“PAASBAAN – Crime Prediction and Classification in Indore City”

A Minor Project Report submitted to

Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal

in partial fulfillment of the requirements for the award of

Degree of

Bachelor of Engineering

in

Information Technology

by

Kunal Diwan (0832IT151017)
Sahitya Nigam (0832IT151038)
Sourabh Tiwari (0832IT151054)
Vikramaditya Singh Bhati (0832IT151060)

Under the guidance of

Ms. Grishma Pandey
(Assistant Professor)

Session: 2017-18

Department of Information Technology

Chameli Devi Group of Institutions, Indore
452 020 (Madhya Pradesh)

DECLARATION

We certify that the work contained in this report is original and has been done by us under the guidance of my supervisor(s).

a. The work has not been submitted to any other Institute for any degree or diploma.
b. We have followed the guidelines provided by the Institute in preparing the report.
c. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
d. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references.

Name and Signature of Project Team Members:

Sr. No.	**Enrollment No.**	**Name of students**	**Signature of students**
1.	0832IT151017	Kunal Diwan	
2.	0832IT151038	Sahitya Nigam	
3.	0832IT151054	Sourabh Tiwari	
4.	0832IT151060	Vikramaditya Singh Bhati	

CHAMELI DEVI GROUP OF INSTITUTIONS, INDORE

CERTIFICATE

Certified that the project report entitled, "Paasbaan – Crime prediction and classification in Indore city" is a bonafide work done under my guidance by Kunal Diwan,Sahitya Nigam,Sourabh Tiwari & Vikramaditya Singh Bhati in partial fulfillment of the requirements for the award of degree of Bachelor of Engineering in Information Technology Engineering.

Date:

(Ms.Grishma Pandey)

Guide

(Prof. Jasvant Mandloi)

Head of the Department

(Dr. K.S. Jairaj)

Dean, CDGI

(Internal Examiner)

(External Examiner)

CHAMELI DEVI GROUP OF INSTITUTIONS

INDORE

ACKNOWLEDGEMENT

We have immense pleasure in expressing our sincerest and deepest sense of gratitude towards our guide Ms. Grishma Pandey for the assistance, valuable guidance and co- operation in carrying out this Project successfully. We have developed this project with the help of Faculty members of our institute and we are extremely grateful to all of them. We also take this opportunity to thank Head of the Department Prof. Jasvant Mandloi, and Dean of Chameli Devi Group of Institutions, Dr. K.S. Jairaj, for providing the required facilities in completing this project. We are greatly thankful to our parents, friends and faculty members for their motivation, guidance and help whenever needed.

Name and signature of team Members:

1. Kunal Diwan
2. Sahitya Nigam
3. Sourabh Tiwari
4. Vikramaditya Singh Bhati

Abstract

To be better prepared to respond to criminal activity, it is important to understand patterns in crime. In our project, we analyze crime data from the city of Indore, scraped from publicly available website of Indore Police.

At the outset, the task is to predict which category of crime is most likely to occur given a time and place in Indore.
The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand.

The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

We also attempt to make our classification task more meaningful by merging multiple classes into larger classes. Finally, we report and reflect on our results with different classifiers, and dwell on avenues for future work.

List of Figures

List of Tables

TABLE OF CONTENTS

Chapter-1
Introduction

Paasbaan which is an Urdu word meaning protector, many important questions in public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and more concerted efforts by citizens and authorities to create healthy neighborhood environments.

With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

The inputs to our algorithms are time (hour, day, month, and year), place (latitude and longitude), and class of crime:

- Act 379 - Robbery
- Act 13 - Gambling
- Act 279 - Accident
- Act 323 - Violence
- Act 302 - Murder
- Act 363 - Kidnapping

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbors), Decision Trees, and Random Forests.

We also perform multiple classification tasks – we first try to predict which of 6 classes of crimes are likely to have occurred, and later try to differentiate between violent and non-violent crimes.

1.1 Rationale

Madhya Pradesh's commercial capital Indore has topped the crime record in the country in 2008 followed by Bhopal and Jaipur. Crime rate of Indore was 941.4, which is the highest in the country, according to National Crime Record Bureau's (NCRB) report - "Crime in India 2008".

With the rapid urbanization and development of big cities and towns, the graph of crimes is also on the increase. This phenomenal rise in offences and crime in cities is a matter of great concern and alarm to all of us.

There are robberies, murders, rapes and what not. The frequent and repeated thefts, burglaries, robberies, murders, killings, rapes, shoplifting, pick pocketing, drug- abuse, illegal trafficking, smuggling, theft of vehicles etc., have made the common citizens to have sleepless nights and restless days.

They feel very insecure and vulnerable in the presence of anti-social and evil elements. The criminals have been operating in an organized way and sometimes even have nationwide and international connections and links.

1.2 Goal

Much of the current work is focused in two major directions:

- Predicting surges and hotspots of crime, and
- Understanding patterns of criminal behavior that could help in solving criminal investigations.

1.3 Objective

The objective of our work is to:

- Predicting crime before it takes place.
- Predicting hotspots of crime.
- Understanding crime pattern.
- Classify crime based on location.
- Analysis of crime in Indore.

1.4 Methodology

1.4.1 Machine learning

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets.We are surrounded by a machine learning based technology: search engines learn how to bring us the best results (while placing pro_table ads), anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms.

Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, finedetailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired or re_ned through learning from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to learn and adapt

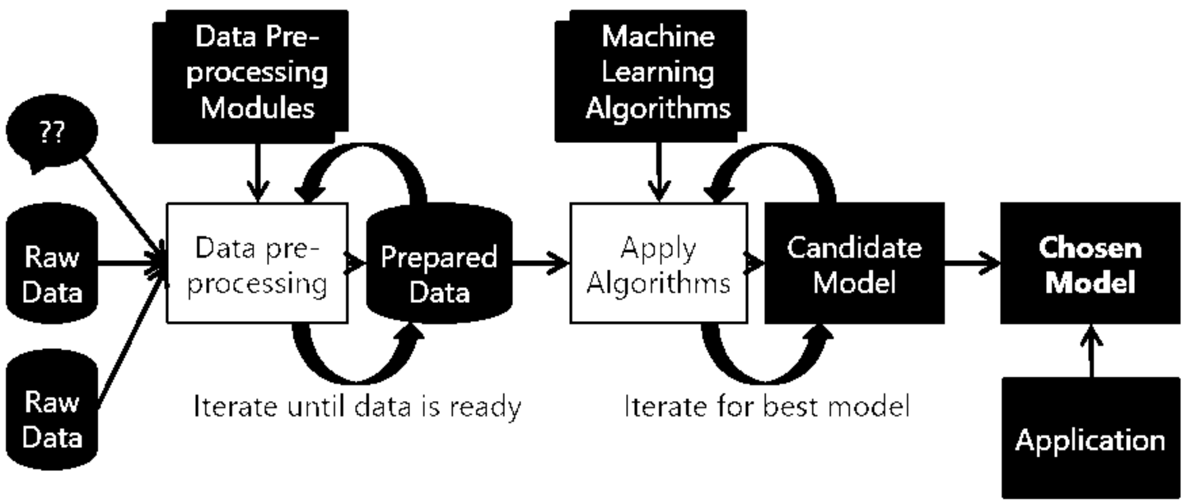

Fig 1.1-Machine learning process

The inputs to our algorithms are time (hour, day, month, year), place (latitude and longitude), class of crime

- Act 379-Robbery
- Act 13-Gambling
- Act 279-Accident
- Act 323-Violence
- Act 302-Murder
- Act 363-Kidnapping

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbors), Decision Trees, and Random Forests.

1.4.2 Our Dataset

Dataset which we are using is scraped daily from website of Indore police which is publically available.

But the dataset is Hindi and in order to perform machine learning this data cannot be used as it is.

Hence the data needs to be processed

Features of this dataset

- थाना : Police Station
- थाना अपराध/मर्ग क्रमांक : Police Station identification number
- धारा : I.P.C. act number
- फरियादी का नाम एवं पता : Complainant name & address
- आरोपी का नाम एवं पता : Accused name & address
- घटना स्थल : Incident place
- घटना दिनांक व समय : Incident date & time
- कायमी दिनांक व समय : Reporting date & time
- विलंब से कायमी का कारण : Reason of Time delay in reporting to police
- घटना के कारण सहित विवरण : Incident information in brief

थाना	थाना अपराध/मर्ग क्रमांक	धारा	फरियादी का नाम एवं पता	आरोपी का नाम एवं पता	घटना स्थल	घटना दिनांक व समय	कायमी दिनांक व समय	विलंब से कायमी का कारण	घटना के कारण सहित विवरण
थाना जूनी इंदौर	89/18	379,	सुनील अन्जाने उम्र २८ वर्ष पिता/पति सुरेश अन्जाने निवासी ५६ टाईप २ पीएनटी कालोनी खातीवालाटैंक इन्दौर	अज्ञात -,	४६ टाईप २ बीएसएनएल क्वार्टर खातीवालाटैंक इन्दौर	08-02-18 11:0 के 12:0 बीच	2/10/2018 12:45:00 PM	फरियादी के थाना आने पर	कोई अज्ञात व्यक्ति फरियादी की बिना नम्बर की मोटर साकयकल को रखे स्थान से चोरी कर ले गया
थाना राऊ	64/18	13 जुआ एक्ट	शासन तर्फे पुलिस सउनि महेश श्रीवास्तव पिता/पति भगवान दास निवासी पुलिस थाना राऊ	दिनेश - केशरसिह , सुरेश -रमेश यादव मुकेश - सत्यनारायण पंवार ,	बाडी मोहल्ला राऊ	10-02-18 19:10 के 20:0 बीच	2/10/2018 8:10:00 PM		घटना दिनाक को आरोपीयो को तास पत्तो सें हारजीत का दाव लगाते हुवे पकडा

Table 1.1: Police Dataset

1.4.3 Preprocessing

Before implementing machine learning algorithms on our data, we went through a series of preprocessing steps with our classification task in mind. These included:

- Dropping features such police station, station number, Complainant name & address ,Accused name & address
- Dropping features such as Resolution, Description and Address: The resolution and description of a crime are only known once the crime has occurred, and have limited significance in a practical, real-world scenario where one is trying to predict what kind of crime has occurred, and so, these were omitted. The address was dropped because we had information about the latitude and longitude, and, in that context, the address did not add much marginal value.

- The timestamp contained the year, date and time of occurrence of each crime. This was decomposed into five features: Year (2018), Month (1-12), Date (1-31), Hour (0-23) and Minute (0-59).

Following these preprocessing steps, we ran some out-of-the box learning algorithms as a part of our initial exploratory steps. Our new feature set consisted of 9 features, all of which were now numeric in nature.

timestamp	act379	act13	act279	act323	act363	act302	latitude	longitude
28-02-2018 21:00	1	0	0	0	0	0	22.73726	75.87599
28-02-2018 21:15	1	0	0	0	0	0	22.72099	75.87608
28-02-2018 10:15	0	0	1	0	0	0	22.73668	75.88317
28-02-2018 10:15	0	0	1	0	0	0	22.74653	75.88714

Table 1.2: Dataset after Preprocessing

1.4.4 Methodology

After the preprocessing described in the previous sections, we had three different classifications problems to solve, which we proceeded to attack with an assortment of classification algorithms. The following are the algorithms which we are using:

- KNN(K- Nearest neighbors)
- Decision Tree
- Random Forests

1.5 Role and Responsibilities

Role	Name	Responsibilities
Data Entry & Tester	Kunal Diwan	• Data Entry • Data Preprocessing • Testing
Data Entry & Tester	Sahitya Nigam	• Data Entry • Data Preprocessing • Testing
Data Scientist & GUI Developer	Sourabh Tiwari	• Data Entry • Data Preprocessing • GUI(Flask) • Documentation • Machine Learing • Data Analysis • Data Mining • Kernel Designing • Data Visualization
Data Scientist & GUI Developer	Vikramaditya Singh Bhati	• Data Entry • Data Preprocessing • GUI(Flask) • Documentation • Machine Learing • Data Analysis • Data Mining • Kernel Designing • Data Visualization

1.6 Contribution of Project

1.6.1 Market potential

The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand.

The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown.The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative.Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes.

For Example: it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

1.6.2 Innovativeness

The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement.This kind of data analysis was technologically impossible a few decades ago, but the hope is that recent developments in machine learning are up to the task.

1.6.3 Usefulness

Public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and more concerted efforts by citizens and authorities to create healthy neighborhood environments. With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

1.7 Report Organization

The remaining section of the report is structured as follows:

- **Chapter 2** provides detailed business and technical requirements
- **Chapter 3** provides analysis and design of this project
- **Chapter 4** provides Construction, implementation details of this project
- **Chapter 5** provides Conclusion and future scope as well as future application of this project

Chapter-2
Requirement Engineering

2.1 Functional Requirement

The functional requirements describe the core functionality of the application.

2.1.1 Interface Requirement:

- Screen 1 to accept user inputs.
- Field 1 accepts numeric data for latitude and longitude.
- Field 2 accepts date & time.
- Button 1 overall analysis.
- Submit button to send data of Field 1 & 2 to Kernel.
- Screen 2 displays predicted values.
- Screen 3 displays analysis.

2.2 Non Functional Requirement

Non function requirement are those requirement of the system which are not directly concerned with specific functional delivered by the system. They may be related to emergent properties such as reliability, extendibility, usability,etc.

- To provide prediction of crime.
- To provide maximum accuracy.
- Provide visualized analysis.
- Ease of use.
- Availability
- Reliability
- Maintainability

Chapter-3

Analysis and Design

3.1 Use case diagram

Use case diagram represent the overall scenario of the system. A scenario is nothing but a sequence of steps describing an interaction between a user and a system.

Thus use case is a set of scenario tied together by some goal. The use case diagram are drawn for exposing the functionalities of the system.

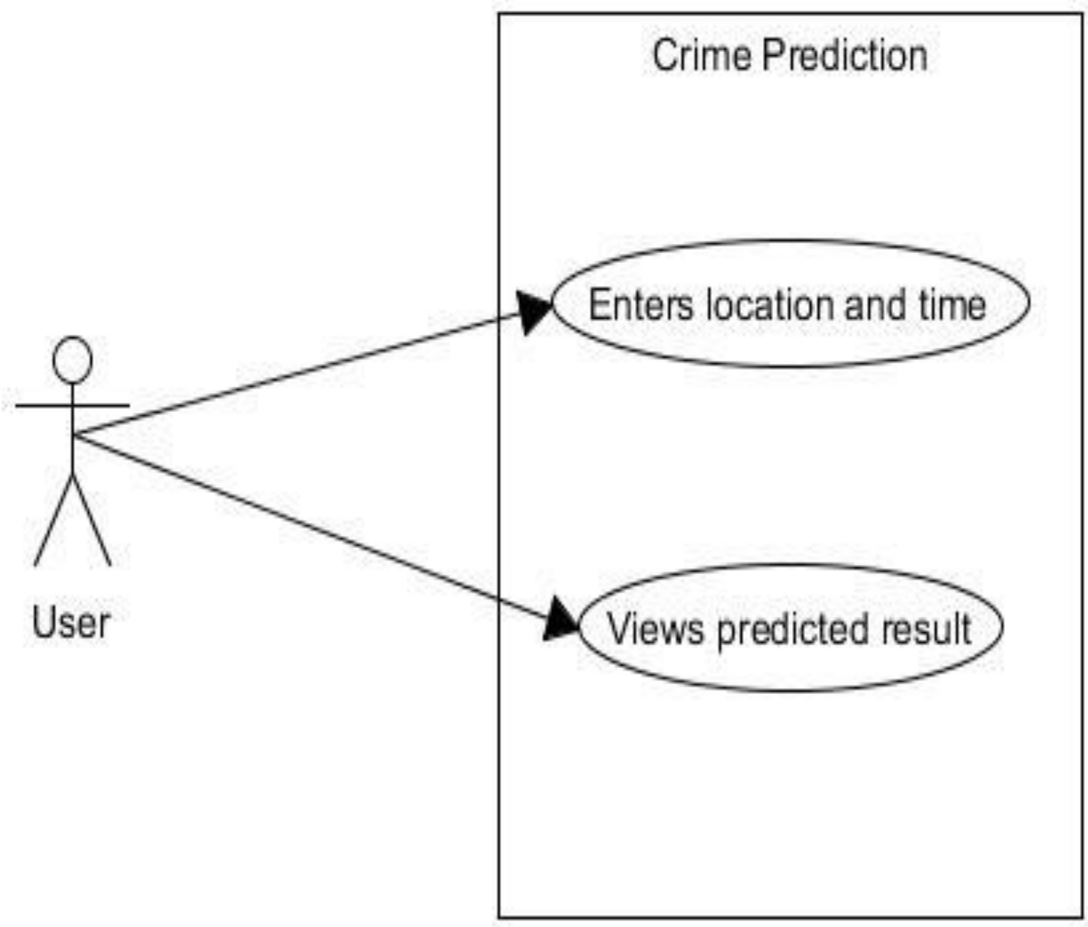

Fig 3.1-Use case diagram of Paasbaan

3.2 Activity diagram

The activity diagram is a graphical representation for representing the flow of interaction within specific scenatios. It is similar to a flowchart in which various activities that can be performed in the system are represented.

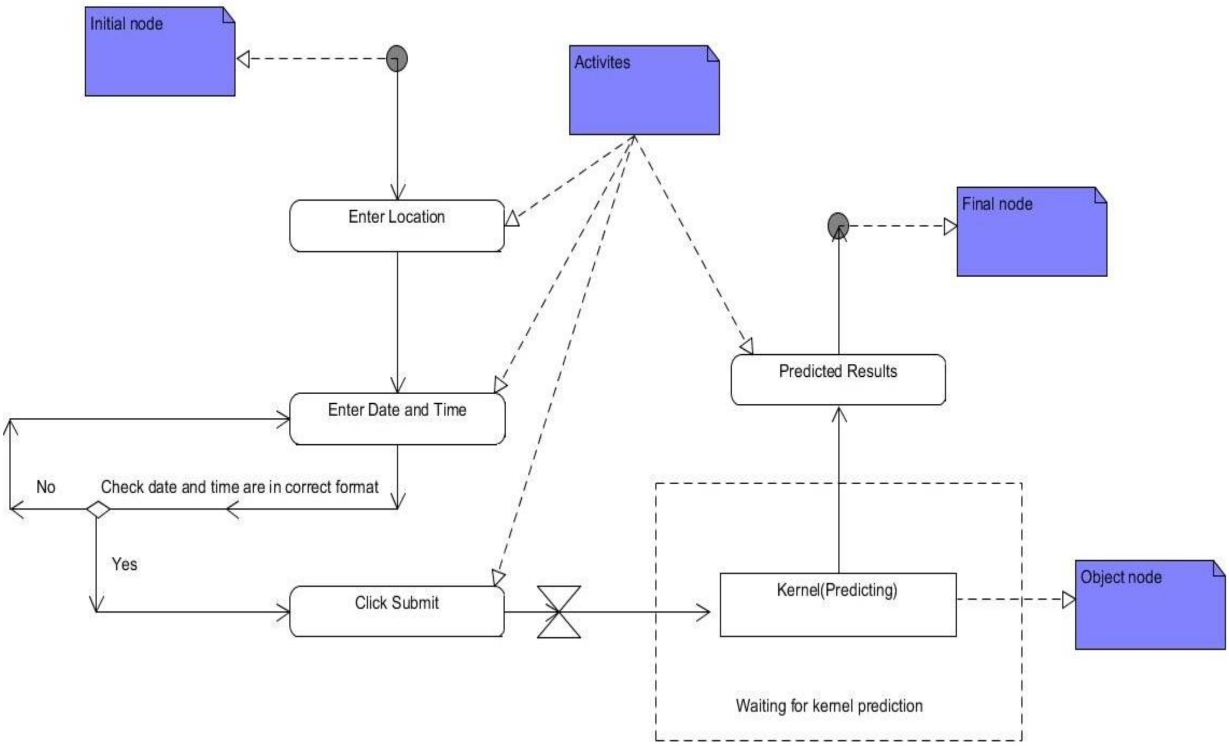

Fig 3.2-Activity diagram of Paasbaan

3.3 Sequence diagram

In the sequence diagram how the object interacts with the other object is shown. There are sequence of events that are represented by a sequence diagram.

It is a time oriented view of the interation between objects to accomplish a behavioural goal of the system.

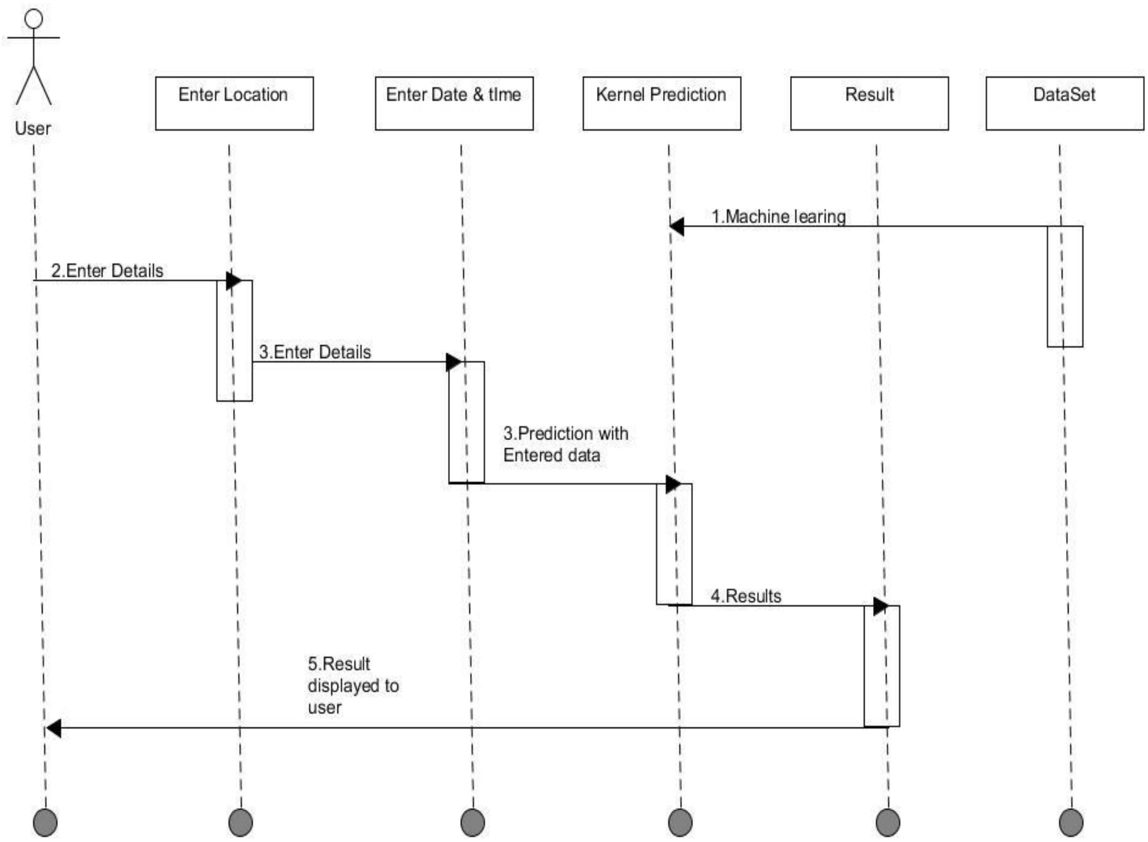

Fig 3.3-Sequence diagram of Paasbaan

3.4 System architecture

The system architectural design is the design process for identifying the subsystems making up the system and framework for subsystem control and communication. The goal of the architectural design is to establish the overall structure of software system.

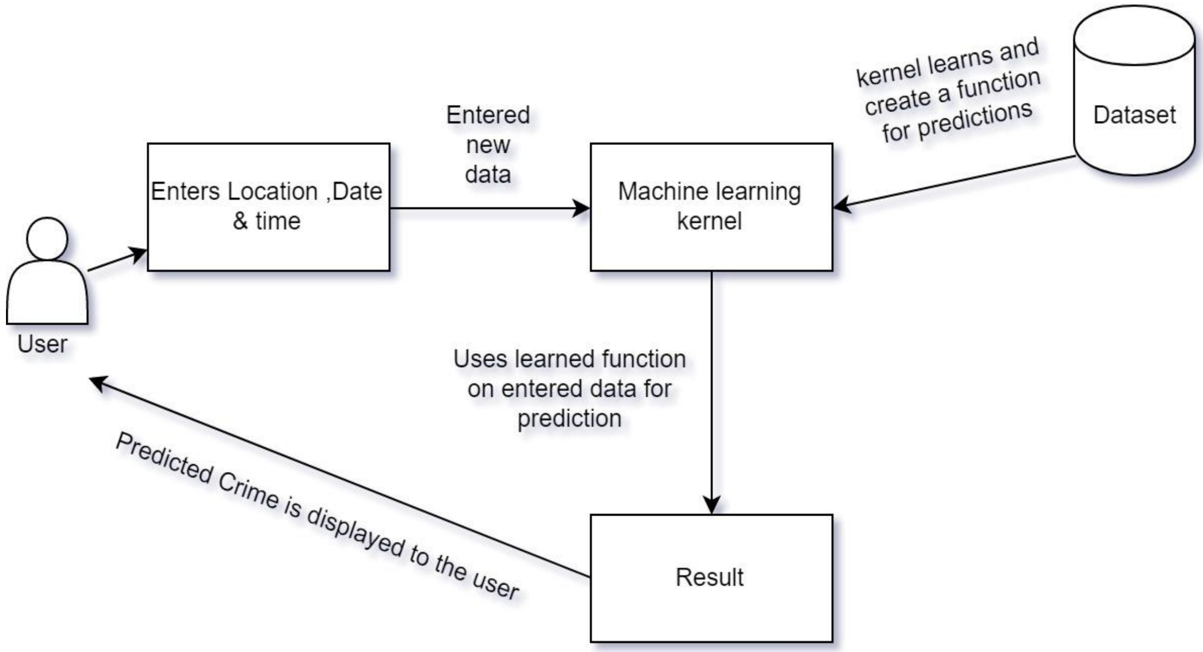

Fig 3.4-System architecture of Paasbaan

Chapter-4
Construction

4.1 Implementation

The implementation of the project is done with the help of python language. To be particular, for the purpose of machine learning Anaconda is being used.

Anaconda is one of several Python distributions. Anaconda is a new distribution of the Python. It was formerly known as Continuum Analytics. Anaconda has more than 100 new packages. Anaconda is used for scientific computing, data science, statistical analysis, and machine learning.

On Python technology, we found out Anaconda to be easier. Since it helps with the following problems:

- Installing Python on multiple platforms.
- Separating out different environments.
- Dealing with not having correct privileges.
- Getting up and running with specific packages and libraries.

This data was scraped from the publically available data from Indore police website which had been made by people in police station of different areas. Implementation of the idea started from the Indore city itself so as to limit an area for the prediction and making it less complex. The data was sorted and converted into a new format of timestamp, longitude, latitude, which was the input that machine would be taking so as to predict the crime rate in particular location or city.

The entries was done just to make the machine learn what all it has to do with the data and what actually the output is being demanded. As soon as the machine learnt the algorithms and the process, accuracy of different algorithms were measured & the algorithm with the most accuracy is used for the prediction kernel i.e. Random forest.

4.2 Implementation Details

For the purpose of proper implementation and functioning several Algorithms and techniques were used. Following are the algorithms used:

4.2.1 KNN (K-Nearest neighbors)

A powerful classification algorithm used in pattern recognition K nearest neighbors stores all available cases and classifies new cases based on a similarity measure (e.g. distance function).One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method).

KNN: Classification Approach

- An object (a new instance) is classified by a majority votes for its neighbor classes.
- The object is assigned to the most common class amongst its K nearest neighbors.(measured by distance function)

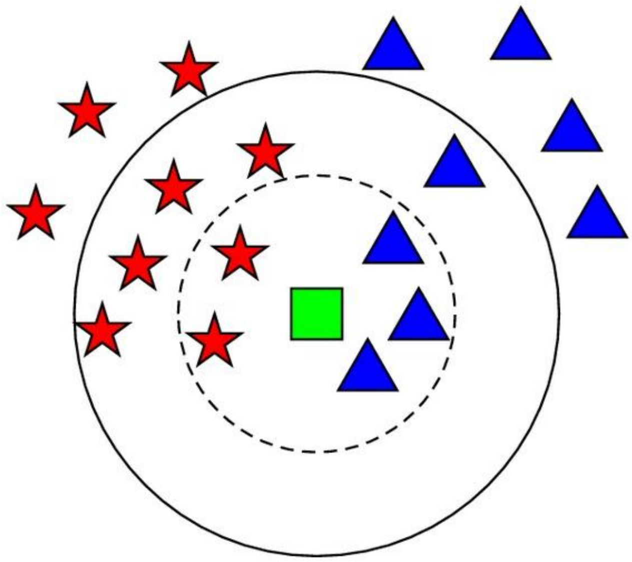

Fig 4.1.1 Principle diagram of KNN

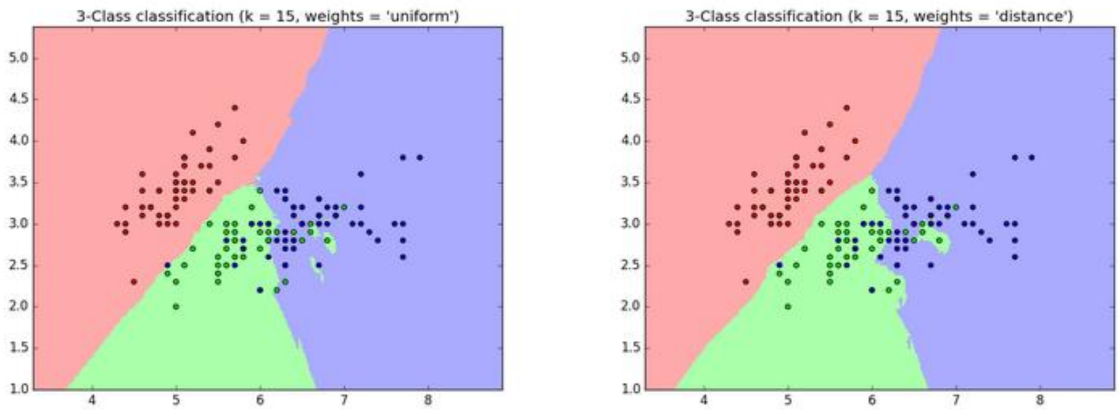

Fig 4.1.2 Shows graphical representation of KNN

Some frequently used distance functions.	
Camberra: $d(x,y)=\sum_{i=1}^{m}\frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$ (2)	Euclidean: $d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$ (5)
Minkowsky: $d(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^r\right)^{1/r}$ (3)	Manhattan / city - block: $d(x,y)=\sum_{i=1}^{m}\lvert x_i-y_i\rvert$ (6)
Chebychev: $d(x,y)=\max_{i=1}^{m}\lvert x_i-y_i\rvert$ (4)	

Fig 4.1.3 Distance functions

4.2.2 Decision Tree

As the name says all about it, it is a tree which helps us by assisting us in decision-making. Used for both classification and regression, it is a very basic and important predictive learning algorithm.

- It is different from others because it works intuitively i.e., taking decisions one-by-one.
- Non-parametric: Fast and efficient.

It consists of nodes which have parent-child relationships:

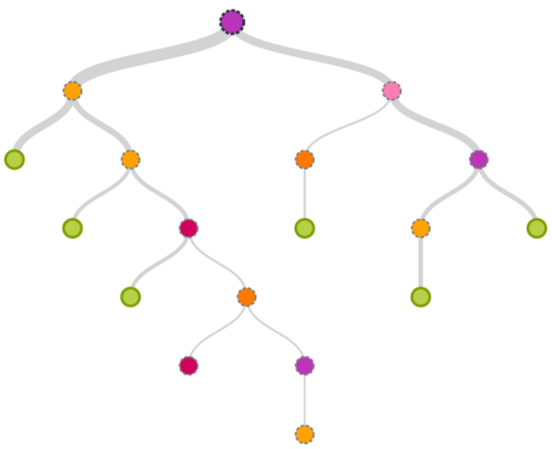

Fig 4.2.1 Decision tree

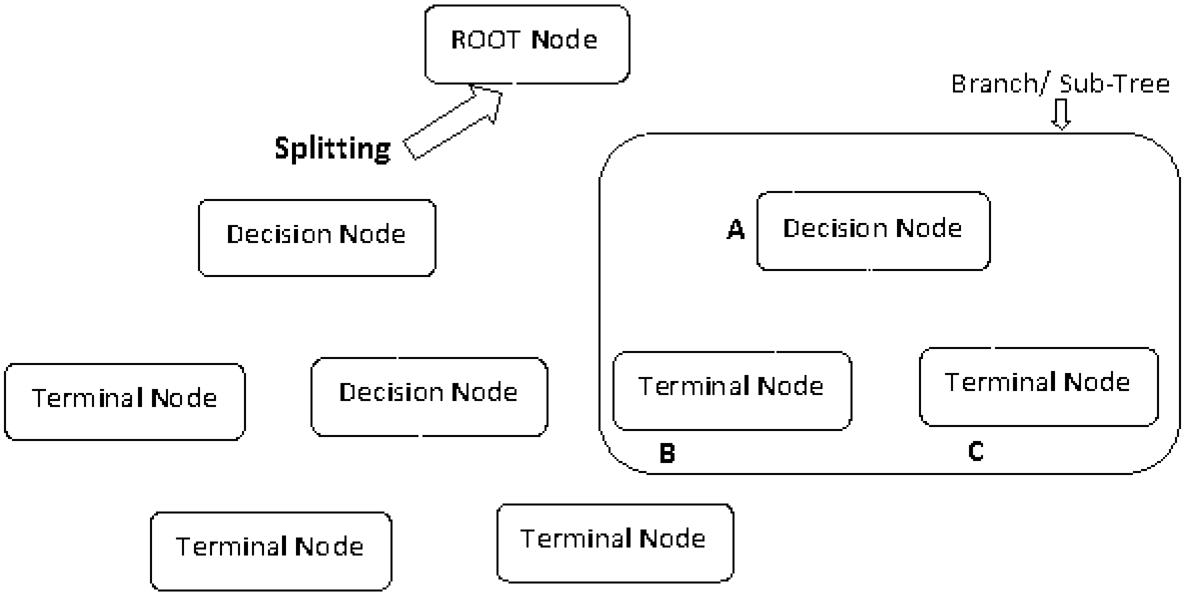

Fig 4.2.2 Decision Tree example

Decision tree considers the most important variable using some fancy criterion and splits dataset based on it. It is done to reach a stage where we have **homogenous subsets** that are giving predictions with utmost surety.

4.2.3 Random forest

Random Forests is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data.

Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making split decision to help avoid overfitting on the training data.

A random forests classifier is an ensemble classifier, which aggregates a family of classifiers $h(x|\theta 1),h(x|\theta 2),..h(x|\theta k)$. Each member of the family, $h(x|\theta)$, is a classification tree and k is the number of trees chosen from a model random vector.

Also, each θk is a randomly chosen parameter vector. If $D(x,y)$ denotes the training dataset, each classification tree in the ensemble is built using a different subset $D\theta k(x,y) \subset D(x,y)$ of the training dataset.

Thus, h(x|θk) is the kth classification tree which uses a subset of features $x\theta k \subset x$ to build a classification model. Each tree then works like regular decision trees: it partitions the data based on the value of a particular feature (which is selected randomly from the subset), until the data is fully partitioned, or the maximum allowed depth is reached. The final output y is obtained by aggregating the results thus:

$$y = \text{argmax}_{p \in \{h(x_1)..h(x_k)\}} \{ \sum_{j=1}^{\kappa} (I(h(x|\theta_j) = p)) \}$$

where I denotes the indicator function.

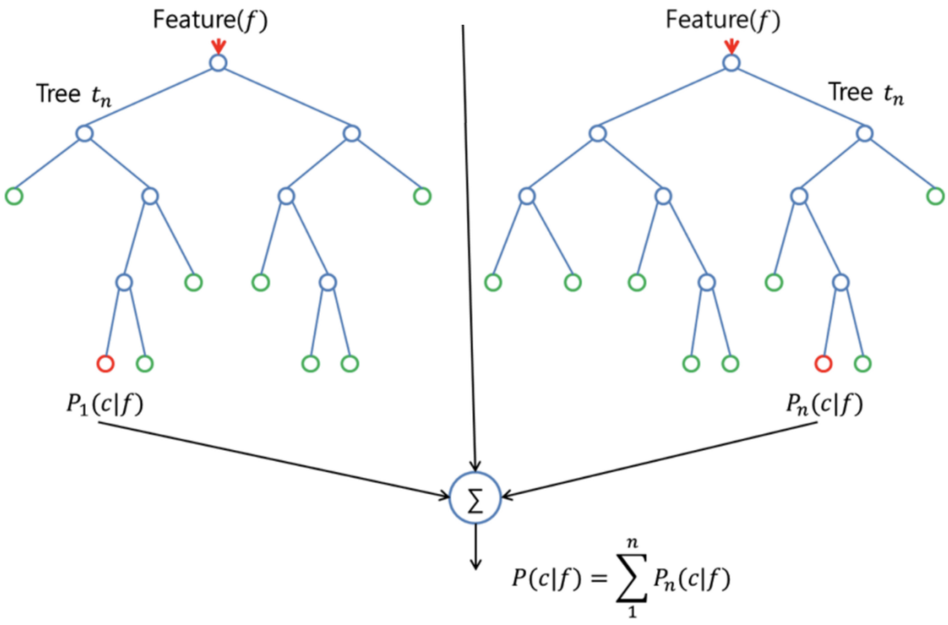

Fig 4.3.1 Random Forest Example

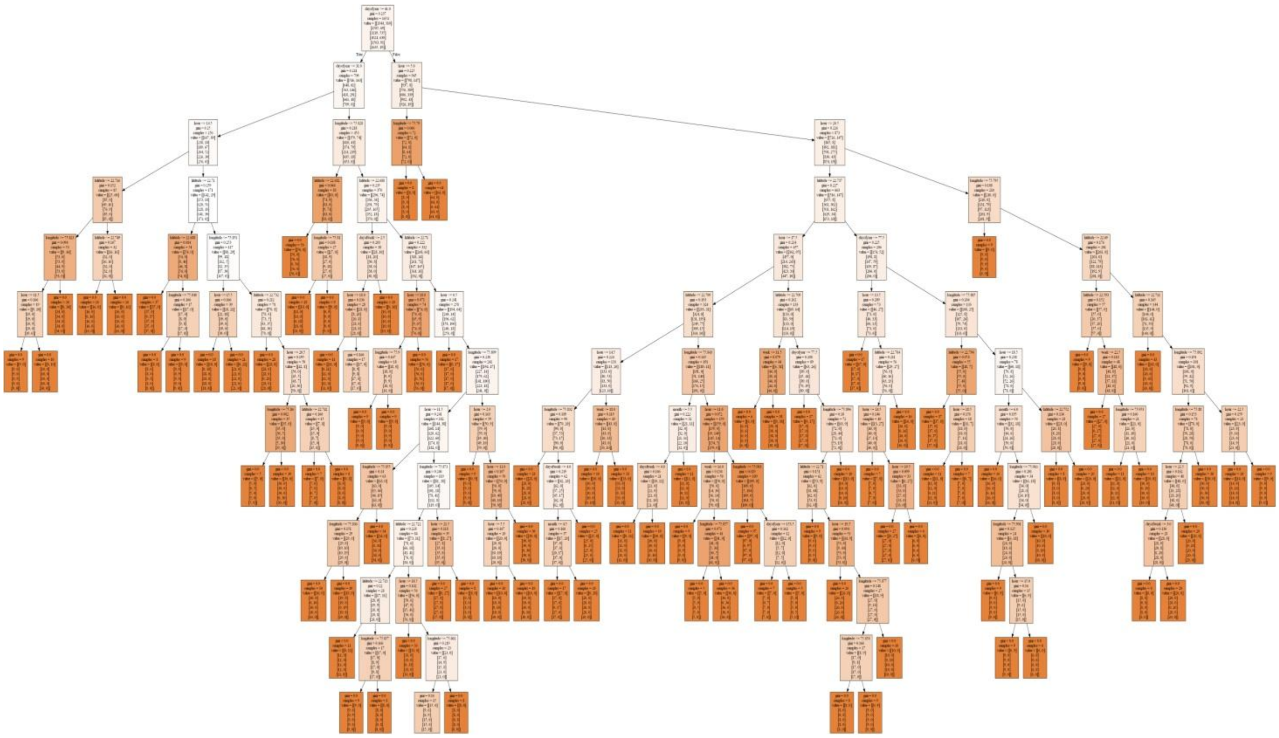

Fig 4.3.2 Decision Tree of Paasbaan

4.2.4 Data Visualization

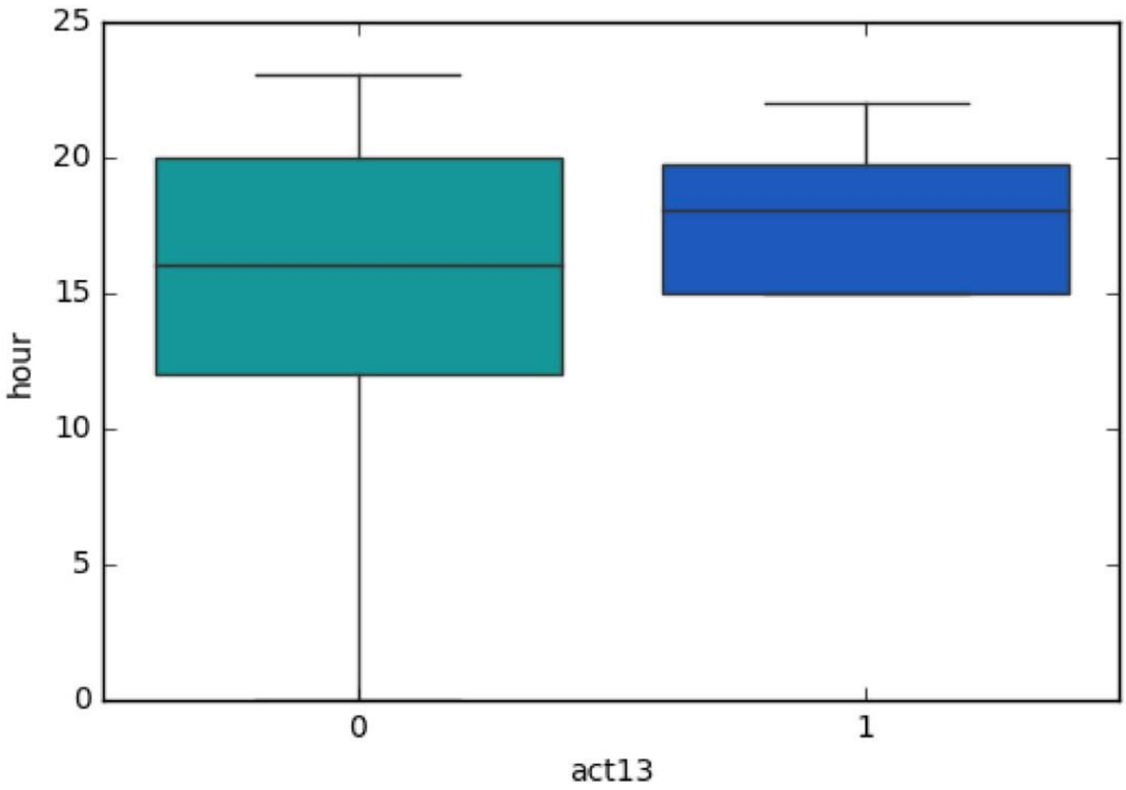

Fig 4.4.1 Act13(Gambling vs Hour)

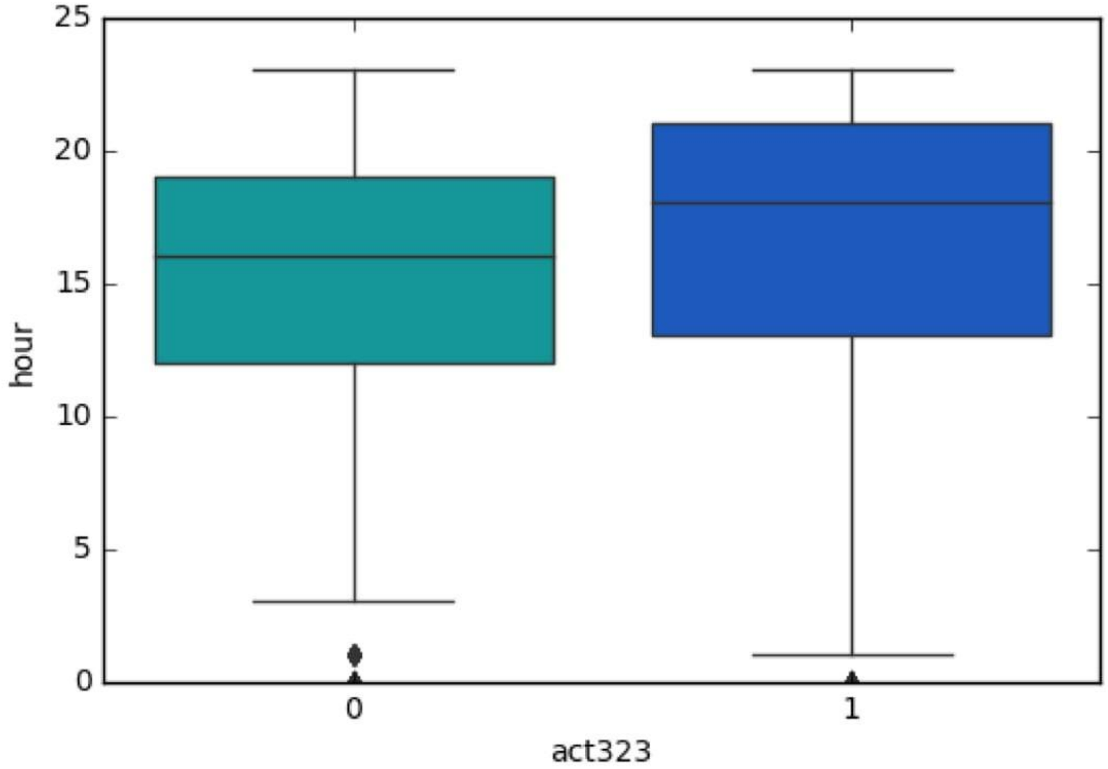

Fig 4.4.2 Act323(Violence vs Hour)

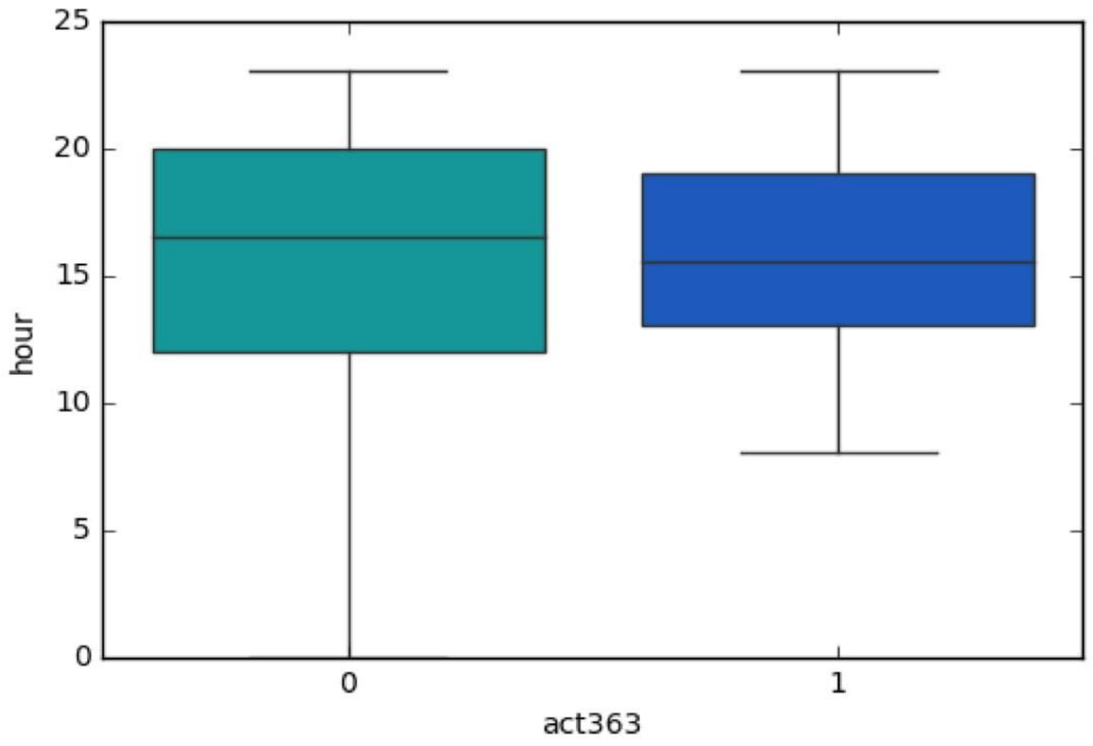

Fig 4.4.3 Act363(Kidnapping vs Hour)

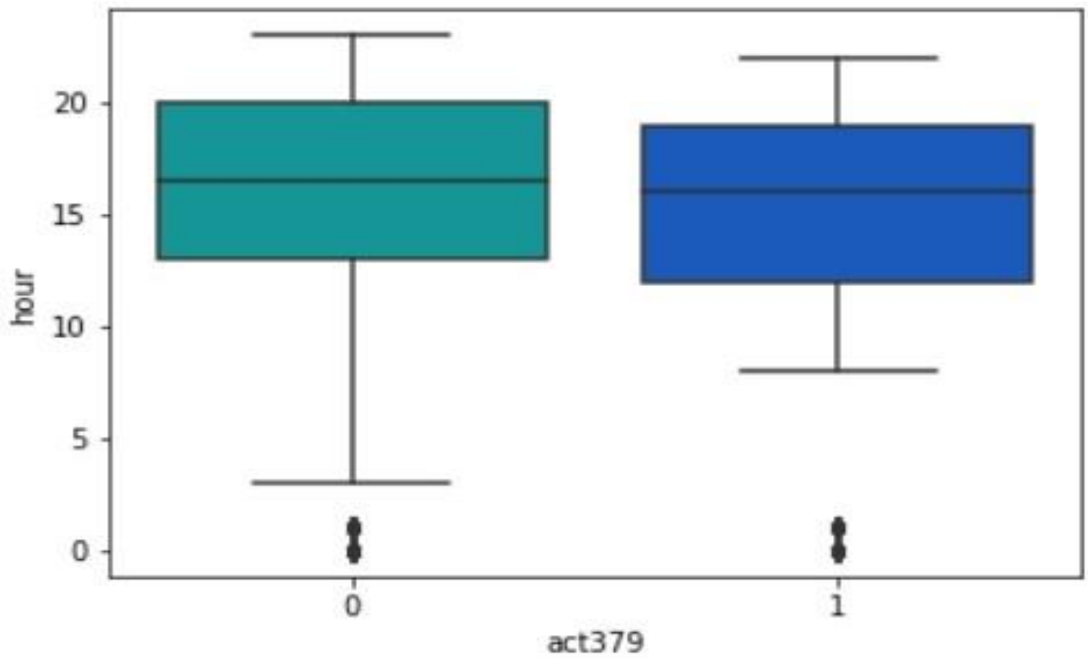

Fig 4.4.4 Act379(Robbery vs Hour)

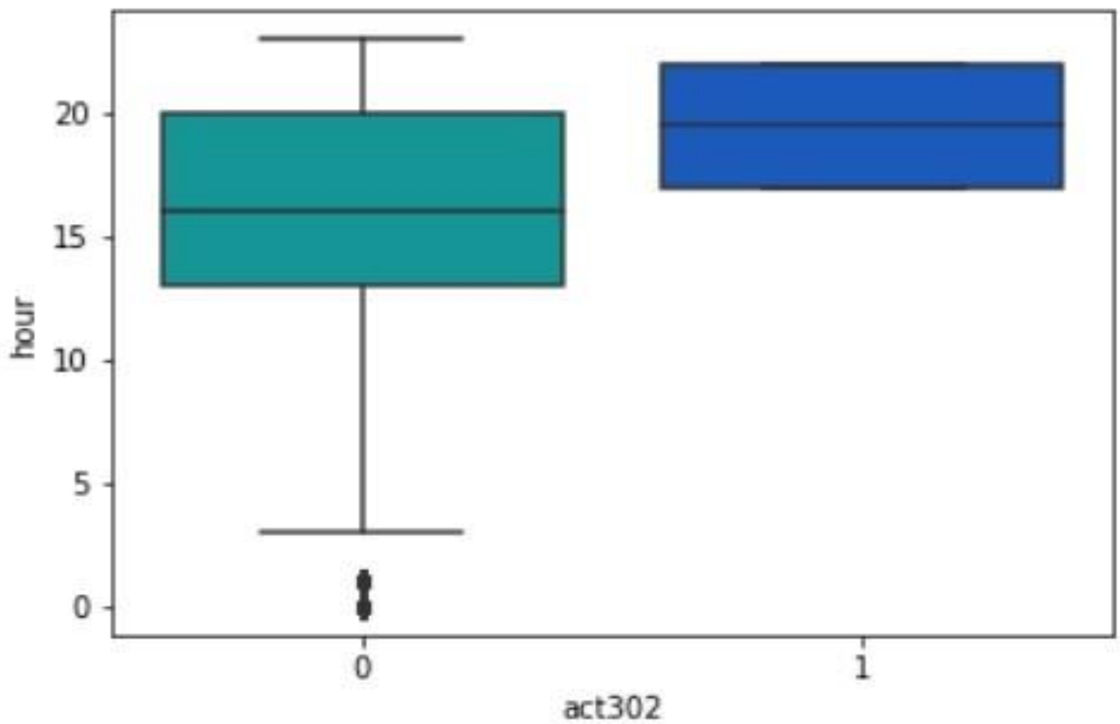

Fig 4.4.5 Act302(Murder vs Hour)

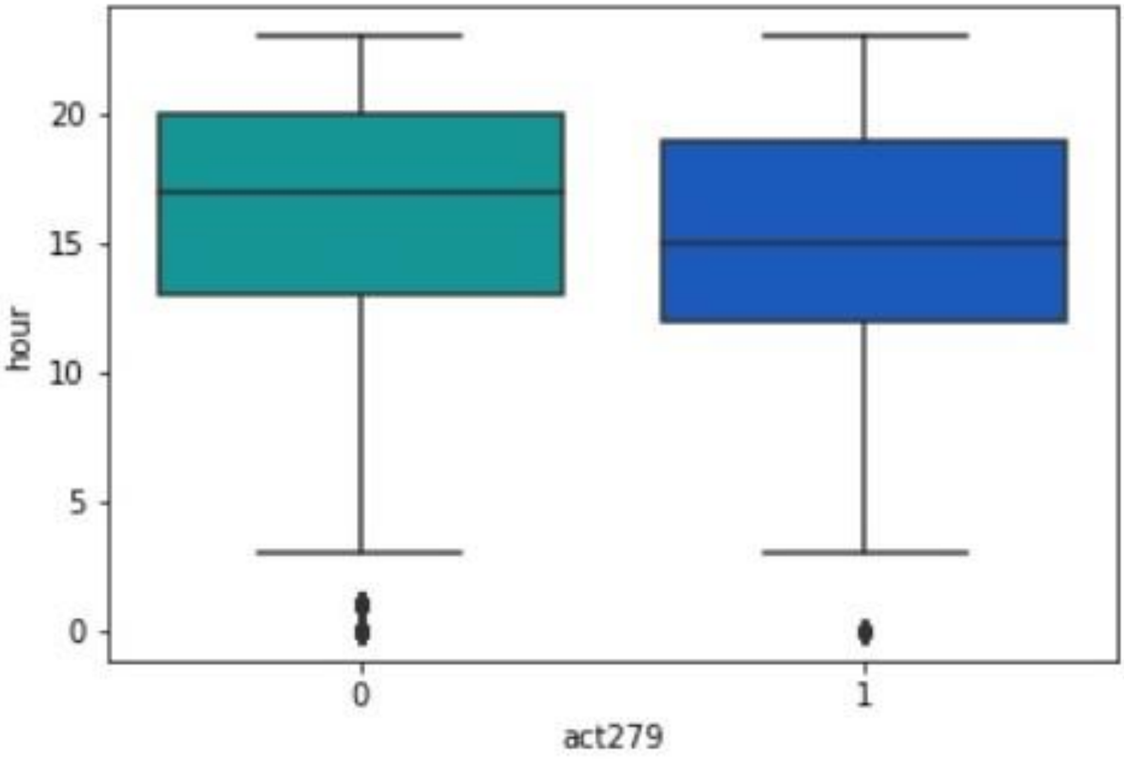

Fig4.4.6 Act279(Accident vs Hour)

4.3 Software Details

- Anaconda Distribution (v5.1)
- Python (3.6.5)
- Packages Used:
  - Flask (0.12.2)
  - Pandas (0.22.1)
  - Numpy (1.14.2)
  - Sklearn (0.19.1)
  - Geopy (1.13.0)
- HTML 5
- CSS 3
- Bootstrap 4
- Java Script 1.8

4.4 Hardware Details

- Operating system: Windows 7 or newer, 64-bit macOS 10.9+, or Linux.
- System architecture: 64-bit x86, 32-bit x86 with Windows or Linux.
- CPU: Intel Core 2 Quad CPU Q6600 @ 2.40GHz or greater.
- RAM: 4 GB or greater.

4.5 Testing

The development of software involves a series of production activities were opportunities for injection of human fallibilities are enormous.

Error may begin to occur at very inspection of the process where the objective may be enormously or imperfectly specified as well as in lateral design and development stage. Because of human inability to perform and communicate with perfection, software development quality assurance activities.

Software testing is a crucial element of software quality assurances and represents ultimate review of specification, design and coding.

4.5.1 White box testing

It focuses on the program control structure. Here all statement in the project have been executed at least once during testing and all logical condition have been exercised.

4.5.2 Black box testing

This is designed to uncover the error in functional requirements without regard to the internal working of the project. This testing focuses on the information domain of the project , deriving test case by partitioning the input and output domain of programming – A manner that provides through test coverage.

Test Case ID	Test Name	Test Description	Steps	Executed result	Actual result	Test case statement
01	Check for correct entered numeric values and date and time.	The entered values are in correct format.	1. Enter details in fields. 2. Click submit.	If format is correct details are sent to kernel successfully.	As expected.	Pass
02	Check for correct entered time.	The entered values are in correct format.	1. Enter details in fields. 2. Click submit.	If format is correct details are sent to kernel successfully	As expected.	Pass
03	Check for correct entered location	The entered values are correct.	1. Enter details in fields. 2. Click submit.	If format is correct details are sent to kernel successfully	As expected	Pass
04	Predicted Result	Output is displayed		If kernel predicts successfully output is then showed to the screen	As expected	Pass
05	Analysis Button	Data visualization is displayed.	1.Click Analysis	Shows the overall analysis on screen 3	As expected	Pass

Table 4.5.3 Tests

Chapter-5

Conclusion and future scope

5.1 Conclusion

The initial problem of classifying 6 different crime categories was a challenging multi-class classification problem, and there was not enough predictability in our initial data-set to obtain very high accuracy on it. We found that a more meaningful approach was to collapse the crime categories into fewer, larger groups, in order to find structure in the data. We got high accuracy and precision on Prediction. However, the Violent/Non-violent crime classification did not yield remarkable results with the same classifiers – this was a significantly harder classification problem. Thus, collapsing crime categories is not an obvious task and requires careful choice and consideration.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes – for example, it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

5.2 Future Scope

The goal of any society shouldn't be to just catch criminals but to prevent crimes from happening in the first place

- **Predicting Future Crime Spots:** By using historical data and observing where recent crimes took place we can predict where future crimes will likely happen. For example a rash of burglaries in one area could correlate with more burglaries in surrounding areas in the near future. System highlights possible hotspots on a map the police should consider patrolling more heavily

Fig 5.1 Predicting Surges

- **Predicting Who Will Commit a Crime**: Using Face Recognition to predict if a individual will commit a crime before it happens. The system will detect if there are any suspicious changes in their behavior or unusual movements. For example if an individual seems to be walking back and forth in a certain area over and over indicating they might be a pickpocket or casing the area for a future crime. It will also track individual over time.

- **Pretrial Release and Parole:** After being charged with a crime, most individuals are released until they actually stand trial. In the past deciding who should be released pretrial or what an individual's bail should be set at is mainly now done by judges using their best judgment. In just a few minutes, judges had to attempt to determine if someone is a flight risk, a serious danger to society, or at risk to harm a witness if released. It is an imperfect system open to bias. The media organization's analysis indicated the system might indirectly contain a strong racial bias. They found, "That black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent)." The report raises the question of whether better AI/ML can eventually produce more accurate predictions or if it would reinforce existing problems. Any system will be based off of real world data, but if the real world data is generated by biased police officers, it can make the AI/ML biased.

5.3 Expected Outcome

The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement. This kind of data analysis was technologically impossible a few decades ago, but the hope is that recent developments in machine learning are up to the task.

The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand. The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative.

Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes – for example, it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

References

[1] Bogomolov, Andrey and Lepri, Bruno and Staiano, Jacopo and Oliver, Nuria and Pianesi, Fabio and Pentland, Alex.2014. Once upon a crime: Towards crime prediction from demographics and mobile data, Proceedings of the 16th International Conference on Multimodal Interaction.

[2] Yu, Chung-Hsien and Ward, Max W and Morabito, Melissa and Ding, Wei.2011. Crime forecasting using data mining techniques, pages 779-786, IEEE 11th International Conference on Data Mining Workshops (ICDMW)

[3] Kianmehr, Keivan and Alhajj, Reda. 2008. Effectiveness of support vector machine for crime hot-spots prediction, pages 433-458, Applied Artificial Intelligence, volume 22, number 5.

[4] Toole, Jameson L and Eagle, Nathan and Plotkin, Joshua B. 2011 (TIST), volume 2, number 4, pages 38, ACM Transactions on Intelligent Systems and Technology

[5] Wang, Tong and Rudin, Cynthia and Wagner, Daniel and Sevieri, Rich. 2013. pages 515-530, Machine Learning and Knowledge Discovery in Databases

[6] Friedman, Jerome H. ”Stochastic gradient boosting.” Computational Statistics and Data Analysis 38.4 (2002): 367-378.sts

[7]Leo Breiman, Random Forests, Machine Learning, 2001,Volume 45, Number 1, Page 5

Appendix-A

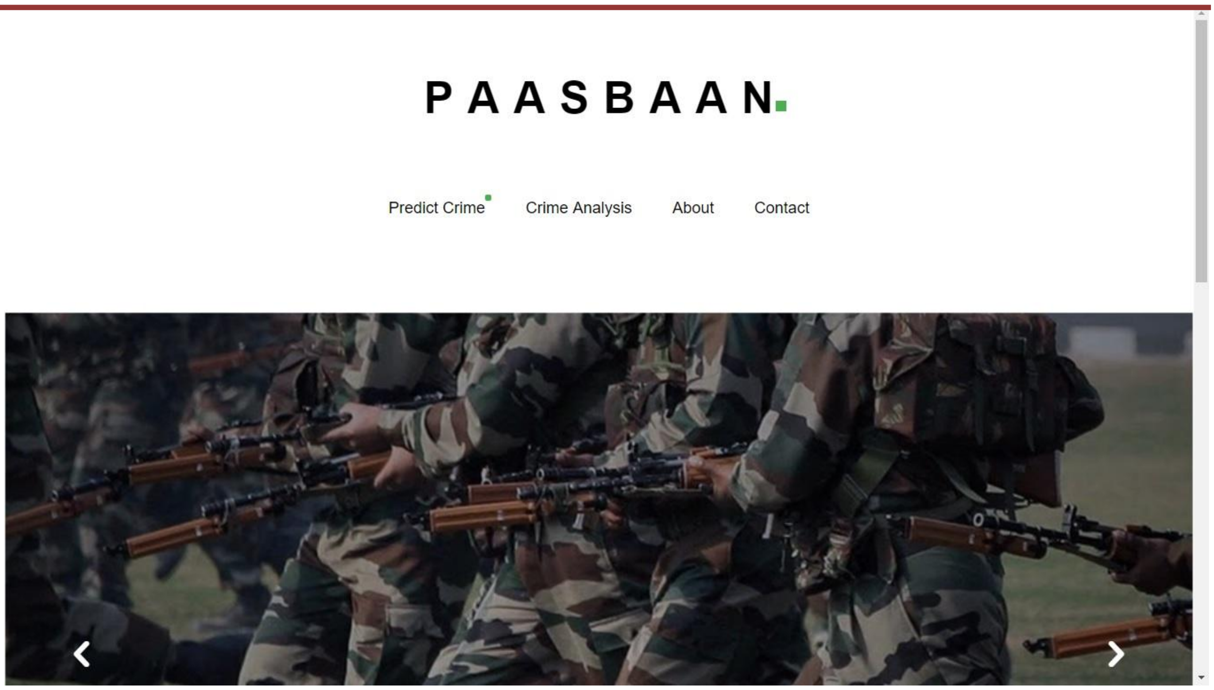

Fig A.1-Snapshot 1

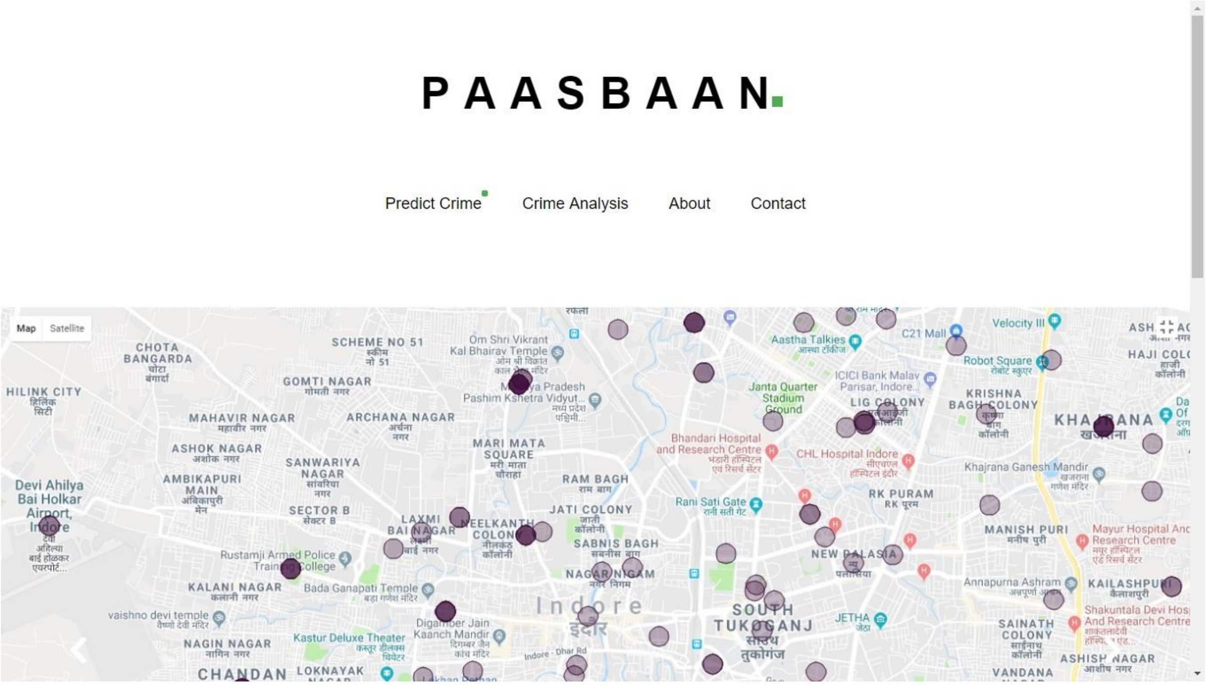

Fig A.2-Snapshot 2

Appendix-B

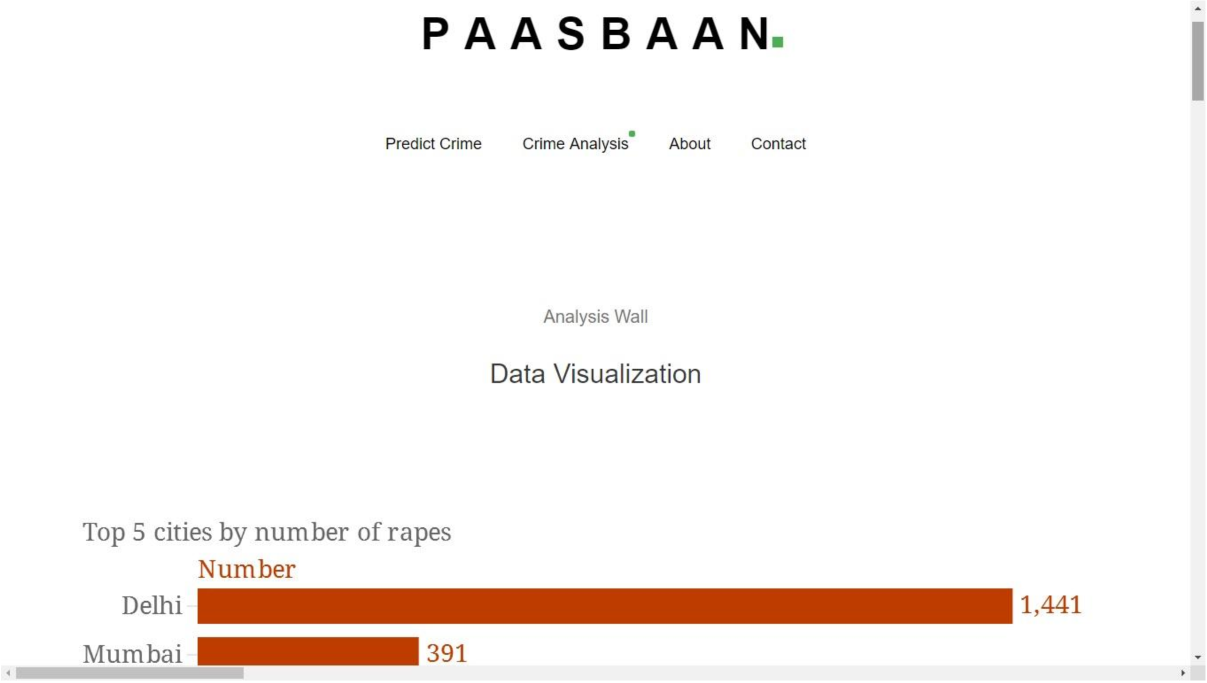

Fig B.1-Snapshot 3

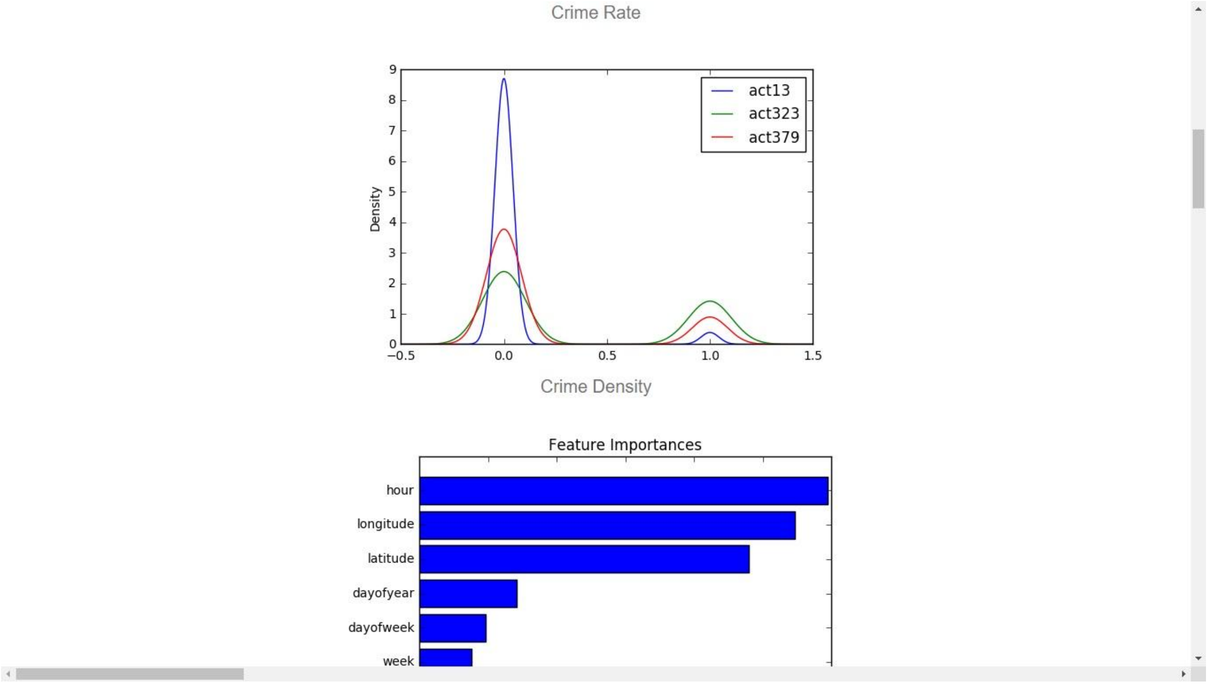

Fig B.2-Snapshot 4

Appendix-C

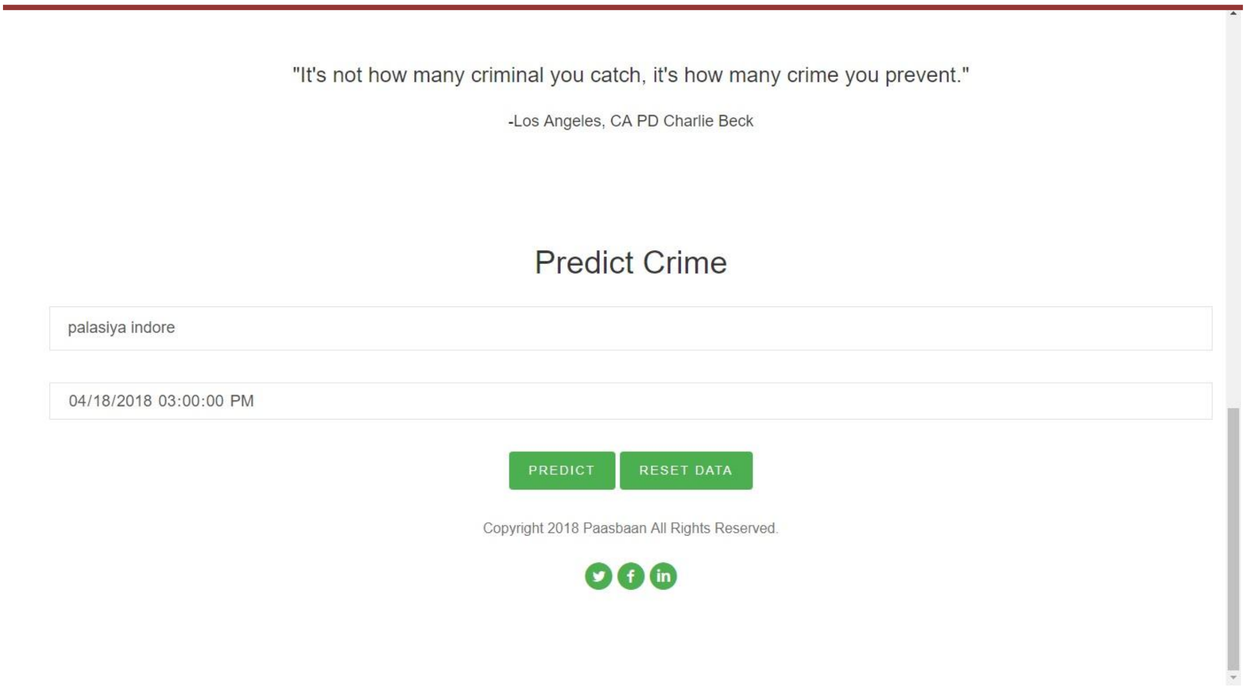

Fig C.1-Snapshot 5

Fig C.2-Snapshot 6