University of Mumbai

Identification of Crime Prone Areas

Submitted in partial fulfillment of requirements

for the degree of

Bachelors in Technology

by

Dhruv Doshi

Roll No: 1814002

Shubham Bhakuni

Roll No: 1814006

Labdhi Jain

Roll No: 1814015

Kunj Gala

1814021

Guide

Dr. Irfan Siddavatam

Department of Information Technology
K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)
Batch 2022

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate

This is to certify that the dissertation report entitled **Identification of Crime Prone Areas** is bona fide record of dissertation work done by **Dhruv Doshi, Shubham Bhakuni Labdhi Jain and Kunj Gala** in the year 2021-22 under the guidance of **Dr. Irfan Siddavatam** of Department of Information Technology in partial fulfillment of requirement for the Bachelors in Technology degree in Information Technology of University of Mumbai.

Guide

Head of the Department

Principal

Date: 20/04/2021

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate of Approval of Examiners

We certify that this dissertation report entitled **Identification of Crime Prone Areas** is bona fide record of project work done by Dhruv Doshi, Shubham Bhakuni, Labdhi Jain and Kunj Gala.

This project is approved for the award of Bachelors in Technology Degree in Information Technology of University of Mumbai.

Internal Examiner

External Examiner

Date: 20/04/2022

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

DECLARATION

We declare that this written report submission represents the work done based on our and / or others' ideas with adequately cited and referenced the original source. We also declare that we have adhered to all principles of intellectual property, academic honesty and integrity as we have not misinterpreted or fabricated or falsified any idea/data/fact/source/original work/ matter in our submission.

We understand that any violation of the above will be cause for disciplinary action by the college and may evoke the penal action from the sources which have not been properly cited or from whom proper permission is not sought.

Signature of the Student **1814002** **Roll No.**	**Signature of the Student** **1814006** **Roll No.**
Signature of the Student **1814015** **Roll No.**	**Signature of the Student** **1814021** **Roll No.**

Date: 20/04/2022

Place: Mumbai-77

Dedicated to

Family, friends and teachers…

Contents

CHAPTER 1 1

Introduction 1

1.1 Problem Definition: 1

1.2 Motivation: 1

1.3 Scope of the project: 2

1.4 Organization of the Thesis: 3

CHAPTER 2 4

Background Work 4

CHAPTER 3 8

Software Project Management Plan 8

3.1 Introduction: 8

3.2 Project Organization: 10

CHAPTER 4 22

Software Requirements Specification 22

4.1 Introduction: 22

4.2 Specific Requirements: 23

4.3 Functional Requirements: 25

4.4 Software System Attributes 27

4.5 Database Requirements 28

CHAPTER 5 29

Software Design Description 29

5.1 Introduction: 29

5.2 System Architectural Design: 29

5.3 Detailed description of components: 32

5.4 User Interface ... 33

CHAPTER 6 ... 36

Implementation ... 36

6.1 Technologies used: ... 36

6.2 Algorithms used: ... 37

6.3 Implementation: ... 40

CHAPTER 7 ... 63

Software Testing Document ... 63

7.1 Introduction ... 63

7.2 Introduction ... 65

7.3 Test Cases ... 67

7.4 Test results: ... 70

CHAPTER 8 ... 72

Conclusions and Scope for Future Work ... 72

8.1 Conclusion ... 72

8.2 Limitations ... 72

8.3 Scope for Future Work ... 73

REFERENCES ... 74

ACKNOWLEDGEMENTS ... 76

List of Figures

Fig 1. Software process model 10

Fig 2. Semester 7 timetable 20

Fig 3. Semester 8 timetable 21

Fig 4. Project timetable 21

Fig 6. Data Flow in Chosen Architecture 30

Fig 7. MVC components 30

Fig 8. RESTful API can serve any application 31

Fig 9. System Architecture 32

Fig 10. Flow chart 33

Fig 11. Dashboard 34

Fig 12. Search by location 34

Fig 13. Search by crime type 35

Fig 14. Spread of data points 40

Fig 15. Data integration results 42

Fig 16. K means clsutering 44

Fig 17. DBSCAN results 45

Fig 18. HDBSCAN results 45

Fig 19. OPTICS results 45

Fig 20. BIRCH results 46

Fig 21. Classification scoring metrics 48

Fig 22. Twitter data results visualization 49

Fig 23. News feed data results visualization 51

Fig 24. Borough-wise clustering 51

Fig 25. Dashboard-1 52

Fig 26. Dashboard-2 53

Fig 27. Dashboard-3 53

Fig 28. Dashboard-4 53

Fig 29. Search by location-1 54

Fig 30. Search by location-2 54

Fig 31. Search by location-3 55

Fig 32. Search by location-4 55

Fig 33. Search by location-5 56

Fig 34. Search by location-6 57

Fig 35. Search by location-7 57

Fig 36. Search by crime type-1 57

Fig 37. Search by crime type-2 58

Fig 38. Search by crime type-3 58

Fig 39. Twitter data source-1 59

Fig 40. Twitter data source-2 59

Fig 41. Twitter data source-3 60

Fig 42. Twitter data source-4 60

Fig 43. News feed data source-1 61

Fig 44. News feed data source-2 61

Fig 45. News feed data source-3 62

Fig 46. News feed data source-4 62

List of Tables

Table 1. Crime type description 411

Table 2. Model results 433

Table 3. Confusion matrix 477

Table 4. Testing tools 677

Table 5. Test results 711

Nomenclature

DBSCAN	Density-based spatial clustering of applications with noise
HDBSCAN	Hierarchical density-based spatial clustering of applications with noise
OPTICS	Ordering Points To Identify Cluster Structure
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
NLTK	Natural Language Toolkit
NYC	New York City
NYPD	New York Police Department

CHAPTER 1

Introduction

This chapter presents a broad introduction to the entire project explaining the problem definition, motivation behind the project, scope of the project and also salient contributors for this project. Finally, overview of the thesis organization is mentioned.

1.1 Problem Definition:

Emergencies can be of different kinds, from fires to road accidents and assaults, to medical emergencies. It is distressing to be faced with an unanticipated incident or emergency. In our country, 112 is the single emergency number, people in distress can call to get immediate assistance – from fire brigade, medical team or from the police. Law enforcement officials across the country would find it incredibly useful if they had an automated system to continually identify areas that are more prone to crimes and policing-related incidents, than others. This would allow them to proactively review and plan suitable resource deployments and patrolling in those areas, and thereby preempt and prevent, as far as possible, untoward incidents that could result in an emergency.

Through this project we aim to utilize the power of Machine Learning to tackle the increasing criminal activity in the society. We can identify the crime prone areas by analyzing the historic data from emergency services and the current data from sources like Twitter and News feeds. Further, we aim to build a solution for policing that is able to filter out crimes based on the type of crime as well as the location. This will enable officials to reduce the crime by taking actions which will be specific to the crimes in each location. It will also help people to know which area is safe for them to visit.

1.2 Motivation:

Crime rate is increasing day by day and even though crimes can occur anywhere, they are more in some places than others. When selecting a neighborhood to move to, crime rate is one of the important factors people consider. For authorities too, knowing the high crime rate areas is crucial to ensure the community's safety. Hence, the identification of crime prone areas will not only benefit the citizens but also the authorities to make better and informed decisions. The authorities can analyze the high crime prone areas using different filters such as location and crime type. This will enable them to take necessary actions in that location. Officials can even impose appropriate rules and resources in locations more prone to crimes in order to prevent it from occurring in future.

The system will also be available to citizens so that they can see the crime prone areas and select the appropriate locations to buy a new house, choose school for their children or even plan their trip. Having necessary information related to crimes about a location will empower people to be alert and in turn save themselves from criminal activities. The advantages of such kind of system motivated us to go ahead with this idea.

1.3 Scope of the project:

The scope of this project mainly includes analysis of crime data from 3 different data sources. The data sources used in this project are as follows:

1. New York Police Department Historic Crime Data
2. Twitter data from New York crime Twitter handles
3. News feeds

Three data sources have been used since the first data source, NYPD Historic Crime Data provides data about the past crimes while the other two data sources provide data about the recent crimes. The main objective of this project is to make the public as well as police personnel aware about the areas where maximum crime occurs and to give other related statistics about that region. Using this information, the police could allot resources accordingly. We also aim to display the results in visuals formats making it simpler for the users to interpret the data. Features such as 'filter by crime type' and 'filter by location' help the users to narrow down their search queries.

1.4 Organization of the Thesis:

The report starts by displaying the certificates and declarations which gives a gist of the project which will be explained in detail ahead. Then comes the introduction section, i.e. Chapter 1, which explains the problem definition, motivation behind the project, scope of the project, salient contribution and organization of report. Then the research done for this project and the concepts learned are explained in short in the section Literature Survey which forms the Chapter 2 of this report. Chapter 3 emphasizes on the Software Project Management Plan which contains the project planning in detail. It would have details about project objectives, project estimates, project schedules, project resources, project staffing, risk management plans, project monitoring, project control and other miscellaneous activities. Next, requirements, expectations, and standards for the project are described in Chapter 4. A detailed plan of how to develop a piece of software is discussed in Chapter 5 of Software Design Document. In Chapter 6, Implementation explains the technologies used, algorithm or the methodology used in this project and the sample images taken which show how to use this system. Chapter 7, The Software Test Document(STD) describes the test preparations, test cases, and test procedures to be used to perform qualification testing of the Computer Software. Conclusion explains what we are concluding from the results we obtained and scope for future work on this project, which is written in Chapter 8 of this thesis. At the end of this thesis, the references used to develop this system are mentioned along with Appendix A, Author’s Publication and Acknowledgments.

Hence, the general idea of the project is portrayed by discussing the motivation behind this topic and the scope of this topic. Next, the background work or the literature survey done before and during the implementation of this project is stated in the next chapter.

CHAPTER 2

Background Work

This chapter presents the summary of research papers that were referred before the beginning of the project. The summary typically consists of the datasets used, algorithms used and the outcome achieved by the authors.

Abrar A. Almuhanna et al. [1] proposes a methodology to predict Spatio-temporal criminal patterns within the New York City neighbourhoods using a dataset from 2006 until 2019 with 2.2M criminal records for 25 different crimes type. The best model out of Support Vector Machine (SVM), Random Forest (RF), and XGboost classifiers is to be found out. After analysis, it is illustrated that XG boost has predicted the highest number of correct classifications out of 25 different crime types it has predicted 22 types of crime accurately, whereas Random Forest has predicted 21 types of crime accurately and SVM predicted accurately 17 types of crimes with lowest accuracy. The methodology is very well illustrated (step-by-step). Data is visualized initially using plots and heatmaps. Locations of crimes on the NYC geographic map have been plotted. Attribute selection process has been mentioned well and attributes like date or time of occurrence of event, X and Y coordinate for New York State Plane, geospatial Location Point, Level of offense, Suspects and victim Age Group, etc are considered.

Wajiha Safat et al. [2] the goal was to improve the predictive accuracy by applying different machine learning algorithms, namely, the logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), and time series analysis by long-short term memory (LSTM) and autoregressive integrated moving average (ARIMA) model to better fit the crime data. Dataset of two cities- Los Angeles and Chicago was used for analysis. XGBoost achieved the maximum accuracy in the Chicago dataset and KNN in the Los Angeles dataset.

Shiju Sathyadevan, Devan et al. [3] the aim was to predict regions which have high probability for crime occurrence and visualize crime prone areas. The authors of the paper tested the accuracy of classification and prediction based on different test sets and they found out that bayes theorem showed more than 90% accuracy. The data mainly consists of news, so the text is the news used to classify them into different types of crimes using Naive Bayes algorithm. Pattern of days on which crime occurred in found using Apriori algorithm. Then sample data of a day is given and then prediction of whether crime will happen on that day or not will be done.

Andrew J. Park et al. [4] the motive was to introduce a three-dimensional (3D) visual analytics framework that interactively visualizes crime data and other relevant datasets on a highly accurate 3D model of the City of Vancouver, Canada. This 3D visualization advances crime analysis activities through a more accurate display of the area under study. Three major visualization techniques - building highlights, street sections, and kernel density were used and in turn gained insights into the data which may assist in developing innovative strategies for crime prevention. The 3D visual analytics can improve identification of temporal and spatial criminal areas and provide strategic and effective plans for pro-active police deployment.

Prabhat Sharma et al. [5] visualized the NY City crime data in form of various plots using crime data which was available on the US government website. They have explained the process of data visualization in various steps starting with Data collection. Next is data selection where they selected only 14 columns which they required for their research and dropped the other 21 columns. Next important step was data cleaning. They treated all the missing values with appropriate methods and further broke down the data cleaning process in 3 stages: setting the environment, preliminary text pre-processing and fuzzy matching. Next they selected Python language and Ubuntu OS for their analysis and visualization. The tool used for data visualization was Spyder since it works best with Python. They performed dimensionality reduction (PCA method) to reduce the number of attributes. PCA technique reduces the dimensionality but also makes sure that significant data is not lost. The parameters chosen for visualization were year, area, premise code and age. Various graphs are displayed in the paper in the results section. Based on these plots they were able to recognize the densities of crimes in different areas.

Mohammad Al Boni et al. [6] focused on Crime prediction models based on hierarchical and multi-task statistical learning. They were able to develop area-specific crime prediction models using hierarchical and multi-task learning. The models mitigated sparseness by sharing information across areas. The area taken into consideration is Illinois, Chicago. Criminogenic factors were also considered to improve model. The difficulties faced during implementation were that many areas have sparse data, thus complicating the direct application of pooled models.

Sumanta Mukherjee et al. [7] the aim was to build a model for news headline classification, crime location extraction, and crime report generation and the text was in Bengali language. The models used were Naïve Bayes and SVM The overall accuracy of the model was 82% and they automated the data for computing district wise crime rate with data in Indian language. Following are the steps they used for the process: News Data Crawling, News Headline Classification and Crime Article Collection, Crime location Tagging, Crime Location Disambiguation and Mapping them to Respective District names and Crime Histogram Generation and Crime Prone Area Identification. Key Challenge was to correctly identifying location of crime and correctly identifying areas that need attention

Romika Yadav et al. [8] aims to locate the offender site in advance with more accuracy. The Auto Regression Techniques to accurately predict the crime with minimum error for time series data by identifying the relationship among crimes attributes was also explored. The experimental result obtained using “R” tool show that formulation work well for all parameters and improves certainty in prediction. They concluded that the Generalized Linear Model (GLM) for Crime Site Selection (CSS) using Big Data delivers better results and forecasts spatio-temporal crime events with certainty.

Hana Anber et al. [9] explains how Twitter forms a good source of data as compared to other social media platforms. Twitter is described as an online networking service that allows users to send short and read short 140-word messages. It is accessible for unregistered users to read and monitor most tweets, unlike Facebook where users can control the privacy of their profiles. The tweets provide massive amount of information which consists the tweet message, number of followers, user profile information, time and date of a tweet, etc. All this information can be used for the purpose of data analysis.

Heidi Cohen [10] points out that 29% of the data/information floated using tweets are rumors or fake news. 51% of the data was generic. The total amount of unique information was limited to about 20%. From this article we concluded that randomly collecting data from tweets would be a wrong technique. Instead, we only selected verified Twitter handles of agencies which were directly related to the aim of our project.

Summarizing the literature survey, we went through the different models used for clustering, classification and prediction of crime. We observed that in none of the research papers a dashboard was made for analyzing the results dynamically. Almost all the researchers had used only historic data which is easily available on the Internet. Hence, we set a hypothesis to use Twitter and news feed as a source of data to get latest data and create our own dataset. We confirmed this hypothesis by reviewing other research work as mentioned above.

CHAPTER 3

Software Project Management Plan

This chapter presents the project deliverables, the software model followed, roles and responsibilities of each team member, breakdown of the project into multiple tasks, assignment of tasks and finally the timeline to be followed for successful completion of the project.

3.1 Introduction:

3.1.1 Project Overview

Through this project we aim to utilize the power of machine learning to tackle the increasing criminal activity in our society. We will identify crime prone areas by analyzing historic data from emergency services and the current data from Twitter and News feeds. Further we aim to build a solution for policing that is able to filter on the basis of crime type as well as the location. This project will help authorities identify areas prone to crimes, making deployments and patrolling in those areas more effective thereby preventing and preempt any untoward incidents.

3.1.2 Project Deliverables

Scope of this topic	September 2021
Requirement specification document	September 2021
Project management plan	September 2021
Testing document	October 2021
Project presentation-1	October 2021
Synopsis	November 2021

Data collection	January 2022
Pre-processed data ready	February 2022
Feature subset selection	February 2022
Model selection and results	February 2022
NLP pipeline for Tweets and News feed	March 2022
Website dashboard to display results	April 2022
Black book/thesis	April 2022

3.2 Project Organization:

3.2.1 Software Process Model

The Waterfall Process Model has been chosen as the software process model for this project.

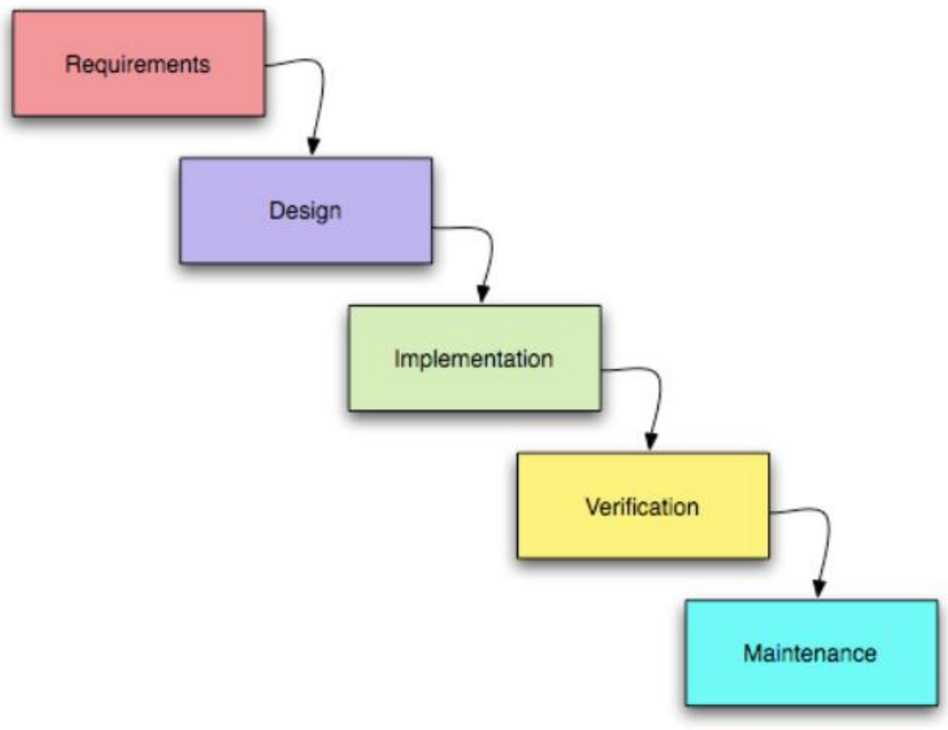


Fig 1. Software process model

First the requirements of the user will be discussed and then the User interface will be designed and presented, next we start with the implementation of the source code, executable code and the databases. Finally, when the implementation is done, the software is tested using different test cases and the flaws are corrected and then the software is regularly maintained from time to time.

3.2.2 Roles and Responsibilities

This project is divided into 2 main parts:

First part includes data collection from three sources, data pre-processing, analysis, model selection, visualization and result analysis. All these different sections will be divided equally between the team members.

Next part includes developing a website for users to help in predicting the crime type based on the location which will be taken as the input. For this part, front-end of the website will be handled by Kunj Gala and Labdhi Jain. Back-end of the website will be handled by Dhruv Doshi and Shubham Bhakuni. With this, all the aspects of the project are covered by the entire team with equivalent responsibilities among the team members.

3.2.2 Tools and Techniques

The project will be implemented using Python Programming language and graphs will be visualized using libraries such as Seaborn, Folium and Matplotlib. The required UML diagrams will be made using StarUML software. We have used popular python libraries such as Numpy, Pandas, Seaborn, Matplotlib, etc to implement and visualize our algorithms. For our web application, we have used Django framework wherein the frontend was made using Bootstrap, HTML, CSS, JavaScript whereas the backend was handled by Django. Moreover, we have also made use of various online platforms such as Google Colab, Github, Google drive for smooth sharing of codes.

3.3 Project Management Plan

3.3.1 Tasks:

1) Data Collection and pre-processing

2) Data visualization and analysis

3) Feature subset selection

4) Model selection

5) Visualizing the model results

6) NLP pipeline for Tweets and News feed

7) Website design, planning and implementation

8) Documentation

3.3.1.1 Data Collection and pre-processing:

The name of the task will be Data Collection and pre-processing and the Unique Id for the same is SPMP-T01.

Description**-**

The entry point to the development cycle of any ML project is the data preparation stage. In this task we will collect freely available data from various official sources like New York Police department(NYPD), National Crime Records Bureau (NCRB), Twitter handles related to crime, news feed related to crime. Real-world raw data is often incomplete, inconsistent and lacking in certain behaviors or trends. They are also likely to contain many errors. So, once collected, they will be pre-processed into a format the machine learning algorithm can use for the model.

Deliverables and Milestones-

Milestone 1: Collecting data from different sources.

Milestone 2:Data Integration and Data cleaning.

Milestone 3: Data Normalization.

Resources needed-

The Project Team, Laptop, Datasets, Access to Twitter API, and Jupyter Notebook as an IDE to pre-process and clean data

Dependencies and Constraints-

Selection of appropriate datasets and data sources is an important step before pre-processing and cleaning data.

Risks and Contingencies-

Data might be inaccurate as the collected data could be unrelated to the problem statement. Some classes or categories in the data may have a disproportionately high or low number of corresponding samples. As a result, they risk being under-represented in the model.

3.3.1.2 Data visualization and analysis:

The name of the task will be Data visualization and analysis and the Unique Id for the same is SPMP-T02.

Description-

In this task we will be plotting graphs of different attributes to analyse their distributions. We can analyse those results and obtain insights about our data from that. These plots will also be helpful while reducing the number of attributes. We could also infer from the plot about the type of distribution of that attribute. Whether it is highly skewed or is it a normal distribution.

Deliverables and Milestones-

Milestone 1: Plotting all the attributes.

Milestone 2: Analysing the plots.

Milestone 3: Working on the analysis results.

Resources needed-

Python programming and python libraries such as Pandas, matplotlib and seaborn.

Dependencies and Constraints-

Data visualization result quality will depend on how well the data has been pre-processed. For example, visualization results would be different if the missing values are not handled properly.

Risks and Contingencies-

There could be some attributes which are difficult to visualize. Dealing with them could become time consuming. Even if the quality of data is poor, visualization results won't be useful.

3.3.1.3 Feature Subset Selection:

The name of the task will be Feature Subset Selection and the Unique Id for the same is SPMP-T03.

Description-

Feature Selection is the most critical pre-processing activity in any machine learning process. We intend to select a subset of attributes or features that makes the most meaningful contribution to a machine learning activity. Feature subset selection will lead to having a faster and more cost-effective (less need for computational resources) learning model, Improved efficacy of the learning model.

Deliverables and Milestones-

Milestone 1: Measure Feature Relevance. We can use mutual information as a measure for feature relevance.

Milestone 2: Measure Feature Redundancy. There a variety of ways we can do this like Correlation-based Measures, Distance-based Measures, etc.

Milestone 3: Elimination of features with low relevance and high redundancy.

Resources needed-

The Project Team, Laptop, Datasets, and Jupyter Notebook as an IDE to pre-process and clean data

Dependencies and Constraints-

Due to complex nature of multi dimensional datasets, feature selection and classifier are typically more expensive or time-consuming. Therefore, we need a robust feature selection technique for selecting the optimum single subset of the features of the data for further analysis or to design a classifier.

Risks and Contingencies-

Feature subset selection sometimes leads to over fitting of model and hence provides inaccurate results.

3.3.1.4 Model selection:

The name of the task will be Model selection and the Unique Id for the same is SPMP-T04.

Description-

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

Deliverables and Milestones-

Milestone 1: Split the data into training and test sets.

Milestone 2: Fit candidate models on the training set.

Milestone 3: Evaluate the performance of each model on test set and select the best model.

Resources needed-

The Project Team, Laptop, Datasets, and Jupyter Notebook as an IDE to pre-process and clean data

Dependencies and Constraints-

Some algorithms require specialized data preparation in order to best expose the structure of the problem to the learning algorithm. Hence each model has a model development pipeline to be implemented before selecting the actual model.

Risks and Contingencies-

The project stakeholders may have specific requirements, such as maintainability and limited model complexity. As such, a model that has lower skill but is simpler and easier to understand will have be preferred.

Alternately, if model skill is prized above all other concerns, then the ability of the model to perform well on out-of-sample data will have to be preferred regardless of the computational complexity involved.

3.3.1.5 Visualizing the model results::

The name of the task will be Visualizing the model results and the Unique Id for the same is SPMP-T05.

Description-

We will visualize location wise and crime category wise clusters of data. Model result visualization gives us proper explanations of what the model is doing, why the results are what they are and finally provides output in a visual form which can be described to the non-technical persons.

Deliverables and Milestones-

Milestone 1: Installing required libraries

Milestone 2: Importing required libraries

Milestone 3: Implementing models and Creating Visualizations.

Resources needed-

Laptop, Datasets, and Jupyter Notebook as an IDE, libraries to implement different algorithms, Tableau or PowerBI.

Dependencies and Constraints-

The result of the model depends on the value of K. Only the correct choice of K will help us in finding crime prone areas accurately.

Risks and Contingencies-

In clustering visualization if the value of K is not chosen correctly in the case of K means clustering then the results could be incorrectly interpreted.

3.3.1.6 NLP pipeline for Tweets and News feed:

Description-

A NLP pipeline will be required to process the data extracted from Twitter handles and from news websites. This pipeline will begin by pre-processing and cleaning the raw data till producing the required results i.e. statistics and graph plots.

Deliverables and Milestones-

Milestone 1: Aquiring access to Twitter API.

Milestone 2: Successfully extracting tweets and feeds.

Milestone 3:Running classification models and plotting the results.

Resources needed-

Laptop, Datasets, and Jupyter Notebook as an IDE, libraries to implement different algorithms, Twitter API.

Dependencies and Constraints-

Tweet extraction is highly dependent on the access level of API aquired. If the basic access is aquired then limited amount of tweets can be retrieved.

Risks and Contingencies-

There is a risk of selecting an incorrect Twitter handle which will result in non-crime related tweet extraction. Another risk is not getting the access to Twitter API at all. In this case, no tweet extraction will be possible. Not finding appropiate news feed to retrieve data is another risk.

3.3.1.7 Website design, planning and implementation:

The name of the task will be Website design, planning and implementation. The the Unique Id for the same is SPMP-T07.

Description-

Planning a website involves deciding the website goal, pulling together information needed on the site to achieve your goal and Organizing information. Web design refers to design of websites that are displayed on the internet. Implementing a website means passing from the design stage to the completion and the launch of the site.

Deliverables and Milestones-

Milestone 1: Setting out website's objectives.

Milestone 2: Measure Feature Redundancy. There a variety of ways we can do this Correlation-based Measures, Distance-based Measures, etc.

Milestone 3: Website Security, especially important when website is collecting crime related data.

Resources needed-

Laptop, Atom IDE, Apache Server, Domain name, Website host.

Dependencies and Constraints-

Website should be secure as it will contain crucial information and will be used by police departments.

Risks and Contingencies-

Complex visual features may well be visually appealing, but if they negatively impact a user's experience, they are likely to be counterproductive.

3.3.1.8 Documentation:

The name of the task will be Documentation. The the Unique Id for the same is SPMP-T08.

Description-

Documentation is a written piece of text that is often accompanied with a software program. This makes the life of all the members associated with the project more easy. It may contain anything from API documentation, build notes or just help content.

Deliverables and Milestones-

Milestone 1: Requirement and Planning Documentation.

Milestone 2: Architectural Documentation

Milestone 3: End-User Documentation.

Resources needed-

Project Team, Laptop, Overleaf LaTeX software.

Dependencies and Constraints-

The documenting code is time consuming. The documentation has no influence on the performance of the an application.

Risks and Contingencies-

Complex visual features may well be visually appealing, but if they negatively impact a user's experience, they are likely to be counterproductive.

3.3.2 Assignments:

Since there are multiple tasks in the project, they are equally assigned as follows:

Data Collection and pre-processing: Team

Data visualization and analysis: Dhruv Doshi and Kunj Gala

Feature subset selection: Shubham Bhakuni and Labdhi Jain

Model Selection: Dhruv Doshi and Kunj Gala

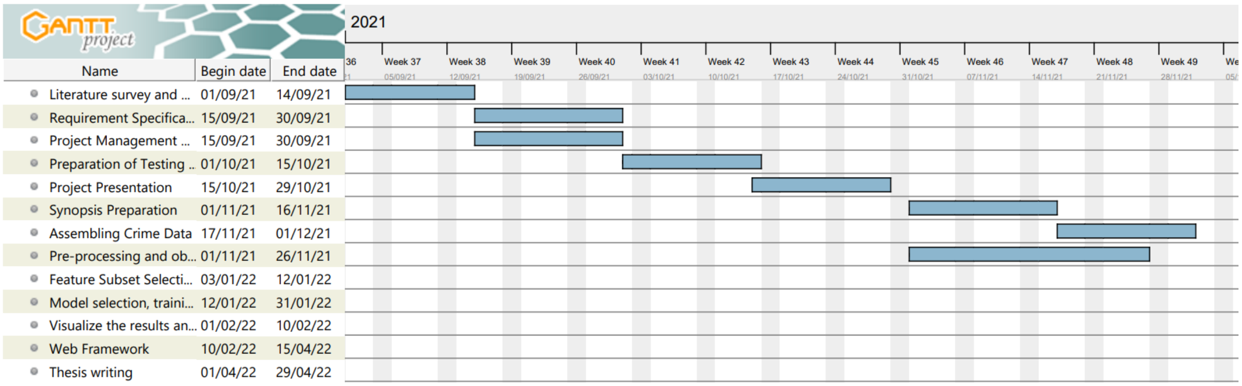
Visualizing the model results: Shubham Bhakuni and Labdhi Jain

NLP pipeline for Tweets and News feed: Dhruv Doshi and Shubham Bhakuni

Website design, planning and implementation: Team

Documentation: Team

3.3.3 Timetable:



Name	Begin date	End date
Literature survey and ...	01/09/21	14/09/21
Requirement Specifica...	15/09/21	30/09/21
Project Management ...	15/09/21	30/09/21
Preparation of Testing ..	01/10/21	15/10/21
Project Presentation	15/10/21	29/10/21
Synopsis Preparation	01/11/21	16/11/21
Assembling Crime Data	17/11/21	01/12/21
Pre-processing and ob...	01/11/21	26/11/21
Feature Subset Selecti...	03/01/22	12/01/22
Model selection, traini...	12/01/22	31/01/22
Visualize the results an...	01/02/22	10/02/22
Web Framework	10/02/22	15/04/22
Thesis writing	01/04/22	29/04/22

Fig 2. Semester 7 timetable

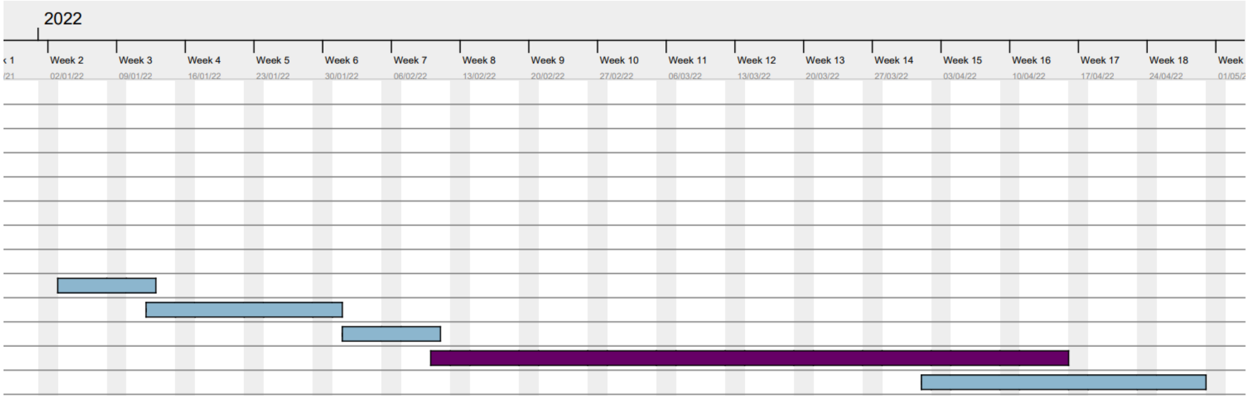
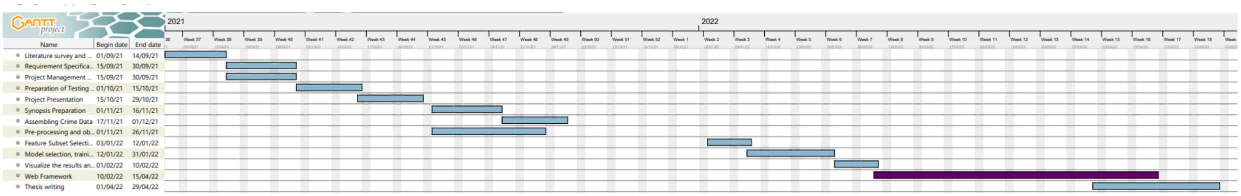


Fig 3. Semester 8 timetable



Name	Begin date	End date
Literature survey and ...	01/09/21	14/09/21
Requirement Specifica...	15/09/21	30/09/21
Project Management ...	15/09/21	30/09/21
Preparation of Testing ...	01/10/21	15/10/21
Project Presentation	15/10/21	29/10/21
Synopsis Preparation	01/11/21	16/11/21
Assembling Crime Data	17/11/21	01/12/21
Pre-processing and ob...	01/11/21	26/11/21
Feature Subset Selecti...	03/01/22	12/01/22
Model selection, traini...	12/01/22	31/01/22
Visualize the results an...	01/02/22	10/02/22
Web Framework	10/02/22	15/04/22
Thesis writing	01/04/22	29/04/22

Fig 4. Project timetable

Hence, the above chapter described the project deliverables, the software model followed, roles and responsibilities of each team member, breakdown of the project into multiple tasks, assignment of tasks and finally the timeline to be followed for successful completion of the project. Next chapter states functional and specific details of the system and in general which features are to be included in the system.

CHAPTER 4

Software Requirements Specification

This chapter presents the requirements of the system to be designed in detail including the specific requirements, functional requirements, database requirements and specifically which features are to be included in the system and what would be their functionality.

4.1 Introduction:

Software Requirement Specification (SRS) as name suggests, is complete specification and description of requirements of software that needs to be fulfilled for successful development of software system. This section will describe what the software will do and how it will be expected to perform.

The purpose of this SRS document is to cover the overall description of the software system, specifically the system's product perspective, hardware, software and communication interfaces, memory constraints, operations, site adaptation requirements, functions, characteristics, some constraints, assumptions and dependencies.

This SRS also detail the systems specific requirements, specifically requirements on performance, logical database, design constraints, standard compliance, software

system attributes such as reliability, availability, security, maintainability and portability, system mode, user class, objects, features, stimulus, and response.

Lastly, this SRS is intended for the law enforcement officials in general, and the systems engineers, analysts, designers, implementers, testers, maintenance, and the users themselves in particular. This SRS will document further changes, revisions and additions to any requirement and functionality of the software system.

4.1.1 Product Overview

Emergencies can be of different kinds, from fires to road accidents and assaults, to medical emergencies. It is distressing to be faced with an unanticipated incident or emergency. In our country, 112 is the single emergency number, people in distress can call to get immediate assistance – from the fire brigade, medical team or from the police. Law enforcement officials across the country would find it incredibly useful if they had an automated system to continually identify areas that are more prone to crimes and policing-related incidents, than others. This would allow them to proactively review and plan suitable resource deployments and patrolling in those areas, and thereby preempt and prevent, as far as possible, untoward incidents that could result in an emergency.

Using the power of Machine learning, the system will identify the Crime prone areas enabling better deployment and utilization of the police force. The data gathered from various authorized resources will be pre-processed, analyzed, classified and visualized using various machine learning and data analysis techniques. The outputs from this will be shared with the users through a web application wherein users can see the future trends in crime and analyze the areas most prone to criminal activities along with the types of crimes and statistics. The data can be visualized by using the API of maps.

4.2 Specific Requirements:

4.2.1 External Interface Requirements:

- **User Interfaces:** The application will have one interface for all the users. The users will be able to visualize and see the results from the model prepared with ML algorithms. The web app will allow people to search for crime prone areas using different filters. The results will be shown in the form of graphical representations as well as statistics and comments on the overall result. This will enable law enforcement officials to better visualize the future trend of crimes in a particular area and prevent incidents. This website can also be used by the public to see which area is more related to crimes and can plan their visit or housing accordingly. The filters such as search by crime type or location can help to query the results according to the user's needs.

- **Hardware Interfaces:** The web app can be run on any browser supporting device provided there is an internet connection. Browsers can be found on various devices like computers, tablets, and mobile phones. The device must have a strong internet connection to browse through our dashboard. Since the web portal does not have any designated hardware, it does not have any direct hardware interfaces over Client Side. The hardware connection to the database server is managed by the underlying operating system on the web server over Server side.

- **Software Interfaces:** The web app will be built using a web framework like Django for providing an interactive user interface and styling our application. It will be a cross platform web application that can run on a device having any operating system and a browser. Our system will use csv or excel files for storing information related to crime. The files must store all data such that it can be used for accounting, as well as accountability.

- **Communications Protocols:**

  The system will interface with a Wide Area Network (WAN) to maintain communication with all its devices.

  It should use a reliable-type IP protocol such as TCP/IP or reliable-UDP/IP for maximum compatibility and stability.

  All devices it will interface with should contain strong internet connection to access the online system.

  Devices that are wireless can also use Ethernet compatible cards, using the IEEE 802.11b/g standard and having support for WPA2-PSK encryption.

4.2.2 Software Product Features:

Crime Data Visualization:

Priority: Important

The system should provide a visual map which would be obtained using various data visualization techniques from the ML model. Different colors will be used to highlight the various densities of criminal activities and its types. This will also help to analyze crimes associated with a particular area.

Crime Tweets Classification:

Priority: High

The system will extract tweets from official police department twitter handles. The tweets will be classified as crime tweets or non-crime tweets. Further the location of each crime tweet will be identified.

Identification of Location with most crimes:

Priority: Important

This System will utilize all the features mentioned above to identify areas with most crime using various filters such as location, crime type etc. This will help us to summarize and draw appropriate conclusions.

Web Portal:

Priority: Important

All the user will access the system using a web portal. All the result, features will be available for every person to use. It will have important functionalities such as user inputs, graphical representation of the results for user query, different filters for location, time, crime type, etc.

4.3 Functional Requirements:

4.3.1 Data collection:

The data used in this project will consist of the information extracted from emergency calls (112 in India) attended by government officials. Another set of data consists of tweets extracted from official police department twitter handles. We aim to collect the data from various government authorized sites where data is publicly available. The key attributes in the data set would be the latitude, longitude, date of crime, time of crime and crime type. Data that we receive can be in any format such as audio, video, text, etc. So appropriate methods can be adopted to convert the data into a uniform format.

4.3.2 Data Pre-processing:

The data collected will be raw and contain many missing values and irregularities which may hamper the performance of models. Thus, various preprocessing techniques will be employed such as handling missing values, stop word removal, stemming, normalization, sampling, etc to make the data more suitable for training various algorithms.

4.3.3 Analyzing the Data:

The preprocessed data will be analyzed using correlation coefficients, histograms, pie charts, bar graphs, heat map, etc to understand which attributes are important for our implementation. Analyzing the data will help us to see the similarity and differences between the data. By using these techniques, features can be selected on which our model will be trained. This helps to improve the efficiency and accuracy of the model.

4.3.4 Crime Classification:

In order to predict and classify the areas with the most number of crimes along with the crime type various machine learning algorithms will be implemented. The results of all algorithms will be compared and the best one will be selected based on their performance and accuracy. The main aim of using the classification algorithms is to construct a model which could identify the crime type and location within a specific time.

4.3.5 Crime Data Visualization:

The predicted crime prone areas can be visualized on maps using various data visualization techniques. This will help us to better view the results of our implementation. Some of the ways by which this can be done are by highlighting the buildings with different colors to demonstrate the varying crime counts or coloring the streets with different colors according to the crime rate.

4.3.6 Identification of location with most crimes:

The results that we will obtain from all the above modules will be summarized and analyzed to draw appropriate conclusions. These conclusions will help us to identify areas with most crimes using various filters such as location, crime type, etc.

4.3.7 Web Portal:

We aim to create a user interface wherein users will be able to access the results of our implementation according to their needs. This interface can be developed using a web framework and will have important functionalities such as user inputs, graphical representation of the results for user query, different filters for location, time, crime type, etc. This web portal can be useful for law enforcement officials to assess the crime prone areas thereby allocating necessary care and resources in the crime prone areas.

4.4 Software System Attributes

4.4.1 Reliability:

Reliability is the extent to which the software system consistently performs the specified functions without failure. Since the information is supposed to be used by the police department for security purposes hence we aim to propose a system that provides reliable information.

4.4.2 Accuracy:

We aim to achieve high accuracy by classifying crimes into the right categories and in predicting crime prone areas. We will be experimenting with different algorithms and will finally select the most accurate one to build the system.

4.4.3 Usability:

Usability is how easily the user is able to learn, operate, prepare inputs and interpret outputs through interaction with a software system. We will be developing a website with a user friendly interface which will be straightforward, providing quick access to common features or commands.

4.5 Database Requirements

The system will run on the Rest API server. The attributes that play an important role in crime analysis will be: incident date and time, type of crime, a brief description, geolocation (latitude and longitude), nearest police station, pin-code. We'll keep updating the above table with latest data to keep improving our prediction and improve analysis. The table will be an important part of our application and will serve as a data-set to continuously train our ML model, which will eventually tell the user, areas which are more prone to criminal activities.

Hence, this chapter discussed the software requirements specifications of the system. The next chapter describes the designing of the system in terms of user interface as well as choice of system architecture and a flow chart for the system's working.

CHAPTER 5

Software Design Description

This chapter presents the frontend design as well as the backend design of the system. The choice of system architecture, description of the components, flow chart for the system's working is shown and few webpage design images are mentioned.

5.1 Introduction:

Software design is a process to transform user requirements into some suitable form. This section will be a detailed blueprint for software design and serve as documentation later on.

5.1.1 Design Overview:

Through this project we aim to utilize the power of machine learning to tackle the increasing criminal activity in our society. We will identify crime prone areas by analysing historic data from emergency services. Further we aim to build a solution for predictive policing that is able to predict the crime type, given the location.

5.2 System Architectural Design:

5.2.1 Chosen architecture design

MVC pattern is a Product Development Architecture. It solves the traditional approach's drawback of code in one file, i.e., that MVC architecture has different files for different aspects of our web application/ website.

The MVC pattern has three components, namely Model, View, and Controller.

Model − The lowest level of the pattern which is responsible for maintaining data.

View − This is responsible for displaying all or a portion of the data to the user.

Controller − Software Code that controls the interactions between the Model and View.

MVC is popular as it isolates the application logic from the user interface layer and supports separation of concerns. Here the Controller receives all requests for the application and then works with the Model to prepare any data needed by the View. The View then uses the data prepared by the Controller to generate a final presentable response.

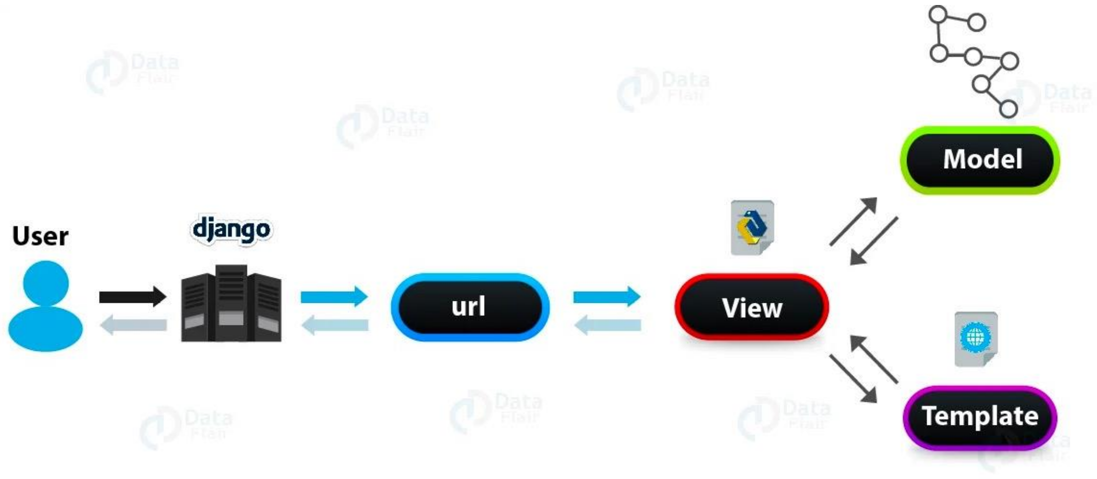


Fig 6. Data Flow in Chosen Architecture

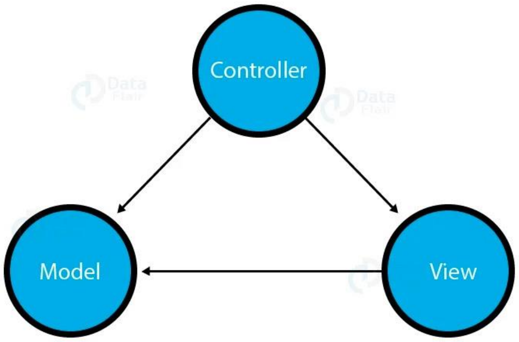


Fig 7. MVC components

5.2.2 Alternate architecture design:

Even though REST Architecture is considered industry standard, there are many great alternatives that work in-place of REST architecture. SOAP (Simple Object Access Protocol) is a popular alternative. SOAP uses XML to transfer data, which uses a lot more bandwidth, whereas in REST Architecture, JSON is the preferred format for transferring data. This is in addition to the fact that RESTful services are easier and flexible in usage.

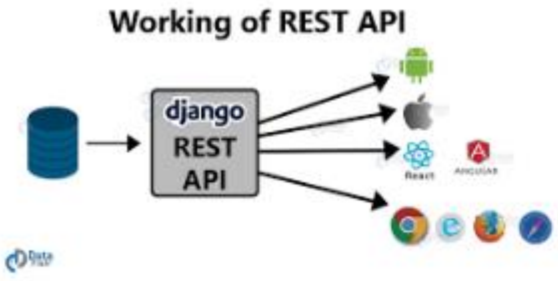


Fig 8. RESTful API can serve any application

5.2.3 System interface design:

The system will work on the tech-stack shown below. On the server side, Django REST framework will serves API requests by fetching data from the PostgreSQL databse. The API returns data based on the request, which is then served to ReactJS, the frontend web framework used. React uses HTML, CSS and javascript to display the information received from the API request.

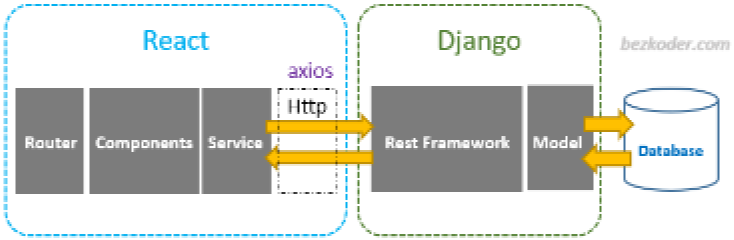


Fig 9. System Architecture

5.3 Detailed description of components:

5.3.1 Primary system components

Crime Data Visualization: Through this component the user is provided with a visual map obtained using data visualization techniques from the ML model. Different colors and opacity determines the various density of criminal activities in that particular area.

Web Portal: All the components described will be accessed using the web portal which will be a website. All results and features will be accessed through the web portal.

Search and Filter: All users can use this component to search and filter through the database the view the results in table or map format. Using this feature users can see the crime prone areas and other cluster related statistics.

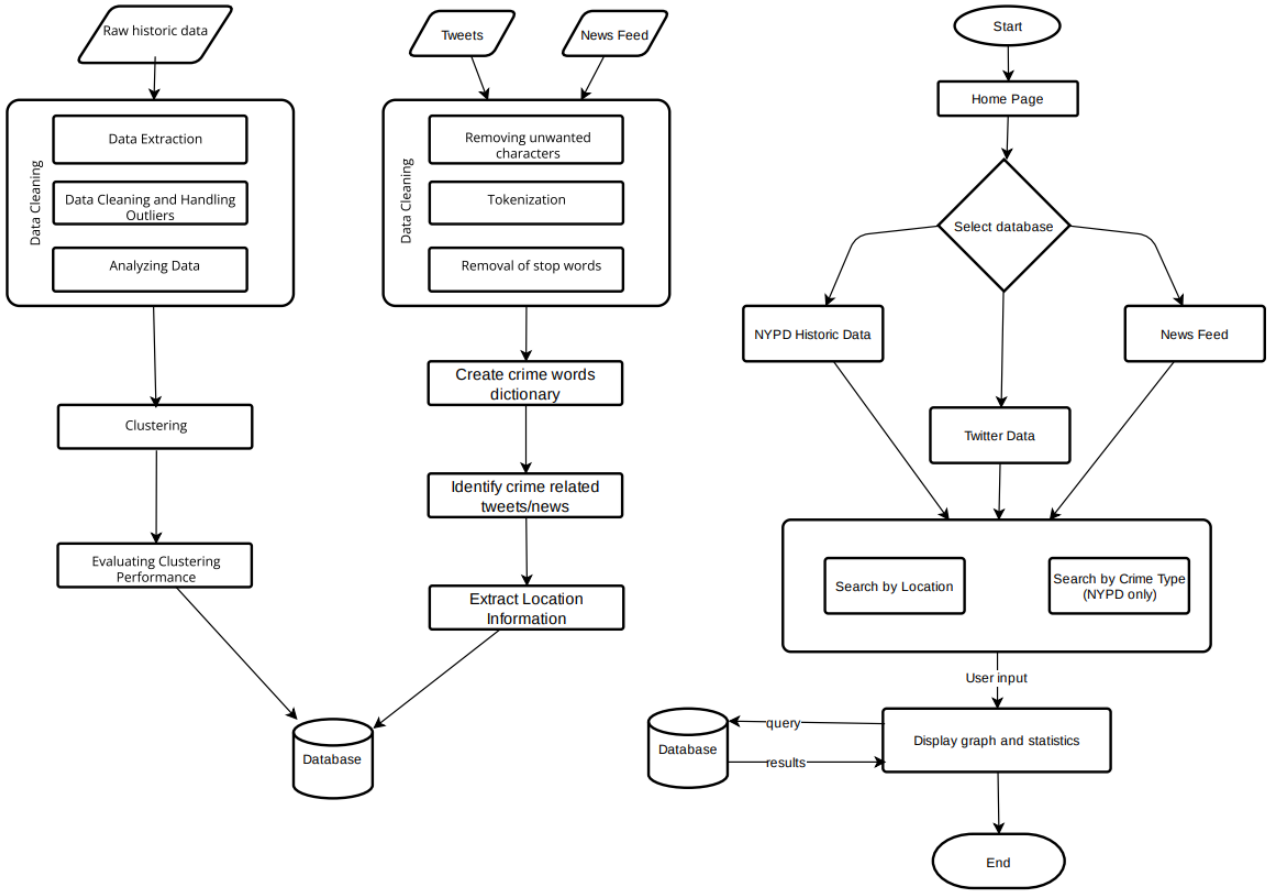
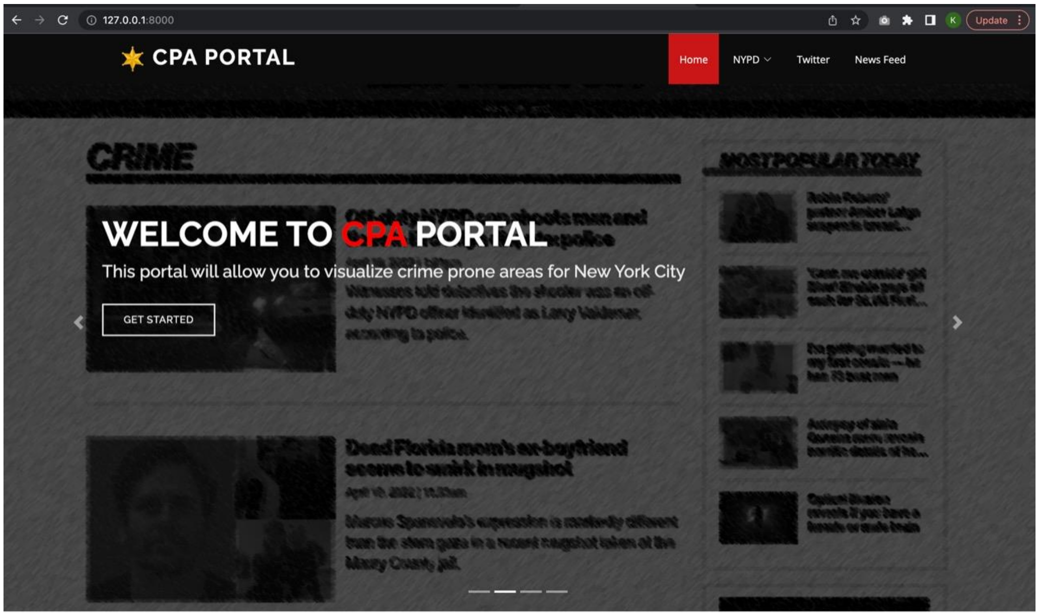


Fig 10. Flow chart

5.4 User Interface

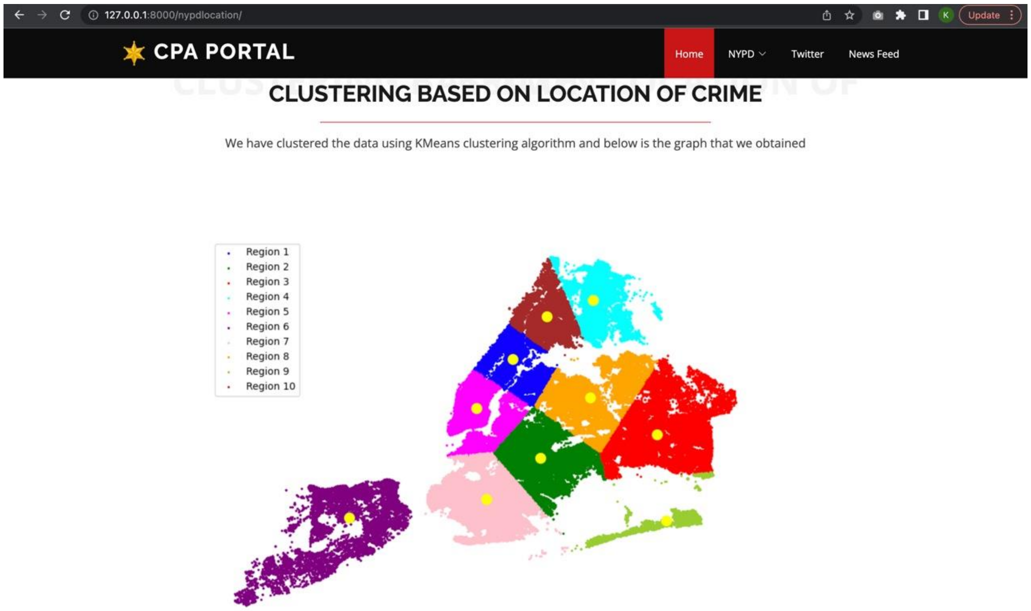
Screen Images: The user interface design will consist of a dashboard which will display all the statistical data.



WELCOME TO CPA PORTAL

This portal will allow you to visualize crime prone areas for New York City

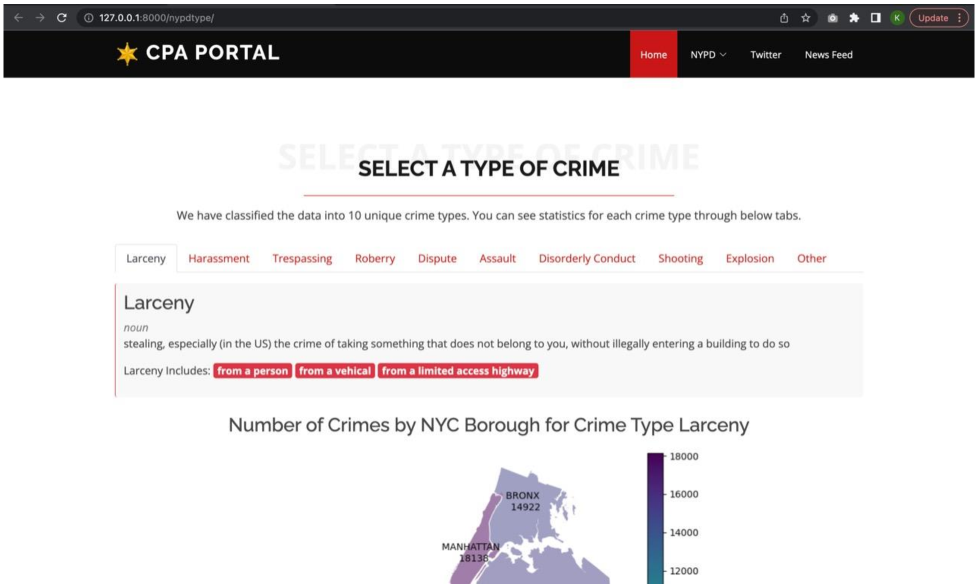
Fig 11. Dashboard



CLUSTERING BASED ON LOCATION OF CRIME

We have clustered the data using KMeans clustering algorithm and below is the graph that we obtained

Fig 12. Search by location



SELECT A TYPE OF CRIME

We have classified the data into 10 unique crime types. You can see statistics for each crime type through below tabs.

Larceny

noun

stealing, especially (in the US) the crime of taking something that does not belong to you, without illegally entering a building to do so

Larceny Includes:

Number of Crimes by NYC Borough for Crime Type Larceny

Fig 13. Search by crime type

In this chapter, software design of the system was discussed which included frontend and the backend design. Next, the main implementation of the system will be discussed, describing all the modules of the system.

CHAPTER 6

Implementation

This chapter presents the core of this project- how was it implemented. A step-by-step procedure is explained for each module in the project along with information about the technologies used and explanation of the algorithms used.

6.1 Technologies used:

6.1.1 Python 3.7

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Since 2003, Python has consistently ranked in the top ten most popular programming languages in the TIOBE Programming Community Index where, as of February 2021, it is the third most popular language (behind Java, and C). It was selected Programming Language of the Year (for "the highest rise in ratings in a year") in 2007, 2010, 2018, and 2020 (the only language to do so four times).

An empirical study found that scripting languages, such as Python, are more productive than conventional languages, such as C and Java, for programming problems involving string manipulation and search in a dictionary, and determined that memory consumption was often "better than Java and not much worse than C or C++".

6.1.2 Django

Django is a high-level Python web framework that enables rapid development of secure and maintainable websites. Built by experienced developers, Django takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It is free and open source, has a thriving and active community, great documentation, and many options for free and paid-for support.

6.1.3 HTML5

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.

6.1.4 CSS3

CSS3 is mainly used to improve the visual elements of a website. CSS3 makes changes to how some visual elements are implemented and rendered by a browser. However, it is not a single hugely unwieldy specification, unlike CSS2. CSS3 is separated into separate modules to facilitate development. This means that the specification comes out in chunks, with more stable modules than others.

6.2 Algorithms used:

6.2.1 Python NLTK- nltk.tokenizer.word_tokenize()

Using nltk.tokenizer.word_tokenize() method, we can extract the tokens from strings using this method. The method returns syllables from the sentences. This method can be also used to extract words out of a sentence.

6.2.2 Create inverted index

This algorithm returns an inverted index which is a data structure that maps words to their locations in the corpus along with a frequency count of the word. It helps to identify the frequently occurring words in the corpus.

6.2.3 Boolean Model

Boolean model is a classical information retrieval model, being among the first and most adopted model. This model is based on set theory and Boolean algebra. It considers whether a word is present in a document or no. If a word is present in a document then it is marked as ‘1’ and if the word is absent in the document then it is marked as ‘0’. Finally, the model produces an output using the Boolean conditions mentioned in the query.

6.2.4 Named Entity Recognition (NER)

NER can automatically scan documents and look for major people, locations or organizations according to the requirement. In the project, location from a crime tweet was extracted using NER algorithm.

6.2.5 Geopy- Nominatim

Nominatim is used to convert textual location into geo coordinates. Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe

6.2.6 K-means clustering algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties. It is a centroid-based algorithm, where each cluster is associated with a centroid.

6.2.7 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

6.2.8 HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise(HDBSCAN) is a clustering algorithm. It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based on the stability of clusters. It Uses a range of distances to separate clusters of varying densities from sparser noise. The HDBSCAN algorithm is the most data-driven of the clustering methods, and thus requires the least user input.

6.2.9 OPTICS

OPTICS Clustering stands for Ordering Points To Identify Cluster Structure. It draws inspiration from the DBSCAN clustering algorithm. It adds two more terms to the concepts of DBSCAN clustering i.e,Core Distance and Reachability Distance. It Uses the distance between neighboring features to create a reachability plot, which is then used to separate clusters of varying densities from noise. The OPTICS algorithm offers the most flexibility in fine-tuning the clusters that are detected, though it is computationally intensive, particularly with a large Search Distance.

6.2.10 BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset. BIRCH is often used to complement other clustering algorithms by creating a summary of the dataset that the other clustering algorithm can now use.

6.3 Implementation:

6.3.1 Module 1- NYPD Historic Data

- Data Collection: Data collection is the process of gathering information on variables of interest. The dataset used in the project is by the New York Police Department(NYPD). The data is collected by the NYPD from the ICAD system. It consists of 5 million rows and 19 columns. Latitude, longitude, incident description, incident data and time, radio code are some of the important columns. The spread of data across New York is shown in the figure below:

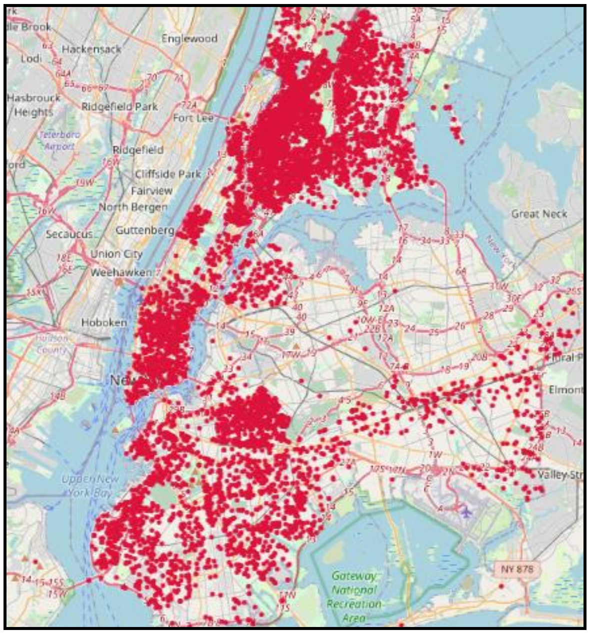


Fig 14. Spread of data points

- Data Pre-processing: Data preprocessing is the process of transforming raw data into an understandable format.

1. **Date Cleaning:** The first step of data preprocessing was Data Cleaning. It is a way of identifying incomplete or inaccurate parts of the data and then modifying, replacing or deleting them based on the necessity. The column 'CAD_EVNT_ID' had 3.2% repeated values. The rows with repeated values

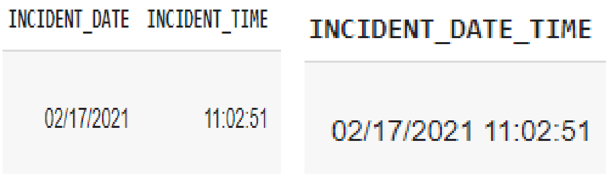
were dropped. Later all the missing values were dropped. After this step, 4,871,204 rows were left.

2. **Data Reduction**: As a part of the Data Reduction step, the OBJECTID column was dropped as it had 68% missing values, CREATE_DATE column was dropped as it was irrelevant for our project, BORO_NM and PATRL_BORO_NM were highly correlated so PATRL_BORO_NM was retained as it provided more detailed information,GEO_CD_X and GEO_CD_Y were also dropped as they were highly correlated to the Latitude and Longitude columns respectively. Next, the RADIO_CODE column had 420 distinct events but not all of them were crimes. Hence, radio codes which did not represent crimes were manually detected and dropped. Also, different radio codes that depicted the same crime were merged into a single radio code. Finally data was reduced to 10 unique crime types including harassment, robbery, assault, etc. as shown in the table below.

	TYP_DESC
22Q2	Larceny
29H1	Harassment
29Q1	Other
39T1	Trespassing
20R	Robbery
53D	Dispute
24Q2	Assault
50G2	Disorderly Conduct
10S2	Shooting
33	Explosion

Table 1. Crime type description

3. **Data Integration:** The third step of data pre-processing was Data Integration. Data integration is the process of combining data from different sources into a single, unified view. In this step INCIDENT_DATE and INCIDENT_TIME were merged into the INCIDENT_DATE_TIME column as shown in figure below.



INCIDENT_DATE	INCIDENT_TIME
02/17/2021	11:02:51

INCIDENT_DATE_TIME
02/17/2021 11:02:51

Fig 15. Data integration results

4. **Data Conversion:** The fourth data pre-processing step is Data Conversion. Data conversion means changing the data in one format to another format. The datatype of INCIDENT_DATE_TIME was originally String. But when in String, date/time operations could not be performed on them. Hence, this attribute was converted into Pandas DateTime (Timestamp) datatype. Similarly ADD_TS, DISP_TS, ARRVD_TS and CLOSNG_TS were also converted into the Pandas DateTime (Timestamp) datatype.

5. **Data Sampling:** The next step in data preprocessing was Data Sampling. Sampling is performed to reduce the size of the dataset by selecting only the samples that would represent the whole dataset. Since the dataset selected was too huge (1064806 rows and 13 columns), a sample containing 25% data using the random sampling method was created which could be used in further model building. After performing this step the dataset obtained had 500459 rows.

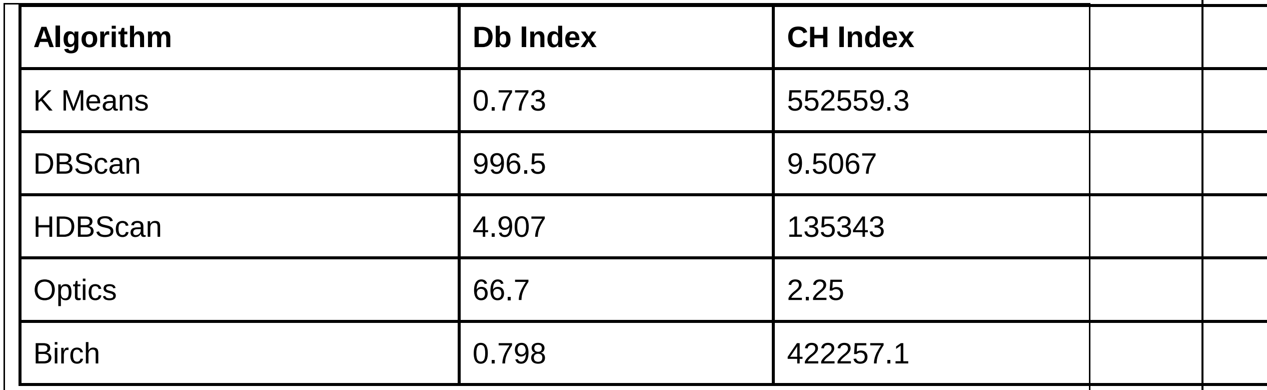
6. **One Hot Encoding:** Then lastly One Hot Encoding was implemented on the dataset to perform binarization of the categories in order to use those columns in our model. ‘PTRL_BORO_CD’ column had 8 categories as there are 8 boroughs in New york. This column was one hot encoded to convert it to 1/0

data for each category. Similarly, ‘CIP_JOBS’ containing 4 categories was also one hot encoded. So after the entire data preprocessing the dataset obtained had 500455 rows and 22 columns.

- Model Building:

Clustering is a technique of grouping data into different clusters, consisting of similar data points. In this section different clustering algorithms or models are used on the dataset to finally find the most optimal one. The different clustering algorithms used were K means clustering, DBSCAN, HDBSCAN, OPTICS and BIRCH, each of which have been explained in detail above.

The above 5 algorithms were compared using two evaluation metrics ‘DB Index’ and ‘CH Index’. Davies-Bouldin Index is the ratio of distances within the clusters to distances between the clusters. Lower the values of db index, better is the clustering performance. Calinski-Harabasz is the ratio of the dispersion within the cluster to the dispersion between the clusters. Higher the values of ch index better is the clustering performance. The values of DB Index and CH Index for all the 5 algorithms are listed in table below.



Algorithm	Db Index	CH Index
K Means	0.773	552559.3
DBScan	996.5	9.5067
HDBScan	4.907	135343
Optics	66.7	2.25
Birch	0.798	422257.1

Table 2. Model results

From the above table it can be seen that K means clustering algorithm has lowest DB Index and highest CH Index followed by Birch algorithm. Thus it can be concluded that K means clustering algorithm performs best among all the other algorithms.

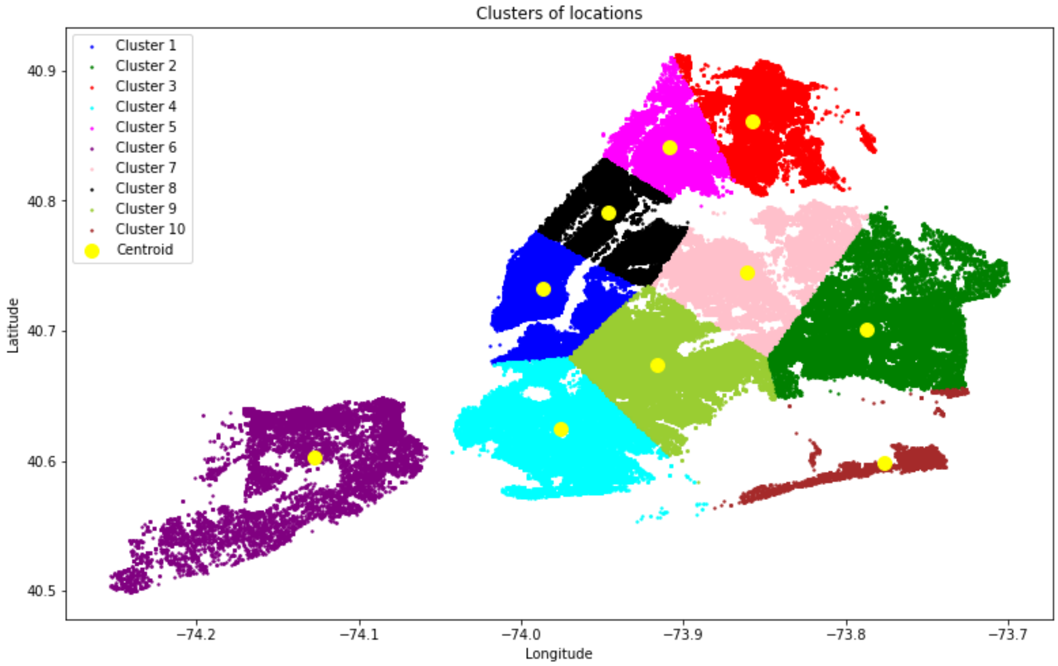


Fig 16. K means clsutering

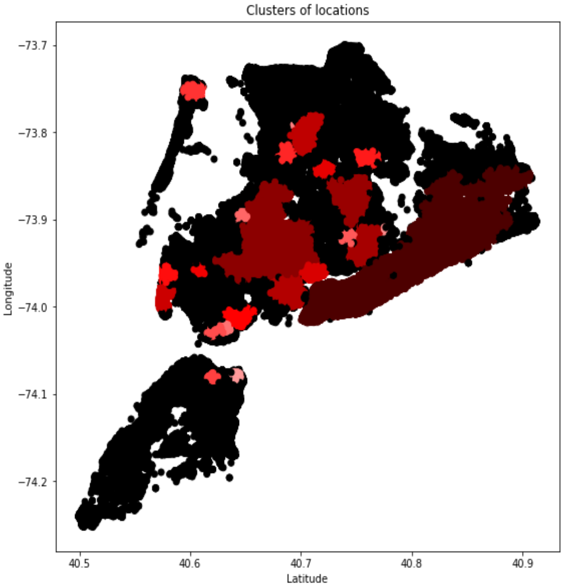


Fig 17. DBSCAN results

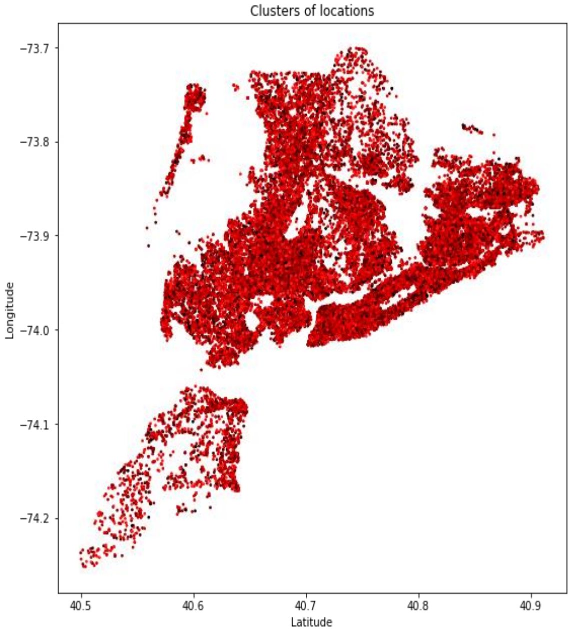


Fig 18. HDBSCAN results

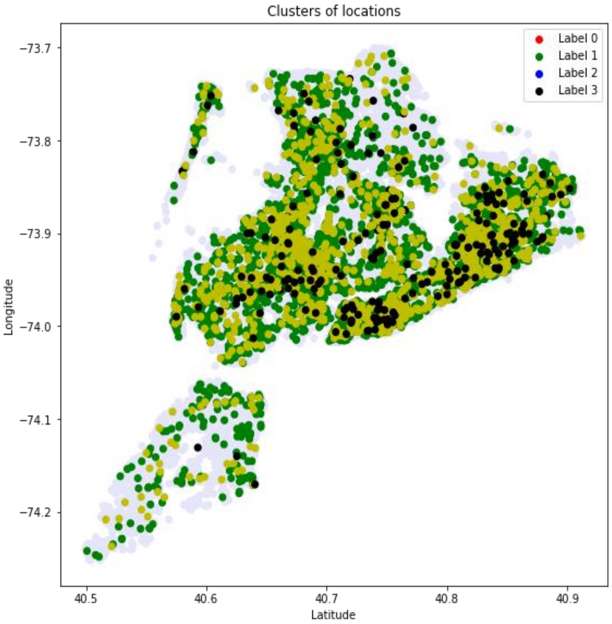


Fig 19. OPTICS results

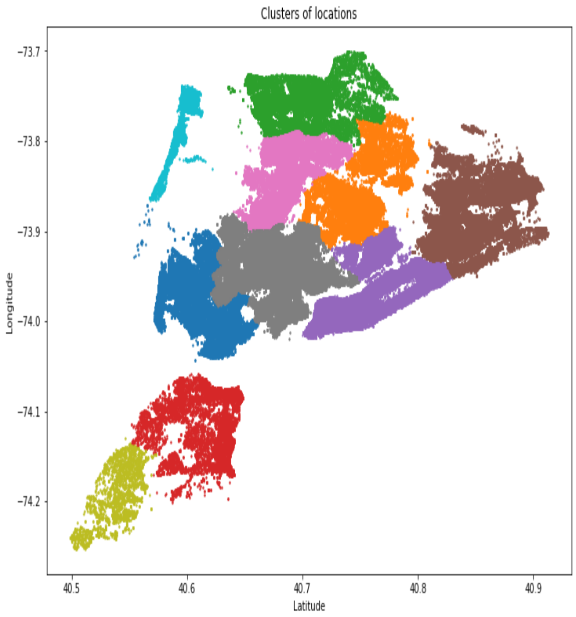


Fig 20. BIRCH results

6.3.2 Module 2: Twitter handles

Step-wise implementation procedure:

- Selection of 2 Twitter handles which regularly update crime news on their feed. @NYPDNews and @NYPDTips were the 2 handles finalized for extraction as they provided latest news updates regarding the criminal activity in New York.

- Extraction of total 200 tweets using Twitter API, from each of the handles and converting them into a Python Pandas Dataframe.

- Cleaning the tweets forms a necessary step since tweets consists of a lot of information which is of no use for further processing. Hence, a procedure must be followed to remove unwanted information. This includes steps like- Removing unwanted characters (like @, # and $) since Twitter users use these symbols to tag other users. Tokenization is the process of separating out words from sentences. Using NLTK library in Python, tokenization can be easily carried out. Next, stop words like ‘and’, ‘at’, ‘or’, etc. need to be removed since they provide no significant impact on the end results. The tweets have been cleaned and now ready for further processing.

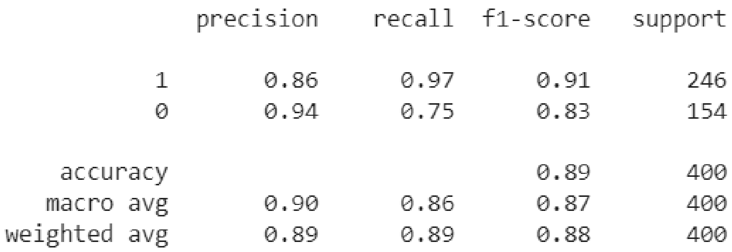
- Next objective is to decide if a tweet has information related to a crime or not. If it does, then it is called a ‘crime tweet’ and if it doesn’t then it is called a ‘non-crime tweet’. To help classify the tweets, a crime dictionary is required, which was not readily available. Due to this, dictionary building approach was chosen. To build a crime words dictionary, an inverted index was created using the tweet corpus to help identify which crime related words are used frequently in the tweets. Manual labelling was carried out to mark the frequently occurring words which were related to crime. After the completion of this step, a crime dictionary was ready to be used for classification.

- A simple yet efficient information retrieval model- Boolean Model is used to classify tweets as crime and non-crime. Using this model only the crime tweets are retrieved and the non-crime tweets are ignored. Other more complex models like TF-IDF could also be used but that level of complexity was uncalled for in this retrieval.

- To verify the classification results, confusion matrix was calculated along with scoring parameters like precision, recall and accuracy. Following are the results:

	Crime	**Non-crime**
Detected	239	39
Not detected	7	115

Table 3. Confusion matrix



```
              precision    recall  f1-score   support

           1       0.86      0.97      0.91       246
           0       0.94      0.75      0.83       154

    accuracy                           0.89       400
   macro avg       0.90      0.86      0.87       400
weighted avg       0.89      0.89      0.88       400
```

Fig 21. Classification scoring metrics

- Extraction of location from the entire tweet message forms an important step in this process. Using NER algorithm, location related keywords are extracted and stored in another column in the dataframe.

- Next, the location keywords need to be converted to geo-coordinates so that it is possible to plot the data points on a New York map and visualize where are the crime prone areas. This conversion is done using the Geocode API. Taking an address as input, this API returns the latitude and longitude as the output.

- One of the best ways to visualize this data is using Folium maps. Folium maps in Python provide us with FastMarkerCluster module which helps to cluster the data points visually.

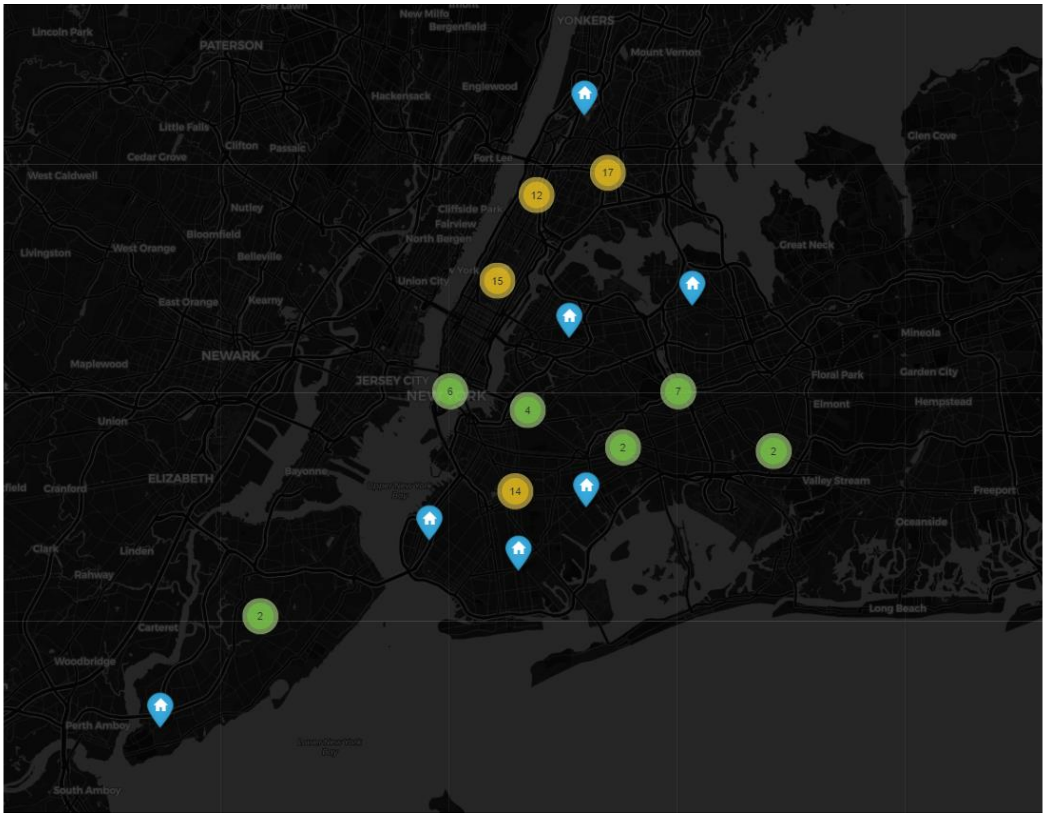


Fig 22. Twitter data results visualization

- To collect statistics of the results obtained, borough-wise clustering was done. To get statistics like which borough has the highest number of crimes, etc.

6.3.3 Module 3: News Feed/ Articles

Step-wise implementation procedure:

- Selection of a news article web page which provides latest news/information about crimes in New York state. New York Post's (NYPost) crime section (in their website) provides exactly the kind of data required for analysis. Following is a link to the website:

  https://nypost.com/tag/crime/

- Extraction of total 160 articles from the web page was possible using RSS feed.

- Cleaning the news articles forms a necessary step since tweets consists of a lot of information which is of no use for further processing. Hence, a procedure must be followed to remove unwanted information. This includes steps like- Removing unwanted characters (like @, # and $) since authors use these symbols to tag other sources or external links. Tokenization is the process of separating out words from sentences. Using NLTK library in Python, tokenization can be easily carried out. Next, stop words like ‘and’, ‘at’, ‘or’, etc. need to be removed since they provide no significant impact on the end results. The new articles have been cleaned and now ready for further processing.

- Since the news articles have been extracted from the crime section of NY Posts, there is no need to classify articles into crime and non-crime. All the articles will be related to crime.

- Extraction of location from the entire news article forms an important step in this process. Using NER algorithm, location related keywords are extracted and stored in another column in the dataframe.

- Next, the location keywords need to be converted to geo-coordinates so that it is possible to plot the data points on a New York map and visualize where are the crime prone areas. This conversion is done using the Geocode API. Taking an address as input, this API returns the latitude and longitude as the output.

- One of the best ways to visualize this data is using Folium maps. Folium maps in Python provide us with FastMarkerCluster module which helps to cluster the data points visually.

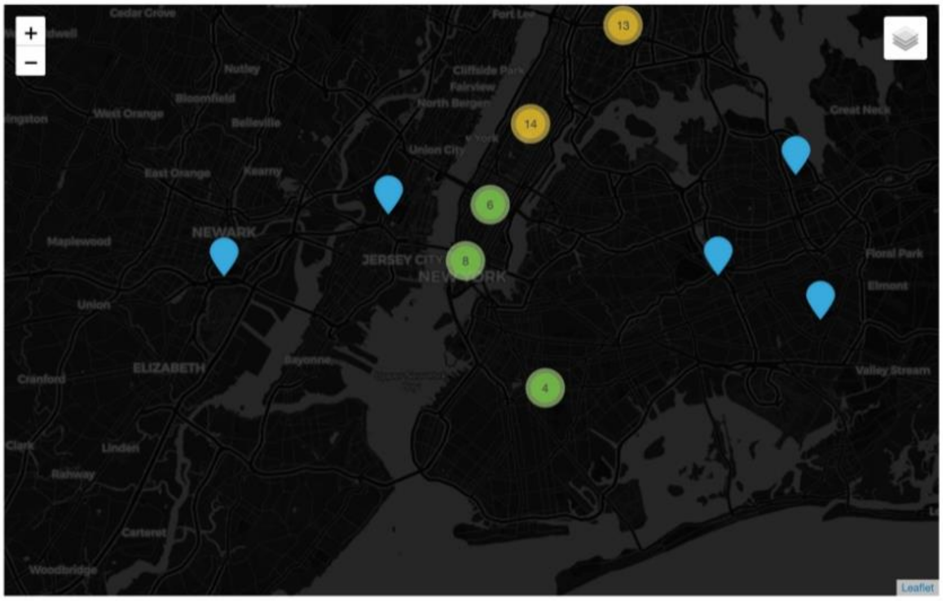


Fig 23. News feed data results visualization

- To collect statistics of the results obtained, borough-wise clustering was done. To get statistics like which borough has the highest number of crimes, etc.

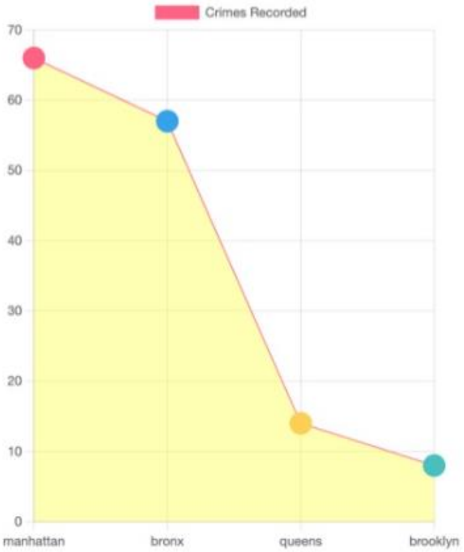
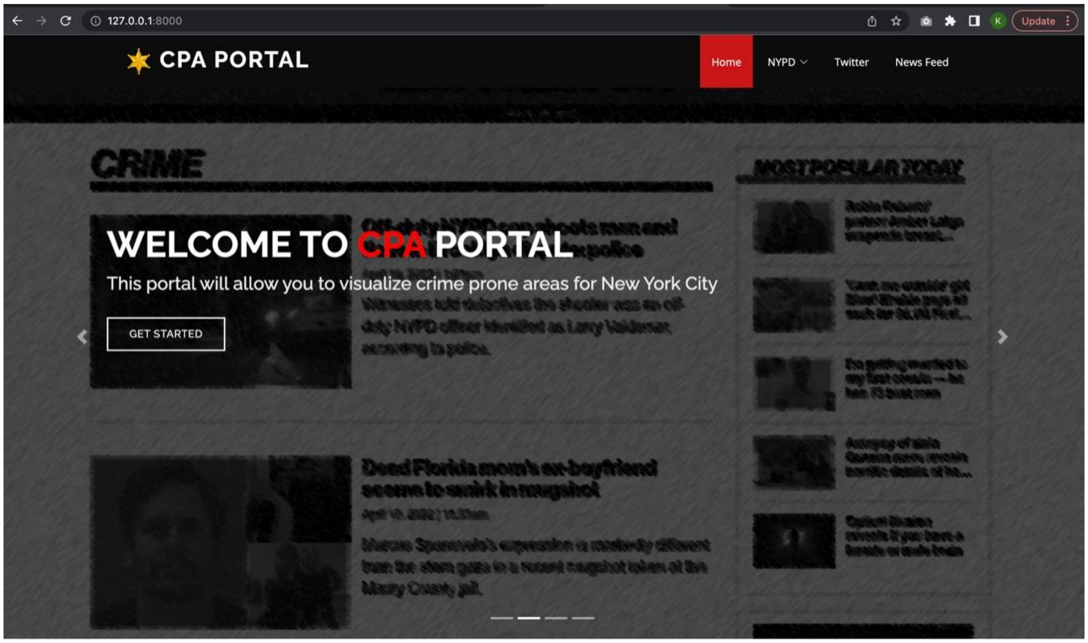


Fig 24. Borough-wise clustering

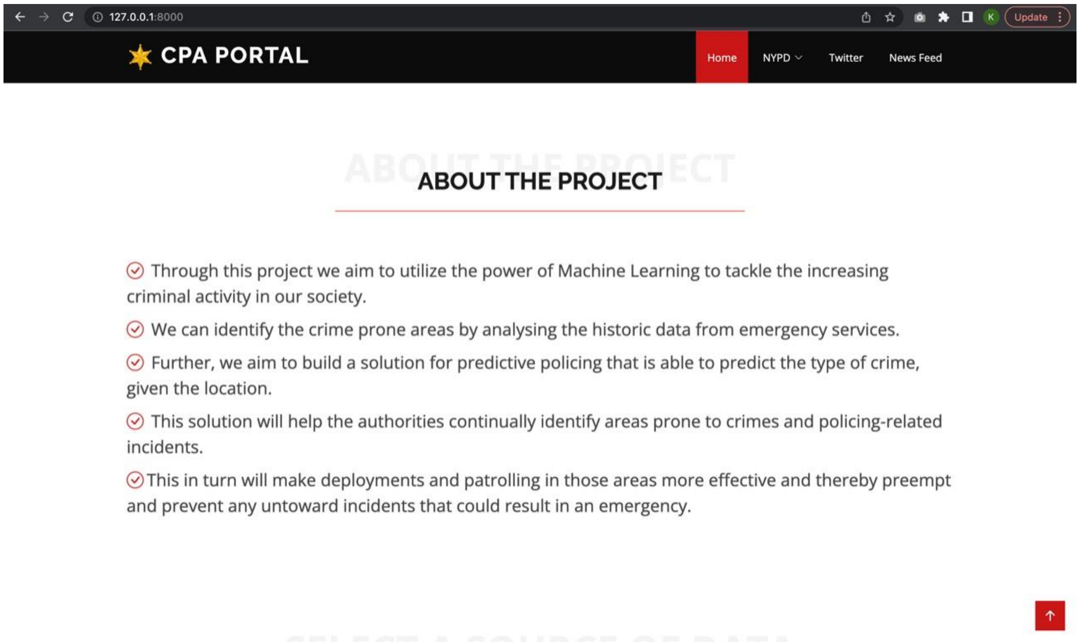
6.3.4 Implementation screen shots:



WELCOME TO CPA PORTAL

This portal will allow you to visualize crime prone areas for New York City

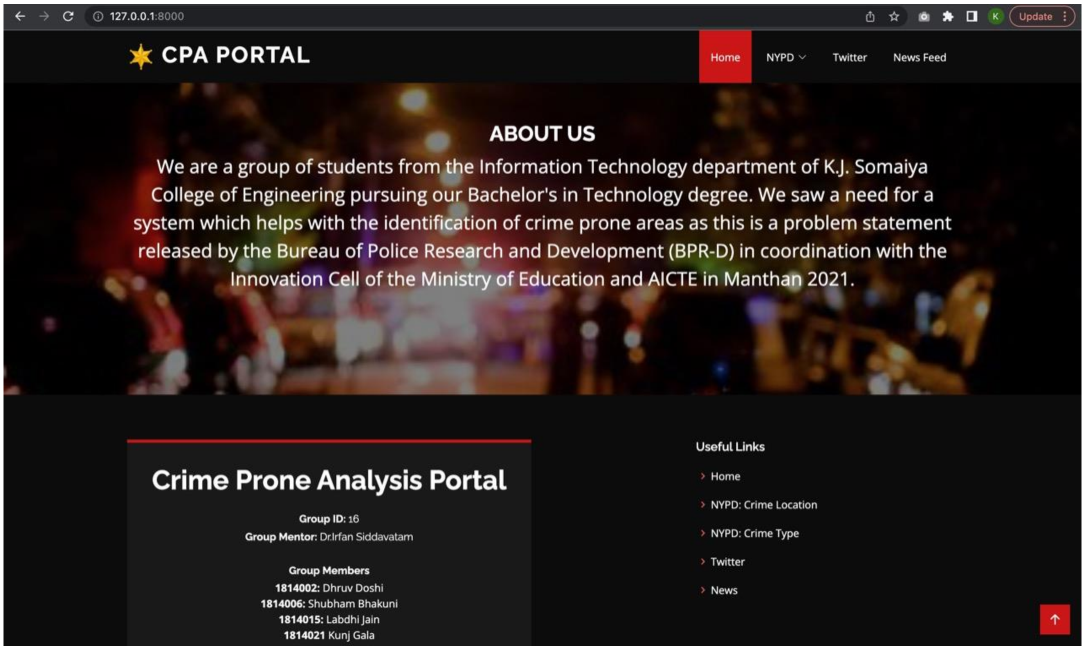
Fig 25. Dashboard-1



ABOUT THE PROJECT

- Through this project we aim to utilize the power of Machine Learning to tackle the increasing criminal activity in our society.
- We can identify the crime prone areas by analysing the historic data from emergency services.
- Further, we aim to build a solution for predictive policing that is able to predict the type of crime, given the location.
- This solution will help the authorities continually identify areas prone to crimes and policing-related incidents.
- This in turn will make deployments and patrolling in those areas more effective and thereby preempt and prevent any untoward incidents that could result in an emergency.

Fig 26. Dashboard-2



ABOUT US

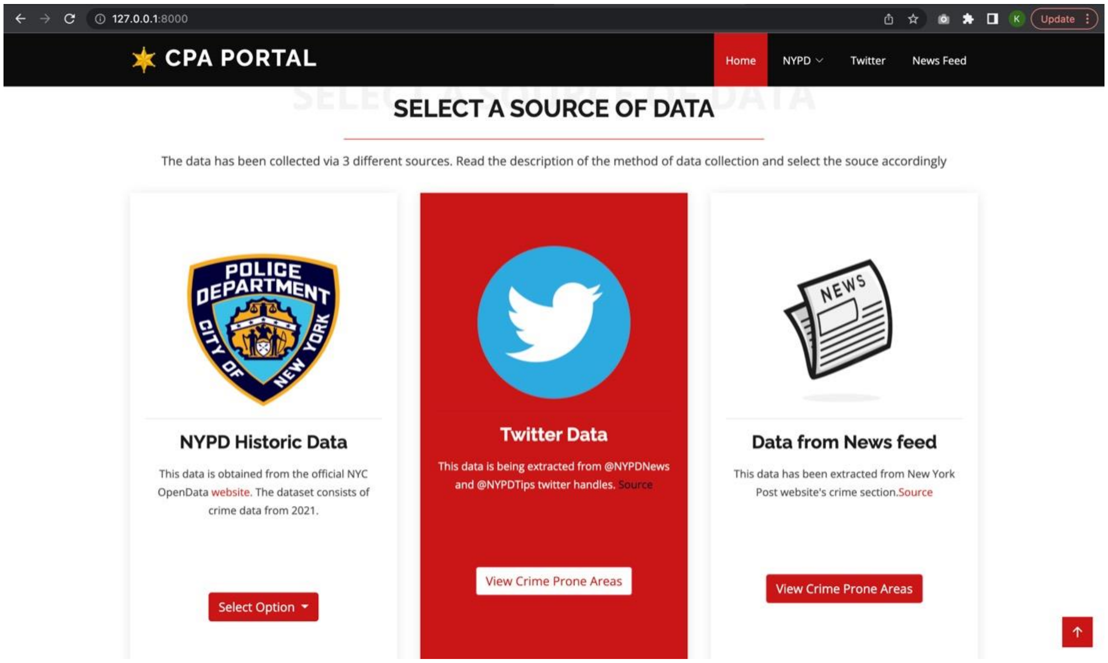
We are a group of students from the Information Technology department of K.J. Somaiya College of Engineering pursuing our Bachelor's in Technology degree. We saw a need for a system which helps with the identification of crime prone areas as this is a problem statement released by the Bureau of Police Research and Development (BPR-D) in coordination with the Innovation Cell of the Ministry of Education and AICTE in Manthan 2021.

Crime Prone Analysis Portal

Group ID: 16
Group Mentor: Dr.Irfan Siddavatam

Group Members
1814002: Dhruv Doshi
1814006: Shubham Bhakuni
1814015: Labdhi Jain
1814021 Kunj Gala

Fig 27. Dashboard-3



SELECT A SOURCE OF DATA

The data has been collected via 3 different sources. Read the description of the method of data collection and select the souce accordingly

NYPD Historic Data

This data is obtained from the official NYC OpenData website. The dataset consists of crime data from 2021.

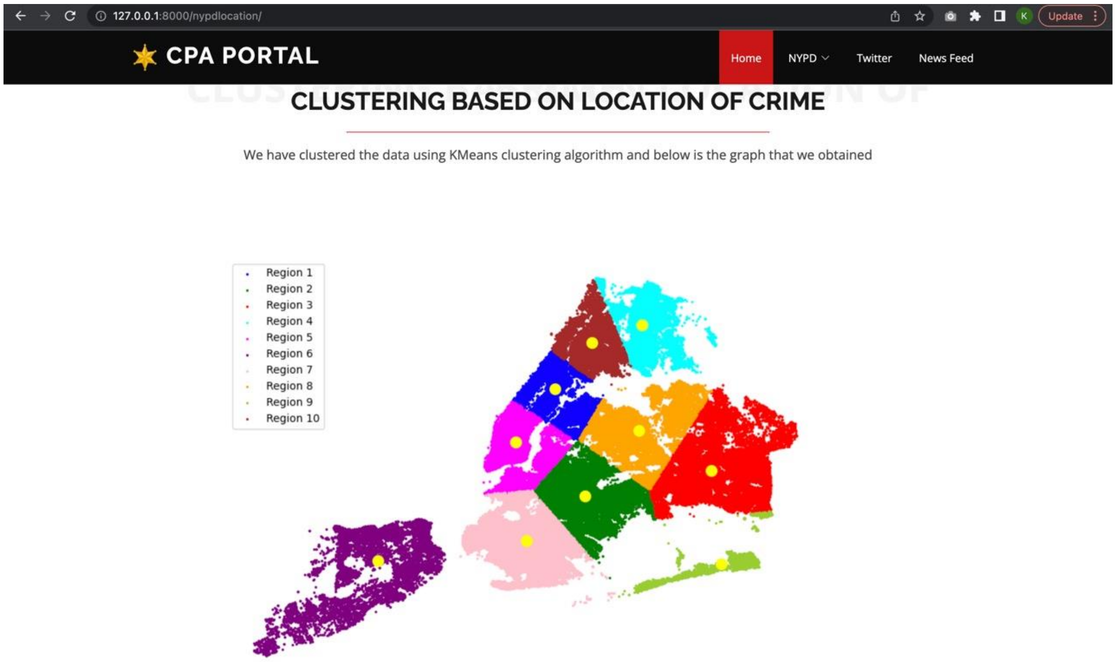
Twitter Data

This data is being extracted from @NYPDNews and @NYPDTips twitter handles. Source

Data from News feed

This data has been extracted from New York Post website's crime section.Source

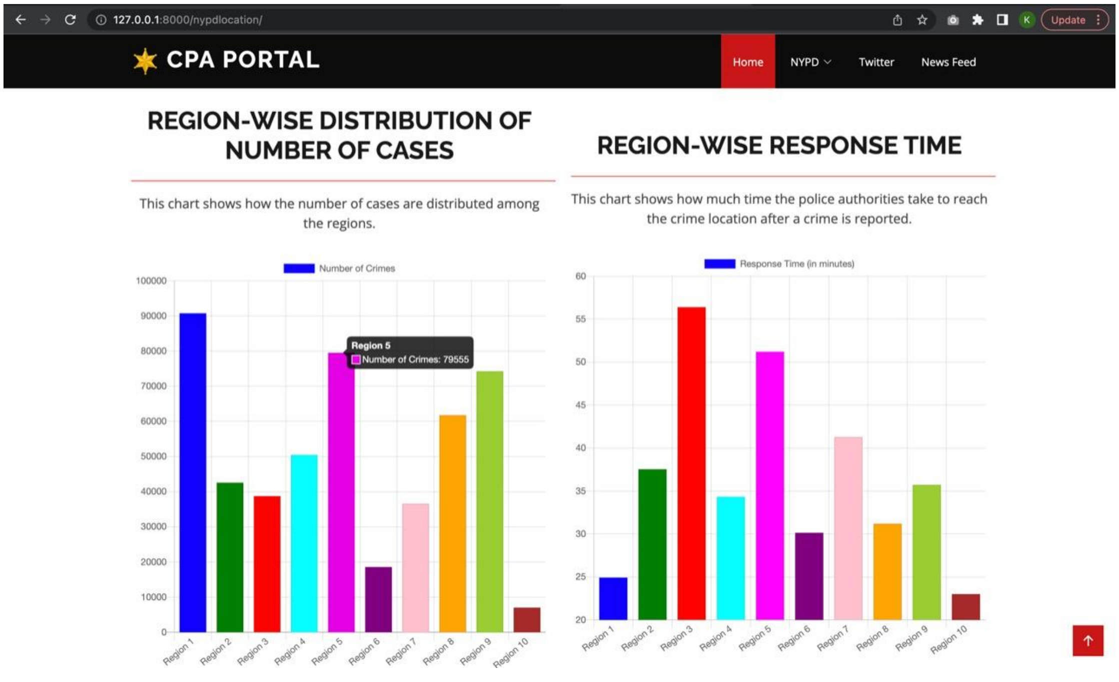
Fig 28. Dashboard-4



CLUSTERING BASED ON LOCATION OF CRIME

We have clustered the data using KMeans clustering algorithm and below is the graph that we obtained

Fig 29. Search by location-1



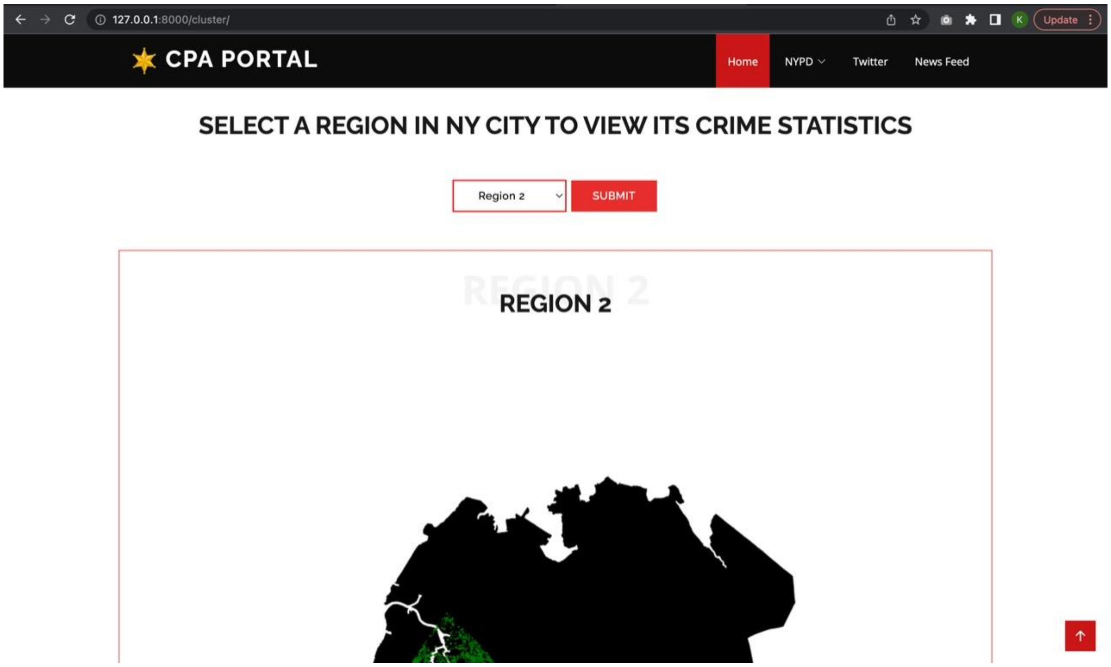
REGION-WISE DISTRIBUTION OF NUMBER OF CASES

This chart shows how the number of cases are distributed among the regions.

REGION-WISE RESPONSE TIME

This chart shows how much time the police authorities take to reach the crime location after a crime is reported.

Fig 30. Search by location-2



SELECT A REGION IN NY CITY TO VIEW ITS CRIME STATISTICS

Fig 31. Search by location-3

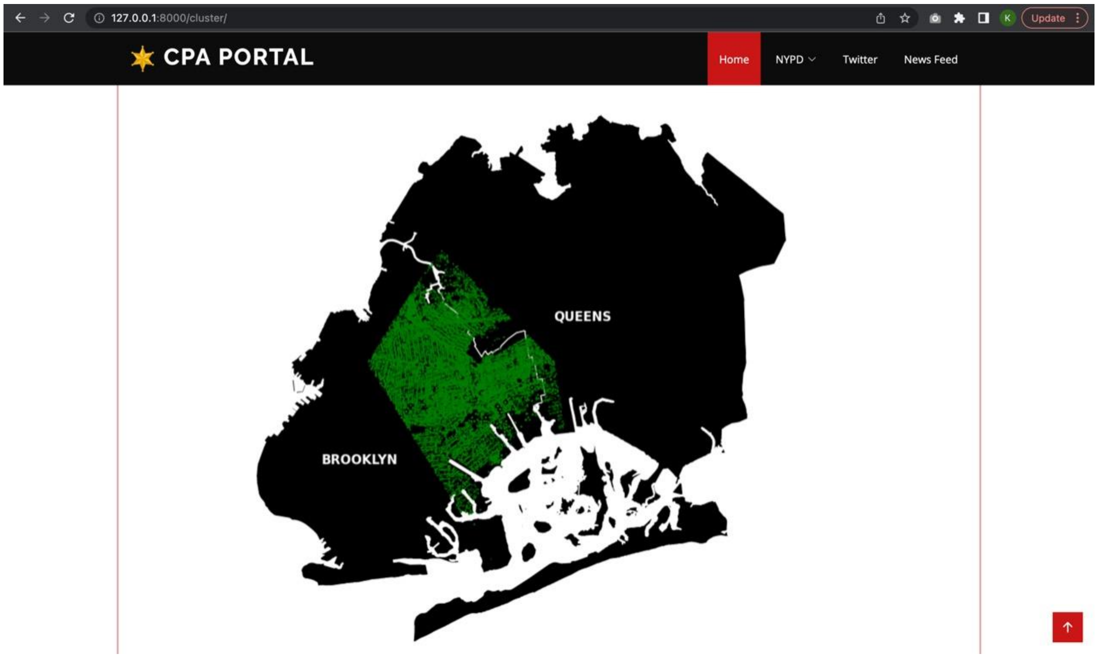
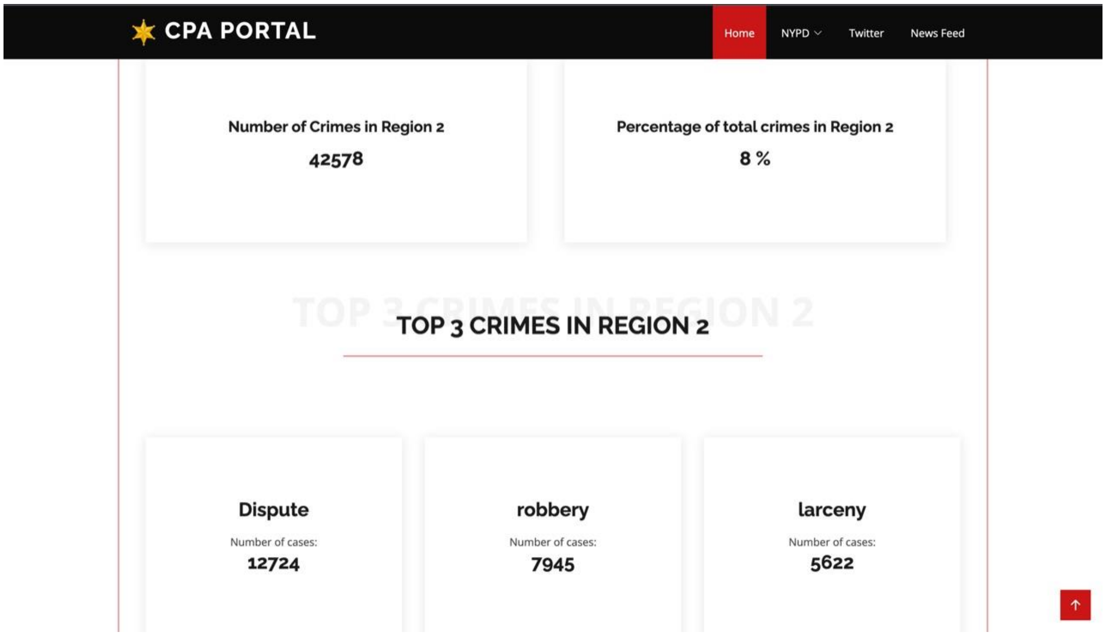


Fig 32. Search by location-4



Number of Crimes in Region 2

42578

Percentage of total crimes in Region 2

8 %

TOP 3 CRIMES IN REGION 2

Dispute

Number of cases:

12724

robbery

Number of cases:

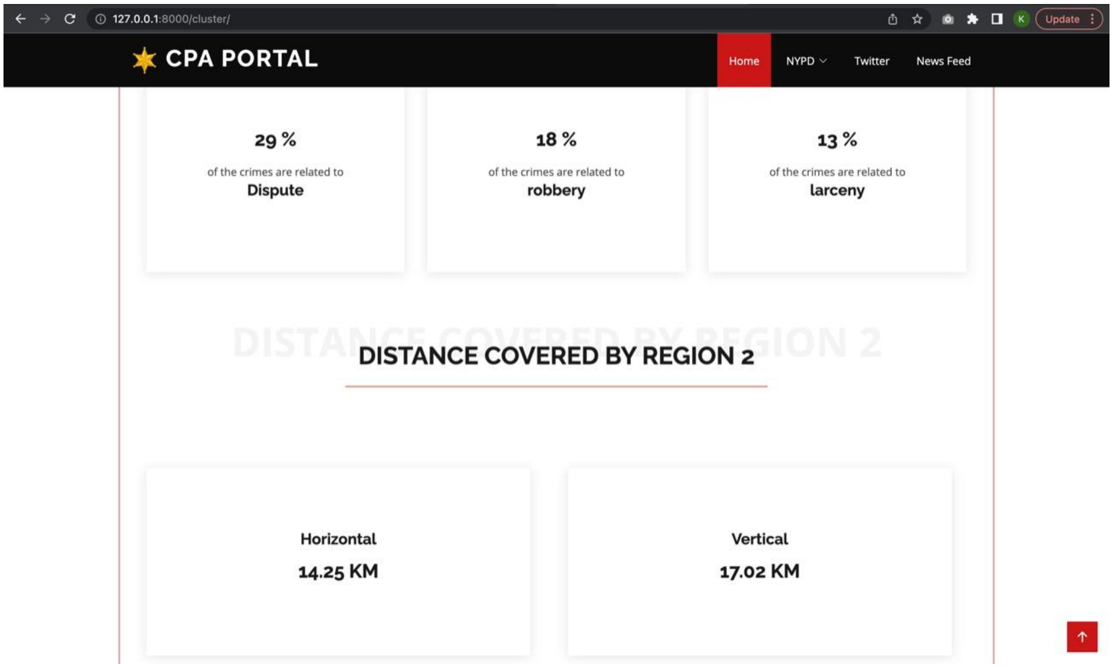
7945

larceny

Number of cases:

5622

Fig 33. Search by location-5



29 %

of the crimes are related to

Dispute

18 %

of the crimes are related to

robbery

13 %

of the crimes are related to

larceny

DISTANCE COVERED BY REGION 2

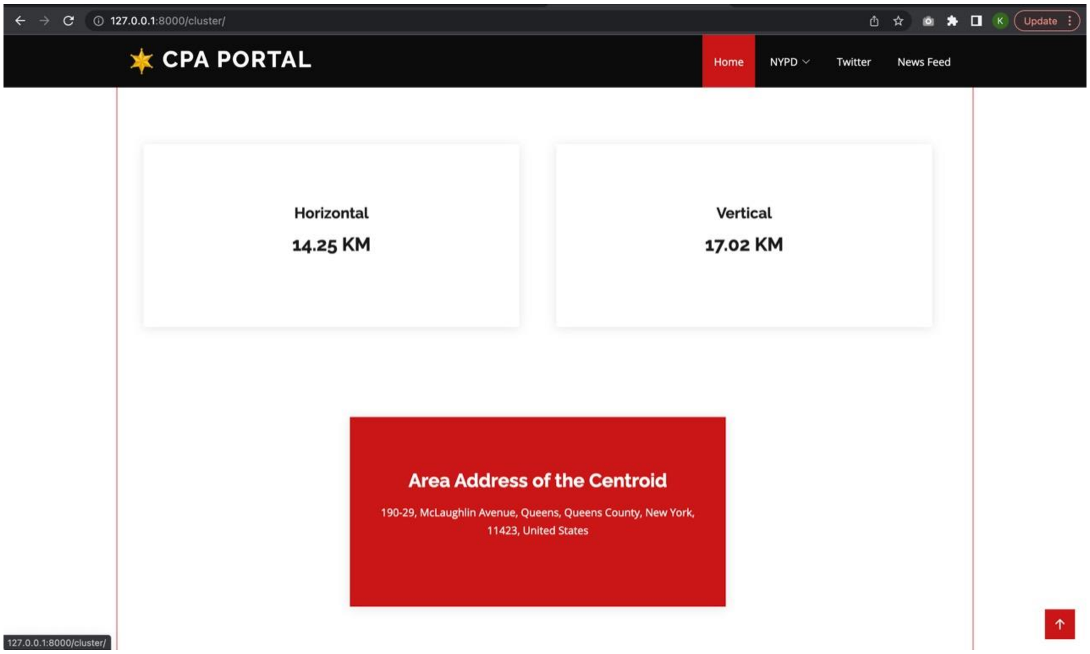
Horizontal

14.25 KM

Vertical

17.02 KM

Fig 34. Search by location-6



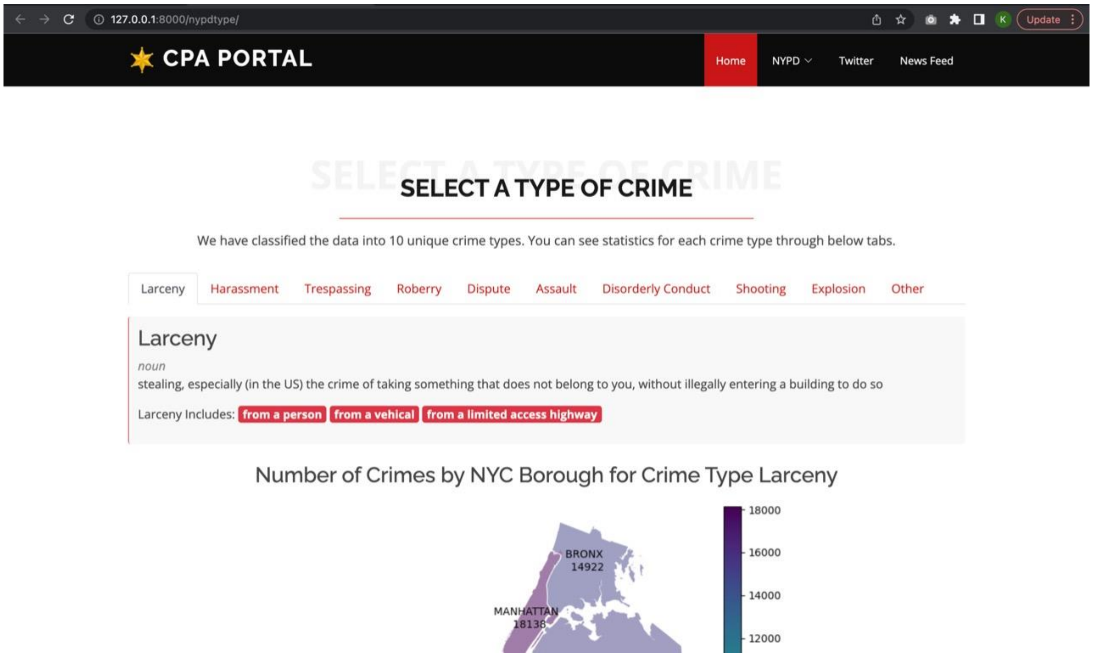
Horizontal
14.25 KM

Vertical
17.02 KM

Area Address of the Centroid

190-29, McLaughlin Avenue, Queens, Queens County, New York, 11423, United States

Fig 35. Search by location-7



SELECT A TYPE OF CRIME

We have classified the data into 10 unique crime types. You can see statistics for each crime type through below tabs.

Larceny

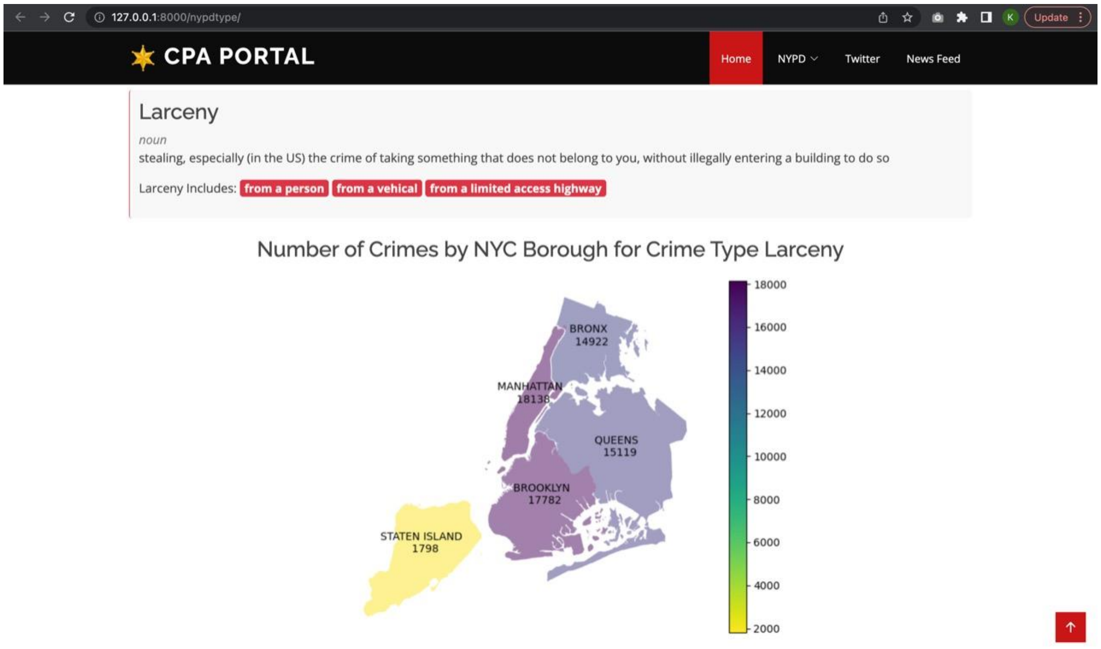
noun

stealing, especially (in the US) the crime of taking something that does not belong to you, without illegally entering a building to do so

Larceny Includes: from a person from a vehical from a limited access highway

Number of Crimes by NYC Borough for Crime Type Larceny

Fig 36. Search by crime type-1



Larceny

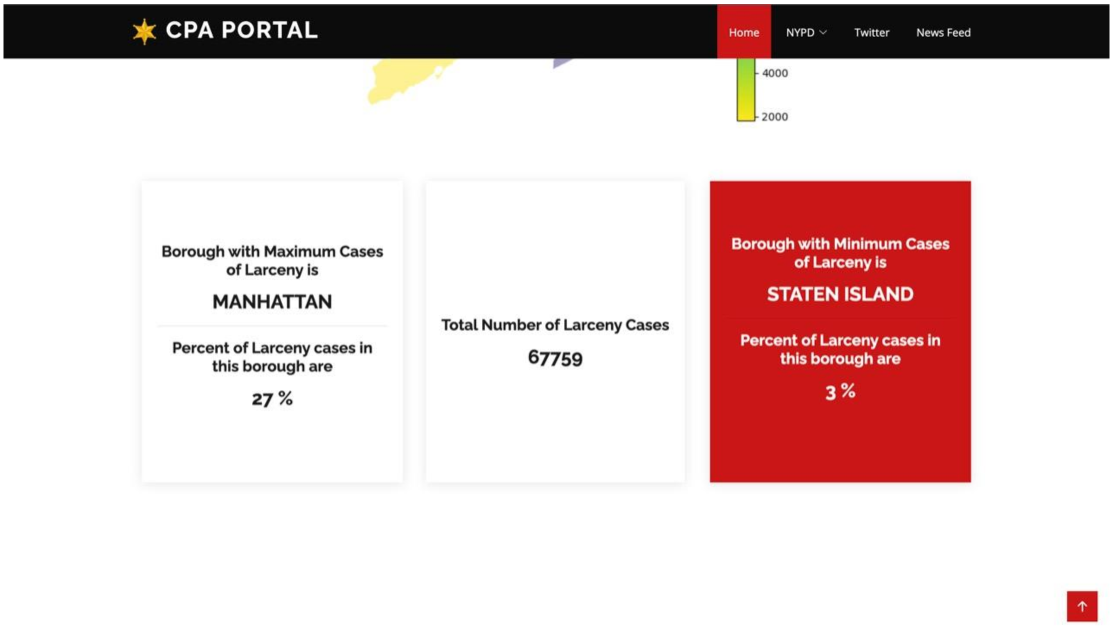
noun

stealing, especially (in the US) the crime of taking something that does not belong to you, without illegally entering a building to do so

Larceny Includes: from a person from a vehical from a limited access highway

Number of Crimes by NYC Borough for Crime Type Larceny

Fig 37. Search by crime type-2



Borough with Maximum Cases of Larceny is

MANHATTAN

Percent of Larceny cases in this borough are

27 %

Total Number of Larceny Cases

67759

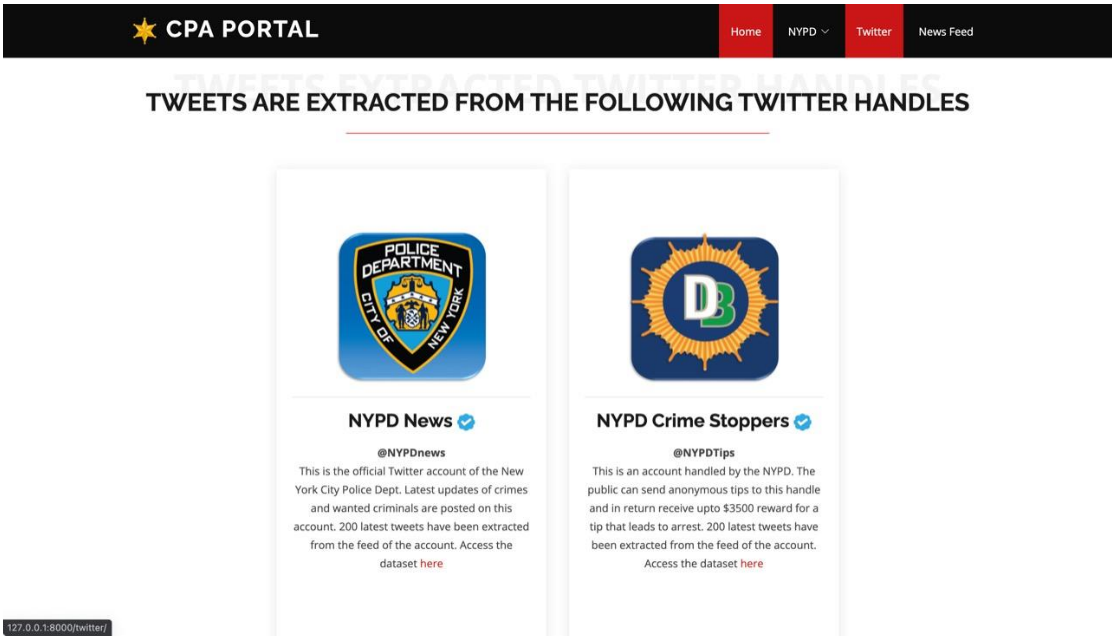
Borough with Minimum Cases of Larceny is

STATEN ISLAND

Percent of Larceny cases in this borough are

3 %

Fig 38. Search by crime type-3



TWEETS ARE EXTRACTED FROM THE FOLLOWING TWITTER HANDLES

NYPD News

@NYPDnews

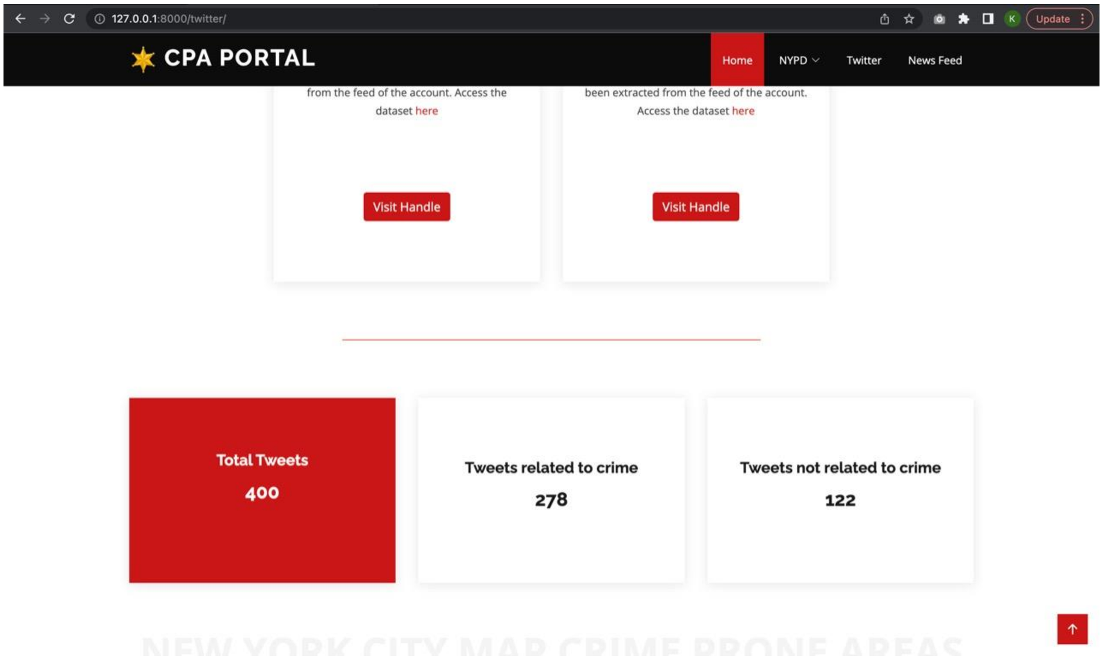
This is the official Twitter account of the New York City Police Dept. Latest updates of crimes and wanted criminals are posted on this account. 200 latest tweets have been extracted from the feed of the account. Access the dataset here

NYPD Crime Stoppers

@NYPDTips

This is an account handled by the NYPD. The public can send anonymous tips to this handle and in return receive upto $3500 reward for a tip that leads to arrest. 200 latest tweets have been extracted from the feed of the account. Access the dataset here

Fig 39. Twitter data source-1



from the feed of the account. Access the dataset here

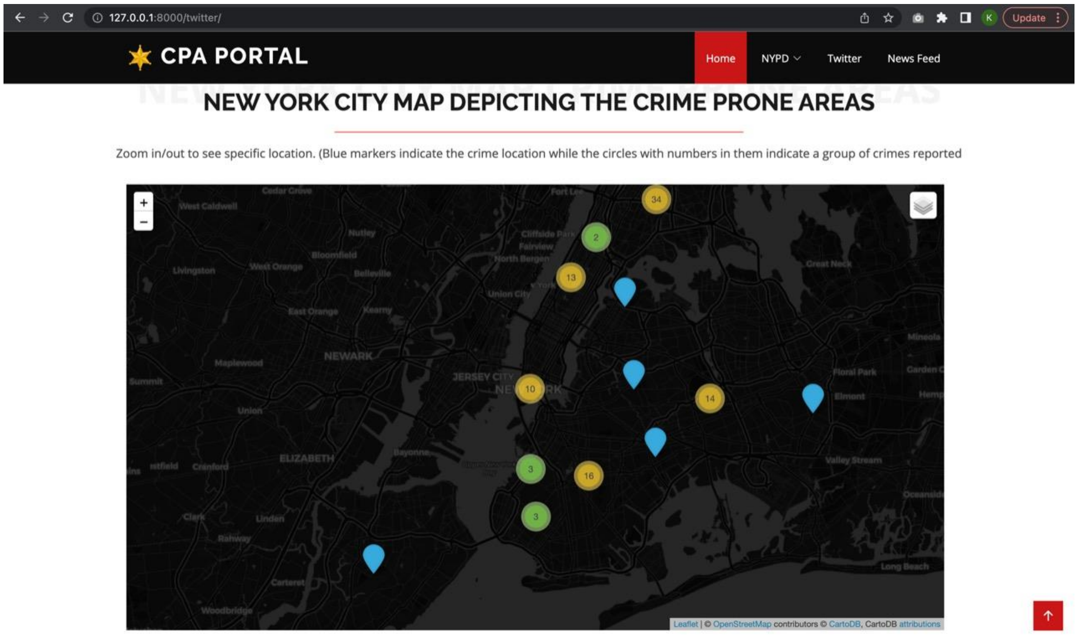
been extracted from the feed of the account. Access the dataset here

Total Tweets
400

Tweets related to crime
278

Tweets not related to crime
122

Fig 40. Twitter data source-2

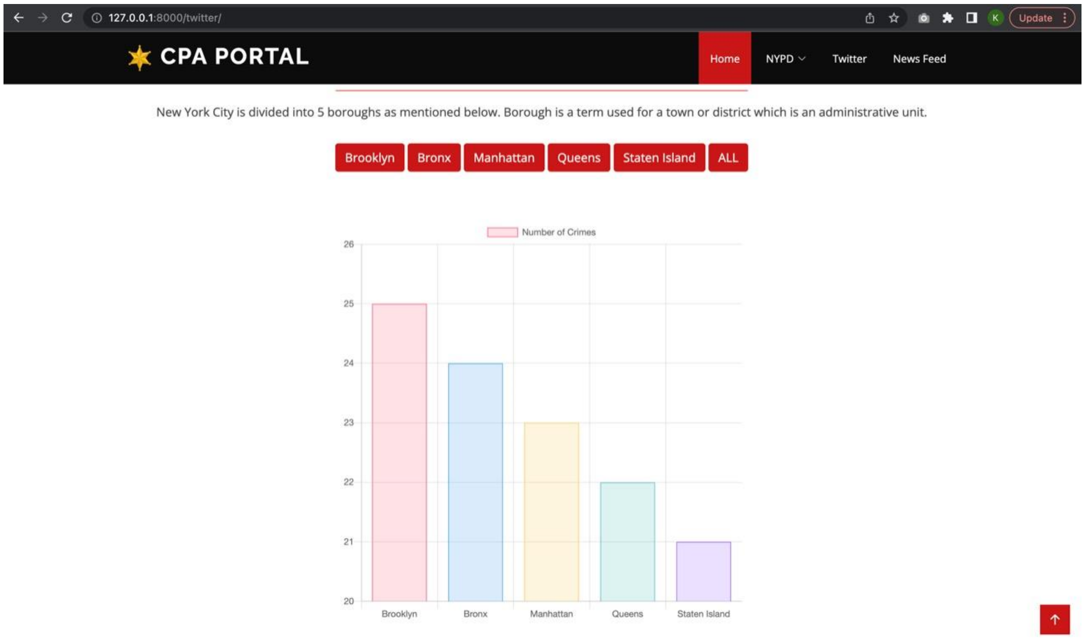


NEW YORK CITY MAP DEPICTING THE CRIME PRONE AREAS

Zoom in/out to see specific location. (Blue markers indicate the crime location while the circles with numbers in them indicate a group of crimes reported

34
2
13
10
14
16
3
3
NEWARK
JERSEY CITY
ELIZABETH
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

Fig 41. Twitter data source-3



New York City is divided into 5 boroughs as mentioned below. Borough is a term used for a town or district which is an administrative unit.

Fig 42. Twitter data source-4

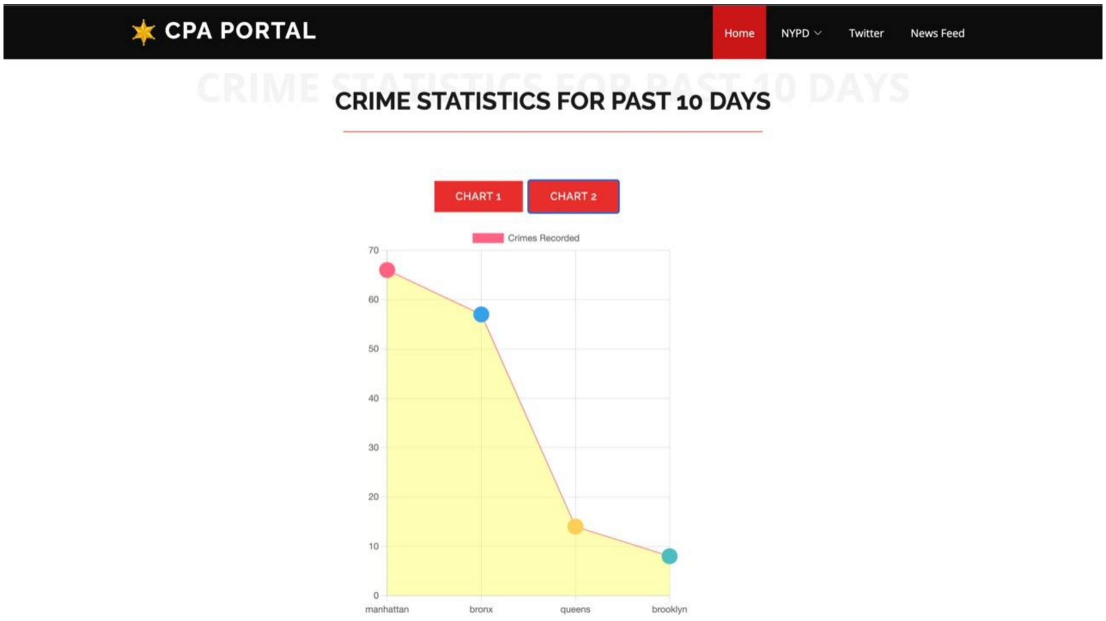


Fig 43. News feed data source-1

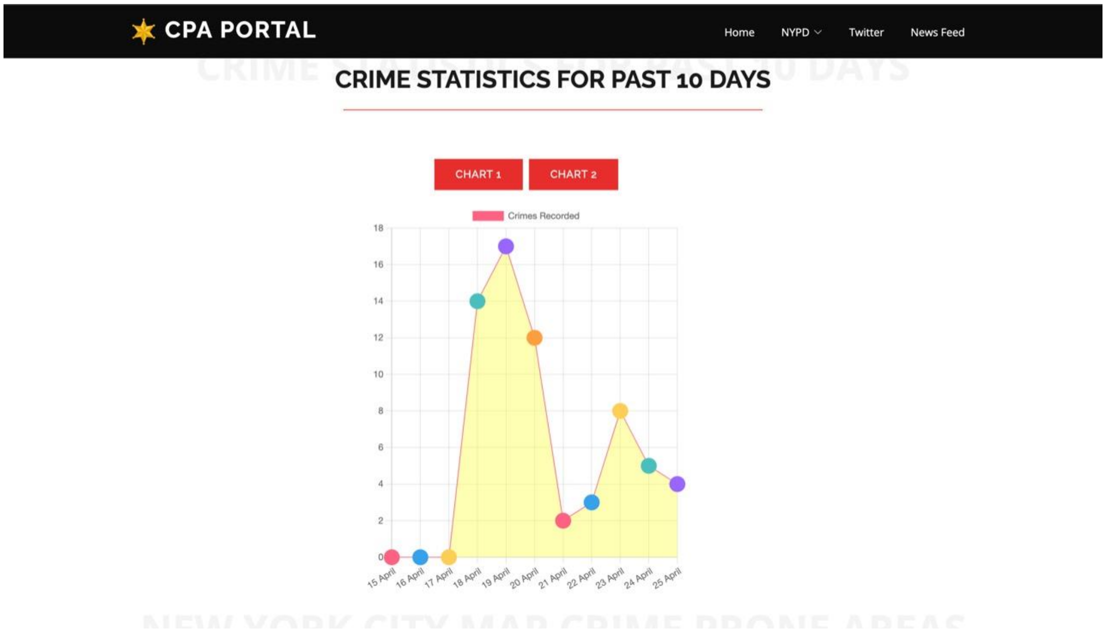


Fig 44. News feed data source-2

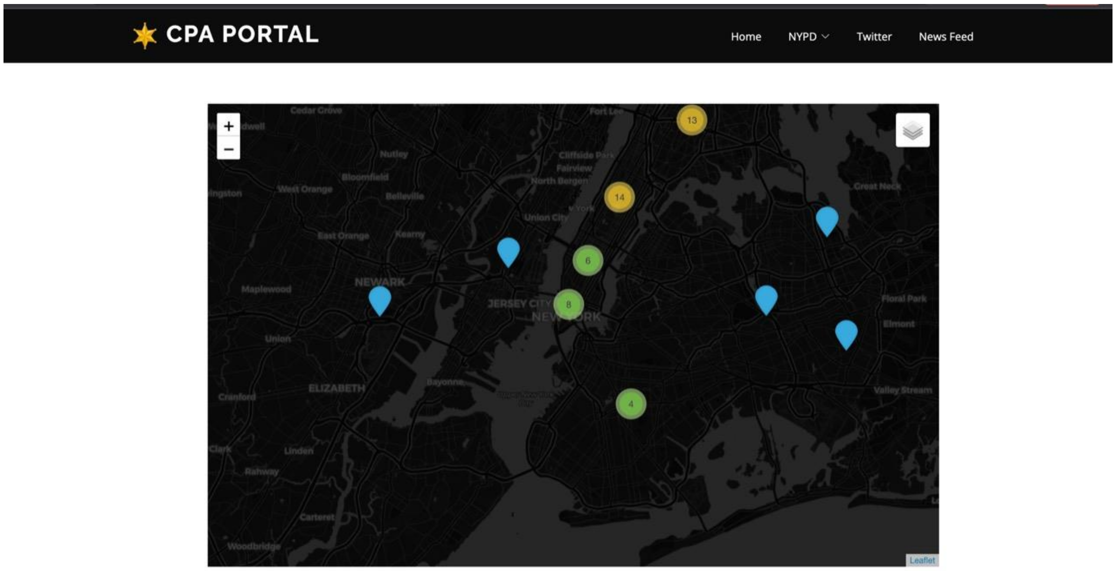
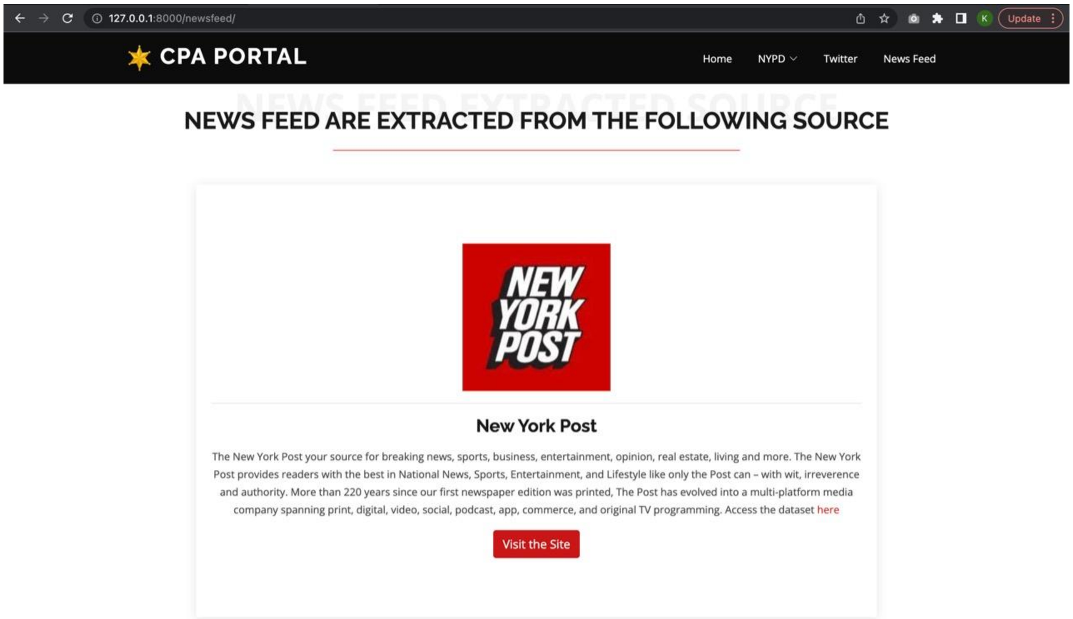


Fig 45. News feed data source-3



NEWS FEED ARE EXTRACTED FROM THE FOLLOWING SOURCE

New York Post

The New York Post your source for breaking news, sports, business, entertainment, opinion, real estate, living and more. The New York Post provides readers with the best in National News, Sports, Entertainment, and Lifestyle like only the Post can – with wit, irreverence and authority. More than 220 years since our first newspaper edition was printed, The Post has evolved into a multi-platform media company spanning print, digital, video, social, podcast, app, commerce, and original TV programming. Access the dataset here

Fig 46. News feed data source-4

CHAPTER 7

Software Testing Document

This chapter presents explanation about the test cases which were carried out in the coding as well as the website designed. Each test case is mentioned with details such as input, output and expected output and then it is summarized in form of a table. Clearing all the test cases adds to the confidence in the system's performance and reliability.

7.1 Introduction

Software Testing is a method to check whether the actual software product matches expected requirements. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements.

7.1.1 System Overview

Emergencies can be of different kinds, from fires to road accidents and assaults, to medical emergencies. It is distressing to be faced with an unanticipated incident or emergency. In our country, 112 is the single emergency number, people in distress can call to get immediate assistance – from the fire brigade, medical team or from the police.

Law enforcement officials across the country would find it incredibly useful if they had an automated system to continually identify areas that are more prone to crimes and policing-related incidents, than others. This would allow them to proactively review and plan suitable resource deployments and patrolling in those areas, and thereby pre-empt and prevent, as far as possible, untoward incidents that could result in an emergency.

Using the power of Machine learning, the system will identify the Crime prone areas enabling better deployment and utilization of the police force. The data gathered from various authorized resources will be pre-processed, analysed, classified and visualized using various machine learning and data analysis techniques. The outputs from this will be shared with the users through a web application wherein users can see the future trends in crime and analyse the area's most prone to criminal activities along with the types of crimes and statistics. The data can be visualized by using the API of maps.

7.1.2 Test Approach

- Crime Classification in Tweets: Using scoring metrics like confusion matrix, precision, recall and accuracy, it is possible to determine how well the classification model results are. Testing this will determine if the model needs to be configured more.

- Clustering of crimes based on location: Unsupervised machine learning algorithms like K means, HDBSCAN, DBSCAN, etc. were used to cluster data points based on the latitude and longitude i.e, location of the crime. Tests on how appropriately the data is clustered will be done.

- Selection of Data Source: This project consists of 3 data sources - Historic NYPD data from call records, data extracted from Tweets and data extracted from news feeds. This test is done to ensure that the user is able to access dashboards of all data sources and is directed to the right data source.

- Selection of Crime Type: Under the NYPD data source the user gets an option to select the crime type among 10 different crime types like disorderly conduct, robbery, larceny, etc. This test is done to check if the right set of statistics are displayed to the user based on the selection criteria.

- Selection of Location: Testing would be carried out on the website showing the map and statistics of the location entered in the Search By Location section of the website.

- Browser Compatibility: Running the website folder in popular browsers like: Google Chrome, Firefox and Microsoft Edge and then checking all whether the required functionalities of the system are working as expected.

7.2 Introduction

This section will provide a detailed description about the features to be tested, features not to be tested and the testing tools and environments to be used.

7.2.1 Features to be tested

- Crime Classification in Tweets: This feature will determine whether a tweet is related to crime or not. Goal is to exclude all the tweets which are not related to crime. Hence, this classification helps in deciding crime relation of a tweet.

- Clustering of crimes based on location: This feature helps in identifying areas that are prone to crime by creating clusters of data points based on the latitude and longitude attributes of that data point. Clustering was done using various unsupervised machine learning algorithms like K means clustering, DBSCAN, HDBSCAN, Birch, and OPTICS. The better the accuracy of clustering, the more accurate will be the identification of crime prone areas. Hence the accuracy of each algorithm was calculated using DB Index and K means algorithm was chosen for clustering as it outperformed the other algorithms.

- Selection of Data Source: The project consists of data taken from three different sources. The sources include. Historic NYPD data taken from call records, data extracted from twitter and data extracted from news feeds. This feature enables users to view the statistics of crime records of the same location but from different sources like social media, news feeds, and emergency call records.

- Selection of Crime Type: This is a feature provided under the NYPD historic data tab in which the user can select from 10 different types of crime like disorderly conduct, robbery, larceny, murder, etc. By selecting a particular crime type the user will be able to see detailed analysis of that crime in the New York area. This might include information on the top 3 locations where the crime occurs, the frequency of that particular crime, etc.

- Selection of Location: This is a very important feature since it lets the user select a location in New York and get the crime statistics of that area. Users get to select one cluster region out of 10 cluster regions marked on the map of New York.
- Browser Compatibility: Users should have the ability to run the website in a browser of their choice. Hence, browser compatibility of the website should be tested on popular browsers.

7.2.2 Features not to be tested

- About Page: The about page will be a static page which will need no access to the database. This page will contain all the details about the system, how it was built, purpose, motivation, etc. This page will be static hence will not be tested. The page will be responsive as well.

7.2.3 Testing tools and environment

Sr. No.	**Resources**	**Description**
1	Web Browser	A web Browser with the latest version for efficiently running the web application.
2	Network	A LAN setup and 1 internet line with at least 5 Mb/s speed.
3	Computer/Laptop	A computer/laptop that has at least windows 7 with more than 2GB RAM and a CPU of 3.4GHz so that application can be tested on a large and extra-large screen size(greater than 1200px).

4	A Tablet	A tablet with the latest version of any browser so that the application can be tested on a medium and small screen size(992px-768px)
5	A Mobile Phone	A mobile phone with the latest version of any browser so that the application can be tested on a small and extra small screen size(smaller than 768px).

Table 4. Testing tools

7.3 Test Cases

7.3.1 Crime Classification in Tweets

Purpose- To correctly classify tweets into crime category or non-crime category. This will help in filtering out the tweets which are non-crime and taking into account only the tweets which are related to crime.

Inputs- Tweet messages

Expected outputs-

Incorrect outputs: Tweet messages separated into 2 categories incorrectly.

Correct outputs: Tweet messages separated into 2 categories correctly.

Test procedure- Using testing metrics such as confusion matrix, precision, accuracy, recall it can be determined that the classification has been done correctly or incorrectly.

7.3.2 Clustering of crimes based on location:

Purpose- This test is done to check if the data points have been accurately clustered because the identification of crime prone areas will be based on how well the points have been clustered.

Inputs- We use Davies-Bouldin(DB) Index hence the dataset with latitude and longitude attributes(x), and the output of k means fit on the x(y_predict) as inputs.

Expected outputs-

Incorrect outputs: Higher values of DB index means bad clustering

Correct outputs: Lower values of DB index means good clustering

Test procedure- Use different clustering algorithms with the same data points as input and compute the DB index. Use the algorithm which provides a low DB index and hence better clustering of data.

7.3.3 Selection of Data Source

Purpose- This project consists of 3 data sources - Historic NYPD data from call records, data extracted from Tweets and data extracted from news feeds. This test is done to ensure that the user is able to access dashboards of all data sources and is directed to the right data source.

Inputs- Click on the button of any data source.

Expected outputs-

Incorrect outputs: The user is directed to the dashboard of another data source.

Correct outputs: The user is directed to the dashboard of the selected data source.

Test procedure- The user has to click on each data source button one after the other. On clicking the Historic data button the user should have 2 more options(search by location and search by crime type). On clicking on the news feed and twitter data source buttons the user should be directed to their respective dashboards.

7.3.4 Selection of Crime Type

Purpose- Under the NYPD data source the user gets an option to select the crime type among 10 different crime types like disorderly conduct, robbery, larceny, etc. This test is done to check if the right set of statistics are displayed to the user based on the selection criteria.

Inputs- Select a crime type from a list of 10 crime types. Press the submit button.

Expected outputs-

Incorrect outputs: No output is returned or statistics of another crime type is displayed.

Correct outputs: On clicking submit, a detailed statistics of the selected crime type is displayed on the same page below the list of options.

Test procedure- The user has to select a crime type from the list of crime types available and click on the submit button. Then the user can check if the details of the selected crime type are displayed on that page or if there is any discrepancy in the dashboard.

7.3.5 Selection of Location

Purpose- Allow users to view the spread of crime and the statistics of a location of their choice out of the location choices provided.

Inputs-

Incorrect inputs: Selecting a location which is not present in the list of locations provided.

Correct inputs: Selecting a location which is present in the list of locations provided.

Expected outputs-

Incorrect outputs: Not displaying the results to the user when a wrong location is selected.

Correct outputs: Displaying the results according to the location that is selected by the user. The map and statistics should match the selection made by the user.

Test procedure- Testing would be carried out on the website showing the map and statistics of the location entered in the Search By Location section of the website.

7.3.6 Browser Compatibility

Purpose- To check whether the website runs on all the popular browsers properly and without any errors.

Inputs- Running the website folder in popular browsers like: Google Chrome, Firefox and Microsoft Edge.

Expected outputs-

Incorrect outputs: Error is shown while loading the home page and dashboard.

Correct outputs: No error is shown and the website runs flawlessly on that particular browser.

Test procedure- Running the website folder in popular browsers like: Google Chrome, Firefox and Microsoft Edge and then checking all whether the required functionalities of the system are working as expected.

7.4 Test results:

Test ID	**Test Case Name**	**Input**	**Expected Results**	**Actual Results**	**Pass/ Fail**
1.1	Crime Classification in Tweets	Tweet messages	More than 95% accuracy	89% accuracy	Pass
1.2	Clustering of crimes based on	Dataset with latitude and longitude	Lower values of DB	0.7728	Pass

	location	attributes(x), and the output of k means fit on the x(y_predict)	index		
1.3	Selection of Data Source	Click on the button of any data source	The user should be directed to the dashboard of the selected data source.	The user is directed to the right data source.	Pass
1.4	Selection of Crime Type	Select a crime type from a list of 10 crime types. Press the submit button.	On clicking submit, a detailed statistics of the selected crime type is displayed on the same page below the list of options.	The selected crime type statistics are displayed.	Pass
1.5	Selection of Location	Location selected by the user	Map and statistics related to the selected location are displayed	Map and statistics related to the selected location are displayed	Pass
1.6	Browser Compatibility	Open website folder on a browser	Website runs smoothly on Chrome, FIrefox and Edge	Website runs smoothly on Chrome, FIrefox and Edge	Pass

Table 5. Test results

CHAPTER 8

Conclusions and Scope for Future Work

This chapter presents the concluding statements, brief about the functionalities completed and limitations in the project. Knowing the limitations of the system is highly important for further improving this system. Also, the scope of this project for future work is stated explaining what can be done further to expand the usability and to increase the scale of this project.

8.1 Conclusion

To tackle the increasing crime rate through our society, we have developed a system that helps in predictive policing by analyzing crimes in locality by using the power of Machine learning and data analysis. The system will help authorities continually identify areas prone to crimes, making deployments and patrolling in those areas more effective. The website developed also will help the users easily understand where the crime rate is higher using visual aids like heat maps and statistics related to it. The system is divided into 3 major sections according to the data sources used for analysis:

1. NYPD Historic data
2. Data from Twitter handles
3. News feed

These 3 different sources provide a different perspective on the spread of crime in a given location since the sources for data collection is different in all sections.

8.2 Limitations

The project has a few limitations which could be improved due to either time constraints or lack of computation power. First, we had a constraint of RAM. Only 12 GB of RAM was available at the time of implementation and the models had to be tweaked to work within this specification. Second, educational access to the Twitter API had an upper limit of 200 tweets per twitter handle. Even if it was possible to process thousands of tweets, due to the access constraint, we were limited to 400 tweets.

Third, since we had extracted raw data directly from Twitter, the data was not labelled. We had to manually label the data, hence the possibility of human errors increases. Forth, crime type detection of data from Twitter and News feeds was also possible but it seemed beyond the scope of this project and also due to time constraints, this feature could not be added.

Fifth, it was beyond our expertise to be able to make the website dashboard more dynamic and responsive as expected by a user. Though all the basic requirements are fulfilled, the dashboard might seem restricted to certain features only till some extent.

8.3 Scope for Future Work

For the NYPD Historic data, more in-depth inferences can be drawn pertaining to the location and time of the crimes reported. More insights can be gained by analyzing crime patterns across the given state i.e., New York and this can be helpful for predicting crimes in that area. Moreover, the amount of data taken into consideration could be much more if there are no computation constraints.

For Twitter and News feed data sources, there is a vast opportunity for future work. Due to API constraints, only 200 tweets per Twitter handle were extracted. If that constraint is removed and all tweets of each handle are extracted, then the analysis results will improve. In this project, total 400 tweets were taken as corpus. Whereas, the handles each had thousands of tweets. If all the tweets are taken and more handles are taken into consideration, then rich results could be obtained.

REFERENCES

[1] A. A. Almuhanna, M. M. Alrehili, S. H. Alsubhi and L. Syed, "Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021, pp. 23-30, doi: 10.1109/CAIDA51941.2021.9425120.

[2] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[3] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, "Crime analysis and prediction using data mining," 2014 First International Conference on Networks Soft Computing (ICNSC2014), 2014, pp. 406-412, doi: 10.1109/CNSC.2014.6906719.

[4] A. J. Park, V. Spicer, H. H. Tsang, K. Behiels and J. Song, "Discovering Crime Trends and Patterns Using Three-Dimensional Visual Analytics," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0545-0549, doi:10.1109/IEMCON.2019.8936251.

[5] P. Sharma, S. Khater and S. Sharma, "Data visualization of crimes in a city using machine learning," 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), 2020, pp. 55-60, doi: 10.1109/INBUSH46973.2020.9392166.

[6] M. A. Boni and M. S. Gerber, "Area-Specific Crime Prediction Models," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 671-676, doi: 10.1109/ICMLA.2016.0118.

[7] S. Mukherjee and K. Sarkar, "Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas," 2020 IEEE Calcutta Conference (CALCON), 2020, pp. 444-450, doi: 10.1109/CALCON49167.2020.9106554.

[8] R. Yadav and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 2018, pp. 1-5, doi: 10.1109/ICRAIE.2018.8710407.

[9] Hana Anber, Akram Salah and A. A. Abd El-Aziz, “A Literature Review on Twitter Data Analysis,” 2016 International Journal of Computer and Electrical Engineering, 2016, doi: 10.17706/ijcee.2016.8.3.241-249

[10] Heidi Cohen, “How reliable is Twitter?”, 2013 Accessed on: 01/12/2022 [Online] Available: https://heidicohen.com/reliable-twitter-research/

[11] N Schwitter, “GOING DIGITAL Web data collection using Twitter as an example”, 2020, Accessed on: 01/12/2022 [Online] Available: www.oxfam.org

[12] “Twitter API documentation” 2022, Twitter Developer Platform, Accessed on: 01/15/2022 [Online] Available: https://developer.twitter.com/en/docs/developer-portal/overview

[13] Zhang AJ, Albrecht L, Scott SD, “Using Twitter for Data Collection With Health-Care Consumers: A Scoping Review”, 2017 International Journal of Qualitative Methods, 2017, doi: https://doi.org/10.1177/1609406917750782

[14] Hana Anber, Akram Salah, A. A. Abd El-Aziz, “A Literature Review on Twitter Data Analysis” International Journal of Computer and Electrical Engineering 2016, doi: 10.17706/ijcee.2016.8.3.241-24

[15] D. M. Raza and D. B. Victor, "Data mining and Region Prediction Based on Crime Using Random Forest," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 980-987, doi: 10.1109/ICAIS50930.2021.9395989.

ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to the successful completion of this project. First, we would like to express our utmost gratitude to our project guide Dr. Irfan Siddavatam for helping us improve the project and bringing it to the stage it is currently at by guiding us throughout the development of this project. We would also like to thank our college K.J. Somaiya College of Engineering and our department of Information technology for providing us with this opportunity.