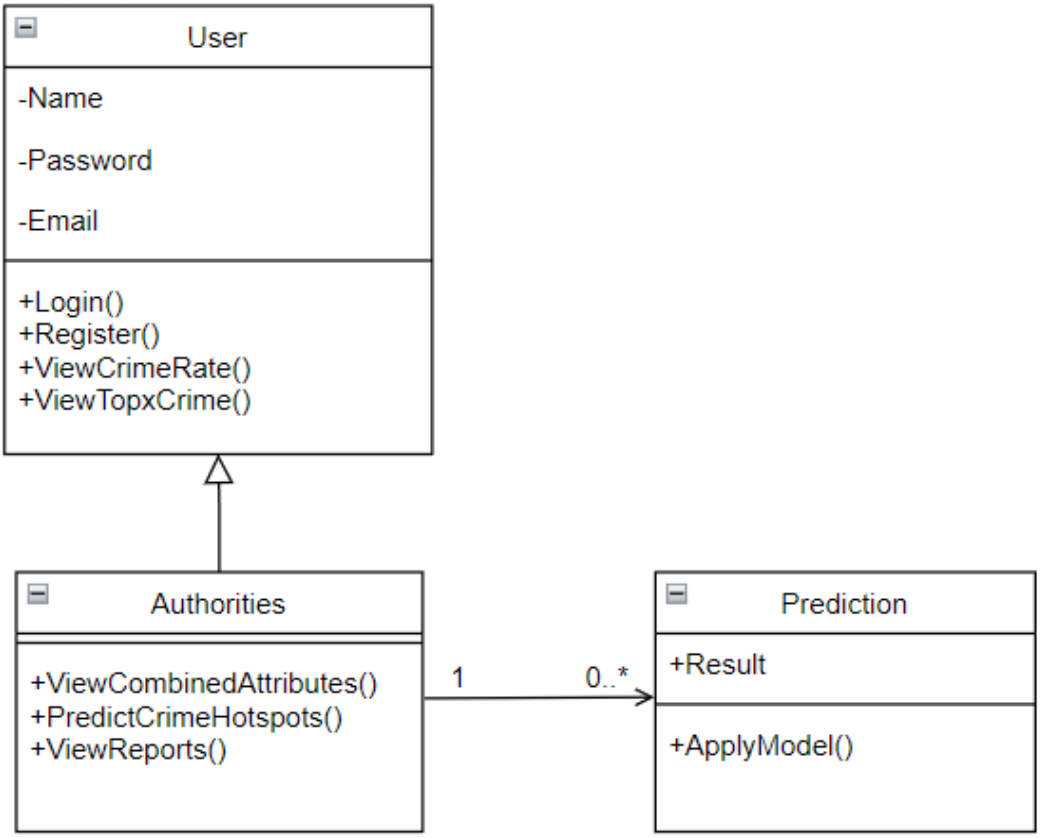

Figure 3.3 Class Diagram

Description

1. User Class: The user has attributes to be able to login/register and has the user's main functions.
2. Authorities: Authorities are extended from the user class, it has the user's main functions in addition to some specific functions that aid authorities in crime pattern detection.
3. Prediction: Apply the model to show prediction results.

3.2.3 Sequence Diagram

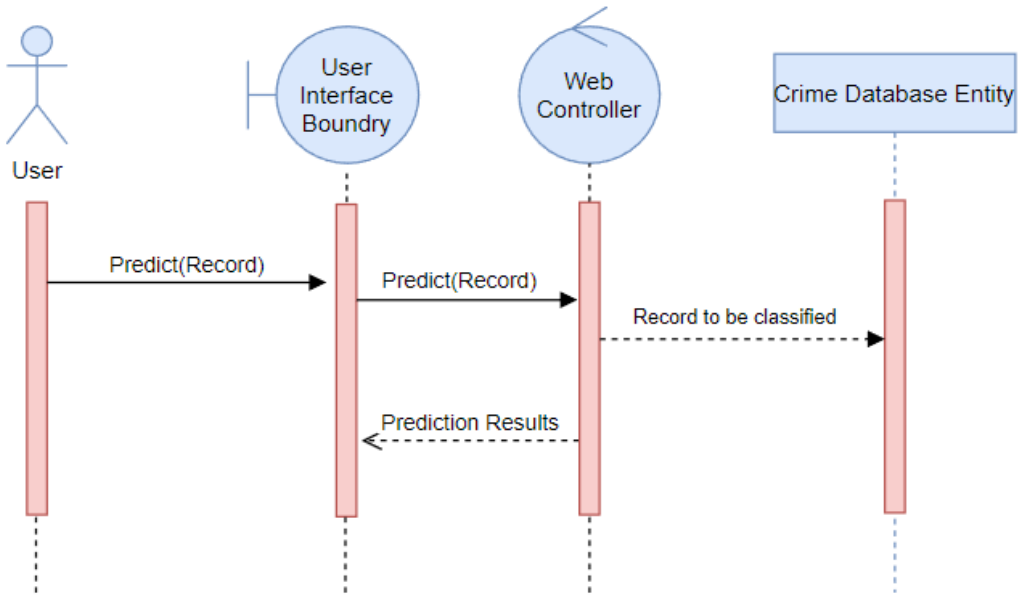

Figure 3.4 Sequence diagram

Chapter 4
Implementation

Chapter 4
Implementation

4.1 Functions

- **Register and Login**
  A user can log in with his email and password that was created while registering to the system. then he can access the available functionality of the system.

- **Text Search**
  In this function, users have the ability to search for crime rate by entering the area name or any other attribute found in the dataset. then the system will display the crime rate that is related to the attribute entered by the user.

- **Visualize Crime Hotspots**
  The system provides a functionality which is the visualization of crime hotspots of each district on a map with different colors that describe crime rate as follows: low as light orange color, medium as dark orange color, high as violet color and finally very high as black color.

- **Visualize Different Attributes Individually:**
  Besides that, the system can visualize the crime rate of each district. Users can choose different attributes individually that the system specifies to visualize on the map with different colors that describe the values of each attribute.

- **Visualize Combinations of Influential Attributes**
  The system provides the user with the ability to enter 2 or 3 attributes and then display the districts that satisfy the combinations of the attributes.

- **Visualize Top X Crime Hotspots**
  This function allows users to enter a number between 1 and 633 which is the number of districts found in the dataset then visualize the districts on the map that have a crime rate that is equal to or greater than the number entered by the user with their crime rates numbers.

- **Prediction**
  The system can predict a future crime hotspot by allowing the user to change an attribute value in the record displayed by the system then reclassifying it and showing the results of the prediction to the user.

- **Statistical Reports**
  The functionality of this function is to generate statistical reports which are the results of the analysis done and can help in decision making. The statistical reports generated are **pie charts** that display the top x districts that have the highest crime rate, **stacked charts** that display the top x districts with their top x districts in crime rate, and a **histogram** displaying the values of crime rate with a number of occurrences of each value.

4.2 Used Technologies

- **Python language**

  Python is the language used in this project, Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.

  Python's simple, easy-to-learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.

- **Flask (Web Framework)**

  Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, and upload handling, various open authentication technologies, and several common framework-related tools.

- **Scikit-learn (sklearn) in Python**

  Scikit-learn is probably the most useful library for Data Mining in Python. The sklearn library contains a lot of efficient tools for Data Mining and statistical modeling including classification, regression, clustering, and dimensionality reduction.

- **Joblib**

  Joblib is a set of tools to provide lightweight pipelining in Python. It provides utilities for saving and loading Python objects that make use of data structures efficiently.

- **ODBC**

  Open Database Connectivity (ODBC) is an open standard application programming interface (API) that allows application programmers to access any database.

  ODBC consists of four components, working together to enable functions. ODBC allows programs to use SQL requests that access databases without knowing the proprietary interfaces to the databases. ODBC handles the SQL request and converts it into a request each database system understands.

- **Geojson**

  GeoJSON is a plain-text format designed for representing vector geometries, with or without non-spatial attributes, based on the JavaScript Object Notation, GeoJSON has become a very popular data format in many GIS technologies and services related to web mapping. It is actually the standard format for passing spatial vector layer data between the client and the server in a web application.

- **Leaflet**

  Leaflet is the leading open-source JavaScript library for mobile-friendly interactive maps. Weighing just about 42 KB of JS, it has all the mapping features most developers ever need.

  Leaflet is designed with simplicity, performance, and usability in mind. It works efficiently across all major desktop and mobile platforms, can be extended with lots of plugins, has a beautiful, easy-to-use, and well-documented API, and a simple, readable source code that is a joy to contribute to.

- **HTML**

  The Hypertext Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as cascading style sheets (CSS) and scripting languages such as JavaScript.

  Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

- **JavaScript**

  JavaScript, often abbreviated as JS, is a programming language that is one of the core technologies of the World Wide Web, alongside HTML and CSS. The majority of websites use JavaScript on the client-side for web page behavior, often incorporating third-party libraries. All major web browsers have a dedicated JavaScript engine to execute the code on users' devices.

- **CSS (Cascading Style Sheet)**

  Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language such as HTML or XML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript.

  CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility; provide more flexibility and control in the specification of presentation characteristics; enable multiple web pages to share formatting by specifying the relevant CSS in a separate .css file, which reduces complexity and repetition in the structural content; and enable the .css file to be cached to improve the page load speed between the pages that share the file and its formatting.

- **SciPy**

  SciPy, a scientific library for Python, is an open-source, BSD-licensed library for mathematics, science, and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. It provides many user-friendly and efficient numerical practices such as routines for numerical integration and optimization.SciPy is predominantly written in Python, but a few segments are written in C.

- **Pandas Library**

  pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

  Pandas is mainly used for data analysis and associated manipulation of tabular data in Data Frames. Pandas allows importing data from various file formats such as comma-separated-values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as

well as data cleaning, and data wrangling features. The development of pandas introduced into Python many comparable features of working with Dataframes that were established in the R programming language. The Pandas library is built upon another library, NumPy, which is oriented to efficiently working with arrays instead of the features of working on Dataframes.

Chapter 5
User Manual

Chapter 5
User Manual

Installation Guide

1. Install Python version 3.10.0

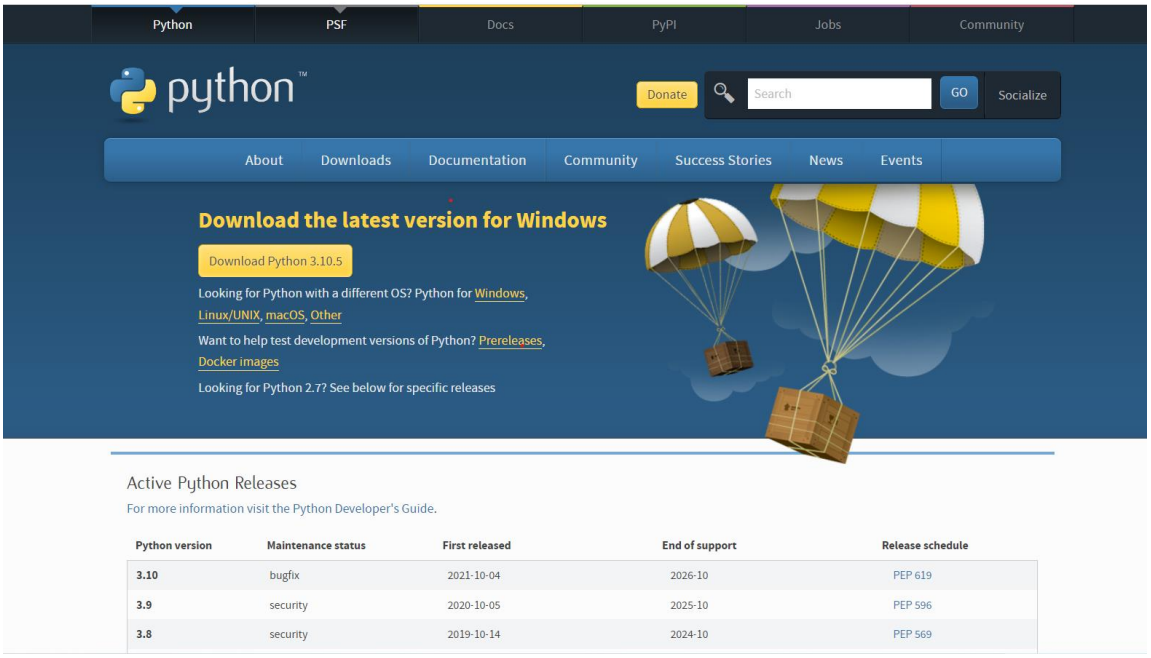

Figure 5.1 Install Python

2. Choose the appropriate version for your operating system

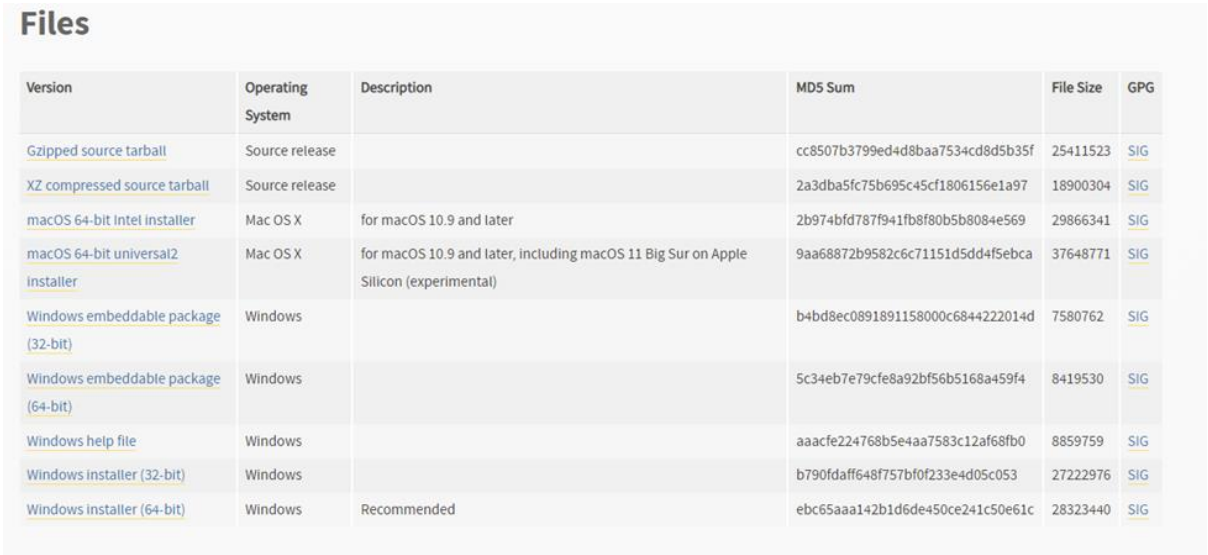

Figure 5.2 Python versions

3. Install Visual Studio Code or Pycharm

Figure 5.3 Install visual studio code

4. Now before you start to run the project create an environment to run and install requirements files that contain dependencies and packages needed to install the project.

5. Now you are ready to run the full project

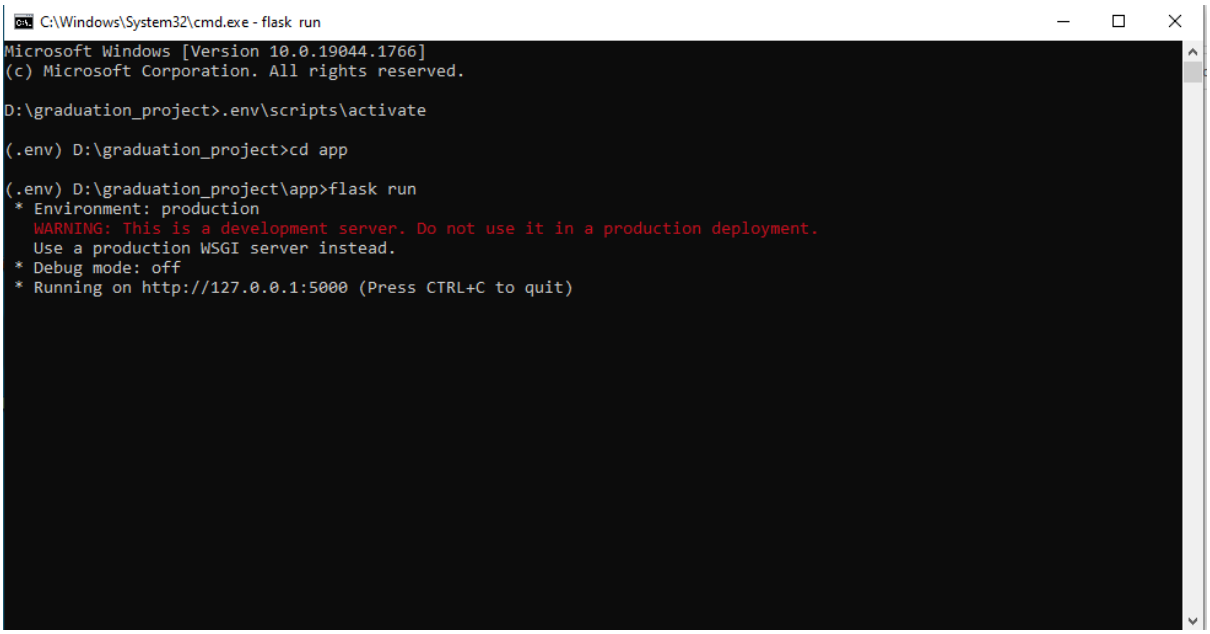

Figure 5.4 Run Flask project

6. Just click on the “ctrl+click” in the terminal or try to open the website http://127.0.0.1:5000/ from the terminal.

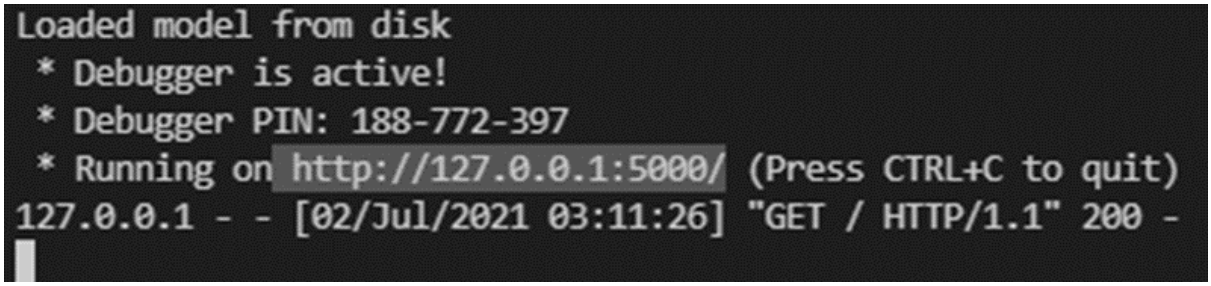

Figure 5.5 URL to open browser

7. First “Login page” will be displayed to the user to login if the user has an account otherwise the user will have to register first by clicking on the register link.

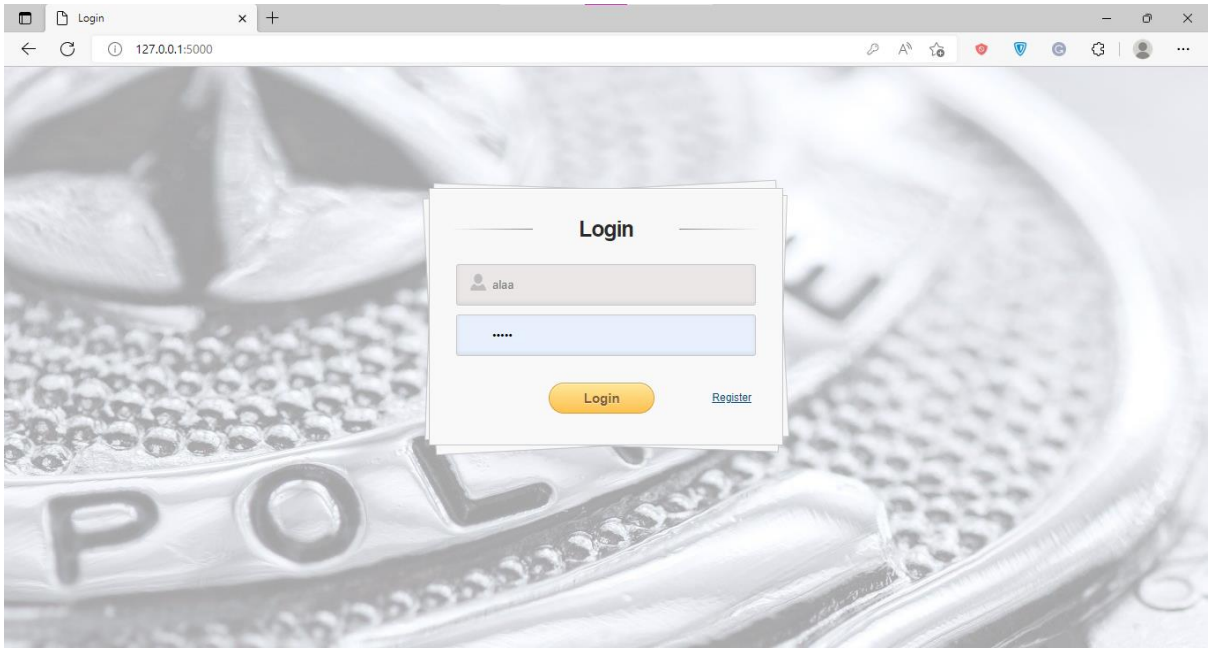

Figure 5.6 Demo screenshot (login)

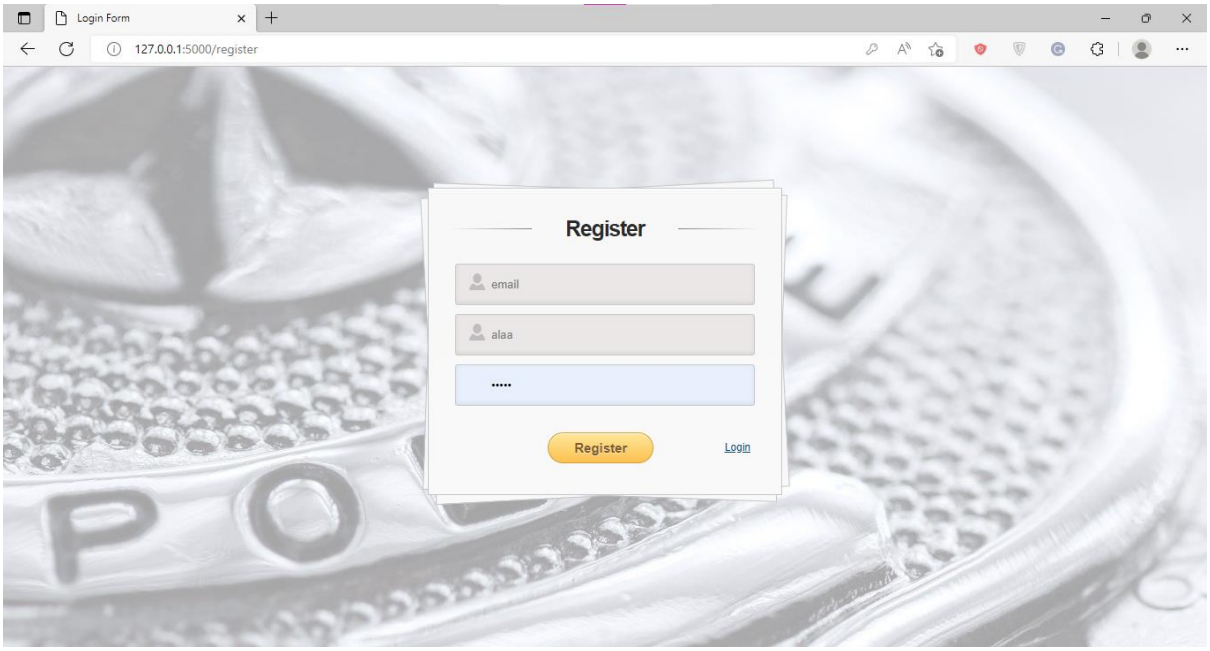

Figure 5.7 Demo screenshot (Register)

8. “Home page” is where the users will go when they successfully log in and it includes an introduction about the objective of the project along with a video talking about the importance of crime prevention.

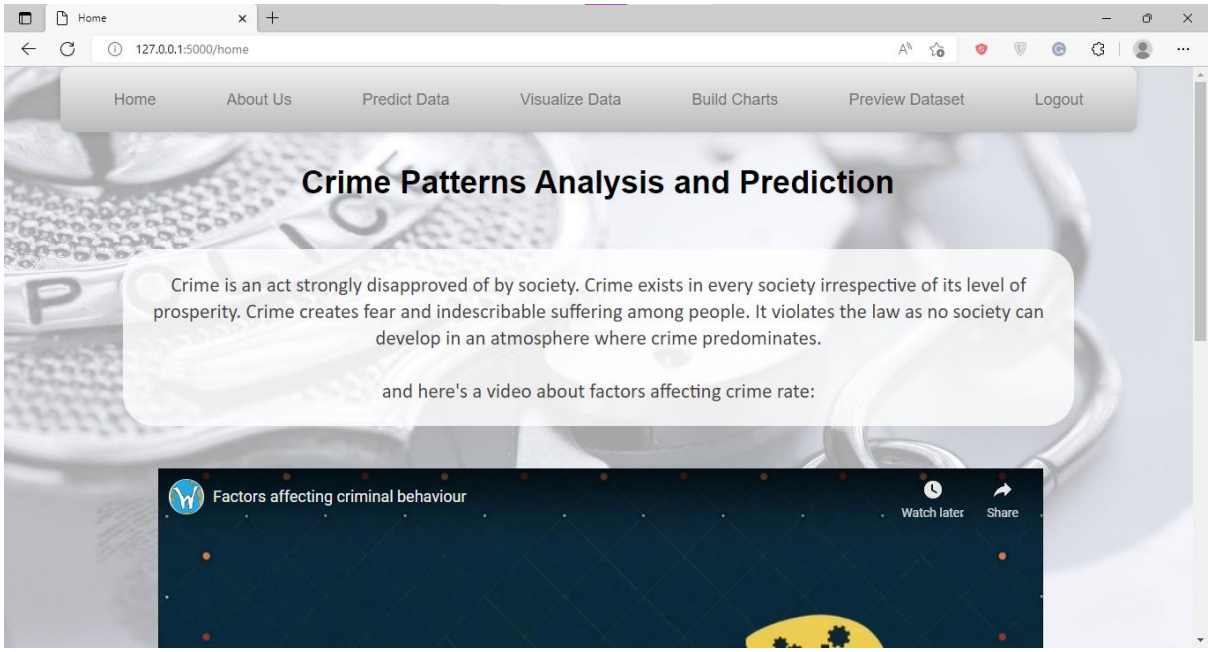

Figure 5.8 Demo screenshot (home page)

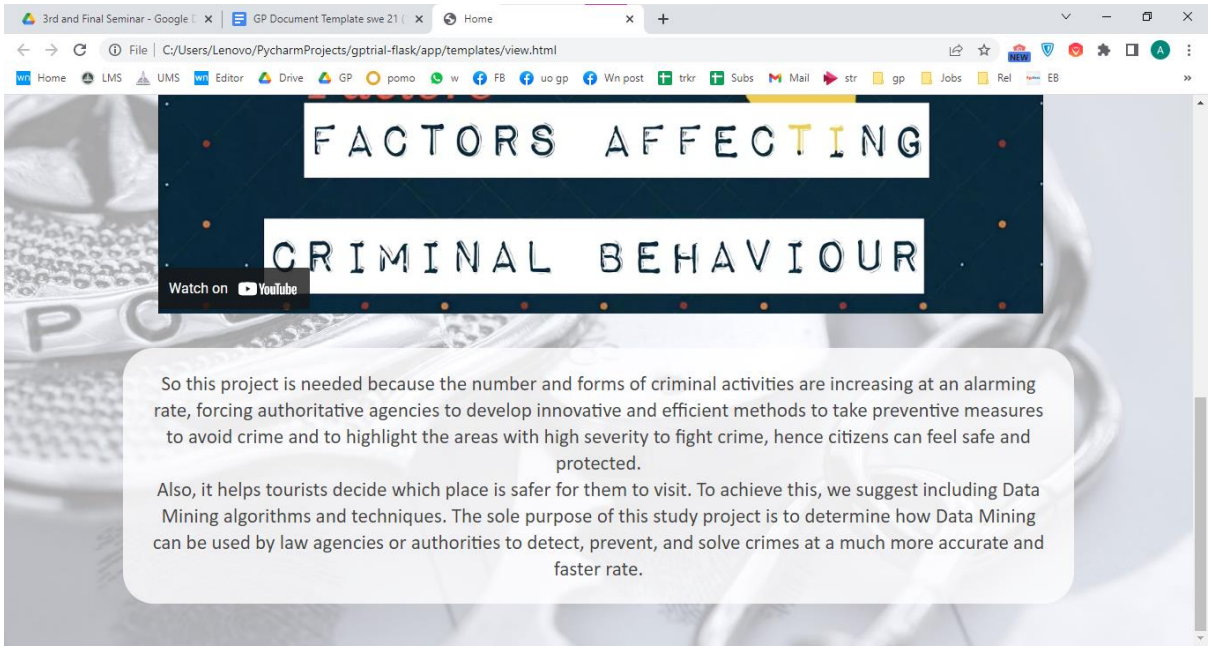

Figure 5.9 Demo screenshot (home page 2)

9. "About Us " page talks about the site and its objective.

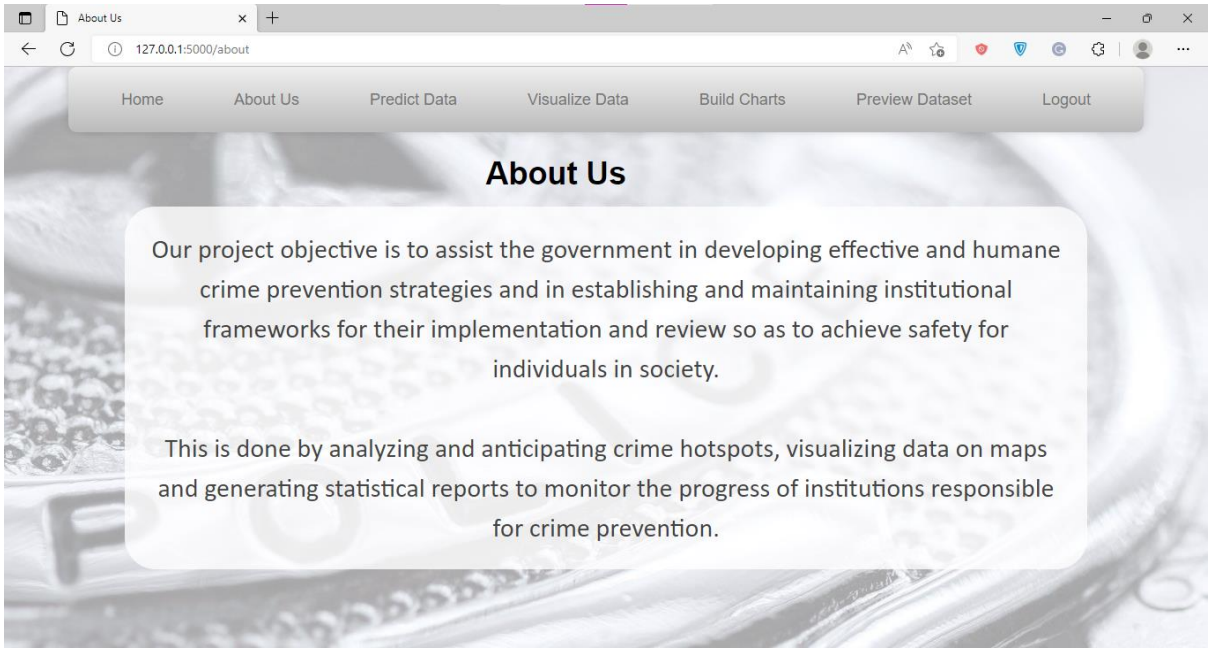

Figure 5.10 Demo screenshot (About Us page)

10. "Preview Data" in this page users have the ability to search for crime rate by entering the area name or any other attribute found in the dataset. then the system will display the crime rate that is related to the attribute entered by the user.

when it first runs, the full dataset is loaded:

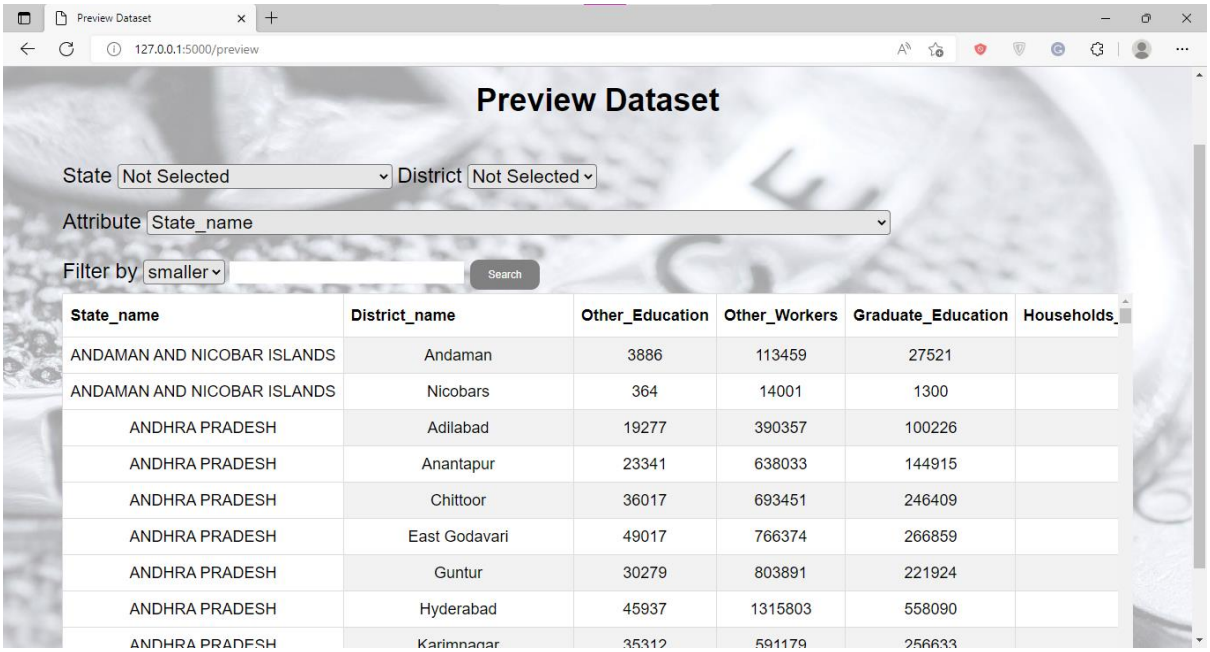

Figure 5.11 Demo screenshot (Preview Dataset page)

since the dataset is large, you can scroll up and down, right and left using scrollbars around the table:

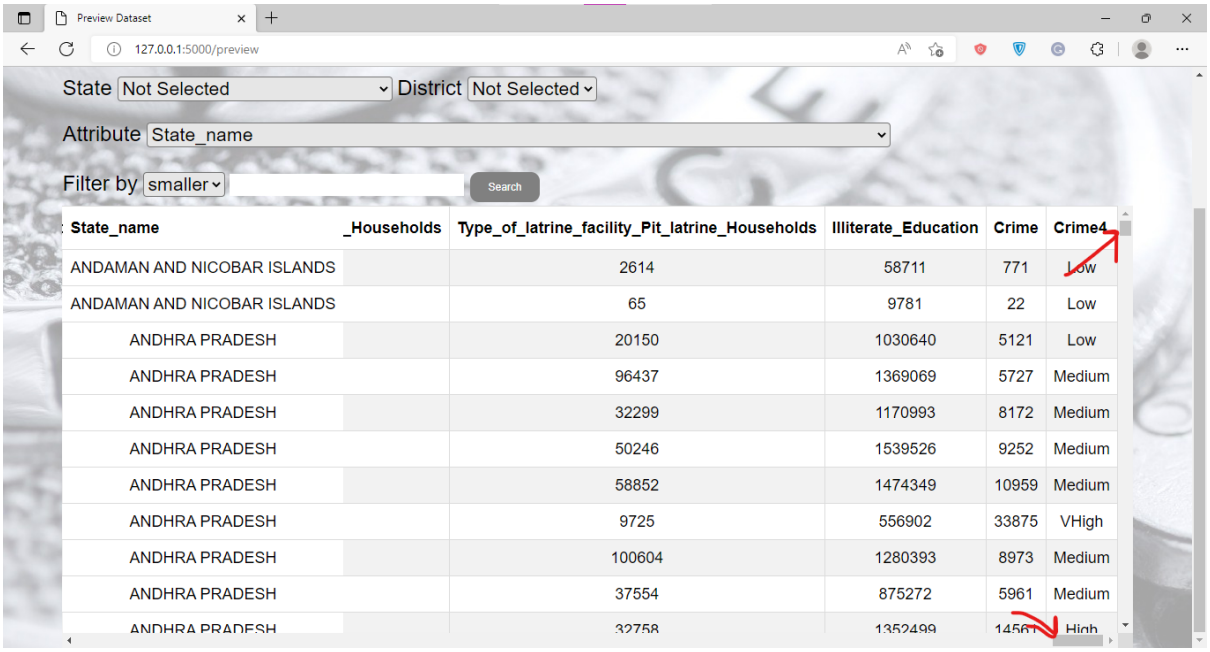

Figure 5.12 Demo screenshot (Preview Dataset page2)

you can also search for any row(s) using filtering by attribute values:

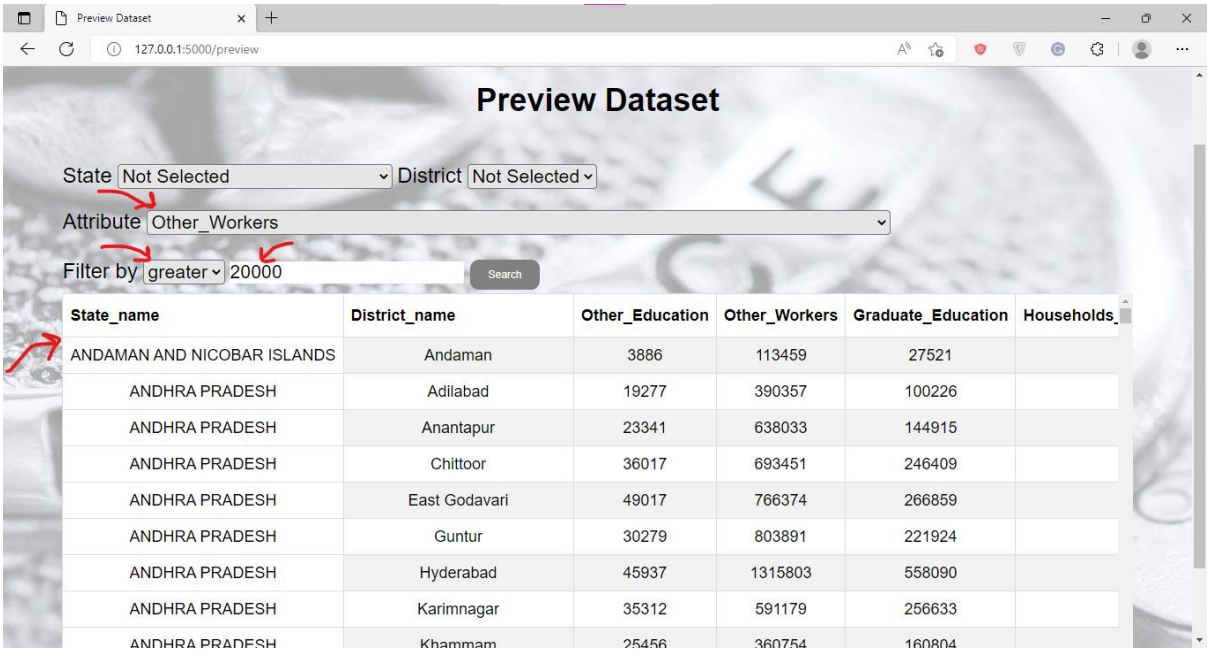

Figure 5.13 Demo screenshot (Preview Dataset page3)

you can also search in a specific state that contains multiple districts:

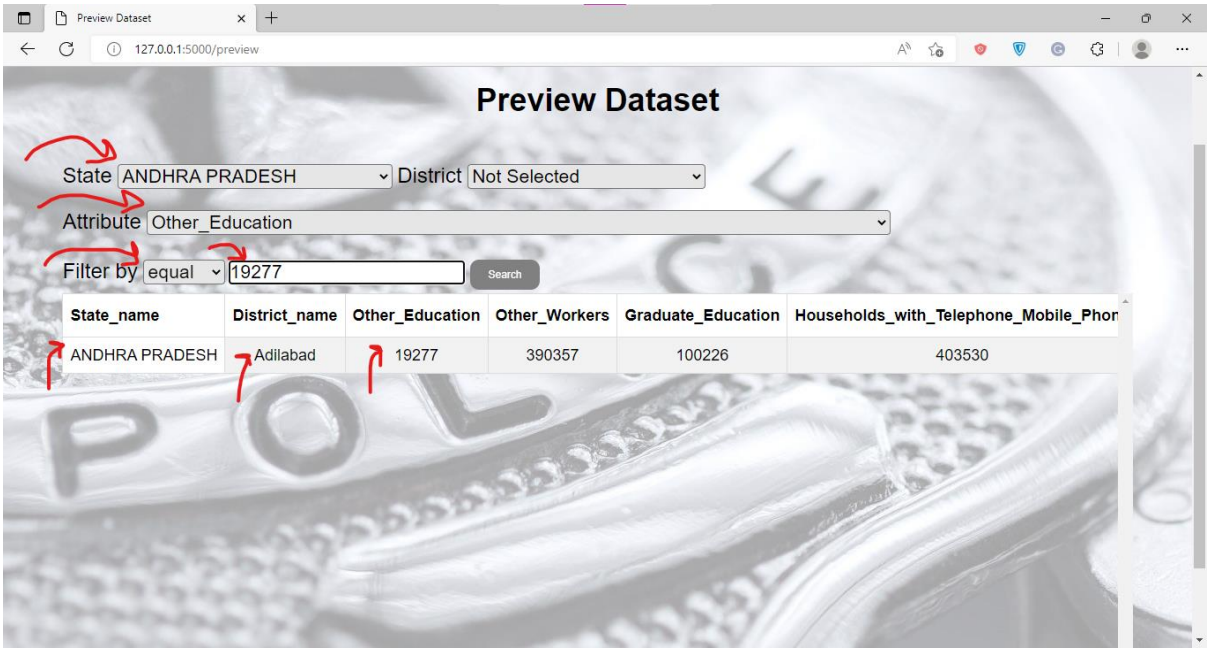

Figure 5.14 Demo screenshot (Preview Dataset page4)

11. “Visualize Individual Attributes” this page provides a functionality which is the visualization of crime hotspots of each district on a map with different colors that describe crime rate as follows: low as light orange color, medium as dark orange color, high as violet color and finally very high as black color.

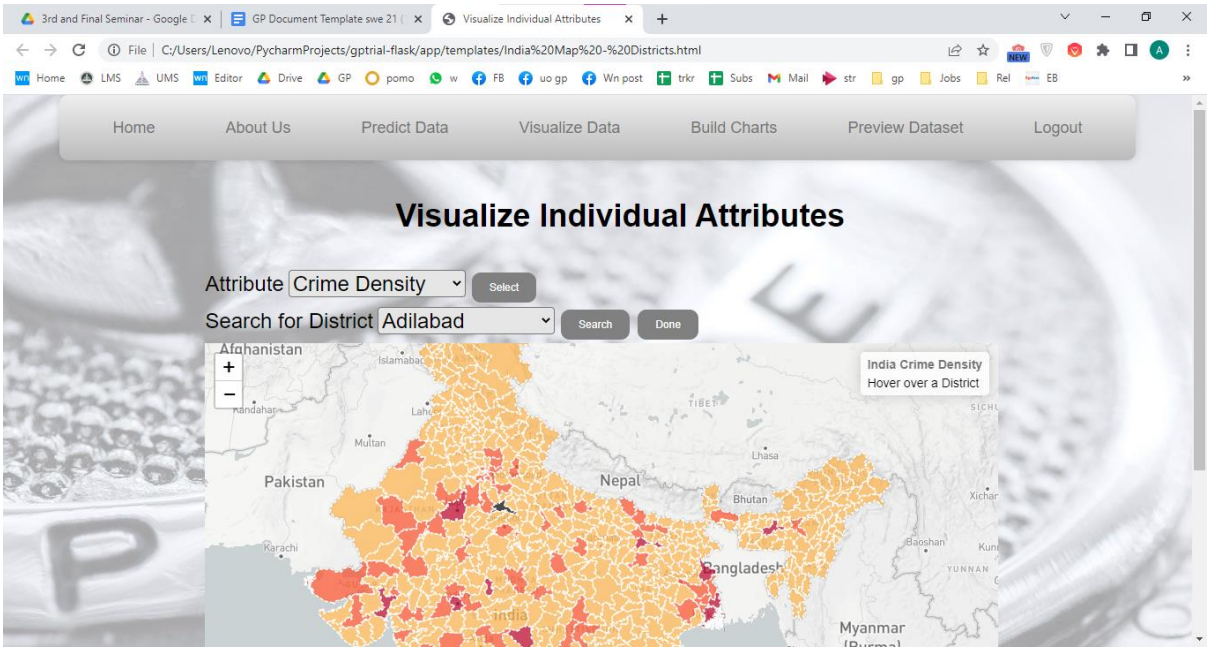

Figure 5.15 Demo screenshot (Visualize Individual Attributes page)

12. “Visualize Attributes Individually” in this page, the user can visualize the crime rate of each district. Users can also choose other attributes individually that the system specifies to visualize on the map with different colors that describe the values of each attribute.

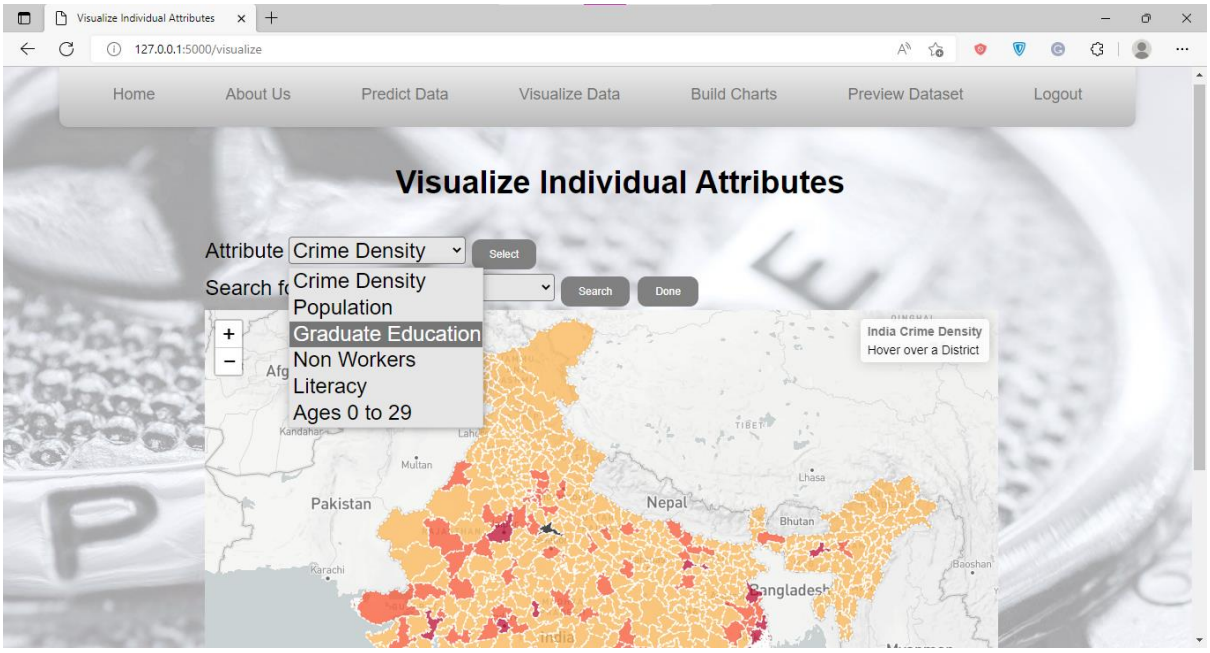

Figure 5.16 Demo screenshot (Visualize Individual Attributes page2)

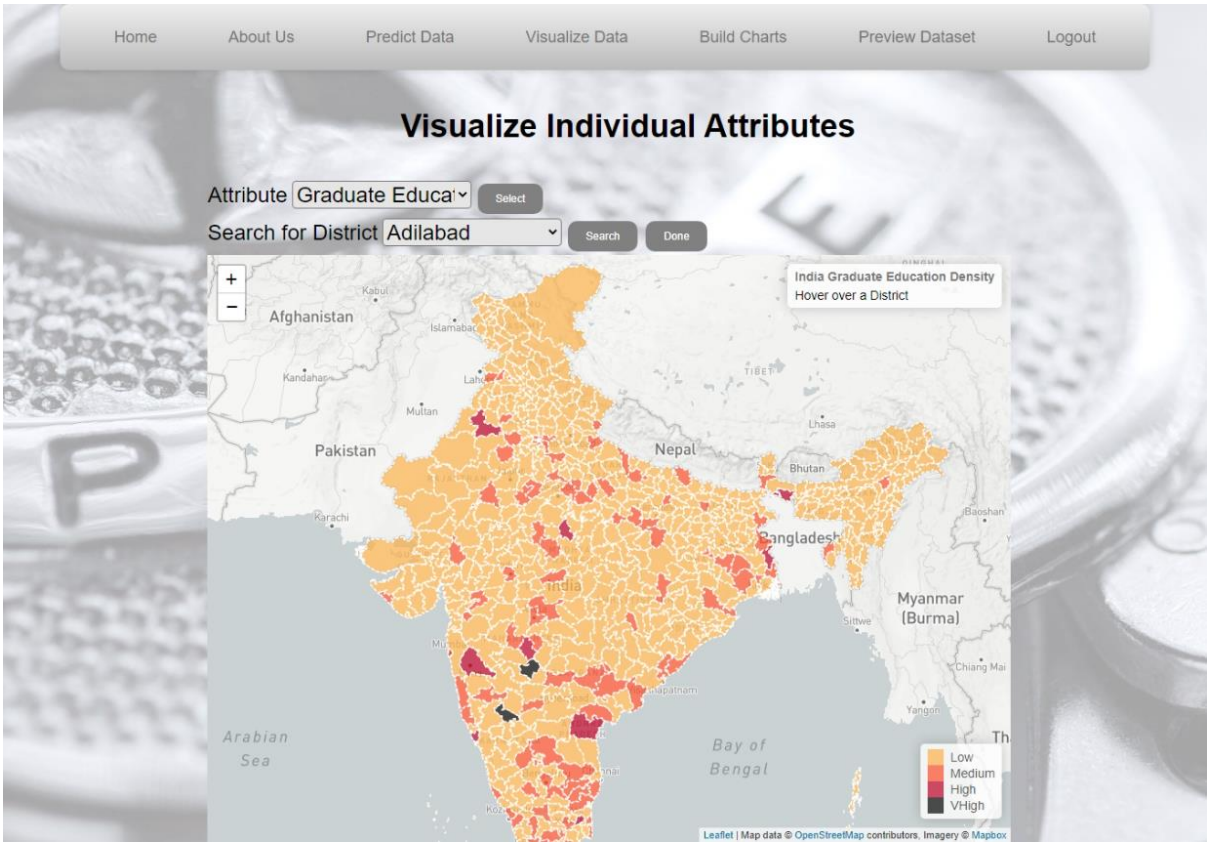

Figure 5.17 Demo screenshot (Visualize Individual Attributes page3)

Also, users can search for any district by choosing the name of the district from the drop down list then it will display on the map with a green color.

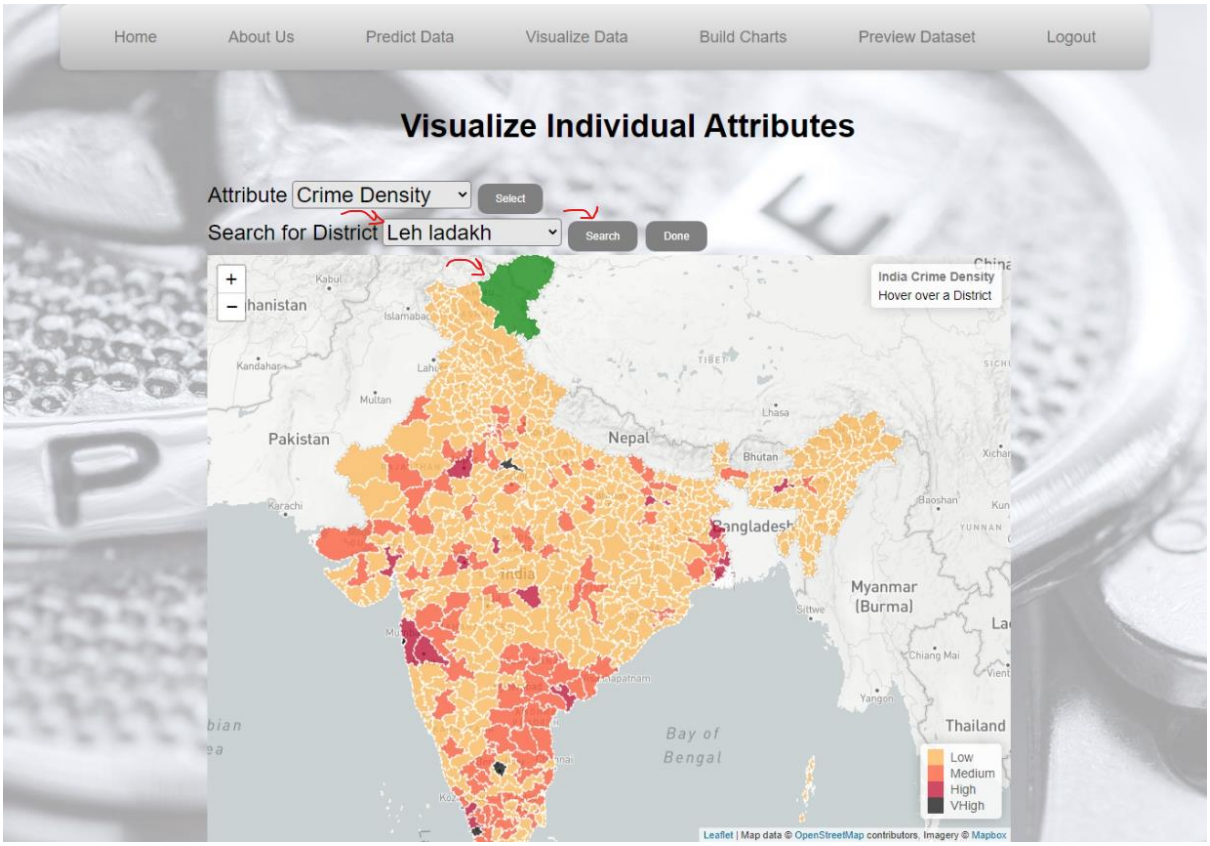

Figure 5.18 Demo screenshot (Visualize Individual Attributes page4)

13. “Visualize combined attributes” in this page,the user can visualize filtered districts based on value of attributes chosen by the user then press add after choosing each attribute and finally press done and submit buttons to see the results

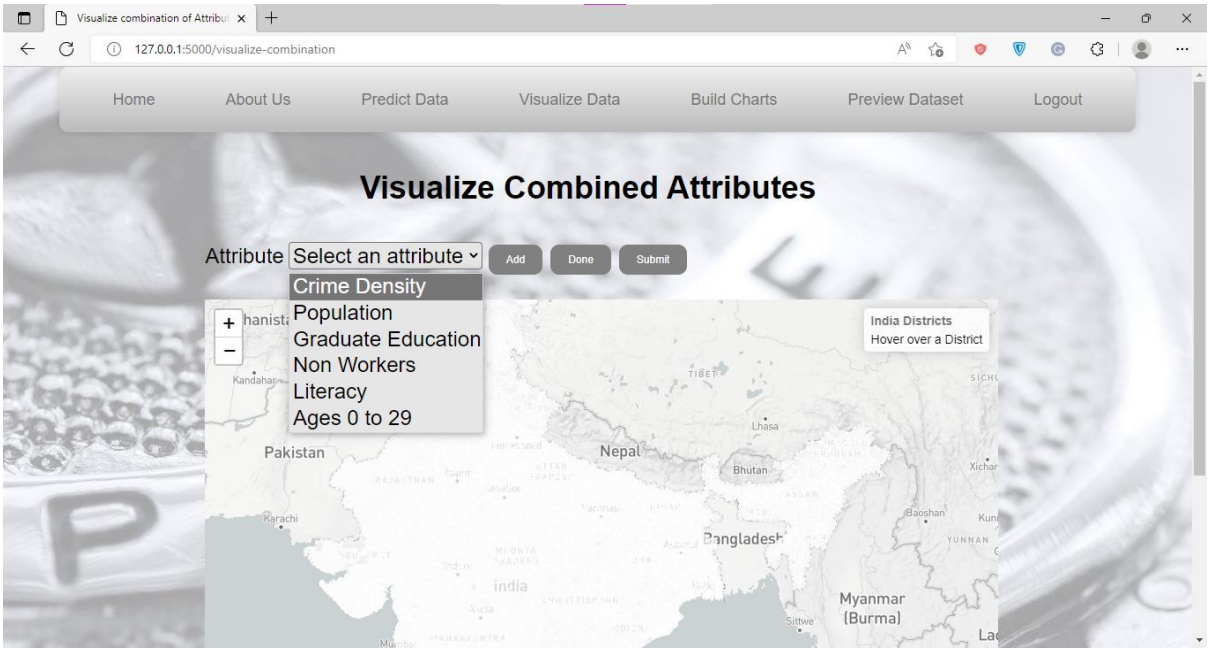

Figure 5.19 Demo screenshot (Visualize Combination Attributes page)

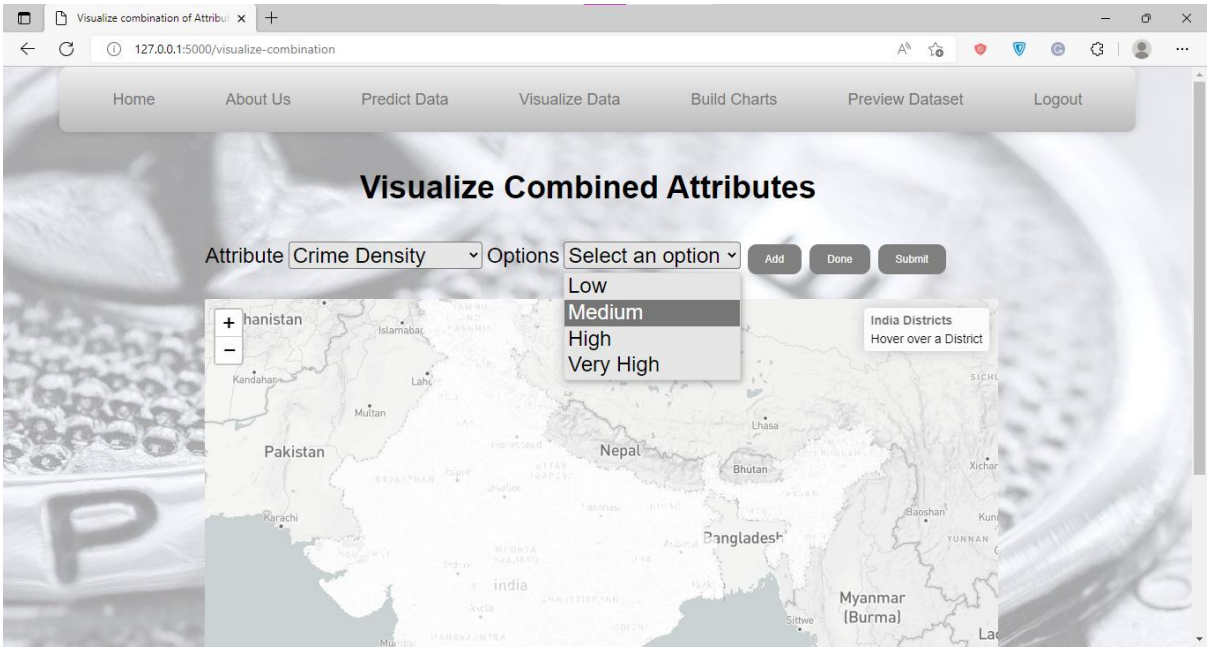

Figure 5.20 Demo screenshot (Visualize Combination Attributes page2)

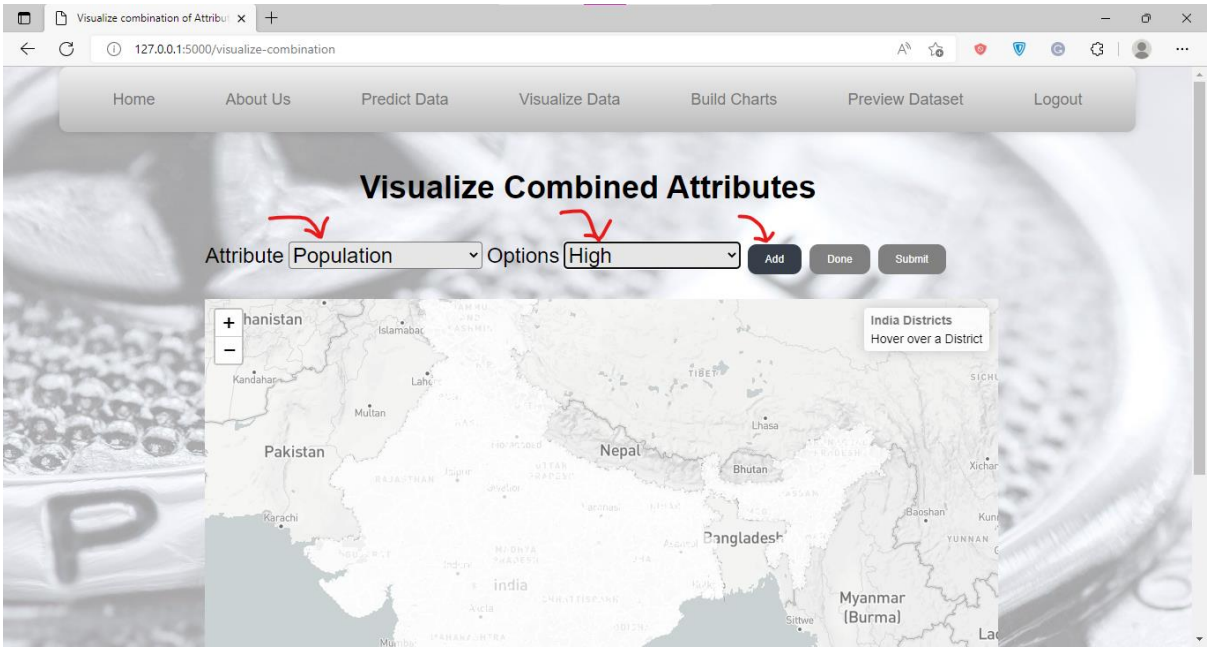

Figure 5.21 Demo screenshot (Visualize Combination Attributes page3)

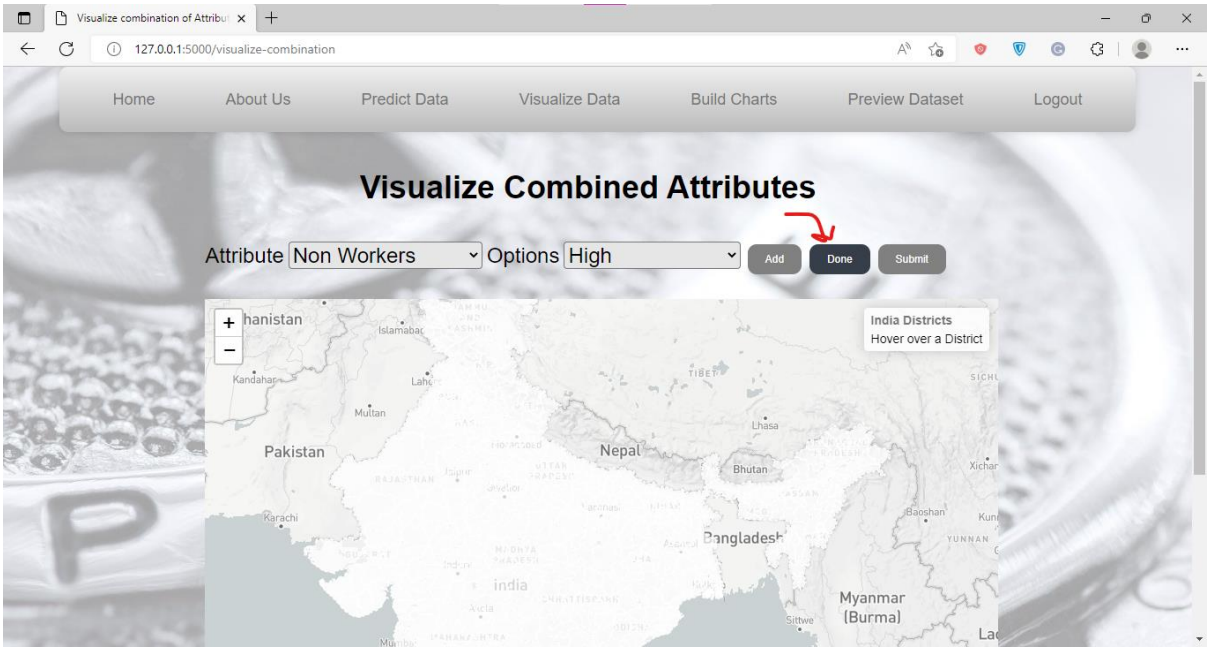

Figure 5.22 Demo screenshot (Visualize Combination Attributes page4)

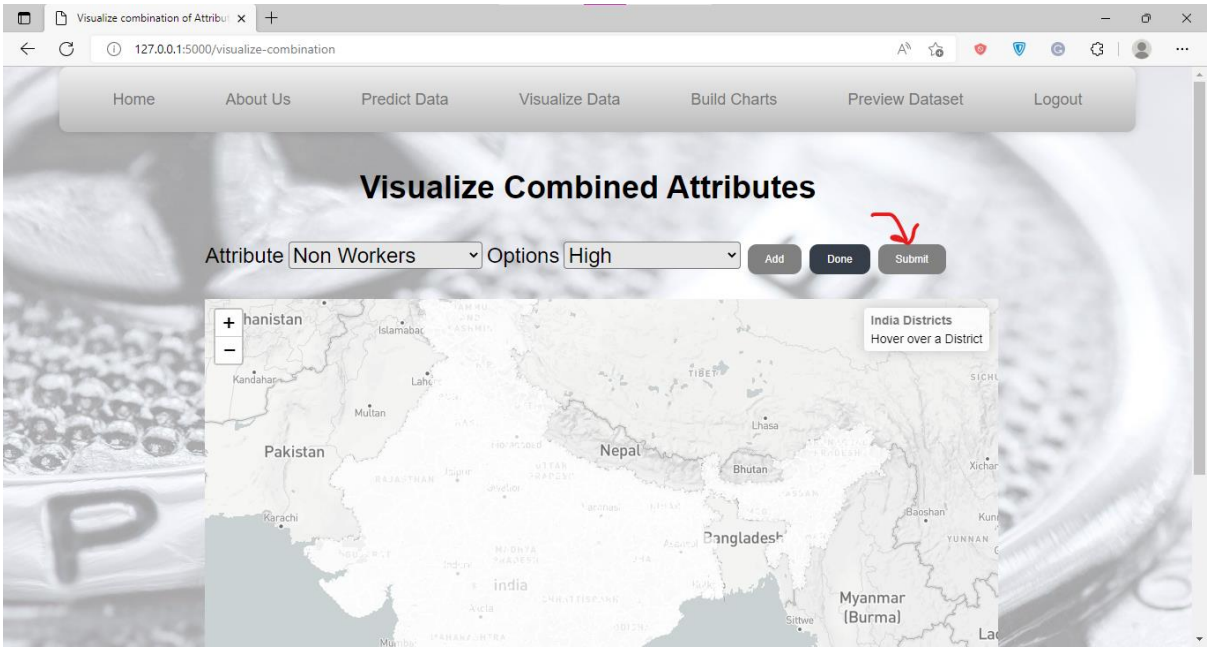

Figure 5.23 Demo screenshot (Visualize Combination Attributes page5)

14. “Visualize Top X Crime Hotspots “ This page allows users to enter a number between 1 and 633 which is the number of districts found in the dataset then visualize the districts on the map that have a crime rate that is equal to or greater than the number entered by the user with their crime rates numbers.

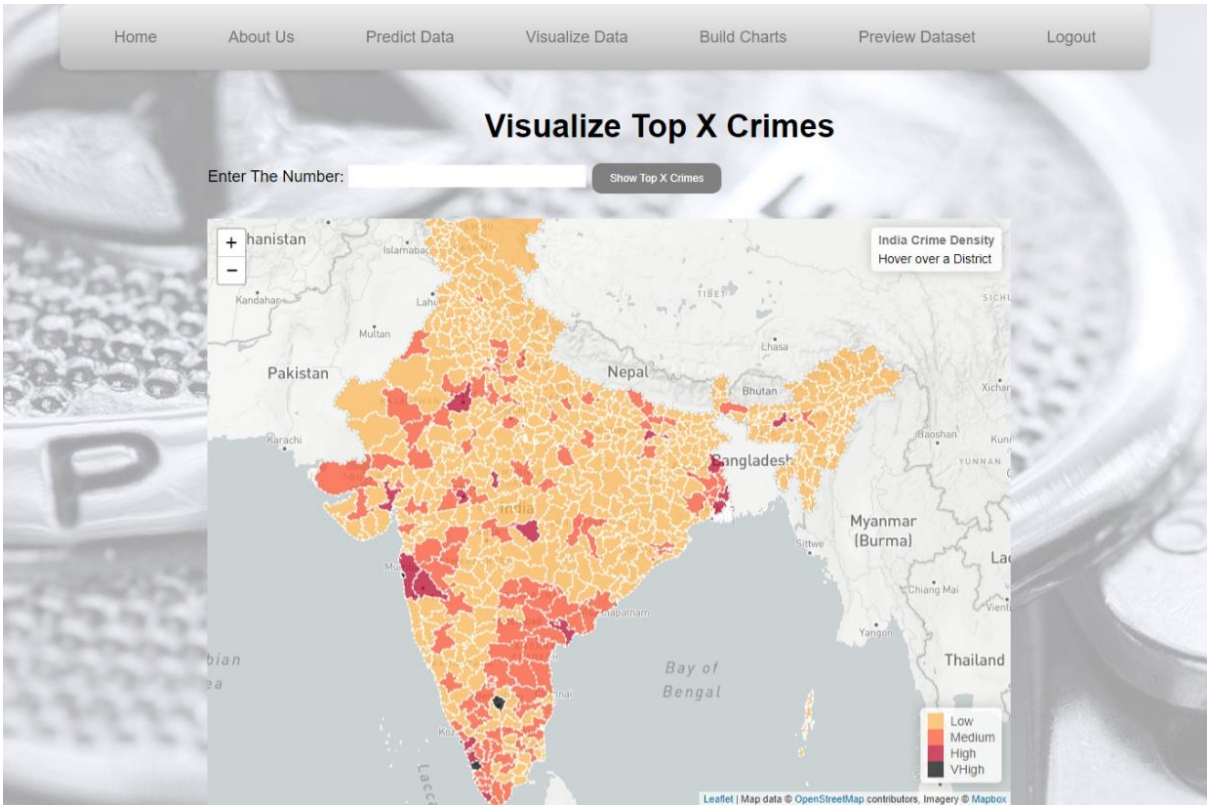

Figure 5.24 Demo screenshot (Visualize Top X Crime page)

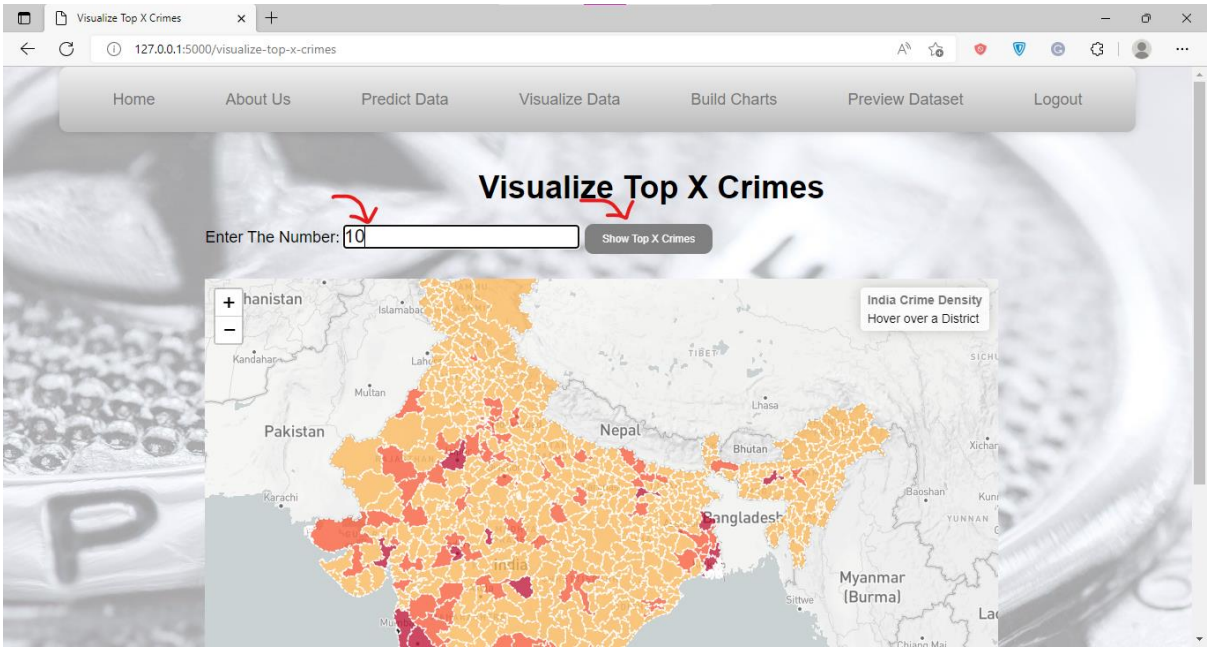

Figure 5.25 Demo screenshot (Visualize Top X Crime page2)

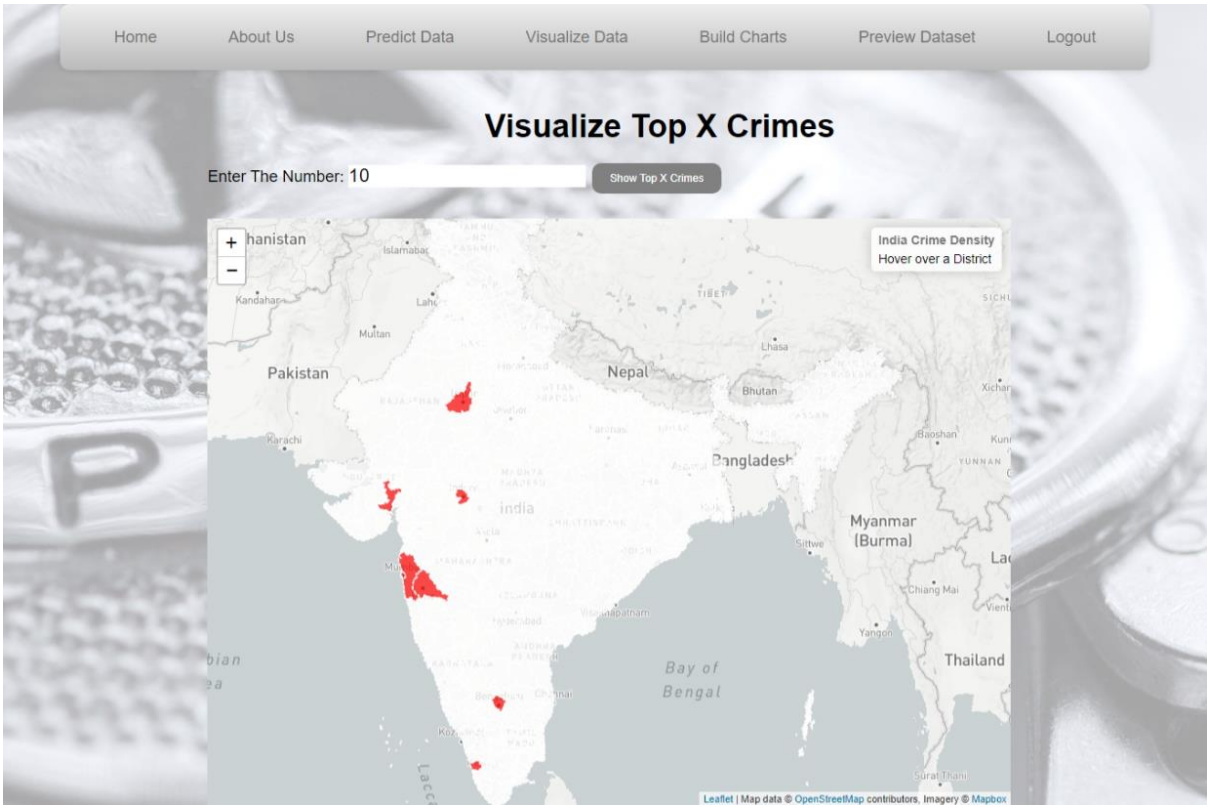

Figure 5.26 Demo screenshot (Visualize Top X Crime page3)

15. “Pie Chart” in this page the user is able to generate pie charts that display the top x districts that have the highest crime rate.

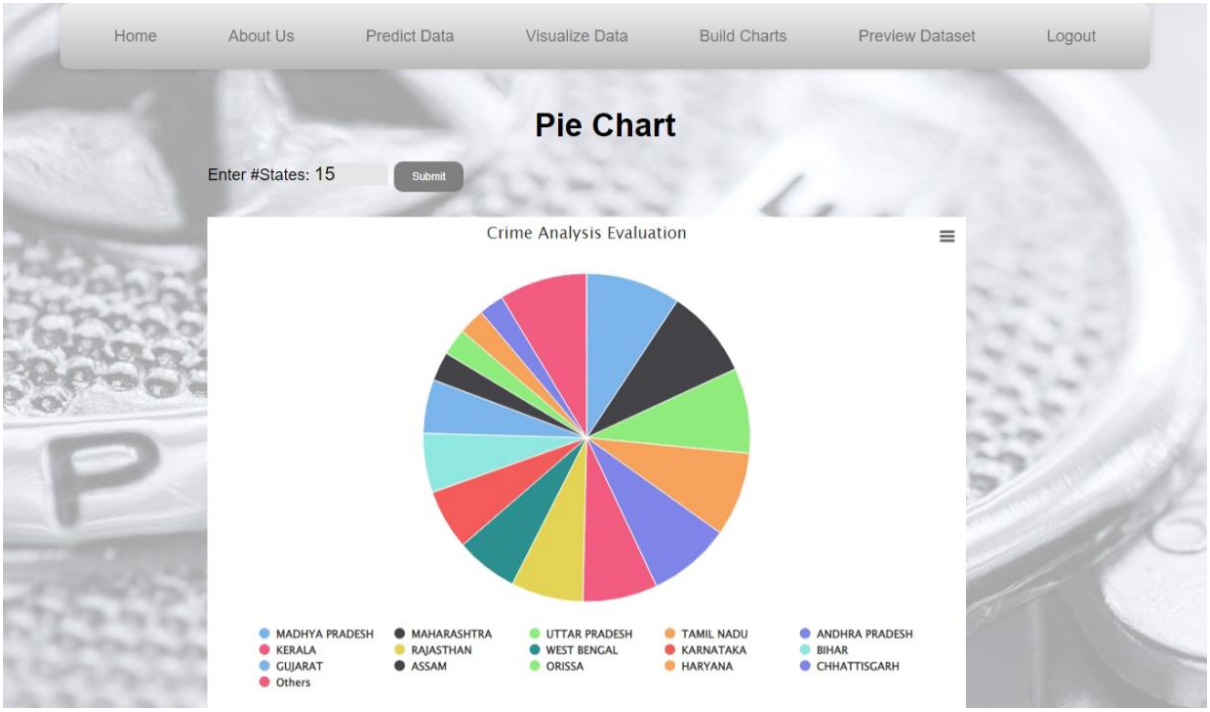

Figure 5.27 Demo screenshot (Pie Chart page)

you can hover on any section and it will show the name of the state:

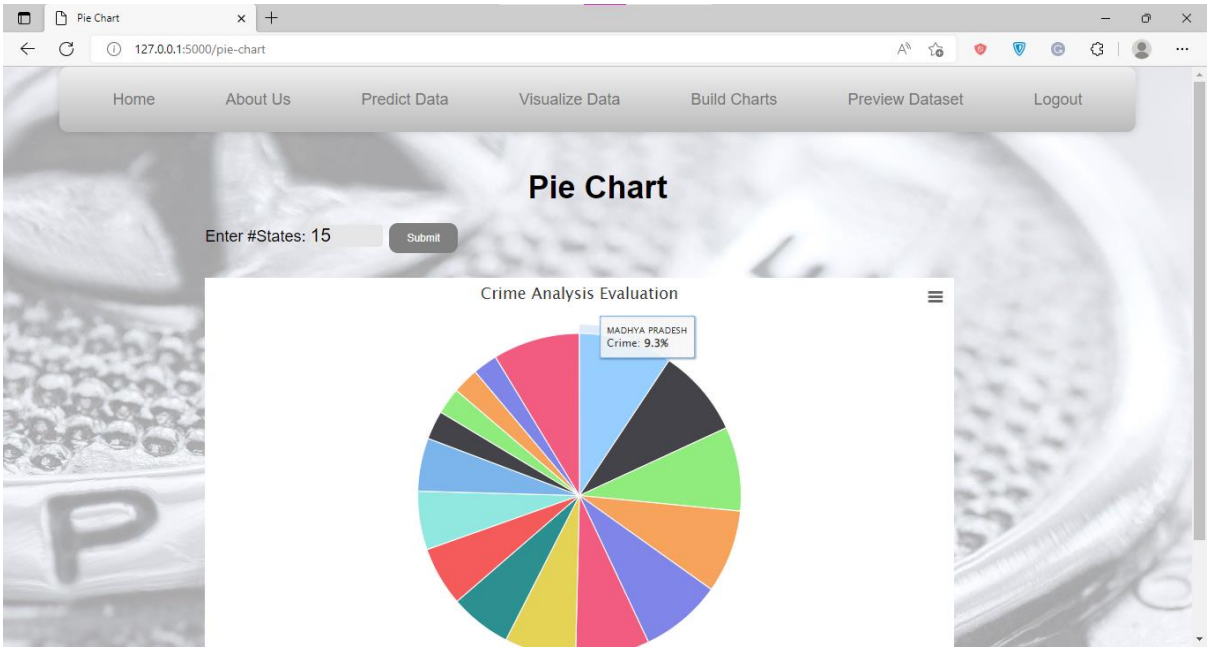

Figure 5.28 Demo screenshot (Pie Chart page2)

16. “Stacked Chart” in this page allows users to generate stacked charts that display the top x districts with their top x districts in crime rate, also users can choose different attributes from the drop down list to visualize.

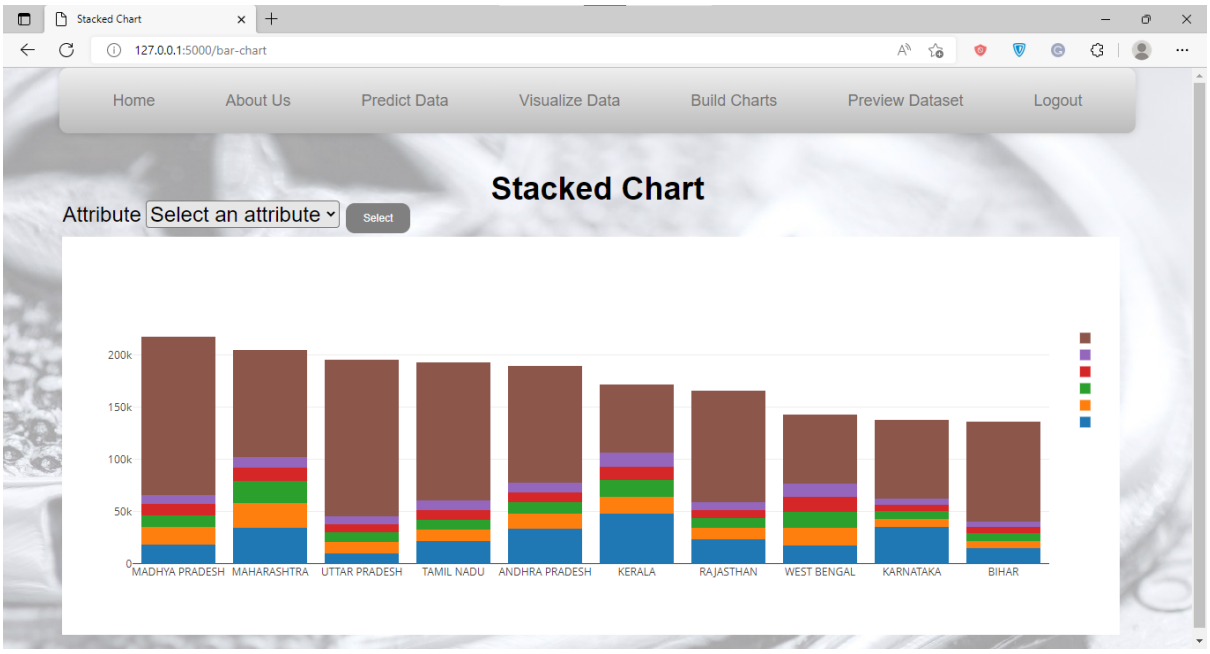

Figure 5.29 Demo screenshot (Stacked Chart page)

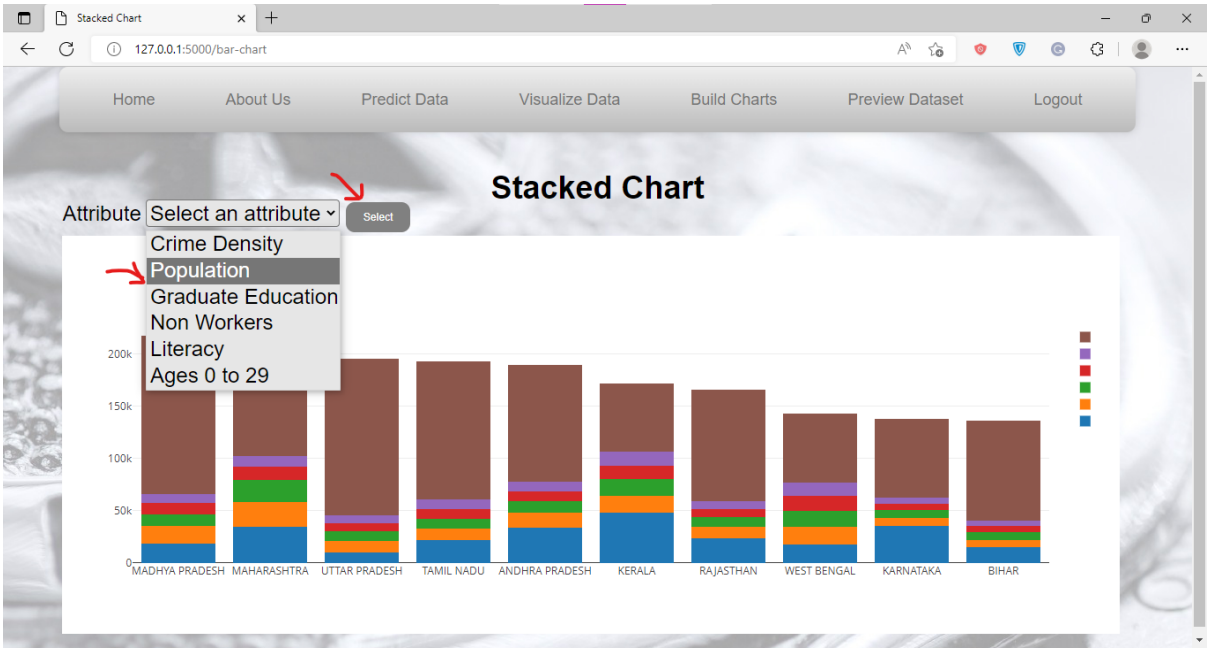

Figure 5.30 Demo screenshot (Stacked Chart page2)

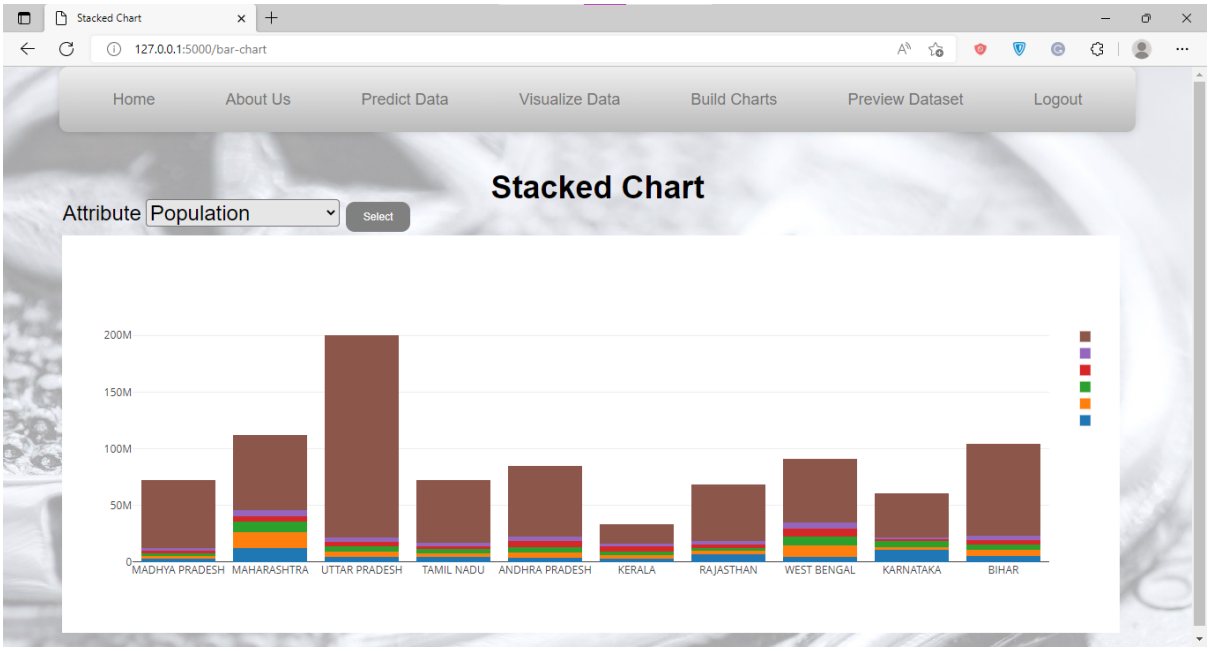

Figure 5.31 Demo screenshot (Stacked Chart page3)

17. **“**Histogram**”** in this page, users are able to generate a histogram displaying the values of crime rate with a number of occurrences of each value, also users are able to choose different attributes from the drop down list to visualize.

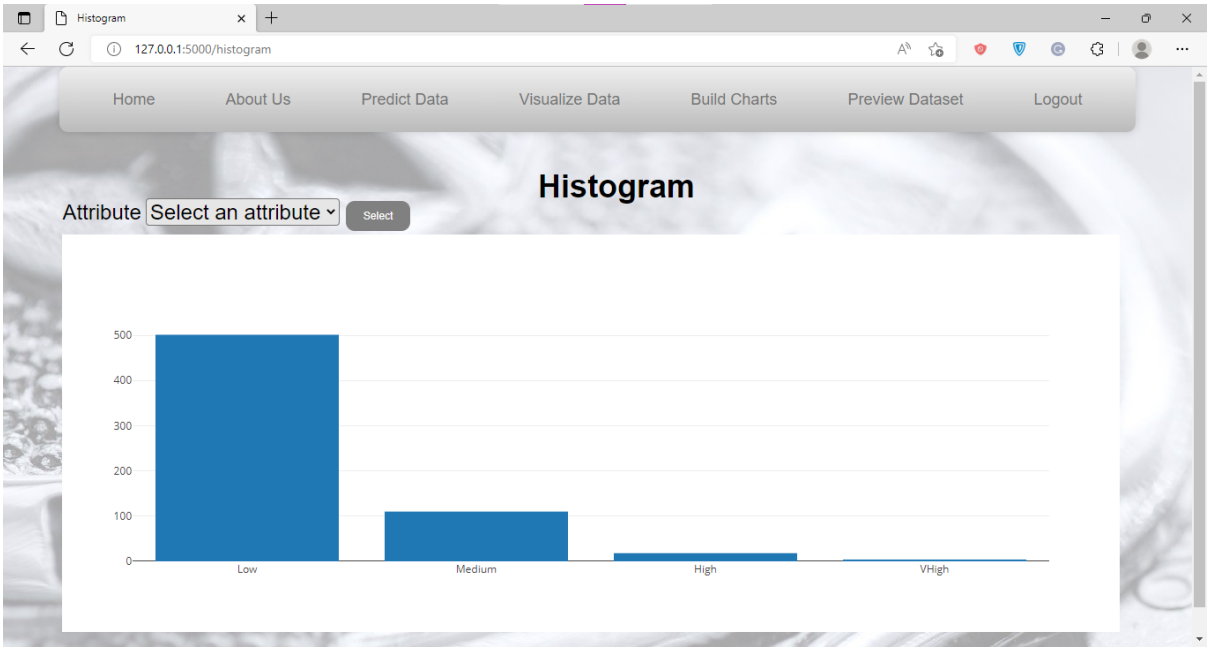

Figure 5.32 Demo screenshot (Histogram page)

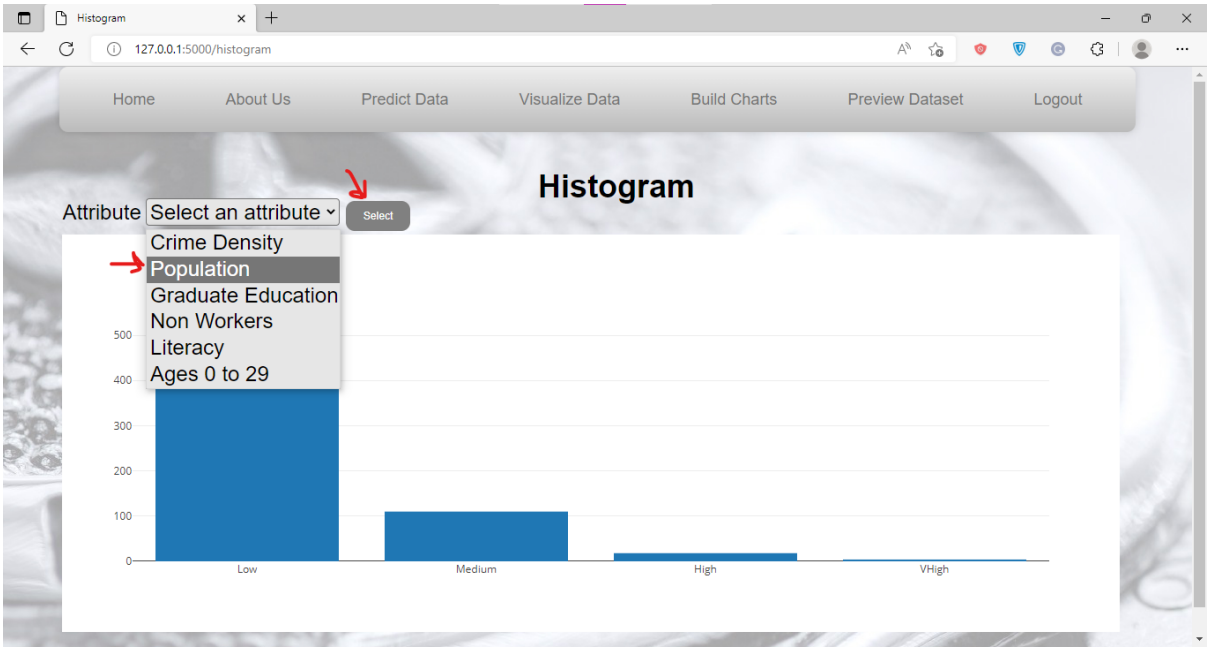

Figure 5.33 Demo screenshot (Histogram page2)

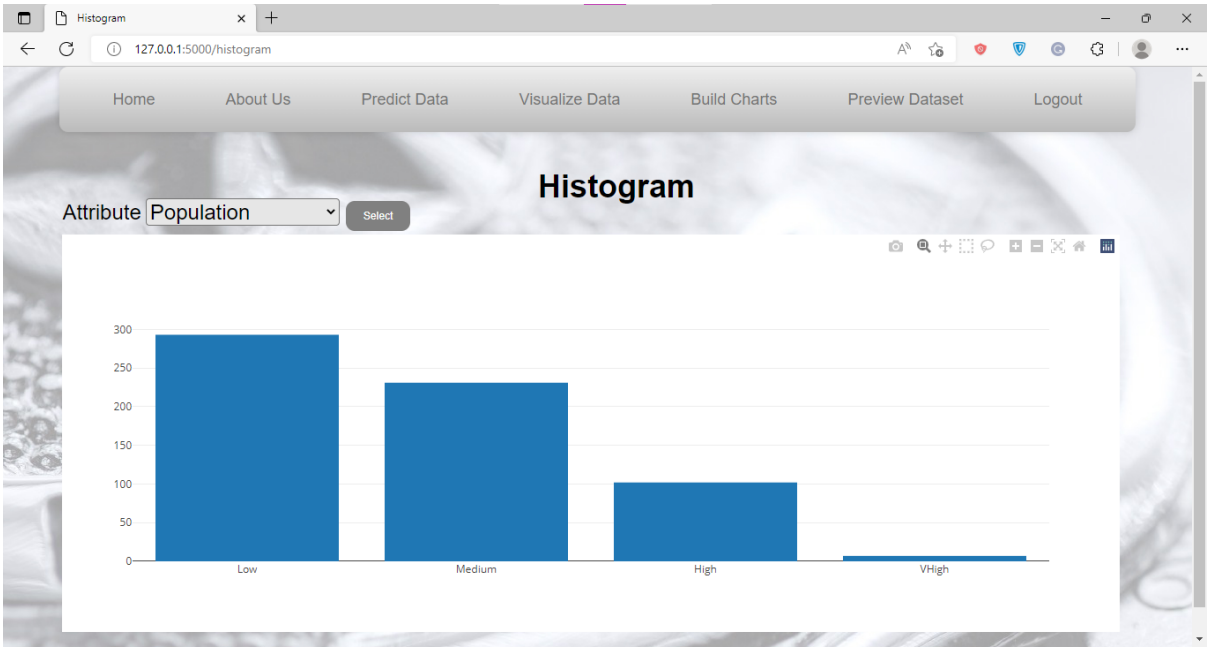

Figure 5.34 Demo screenshot (Histogram page3)

18. “Predict Data“ The system can predict a future crime hotspot by allowing the user to change any attribute value in the record displayed by the system then reclassifying it and showing the results of the prediction to the user.

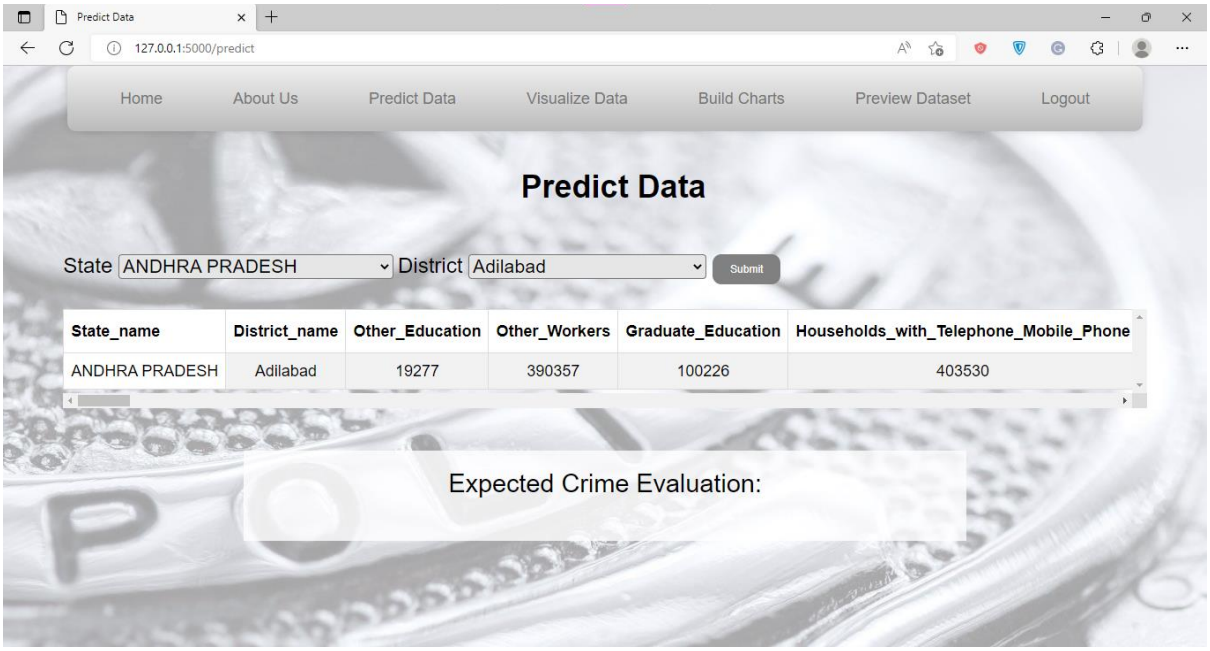

Figure 5.35 Demo screenshot (Predict page)

choose which row by choosing the state and the district, first district in the first state is chosen by default

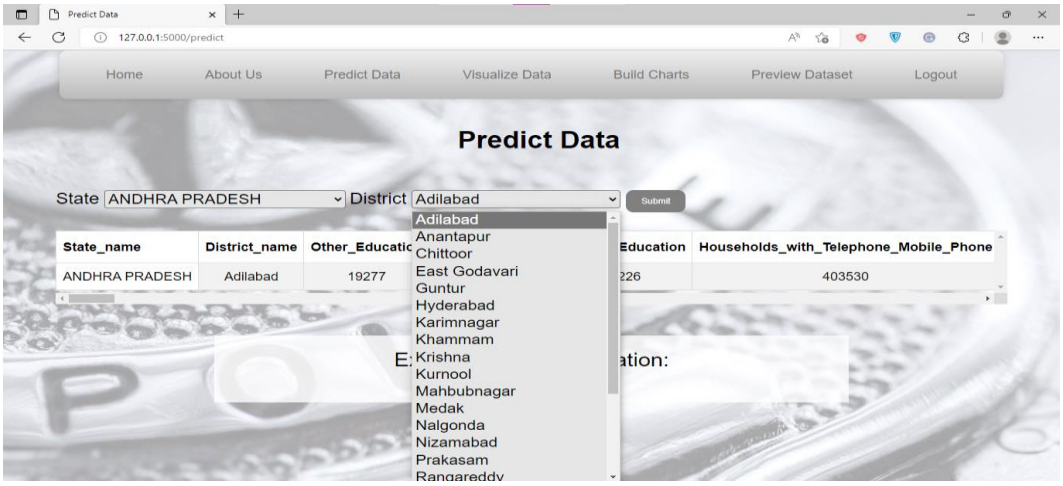

Figure 5.36 Demo screenshot (Predict page2)

change values of any number of attributes, then click submit.

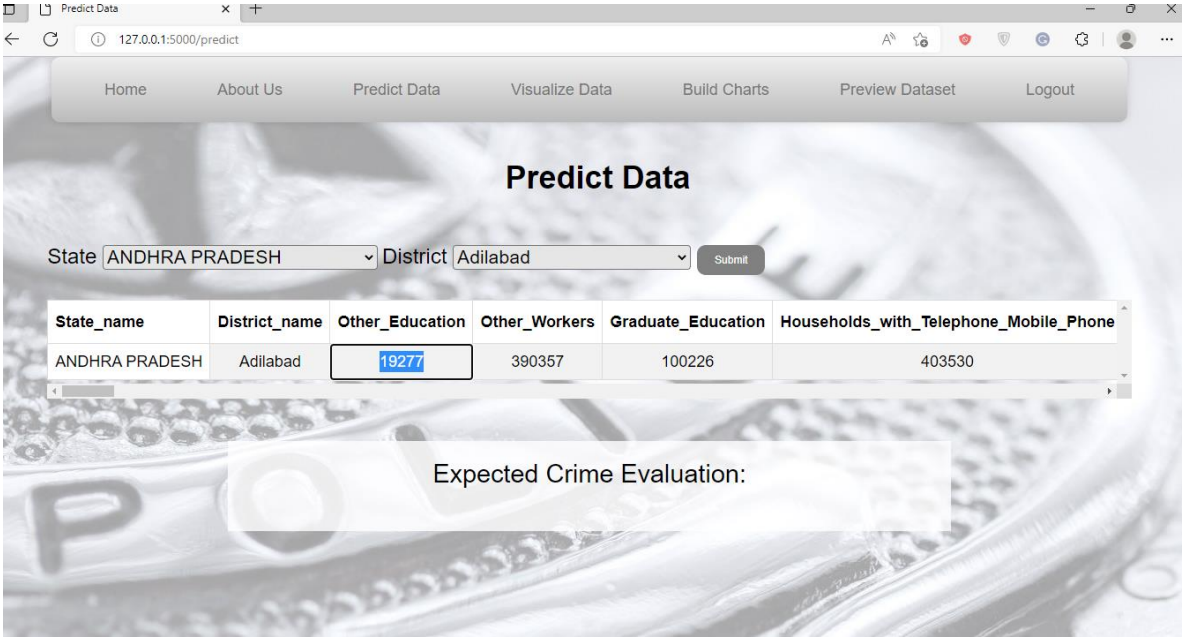

Figure 5.37 Demo screenshot (Predict page3)

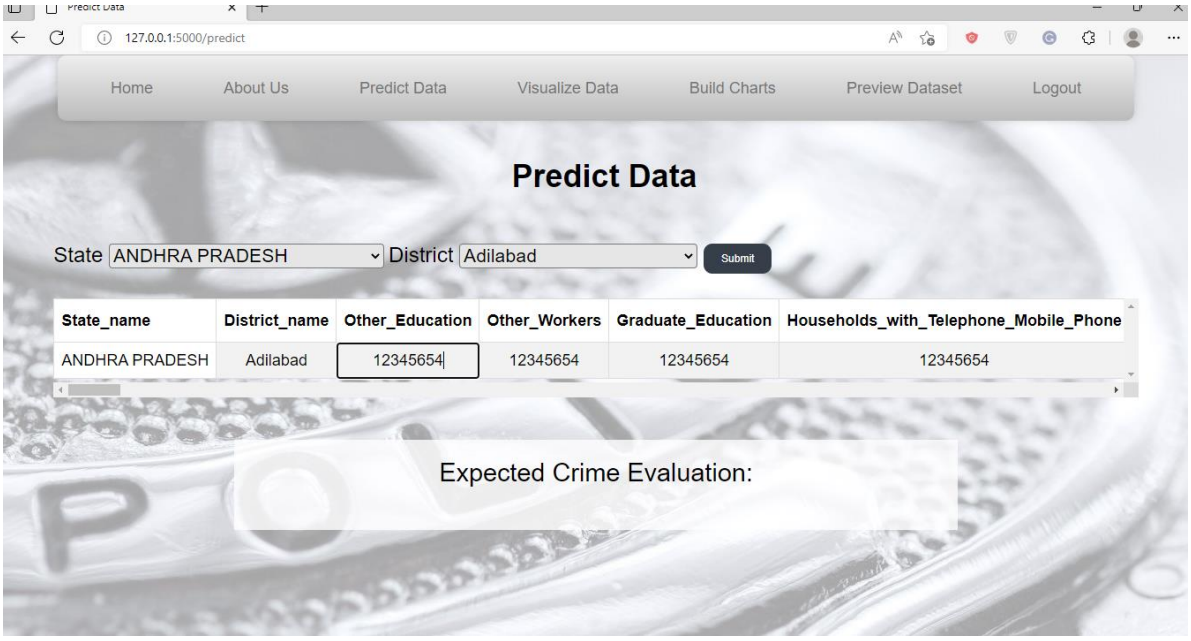

Figure 5.38 Demo screenshot (Predict page4)

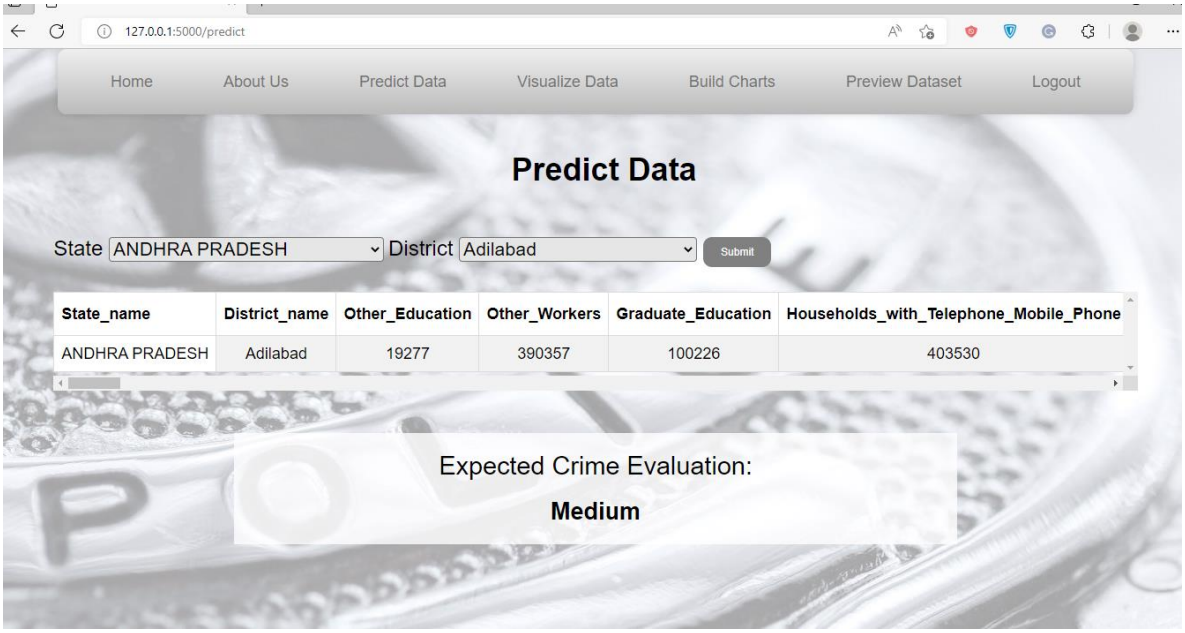

Figure 5.39 Demo screenshot (Predict page5)

Chapter 6
Conclusions and Future Work

Chapter 6
Conclusions and Future Work

6.1 Conclusions

The quality of life is improving due to rapid economic growth and the development of science and technology, but various social problems are also rapidly increasing, especially criminal activities. Crime problems occurring in recent years are becoming more intelligent and diversified than before, and violent crimes are increasing rapidly.

Crimes that have already occurred are irreversible, and even if a criminal is punished for the crime they committed, the pain of the victim and their family cannot be healed. Therefore, in this project, a job was conducted to prevent and respond to crimes by predicting crimes based on data mining techniques.

In order to predict crime rates, we used Data Mining algorithms to use demographics and crime datasets for the prediction of crime. In that function, we tested the following Data Mining algorithms: Logistic regression, Naïve Bayes, Random Forest, and Decision Tree.

Among those techniques, Random Forest provided the highest accuracy when we used it to process our dataset which is why we used Random Forest for the prediction functionality. We use this prediction technique to predict the future crime hotspots.

6.2 Future Work

This paper presented the techniques and methods that can be used to predict crime and help law agencies. The scope of using different methods for crime prediction and prevention can change the scenario of law enforcement agencies. Using DM can substantially impact the overall functionality of law enforcement agencies. In the near future, by combining DM, along with security equipment such as surveillance cameras and spotting scopes, a Machine can learn the pattern of previous crimes, understand what crime actually is, and predict future crimes accurately without human intervention.

Possible automation would be to create a system that can predict and anticipate the zones of crime hotspots in a city. Law enforcement agencies can be warned and can prevent crime from occurring by implementing more surveillance within a zone. This complete automation can overcome the drawbacks of the current system, and law enforcement agencies can depend more on these techniques in the near future.

Designing a Machine to anticipate and identify patterns of such crimes will be the starting point of our future study. Although the current systems have a large impact on crime prevention, this could be the next big approach and bring about a revolutionary change in the crime rate, prediction, detection, and prevention.

References

[1] GONZALEZ, Joana & Leboulluec, Aera. (2019). Crime Prediction and Socio-Demographic Factors: A Comparative Study of Machine Learning Regression-Based Algorithms. Journal of Applied Computer Science & Mathematics. 13. 13-18. 10.4316/JACSM.201901002.

[2] Ramasubbareddy, S., Aditya Sai Srinivas, T., Govinda, K., Manivannan, S.S. (2020). Crime Prediction System. In: Saini, H., Sayal, R., Buyya, R., Aliseri, G. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 103. Springer,Singapore.

[3] De Nadai, M., Xu, Y., Letouzé, E. *et al.* Socio-economic, built environment, and mobility conditions associated with crime: a study of multiple cities. *Sci Rep***10,** 13871 (2020).

[4] Mittal, Mamta & Goyal, Lalit & Sethi, Jasleen & D, Jude. (2019). Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning. Computational Economics. 53. 10.1007/s10614-018-9821-x.

[5] Mahmud, Sakib & Nuha, Musfika & Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. 10.1007/978-981-15-7394-1_5.

[6] Ali, Wasim A., Husam Alalloush, and K. N. Manasa. "CRIME ANALYSIS AND PREDICTION USING
K-MEANS CLUSTERING TECHNIQUE." EPRA International Journal of Economic and Business Review
(August 2020)(EPRA IJRD Volume: 5| Issue: 7| July 2020).

[7] Prathap, Boppuru Rudra, and K. Ramesha. "Geospatial crime analysis to determine crime density using
Kernel density estimation for the Indian context." J. Comput. Theor. Nanosci 17.1 (2020): 74-86.

[8] Albahli, Saleh, et al. "Predicting the type of crime: Intelligence gathering and crime analysis." Computers,
Materials & Continua 66.3 (2021): 2317-2341.

[9] Hajela, Gaurav, Meenu Chawla, and Akhtar Rasool. "A multi-dimensional crime spatial pattern analysis
and prediction model based on classification." ETRI Journal 43.2 (2021): 272-287.

[10] De Nadai, Marco, et al. "Socio-economic, built environment, and mobility conditions associated with crime:
a study of multiple cities." Scientific reports 10.1 (2020): 1-12.

[11] Ramasubbareddy, Somula, et al. "Crime prediction system." Innovations in Computer Science and Engineering. Springer, Singapore, 2020. 127-134.