CRIME ANALYSIS AND PREDICTION SYSTEM PROJECT USING MACHINE LEARNING TO SOLVE CURRENT TRENDS OF VIOLENT CRIMES

BSCLMR230620

October 4, 2021

Declaration

The project is a product of my own work and It has not been previously presented to any other university for a degree. this project would not be a possible success without the guidance from Mr Samuel Karuga, who has continuously supported me through the project. His passion, knowledge and constant advice in his class made it motivational for me to expound more in machine learning and give my best in this project.

Abstract

Machine Learning as a tool has evolved to be used as the latest technology to solve crime related issues such as homicides, theft, cybercrimes, burglary and other violent crimes. As a component of the Artificial Intelligence and big data analysis, it provides certain algorithms to identify patterns, understand how the data is classified and get to perform a task. Before Allan Turing discovered artificial intelligence or even evolution of technology, it was easy for the criminals to get away with crime, erase evidence and no justice was taken. spree of murders were in the neighbourhood, burglaries in the stores and this made neighbourhoods insecure, women and children as well.

This project intends to predict when crime is likely to happen and explain what type is and what can be used as well and solve the data using different algorithms. For example the K-Nearest Neighbour(K-NN)algorithm is used to classify the similarity between the new case data between the new case and the available cases, which then puts the new case to the similar category.

Contents

1 INTRODUCTION

1.1 Background

The concept of crime as examined as a reason for predictive policing which was aimed to stop crime before it occurred. Crimes committed back then became widespread since there we no prediction tools that could be used to automatically identify where the crimes were committed. Not even trackers were developed and this mortified people in the neighbourhood. Technology that was first developed as an evolution of computers (super computers) which would simply store data but not a large amount of data. [16] There were murder sprees [2] in the neighbourhood, burglaries, homicide and rape crimes and yet the hotspots that these crimes were committed were not identified. [3].

These crimes would occur again and as for such reasons, later in the early 1900s there was developments of the models which would show where the crimes occurred. A major disadvantage to this was brought about by the computers that were used; they could not input big data becoming difficult to handle. Ada Lovelace [4], a mathematician described a sequence of operations and Charles Babbage punch-card machine and used equations to develop the patterns of crime to solve such activity that have became a thoughtful way to simulate the activity. In 1950, Allan Turing proposed a hypothesis imposing that machines can be used in convincing humans if a machine can succeed and achieve artificial intelligence. [5] Bernard Wodrow and Marcian Holf [6] developed neuron models that would be used to detect the binary patterns and this would be used in world application.

Later in 1967 the nearest neighbour algorithm was written which would allow the computer to use pattern recognition. The K-NN algorithm was introduced by Fix and Hodges in 1951 which was used for classification and regression [7]. In the 1990s, work on machine learning was shifted from knowledge based to a a data driven approach and scientists later developed programs for computers to analyse large amount of data. Over the decades there has been many improvements of this algorithm there are new approaches that have emerged making it more improved and efficient. Later on in the current technology using this algorithm, alongside ordinal regression, logistic regression and decision stamp are implemented.

A crime predictive software developed was the CrimeScan software to predict violent crime [8].The first trial of this software to test its effectiveness was when the researchers tackled violence and identify the individuals within the crime hotspots.

For example my case study in Atlanta could be in possible causes. If there was a robbery in the neighbourhood, a high probability was that they would use non-lethal weapons such as a baseball bat but later a similar crime would happen they would use a gun. Many cases would arise and as a result mathematical models would be used to predict crime and research has shown that the equations once developed, they can pinpoint the crime hotspots. Later on research done by Neill and Gorr [9] to develop a software tool for crime prediction which

would use the former dataset(historical dataset) which would show where the violent crimes would occur. [10]

As technology has evolved so has the artificial intelligence. How we face crime today using machine learning has made it easier to identify the criminal hotspots and the patrol to act right away. From the beginning of the fuzzy K-NN algorithm to the modern and more developed K-NN algorithm that uses other algorithms such as additive regression and linear regression algorithm that have been tested thus improving efficiency. [11] Machine learning as it has recently trend with its improvement of other algorithms such as logistic regression that has helped us understand dependent and independent variables have become tremendous tool for crime pattern detection whereby these crime patterns are automatically identified. Also we have the PredPol software was created to show the scientific analysis with the goal of data that could spot criminal behaviour and show crimes that may likely occur during a particular period. [1]

1.2 Problem Statement

The main problem addressed in this project is to discover the crime hotspots in the area using the dataset. Through documentation cases and investigation discovery channels, machine learning and data science has made it easier to predict crime and make the work easier. At times, many cases go cold and unsolved,while others have remained a mystery for even us as humans to contemplate the occurrence of events especially in a crime event.People in the society would want answers to know what really happened at the scene of the crime, parents want justice for their children and vice versa. This has further led to the criminals run free because of insufficient evidence and later on for such cases they remain unsolved. For example if a crime has occurred in an area that involves both murder and burglary, it can be assumed that maybe two suspects are involved and this makes have separate cases while it turns out that one suspect has committed both of the crimes.

1.3 Objectives

1.3.1 General Objectives

The aim of this project is to utilise and make the crime prediction features present in the main dataset extracted from the main link and run it with the help of the machine algorithm.

1.3.2 Specific objectives

The main objective is to train a model for prediction which will be trained and tested using the dataset. The enhancement and improvement of the model will be done using the K-NN algorithm(K-Nearest-Neighbour) and visualization for better accuracy. Visualisation of the dataset is done to analyse the crimes done while the algorithm will be used for the crime prediction. These are the procedures that will help achieve the main goal. First is data collection and in

this case the project uses the dataset provided for a certain area.The data will indicate occurrence of events(time, day and week). Secondly is to classify which will rely on the solved crimes. Then implement the pattern identification which will show the trends in crime and the patterns as well. The prediction will use a decision tree to test on the attributes and will help the algorithm make a better decision. Lastly, the visualization whereby the crime hotspot area is represented geographically using the heat map to show the levels of activity based on the color images and analyse the data that is wanted.

1.4 Justification

This project will benefit the society and even the state at large since criminal activities will be controlled. Handling and solving of this cases will make the society feel safe and make the socio-economic lifestyle easier and safer. Not only will it benefit the state but even the patrol making it easier to know the criminal events taking place at a specific time and act fast. Data included in this case will be used to find the unknown patterns from known data and facts and the approach will be automated by learning of the crime series.

1.5 Scope

The scope of the project is to improve the efficiency and accuracy of the machine learning algorithms in the prediction of the violent crimes and identifying the crime patterns when determining crime hotspots and learning the criminal trends Law enforcement agencies can be warned and prevent crime fro occurring by implementing more surveillance within the prediction zone. Furthermore, with complete automation, it overcomes the drawbacks of the current system and the law enforcement can depend on more of these techniques even in the future. Designing the machine learning algorithm to identify patterns of such crimes can be the start of the future of this study. Despite the systems having a big influence on the current crime prevention, with the start of this project, this could be the next revolutionary change in crime rate.

1.6 Limitation

The project will face the following limitations: analysis of data which may be difficult due to inconsistency and incomplete, limitation to getting crime data records from the law enforcement which they may be reluctant to give detailed data and lastly the accuracy of the program which mainly depends if the training set is accurate.In addition to this,finding the patterns and trends can be a challenging factor since it takes a lot of crime, scanning it to see if it fits in a particular pattern. Some of these trends and patterns are difficult in an unspecified geographical area. To expound on this, there are areas whereby the system may not predict since it may not be available or marked on the heat maps.

2 Literature Review

2.1 Theoretical Review

The first related machine learning work was done in 1950 whereby the related work used mainly symbolic data and algorithm that was based on logic done by Alan Turing [5] and many of these crimes have been obtained by the authorities or even a survey. The data collected would be transferred time and location as well as the crime committed.to the police departments and the crime data using the websites. The information later on would take a specific time period to analyse the

Tools used to predict time and location were violated and even sharing the datasets could vary. Grouping of the datasets back then to understand criminal behaviour and analysis took overtime because there were no developed algorithms back then. Trying to verify or even know where the crime was committed was difficult. At some point the coverage area was too large [13] to even know the ground cover and machine learning by then was under-developed.

The traditional method to detect crime hotspot in the area from the past data in the distribution of crime cases and as for others to assume the past pattern to be repeated in the future. A common method used was the KDE method which used to identify stable hotspot areas.This method was used based on the on trends to outperform the method and found that the forest algorithm was more efficient than the traditional KDE. To analyse and understand the crime pattern techniques there were data mining techniques and tools to understand knowledge from them. One of the methods or technique to discover the crime patterns. [14] High number of crimes over the years have forced the government to use technology to predict and control the crime. That was then the start of using machine learning algorithm to detect these crime patterns.

The first computer learning program and it was improved whereby it simulates the thought of the human brain. The Nearest Neighbour allowed the computer to identify pattern recognition. Gerald Dejong (1981)explained the concept on the based learning whereby the computer analysis trained data and the general rule disposing unimportant data. The machine Learning shifts from knowledge driven to the data approach that the computer programs analyse data.

The nearest neighbour algorithm developed at the beginning was developed as an algorithm for mapping of routes before it later advanced before being advanced in crime. Machine learning models are created and have prompted a new array of concepts. As used in the study, the machine learning repository have used different classes of algorithms that are used to identify the dataset which include the regression, classification and clustering. [15]

K-NN algorithm uses the similarity feature to predict the values of any data whereby the value assigned is based on how they closely related hence the average values are taken to be the final prediction and forecast in the crime area. There are techniques used which is the supervised learning technique and assumes the similarity of cases and aligns them into the category which are

similar or available. As a classification technique it evaluates the value of the function hence in the varsity of different scales which normalizes the training of data thus improving the data accuracy. [13] There are steps when conducting the crime analysis:

The system design of the proposed approach using the data mining. Data scientist and analyst have related to machine learning ordinary by use of the K-NN classifier to enhance its exactness with the relative computational time. The K-NN rule has been the most of the most regulated calculation. It holds the preparation set during learning and its easiest type.The key component of the algorithm is that they have similar qualities if grouped in a similar classification. The architecture explains the crime examination framework,structure and test the new framework. [16]

It shows the result of the algorithm of crime prediction system consisting of a collection of thick crime areas and a set of related crime forecaster. This operate primarily for the large region where the large amounts of individuals live and demonstrates that the suggested strategy achieves excellent precision over rolling time horizons in spatial and temporal crime forecasting .This paper's working process collects raw data the hotspot uses after splitting the data to create the new hotspot model and shows the predictive crime rate.

The Neural Network is used for the prediction's precision. The precision rate by using Neural Network the model accuracy is 60, 93 and 97 . It presents a geographical analysis-based and self-regressive approach to automatically identify large danger urban crime areas and to represents crime patterns in each region reliably.Show the result of the algorithm of crime prediction system consisting of a collection of thick crime areas and a set of related crime forecaster. [14]

2.2 Similar Projects

These are some of the projects that have been implemented in this field

2.2.1 Crime Prediction System using K-means clustering

The crime analysed and considered homicide taking into account and the trend has been descending from 1990-2011 and the clustering technique extracts information in form of the crime data setting using the Rapid Miner Tool since its a solid and complete package with the flexible support option and the figure below shows the proposed architecture and flow diagram. [17].

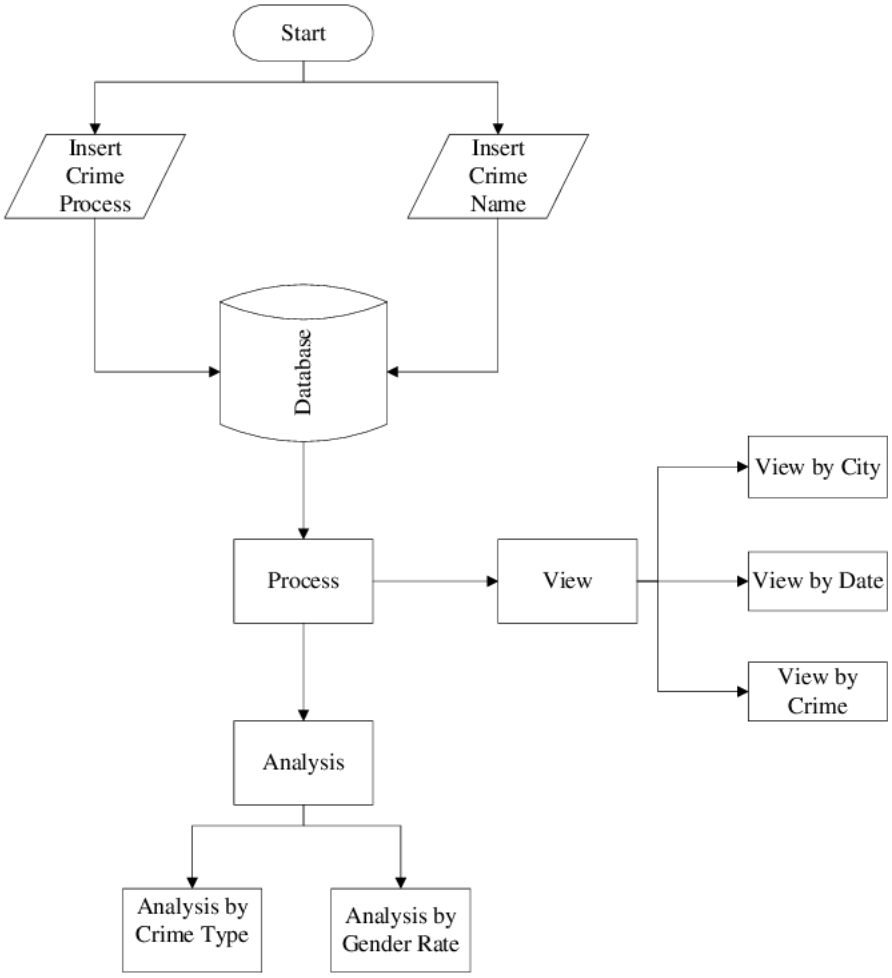


Figure 1: K-means clustering proposed architecture

The system in this case can predict regions with high probability of crime occurrence and can visualize the crime prone areas. The modules explain this process: Start The system starts : we input the crime process and the crime name which takes the data and information to the database. The system inserts the crime process and the crime name to the database. Database The database section :in this case consist of the data retrieved from the sources and at the process stage its preprocessed to make sure that the data is improved and later viewed. View Once viewed: the data is categorised by the place(city) when it occurred (date), specific duration it happened(time) and the type of crime. Analysis After the viewing of the data the collected data is now analysed into two: By the crime type whereby crime is categorised into different types for example rape, robbery, kidnapping and homicide and also by the gender rate(male or female) The machine algorithm will now train the dataset.

2.2.2 Crime Prediction System using Linear regression

Using two separate categories of the visualisation tools whereby one provides the geographical view of crimes and the visualisation ability of the social networks. [**?**]. Giving description of crime which includes the location and personal characteristics as input. By using this technique it involves some of the integrated components:geographical profiling, social network analysis,crime patterns and physical matching. Figure 2 shows the system design of the Linear regression technique.

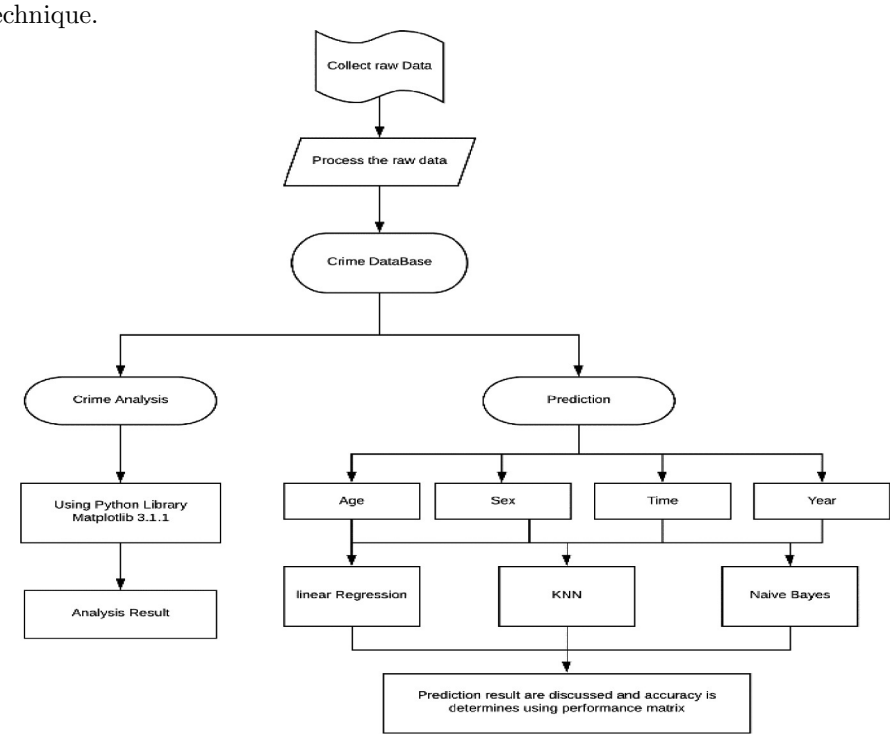


Figure 2: System Architecture: Linear Regression

From the flow diagram above data collected by use of reports and surveys is processed and stored in the crime database. Data is then classified whereby it involve crime analysis and uses python library and if use by machine learning prediction, the dataset used to train consist of the age, sex(gender), time when the crime happened and the year as well. the data is now trained using different algorithms and prediction models and give the prediction result.

2.2.3 Crime Prediction System using Kernel Density Estimation method

This focuses on the criminal hotspot of criminal events in a geographical space which might be related to the criminal theories and provide basic mechanism for the events and the police to use information. It may include criminal theories and are considered as the theoretical basis of crime prevention. Crime pattern

theories mostly explain the distribution of criminal events through the KDE method, a network approach that increases prediction accuracy. It improves the high-dimensional data and extracts the characteristics of data. In this case figure 3 shows a crime map from 2003-2013 with the crime rate of crime hotspot location. [19]

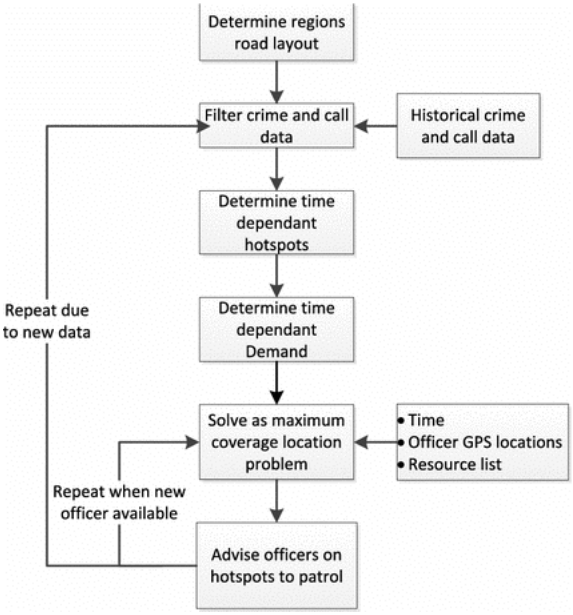


Figure3:KDE graph spatial mapping technique

Initially we identify the crime hotspot and this is achieved by taking historical crime data,lettering it in order to discard data not relevant to the problem, and performing kernel density estimation. Secondly, we use these hotspots as possible locations to send officers who are in patrol and identify which hotspots are chosen and what is determined by finding the configuration with maximum coverage of possible demand, using historical call data to predict the demand. The method adopted to calculate the optimal configuration of hotspots to patrol a version of the method Performing this analysis once however does not give a long-term solution to the patrolling problem as it is a dynamic problem. The location of hotspots and demand are time dependant, as is officer availability. Also once a hotspot has been patrolled the effect of deterring crime in the area has an unpredicted lifetime and hence the area may need to be visited again

2.3 Conceptual Framework

The figure below shows a conceptual framework for the crime prediction system. It will consist of the following: The data set that we will train our data. Secondly,the raw data will be improved at the data preprocessing which then the data train test will split. In this case, there are different techniques that can be used which include; Linear regression, K-NN algorithm, random forest

classifier and the neural network. Once data is ready, it will be trained using machine learning models and the data will be classified.

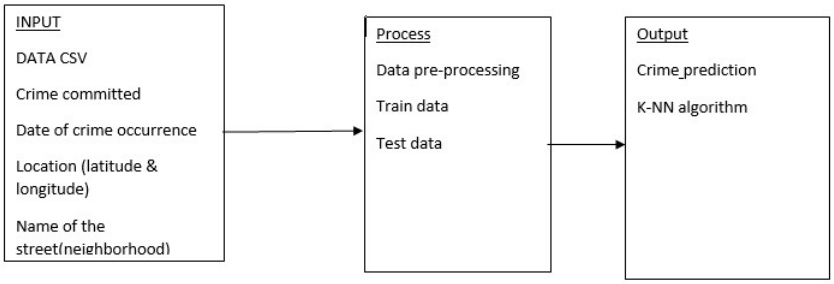


Figure 4: Conceptual framework diagram

The flow begins extracting data from the data collection on different roles of the data repository and the primary data is preprocessed and processed to criminal data. Data is preprocessed to make sure enhancement and improve experimental data. Once at the crime database the crime is classified to different models by the KNN classification.The crime is classified and identified by K-NN classification and crime prediction classified in Kernel and weighed K-NN.

3 Methodology

3.1 Data Source

The dataset is obtained from Atlanta criminal records and included; classification of different types of crime in Atlanta; robbery, domestic violence, homicide, kidnapping, burglary, domestic violence and rape. Also,the records have the crime hotspots (location) and time. Furthermore, the project used one main algorithm which is the K-NN algorithm in crime identification and prediction. A prediction model was used to take the feature vector of the instance as input.

3.2 Data collection tools

3.2.1 Reports

The reports used in this project reflects on the crimes recorded. The reports includes the crimes reported to the police by the public and also the crimes the patrol discover through other sources. the reports will be divided into two categories; the violent crimes such as homicide, domestic violence, rape and murder and property crimes such as burglary, car theft amongst other crimes involved. These reports have helped create a the dataset because it will consist of the type of crime, where the crime has occurred as reported(location of the avenue or street) and when the crime occurred. An example from the dataset that i have homicide crime reference number that occurred in 10/31/2010 in 252 W LAKE DR NW, 105 WestLake

3.2.2 Surveys

The project has also used survey to give insights if the crime is reported or not. Officers can sometimes regard crime as no trouble because some are petty hence not considered as a major crime. Also, with the survey, it has give a clear insight to the crime statistics for example in this project we have domestic violence which can be insignificant since there are relationships involved making it hard for such a case to be reported.

3.3 System design/ Architectural framework

3.3.1 Architectural Framework

Architectural framework also known as the system design will have the user interface which connects with the user system of its Graphical User Interface nature, whereby the administrator will log in and have their details outlined. In this case since it will involve the authority, it will consist of the user administrator and the password and this will be saved in the system database. The database system will contain the log in details: administrators name and password, the k-means cluster algorithm which will highlight the state, district, type of crime, latitude and longitude.The system will require to upload the CSV file which entails the dataset to train it using machine learning algorithm. Once

the data has been uploaded, the training algorithm model will use the dataset which was imported and train it which should give the result of the prediction. Prediction in this case will highlight if its a high or low crime rate area, the number of cases involved for example the number of homicide cases in is 13, and show the analysis which can be in form of a pie chart.

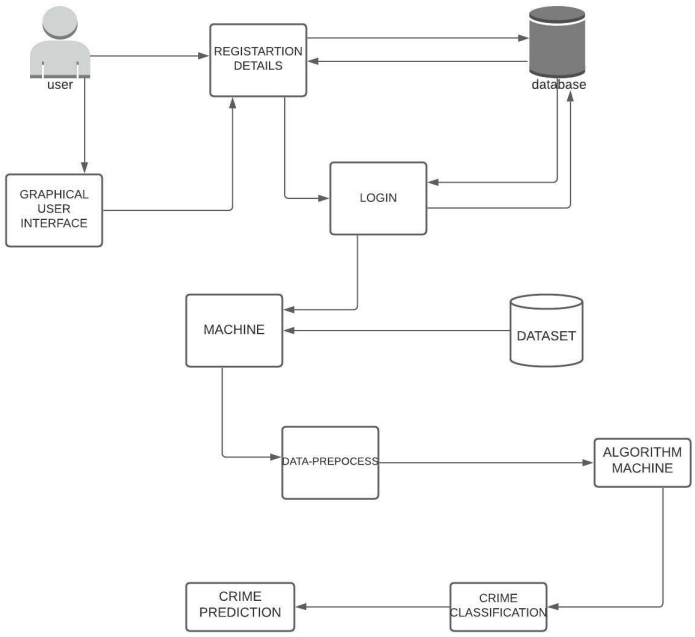


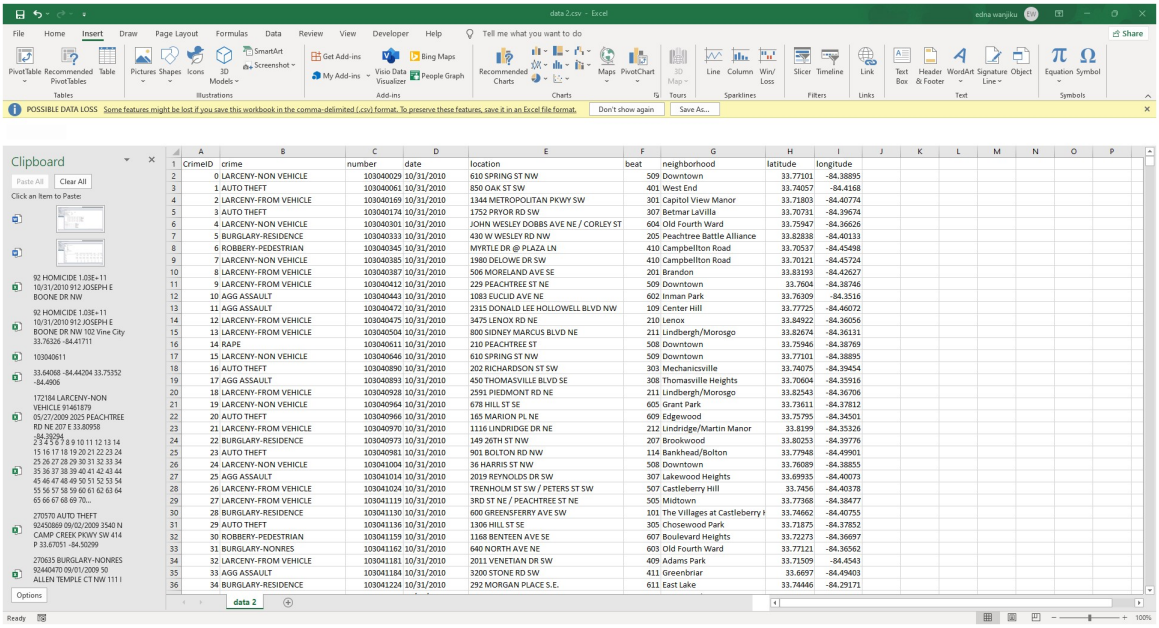
Figure 5: Architectural framework

The framework embraces the crimes committed and different types. We primarily use machine learning algorithms for prediction and find the patterns of crime.The machine learning algorithms include linear regression, k-nn means classification algorithms as well as he apriori algorithms,but in this case k-nn classification algorithm for the prediction has been used. This technique tends to identify the crime patterns by use of the dataset that will be trained and give prediction of the crime areas, if its a high crime rate area or a low crime rate area.

3.4 Implementation

3.4.1 Crime Dataset

This dataset in this case shows a record of different types of crimes committed . It contains information on the types of crime committed, neighbourhood, longitude and latitude. The dataset is obtained from Atlanta Crime records in Kaggle.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	CrimeID	crime	number	date	location	beat	neighborhood	latitude	longitude							
2	0	LARCENY-NON VEHICLE	103040029	10/31/2010	610 SPRING ST NW	509	Downtown	33.77101	-84.38895							
3	1	AUTO THEFT	103040061	10/31/2010	850 OAK ST SW	401	West End	33.74057	-84.4168							
4	2	LARCENY-FROM VEHICLE	103040169	10/31/2010	1344 METROPOLITAN PKWY SW	301	Capitol View Manor	33.71803	-84.40774							
5	3	AUTO THEFT	103040174	10/31/2010	1752 PRYOR RD SW	307	Betmar LaVilla	33.70731	-84.39674							
6	4	LARCENY-NON VEHICLE	103040301	10/31/2010	JOHN WESLEY DOBBS AVE NE / CORLEY ST	604	Old Fourth Ward	33.75947	-84.36626							
7	5	BURGLARY-RESIDENCE	103040333	10/31/2010	430 W WESLEY RD NW	205	Peachtree Battle Alliance	33.82838	-84.40133							
8	6	ROBBERY-PEDESTRIAN	103040345	10/31/2010	MYRTLE DR @ PLAZA LN	410	Campbellton Road	33.70537	-84.45498							
9	7	LARCENY-NON VEHICLE	103040385	10/31/2010	1980 DELOWE DR SW	410	Campbellton Road	33.70121	-84.45724							
10	8	LARCENY-FROM VEHICLE	103040387	10/31/2010	506 MORELAND AVE SE	201	Brandon	33.83193	-84.42627							
11	9	LARCENY-FROM VEHICLE	103040412	10/31/2010	229 PEACHTREE ST NE	509	Downtown	33.7604	-84.38746							
12	10	AGG ASSAULT	103040443	10/31/2010	1083 EUCLID AVE NE	602	Inman Park	33.76309	-84.3516							
13	11	AGG ASSAULT	103040472	10/31/2010	2315 DONALD LEE HOLLOWELL BLVD NW	109	Center Hill	33.77725	-84.46072							
14	12	LARCENY-FROM VEHICLE	103040475	10/31/2010	3475 LENOX RD NE	210	Lenox	33.84922	-84.36056							
15	13	LARCENY-FROM VEHICLE	103040504	10/31/2010	800 SIDNEY MARCUS BLVD NE	211	Lindbergh/Morosgo	33.82674	-84.36131							
16	14	RAPE	103040611	10/31/2010	210 PEACHTREE ST	508	Downtown	33.75946	-84.38769							
17	15	LARCENY-NON VEHICLE	103040646	10/31/2010	610 SPRING ST NW	509	Downtown	33.77101	-84.38895							
18	16	AUTO THEFT	103040890	10/31/2010	202 RICHARDSON ST SW	303	Mechanicsville	33.74075	-84.39454							
19	17	AGG ASSAULT	103040893	10/31/2010	450 THOMASVILLE BLVD SE	308	Thomasville Heights	33.70604	-84.35916							
20	18	LARCENY-FROM VEHICLE	103040928	10/31/2010	2591 PIEDMONT RD NE	211	Lindbergh/Morosgo	33.82543	-84.36706							
21	19	LARCENY-NON VEHICLE	103040964	10/31/2010	678 HILL ST SE	605	Grant Park	33.73611	-84.37812							
22	20	AUTO THEFT	103040966	10/31/2010	165 MARION PL NE	609	Edgewood	33.75795	-84.34501							
23	21	LARCENY-FROM VEHICLE	103040970	10/31/2010	1116 LINDRIDGE DR NE	212	Lindridge/Martin Manor	33.8199	-84.35326							
24	22	BURGLARY-RESIDENCE	103040973	10/31/2010	149 26TH ST NW	207	Brookwood	33.80253	-84.39776							
25	23	AUTO THEFT	103040981	10/31/2010	901 BOLTON RD NW	114	Bankhead/Bolton	33.77948	-84.49901							
26	24	LARCENY-NON VEHICLE	103041004	10/31/2010	36 HARRIS ST NW	508	Downtown	33.76089	-84.38855							
27	25	AGG ASSAULT	103041014	10/31/2010	2019 REYNOLDS DR SW	307	Lakewood Heights	33.69935	-84.40073							
28	26	LARCENY-FROM VEHICLE	103041024	10/31/2010	TRENHOLM ST SW / PETERS ST SW	507	Castleberry Hill	33.7456	-84.40378							
29	27	LARCENY-FROM VEHICLE	103041119	10/31/2010	3RD ST NE / PEACHTREE ST NE	505	Midtown	33.77368	-84.38477							
30	28	BURGLARY-RESIDENCE	103041130	10/31/2010	600 GREENSFERRY AVE SW	101	The Villages at Castleberry H	33.74662	-84.40755							
31	29	AUTO THEFT	103041136	10/31/2010	1306 HILL ST SE	305	Chosewood Park	33.71875	-84.37852							
32	30	ROBBERY-PEDESTRIAN	103041159	10/31/2010	1168 BENTEEN AVE SE	607	Boulevard Heights	33.72273	-84.36697							
33	31	BURGLARY-NONRES	103041162	10/31/2010	640 NORTH AVE NE	603	Old Fourth Ward	33.77121	-84.36562							
34	32	LARCENY-FROM VEHICLE	103041181	10/31/2010	2011 VENETIAN DR SW	409	Adams Park	33.71509	-84.4543							
35	33	AGG ASSAULT	103041184	10/31/2010	3200 STONE RD SW	411	Greenbriar	33.6697	-84.49403							
36	34	BURGLARY-RESIDENCE	103041224	10/31/2010	292 MORGAN PLACE S.E.	611	East Lake	33.74446	-84.29171							

dataset1

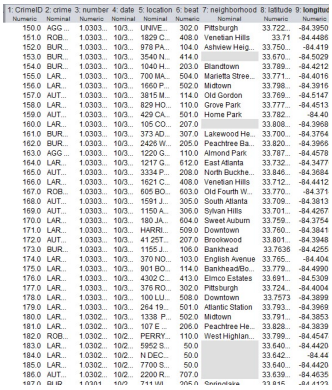
The diagram above shows the dataset used in this project.

3.4.2 Data Processing

In this case, there were missing values in the dataset involved as a result of importing the dataset

3.4.3 Missing Values

With the dataset used for this project, there are missing values which might have been caused by data transfer that cut some of the values and changed the results. This is because, from the dataset we have a small number of the feedback but the missing values are numerous. In this case when using the two algorithms when training the data, the results are not much significant because of the small data missing. For example, in table 1 shows the missing value in the dataset in the neighborhood column.



1: CrimeID	2: crime	3: number	4: date	5: location	6: beat	7: neighborhood	8: latitude	9: longitud
Numeric	Nominal	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Numeric
150.0	AGG ...	1.0303...	10/3...	UNIVE...	302.0	Pittsburgh	33.722...	-84.3950
151.0	ROB...	1.0303...	10/3...	1829 C...	408.0	Venetian Hills	33.71	-84.4486
152.0	BUR...	1.0303...	10/3...	978 PA...	104.0	Ashview Heig...	33.750...	-84.419
153.0	BUR...	1.0303...	10/3...	3540 N...	414.0		33.670...	-84.5029
154.0	BUR...	1.0303...	10/3...	1040 H...	203.0	Blandtown	33.789...	-84.4212
155.0	LAR...	1.0303...	10/3...	700 MA...	504.0	Marietta Stree...	33.771...	-84.4016
156.0	LAR...	1.0303...	10/3...	1660 P...	502.0	Midtown	33.798...	-84.3916
157.0	AUT...	1.0303...	10/3...	3815 M...	114.0	Old Gordon	33.769...	-84.5147
158.0	LAR...	1.0303...	10/3...	829 HO...	110.0	Grove Park	33.777...	-84.4513
159.0	AUT...	1.0303...	10/3...	429 CA...	501.0	Home Park	33.782...	-84.40
160.0	LAR...	1.0303...	10/3...	105 CO...	207.0		33.808...	-84.3968
161.0	BUR...	1.0303...	10/3...	373 AD...	307.0	Lakewood He...	33.700...	-84.3764
162.0	BUR...	1.0303...	10/3...	2426 W...	205.0	Peachtree Ba...	33.820...	-84.3966
163.0	AGG ...	1.0303...	10/3...	1220 G...	110.0	Almond Park	33.787...	-84.4578
164.0	LAR...	1.0303...	10/3...	1217 G...	612.0	East Atlanta	33.732...	-84.3477
165.0	AUT...	1.0303...	10/3...	3334 P...	208.0	North Buckhe...	33.846...	-84.3684
166.0	LAR...	1.0303...	10/3...	1621 C...	408.0	Venetian Hills	33.712...	-84.4412
167.0	ROB...	1.0303...	10/3...	605 BO...	603.0	Old Fourth W...	33.770...	-84.371
168.0	AUT...	1.0303...	10/3...	1591 J...	305.0	South Atlanta	33.709...	-84.3813
169.0	AUT...	1.0303...	10/3...	1150 A...	306.0	Sylvan Hills	33.701...	-84.4267
170.0	LAR...	1.0303...	10/3...	180 JA...	604.0	Sweet Auburn	33.759...	-84.3754
171.0	LAR...	1.0303...	10/3...	HARRI...	509.0	Downtown	33.760...	-84.3841
172.0	AUT...	1.0303...	10/3...	41 25T...	207.0	Brookwood	33.801...	-84.3948
173.0	BUR...	1.0303...	10/3...	1155 J...	106.0	Bankhead	33.7636	-84.4255
174.0	LAR...	1.0303...	10/3...	370 NO...	103.0	English Avenue	33.765...	-84.404
175.0	LAR...	1.0303...	10/3...	901 BO...	114.0	BankheadBo...	33.779...	-84.4990
176.0	LAR...	1.0303...	10/3...	4302 C...	413.0	Elmco Estates	33.691...	-84.5309
177.0	LAR...	1.0303...	10/3...	376 RO...	302.0	Pittsburgh	33.724...	-84.4004
178.0	LAR...	1.0303...	10/3...	100 LU...	508.0	Downtown	33.7573	-84.3899
179.0	LAR...	1.0303...	10/3...	264 19...	501.0	Atlantic Station	33.793...	-84.3969
180.0	LAR...	1.0302...	10/3...	1338 P...	502.0	Midtown	33.791...	-84.3853
181.0	LAR...	1.0302...	10/3...	107 E ...	206.0	Peachtree He...	33.828...	-84.3839
182.0	ROB...	1.0302...	10/2...	PERRY...	110.0	West Highlan...	33.799...	-84.4547
183.0	LAR...	1.0302...	10/2...	5952 S...	50.0		33.640...	-84.4420
184.0	LAR...	1.0302...	10/2...	N DEC...	50.0		33.642...	-84.44
185.0	LAR...	1.0302...	10/2...	7700 S...	50.0		33.640...	-84.4420
186.0	AUT...	1.0302...	10/2...	2200 R...	707.0		33.639...	-84.4635

diagram 1 We solve this by use of the ReplaceMissingValues filter which replaces all the missing values for nominal and numeric attributes in a dataset with the modes and means by the data.

155.0	LAR...	1.0303...	10/3...	700 MA...	504.0	Marietta Stree...	33.771...	-84.40168
156.0	LAR...	1.0303...	10/3...	1660 P...	502.0	Midtown	33.798...	-84.39164
157.0	AUT...	1.0303...	10/3...	3815 M...	114.0	Old Gordon	33.769...	-84.51471
158.0	LAR...	1.0303...	10/3...	829 HO...	110.0	Grove Park	33.777...	-84.45133
159.0	AUT...	1.0303...	10/3...	429 CA...	501.0	Home Park	33.782...	-84.401
160.0	LAR...	1.0303...	10/3...	105 CO...	207.0	Downtown	33.808...	-84.39686
161.0	BUR...	1.0303...	10/3...	373 AD...	307.0	Lakewood He...	33.700...	-84.37645
162.0	BUR...	1.0303...	10/3...	2426 W...	205.0	Peachtree Ba...	33.820...	-84.39664
163.0	AGG ...	1.0303...	10/3...	1220 G...	110.0	Almond Park	33.787...	-84.45789
164.0	LAR...	1.0303...	10/3...	1217 G...	612.0	East Atlanta	33.732...	-84.34776
165.0	AUT...	1.0303...	10/3...	3334 P...	208.0	North Buckhe...	33.846...	-84.36844
166.0	LAR...	1.0303...	10/3...	1621 C...	408.0	Venetian Hills	33.712...	-84.44122
167.0	ROB...	1.0303...	10/3...	605 BO...	603.0	Old Fourth W...	33.770...	-84.3714
168.0	AUT...	1.0303...	10/3...	1591 J...	305.0	South Atlanta	33.709...	-84.38139
169.0	AUT...	1.0303...	10/3...	1150 A...	306.0	Sylvan Hills	33.701...	-84.42678
170.0	LAR...	1.0303...	10/3...	180 JA...	604.0	Sweet Auburn	33.759...	-84.37548
171.0	LAR...	1.0303...	10/3...	HARRI...	509.0	Downtown	33.760...	-84.38418
172.0	AUT...	1.0303...	10/3...	41 25T...	207.0	Brookwood	33.801...	-84.39488
173.0	BUR...	1.0303...	10/3...	1155 J...	106.0	Bankhead	33.7636	-84.42558
174.0	LAR...	1.0303...	10/3...	370 NO...	103.0	English Avenue	33.765...	-84.4042
175.0	LAR...	1.0303...	10/3...	901 BO...	114.0	BankheadBo...	33.779...	-84.49901
176.0	LAR...	1.0303...	10/3...	4302 C...	413.0	Elmco Estates	33.691...	-84.53091
177.0	LAR...	1.0303...	10/3...	376 RO...	302.0	Pittsburgh	33.724...	-84.40041
178.0	LAR...	1.0303...	10/3...	100 LU...	508.0	Downtown	33.7573	-84.38992
179.0	LAR...	1.0303...	10/3...	264 19...	501.0	Atlantic Station	33.793...	-84.39692
180.0	LAR...	1.0302...	10/3...	1338 P...	502.0	Midtown	33.791...	-84.38534
181.0	LAR...	1.0302...	10/3...	107 E ...	206.0	Peachtree He...	33.828...	-84.38396
182.0	ROB...	1.0302...	10/2...	PERRY...	110.0	West Highlan...	33.799...	-84.45474
183.0	LAR...	1.0302...	10/2...	5952 S...	50.0	Downtown	33.640...	-84.44204
184.0	LAR...	1.0302...	10/2...	N DEC...	50.0	Downtown	33.642...	-84.447
185.0	LAR...	1.0302...	10/2...	7700 S...	50.0	Downtown	33.640...	-84.44204
186.0	AUT...	1.0302...	10/2...	2200 R...	707.0	Downtown	33.639...	-84.46358

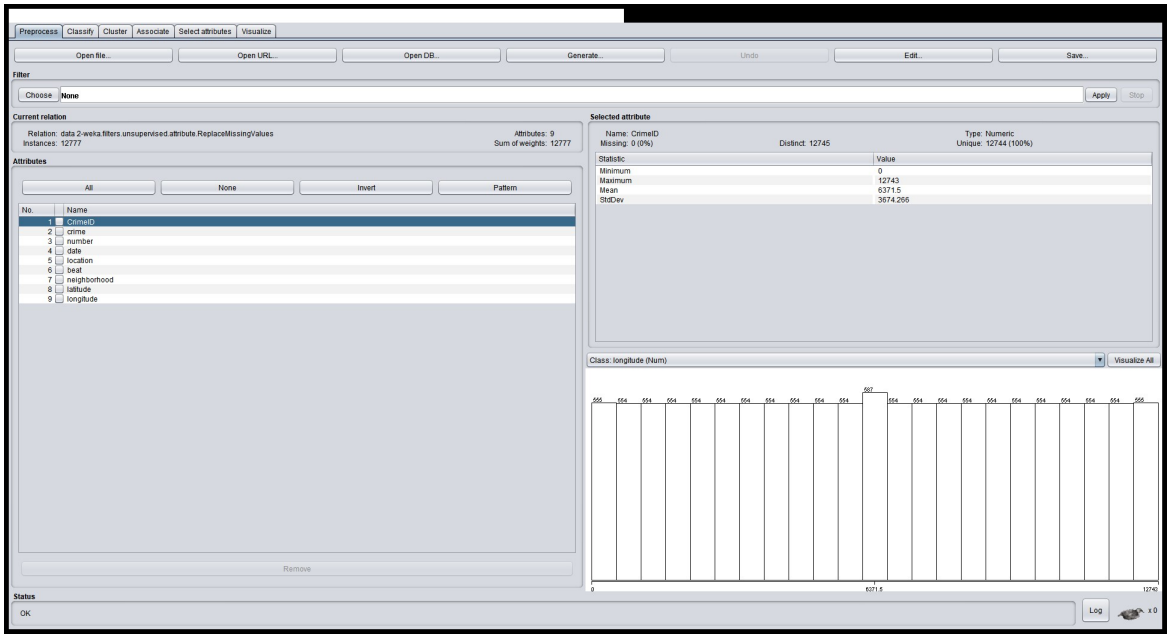
diagam 2

The table above shows the missing values from the neighborhood have been replaced

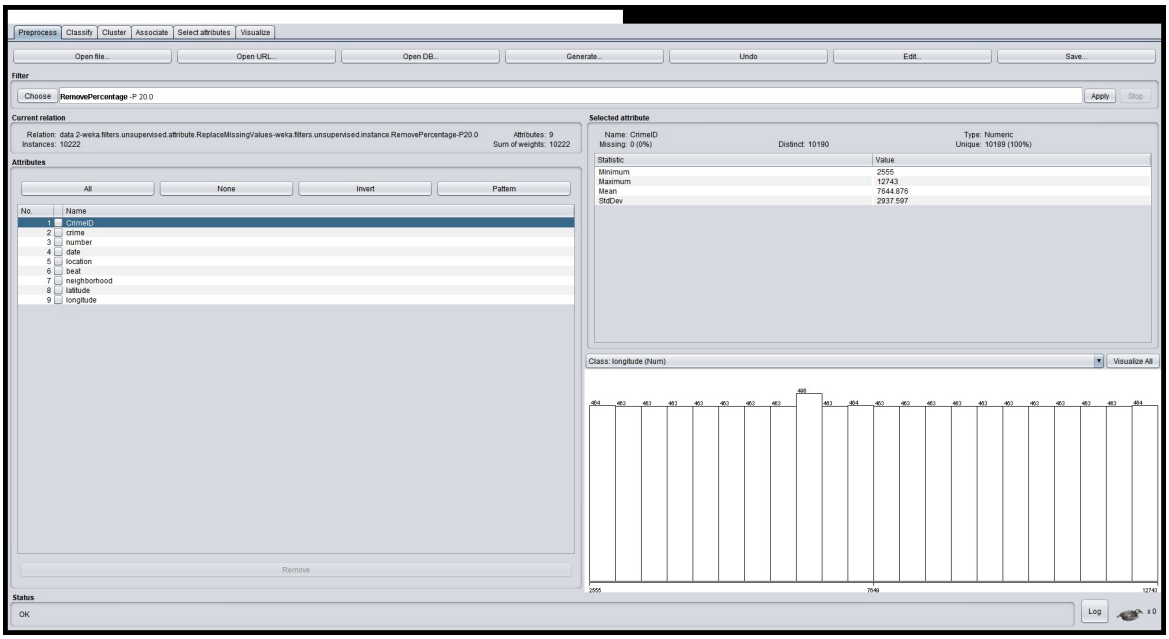
3.4.4 Data Splitting

In table below we have the full data set

3.4.5 Training Dataset

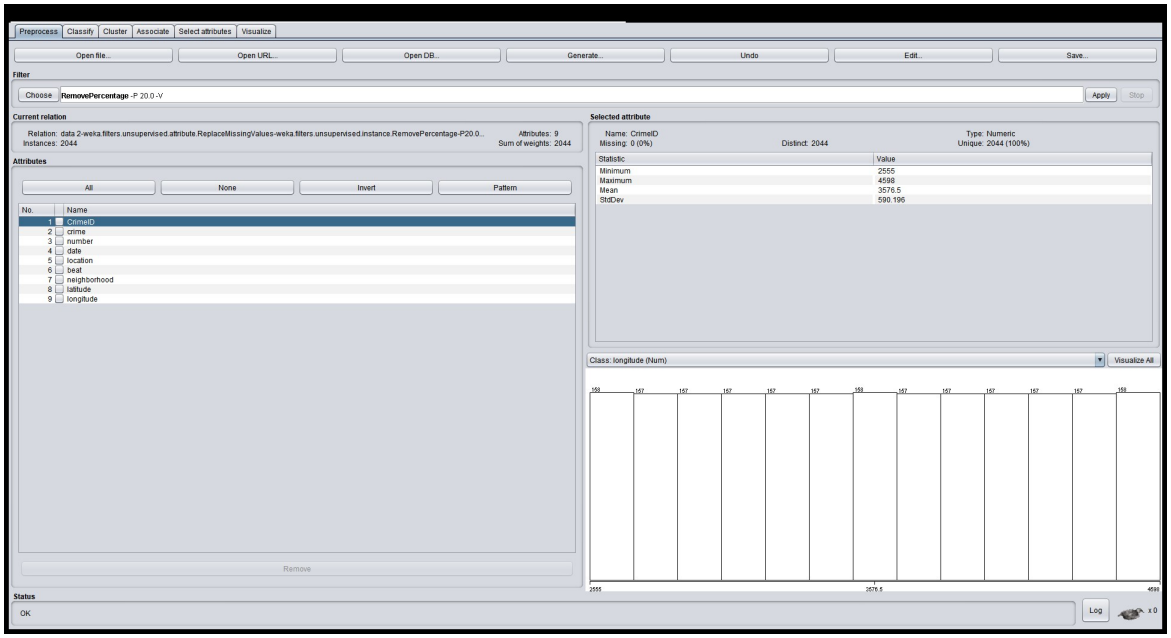


The dataset above has 12,777 instances and we split the data to 80-20. We split this in order to find the model parameter and estimate the general performance. When training the dataset with the algorithm in this case a classification algorithm its often not advised to train using the whole dataset because there are high chances of overfitting.



Therefore, we split the data and the instances reduce to 10,222 and training dataset. We use the RemovePercentage filter for the training dataset and testing set

3.4.6 Testing Dataset



The table above represents the test set trained. The testing dataset in this case has been used to test the accuracy of the classifier. Furthermore, it has helped us come up with the evaluation metrics analysis.

3.5 Relationship between conceptual framework and architectural framework

The architectural model is key to ensure that the user or administrator has a good experience in terms of the navigation and this should not cross the experienced computer users , and the system should also perform executions faster and give up-date information thus making it interactive with the user. In this case, the architecture of the proposed system shows the connection in each of the modules from the user to the crime prediction creating a visualisation of the design of the proposed system. The conceptual framework on the other hand simply shows the input of the data, how its classified and the output used for prediction.

3.6 Tools

3.6.1 Weka

Weka as open free source software for machine learning which contain tools for data mining tasks. Using this technique or framework is that it allows the use of different algorithms.

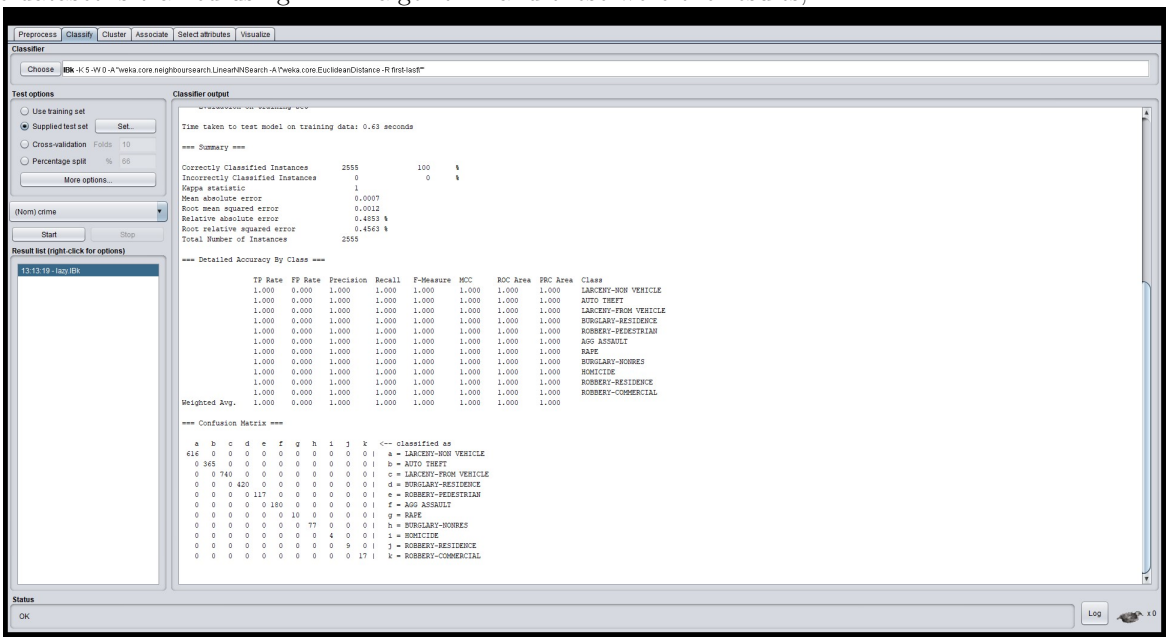
3.6.2 K-NN Algorithm

This algorithm uses classification technique whereby in Weka it uses IBk Instance Based Learner whereby it does not build a model rather it generates a prediction of test, and k represents the number of nearest neighbors to use. The IBk in this case uses distance to measure and locate the k close instance and the selected to make a prediction. This algorithm stores the available cases and classifies new cases based on their similarity. In weka, we use the non-parametric lazy learning algorithm. We use the Euclidian distance and normalize the nominal attributes between 0 and 1.

3.7 Testing

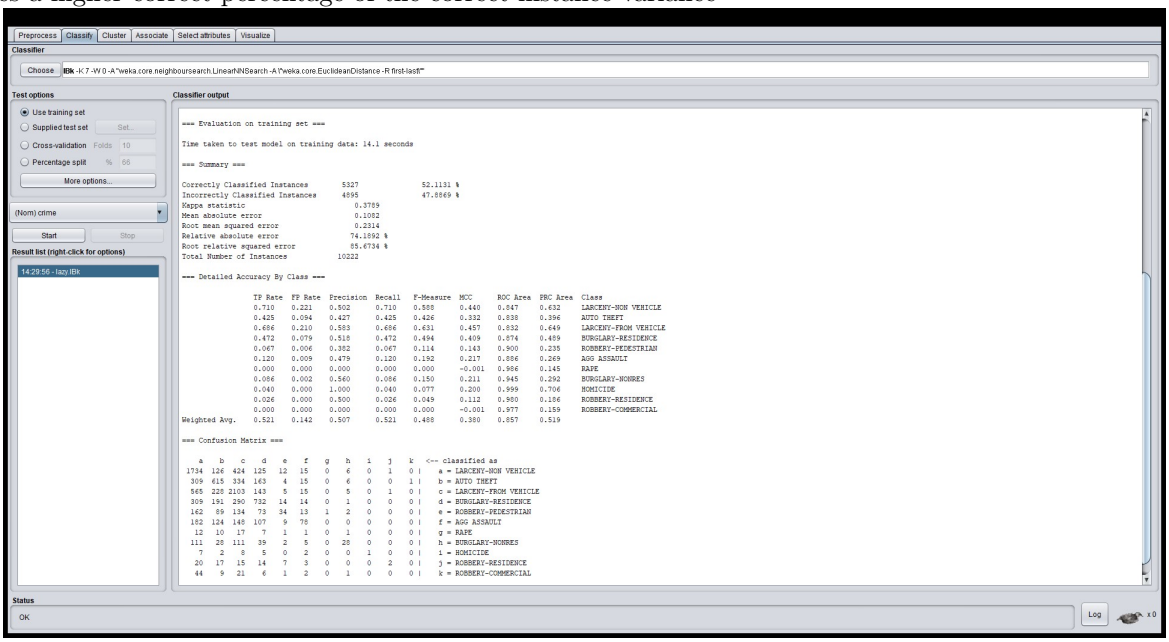
3.7.1 Training Dataset

The dataset is trained using K-NN algorithm and these were the results;



In the table above using k instance as k=5 gives us a correct instance as 100 which indicates that there is overfitting. Training the dataset with a lower or the smallest k instance makes it gives the correct instance as perfect fitting

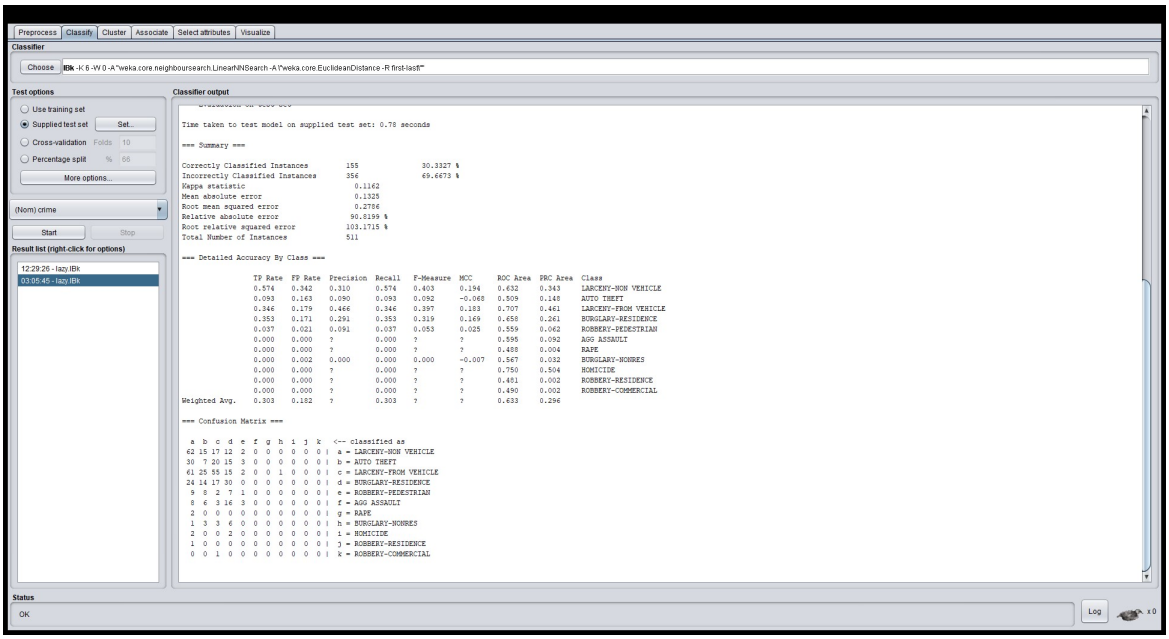
dataset however, the accuracy on this makes it have a poor performance on the new data. We use the default number of nearest neighbors as k=7 because it gives a higher correct percentage of the correct instance variance



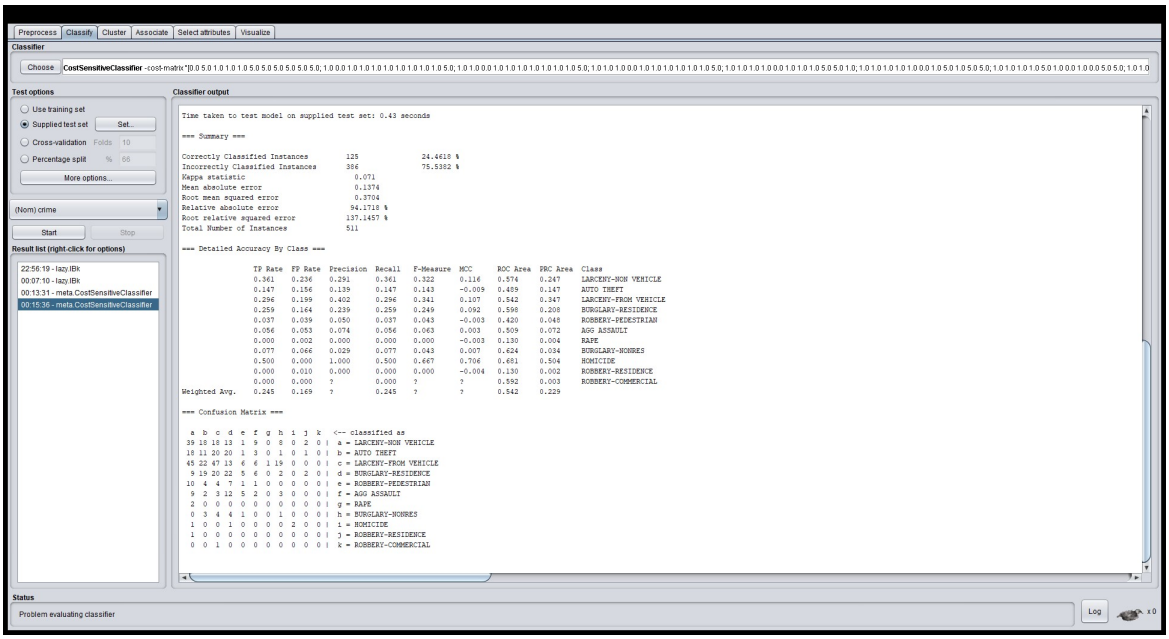
The results in the above changes the correct instances and we have an accuracy of 52.1131 indicating the percentage put in the right class and incorrect classified instance of 47.8869. to the percentage in the wrong class. However, we cannot increase the number of k neighbors because also it results to underfitting.

3.7.2 Testing Dataset

In this case the test dataset used shows the accuracy of the classifier and these are the results.



The results in this case at the confusion matrix show that some of the classes were no classified. The confusion matrix in this case shows how best the classifier is doing. The y-axis on the right shows data belongs. For example, in this case we have 11 rows and if we add up the elements in each case shows the elements that Some of the statistics were not computed and shows that some of these classes are under-represented by (?) symbol. This certainly shows the imbalanced classes and to reduce this or balance the classes we use the CostSensitiveClassifier which improves the performance of the model and probability of the classifier. We set the costmatrix as per the number of classes given which in this case, we have 11 classes and penalize on the less classified classes. For example, we can penalize from.



The results as from using CostSensittiveClassifier shows that now it improves the performance by having some of the classes presented which were not classified but even as we apply this method it changes the percentage of the classified instances. belong to each class.

3.7.3 Evaluation Metrics

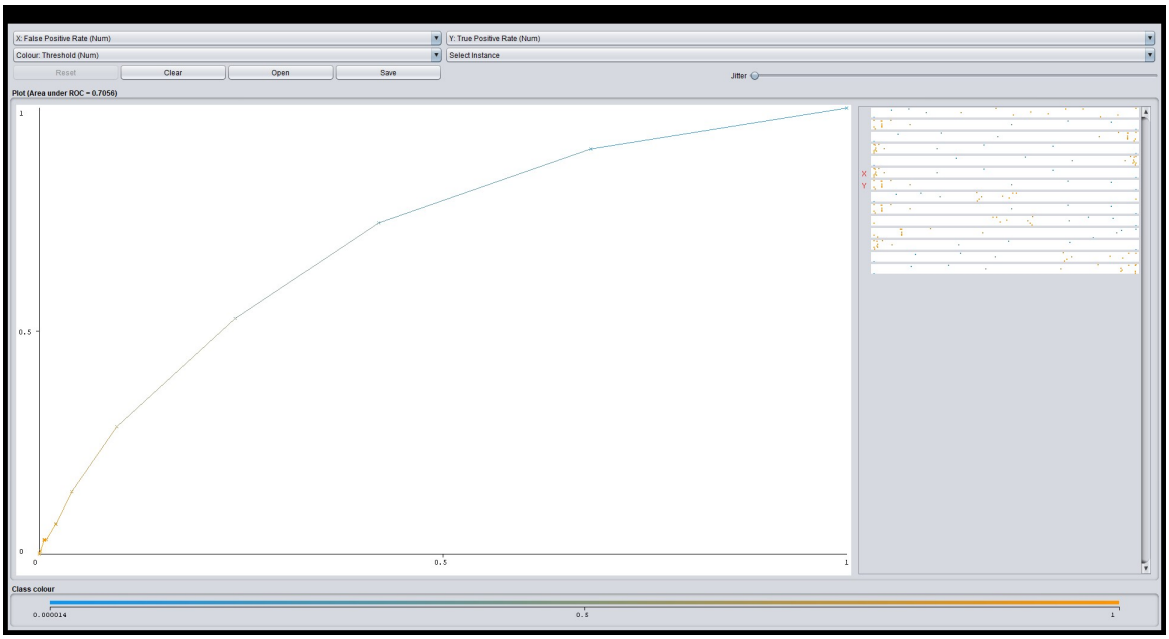
On the database back-end testing technique, map testing will be done to ensure the database components store the right information and match the insertion fields on the user's end Functional testing will be conducted to ensure the queries run properly and the functions run from the front-end maps if the functions require the database queries.

Kappa statistics which are the Cohens Kappa which takes in the value we can randomly guess and have the right value of the class and in this case the Kappa statistic value at 0.11 shows a slight agreement exists.

Accuracy this is detailed and gives a representation of the data and these include;

PRC area which represents the relationship between precision and recall that represents the measure of unbalanced datasets.

ROC area Under the curve which basically describes the performance of the classifiers in general and weka prints this area under the curve which contains the true positive and the false positive and mainly set to be ¿0.5 that shows a good classifier.



The graph above shows that the ROC curve is ¿0.5-1.0 indicating that this is an efficient classifier. The confusion matrix in this case shows how best the classifier is doing. The y-axis on the right shows data belongs. For example, in this case we have 11 rows and if we add up the elements in each case shows the elements that belong to each class.

3.8 Conclusion and Future Work

3.8.1 Conclusion

The main problem having classified different crime categories was challenging because there was not enough predictability in the data to obtain a high accuracy. We found out a meaningful approach of splitting the data to have the final structure of the data to obtain the high accuracy and precision in prediction. Possible areas whereby extended work should include time-series modeling of the data to understand temporal correlations in it, which be used to predict surges in different categories of crime. Also, this would 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.

3.8.2 Future Scope

The main goal shouldn't be only having the criminals arrested but prevent crimes from happening; 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.

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.

;

References

[1] H. U. Ay, A. Aysu Öner, and N. Yıldırım, "Can tech4good prevent domestic violence and femicides? an intelligent system design proposal for restraining order implementations," in *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 2021, pp. 34–45.

[2] M. Vaquero Barnadas, "Machine learning applied to crime prediction," B.S. thesis, Universitat Politècnica de Catalunya, 2016.

[3] M. McGuire, "Technology crime and technology control: Contexts and history," in *The Routledge handbook of technology, crime and justice*. Routledge, 2017, pp. 35–60.

[4] C. Hollings, U. Martin, and A. C. Rice, *Ada Lovelace: The making of a computer Scientist*. Wiley Online Library, 2018.

[5] S. Muggleton, "Alan turing and the development of artificial intelligence," *AI communications*, vol. 27, no. 1, pp. 3–10, 2014.

[6] F. Lara, "Artificial neural networks: An introduction," *Instrumentation and Development*, vol. 3, no. 9, 1998.

[7] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[8] G. C. Oatley and B. W. Ewart, "Crimes analysis software:'pins in maps', clustering and bayes net prediction," *Expert Systems with Applications*, vol. 25, no. 4, pp. 569–588, 2003.

[9] T. Cheng and M. Adepeju, "Detecting emerging space-time crime patterns by prospective stss," in *Proceedings of the 12th international conference on geocomputation*, 2013.

[10] D. J. Fitzpatrick, W. L. Gorr, and D. B. Neill, "Policing chronic and temporary hot spots of violent crime: A controlled field experiment," *arXiv preprint arXiv:2011.06019*, 2020.

[11] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, pp. 1–14, 2021.

[12] R. Hartman, "Ai is watching you."

[13] H. Yu, L. Liu, B. Yang, and M. Lan, "Crime prediction with historical crime and movement data of potential offenders using a spatio-temporal cokriging method," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 732, 2020.

[14] X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, pp. 181 302–181 310, 2020.

[15] N. Abdulrahman and W. Abedalkhader, "Knn classifier and naïve bayse classifier for crime prediction in san francisco context," *International Journal of Database Management Systems (IJDMS)*, vol. 9, no. 4, pp. 1–9, 2017.

[16] A. Almaw and K. Kadam, "Survey paper on crime prediction using ensemble approach," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 133–139, 2018.

[17] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering," *International Journal of Computer Applications*, vol. 83, no. 4, 2013.

[18] R. Kiani, S. Mahdavi, and A. Keshavarzi, "Analysis and prediction of crimes by clustering and classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 8, pp. 11–17, 2015.

[19] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using twitter sentiment and weather," in *2015 Systems and Information Engineering Design Symposium*. Citeseer, 2015, pp. 63–68.