Crime Pattern Clustering and Analysing Model

A dissertation submitted for the Degree of Master of Computer Science

Illamaran. B

University of Colombo School of Computing

2019

Abstract

The details about hotspot and the implicational uses of K-Means clustering for crime pattern cluster creation have been clearly discussed here. This approach initially identifies the significant attributes in the dataset. When compared to other papers, the added weight is given to the mentioned attributes in the data set. The criminal activity is the most important attribute and this is given the highest priority when compared to the other attributes. In this project, there is a significant usage of this feature from the relevant research paper. The crime analyst who does the analysis selects the number of clusters what he/she wants. The crime clusters are created based on that selection.

The spatial-temporal analysis of the crime is conducted using the geographical analysis machine (GAM) which is the GAM clustering techniques. Here using the relevant K-Means clustering with pattern detection techniques, the crime clusters are created. Even though this, the crimes have occurred in various locations and times. As a result of this, the clusters must be analyzed with the guidance of the GAM clustering techniques that are within and across the clusters and that particularly identifies the diversity of the particular criminal activity, and that forecasts on the suspected crime location for the crime clusters. Along with the foreordained qualities and the weights, these obtained records are grouped in light. The bunch of conceivable crime is related to designs in a subsequent way. Then these subsequent bunches are plotted on the geo-spatial plot.

Keywords:

Spatial-Temporal analysis, Geographical analysis machine (GAM), Geo-spatial plot, K-Means clustering, Hotspot, Data mining

Acknowledgment

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this MCS project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

I express my gratitude to Dr. Kasun de Zoysa for support and guidance he provided throughout the research. Without his valued comments and support, this project won't be completed to fruition.

Thank you,

Illamaran. B

Table of Contents

Abstract.....

Acknowledgment ii

Table of Contents iii

List of Figures..... vi

List of Tables..... vii

List of Abbreviations viii

Chapter – 1 | Introduction..... 1

1. Introduction 2

1.1. Project Overview 2

1.2. Problem definition..... 3

1.3. Project Objectives..... 4

1.4. Project Scope 4

1.5. Case Study..... 5

Chapter – 2 | Project Background 6

2. Project Background..... 7

2.1. Background Analysis..... 7

2.2. Background Study..... 7

2.3. Existing Crime Analysis Approach 17

2.3.1 Statistical Approach 17

2.3.2 Expert Knowledge Approach 18

2.3.3 Collective Data Mining Techniques 19

2.3.4 Clustering Approach..... 20

2.3.5 Machine Learning Approach..... 22

Chapter – 3 | Requirement Analysis..... 24

3. Requirement Analysis..... 25

3.1. Introduction..... 25

3.2. Requirement Gathering Techniques..... 26

3.3. Brainstorming 26

3.4. Document Analysis 27

3.5. Document Use Cases 27

3.6. Interview..... 28

3.7. Survey Method 29
3.7.1 Observations 29
3.7.2 Online Survey 30
3.7.3 Limitation of the Online Survey 32
3.8. Requirement Analysis 32
3.9. Functional Requirements 33
3.10 Non - Functional Requirement 34
3.11 Constraints and Dependencies 35
3.12 Use Case Diagram 36
Chapter – 4 | Project Plan 37
4. Project Plan 38
4.1 Project Management Approaches 38
4.2 Project Planning Tools and Techniques 38
4.3 Risk Management 39
4.4 Work Breakdown Chart 39
Chapter – 5 | Design 40
5. Design 41
5.1 High-level Diagram 41
5.2 Data Flow Diagram 42
Chapter – 6 | Methodology & Implementation 45
6. Methodology & Implementation 46
6.1. Project Development Approach 46
6.2. Project Development Approach 48
6.2.1 Clustering 48
6.2.2 K-Means Clustering 48
6.3 System Model 50
6.3.1 Creating target dataset 53
6.3.2 Data cleaning and Data preprocessing 53
6.3.3 Data Reduction and Data projection 54
6.3.4 Data Mining Task for Crime Analysis 54
6.4 System Implementation 55

Chapter – 7 | Results & Discussion 61
7. Results & Discussion 62
7.1. Cluster Creations 62
7.2. Cluster Analysis Using Weka API 68
7.3. Cluster Analysis Using Weka Tool 71
Chapter – 8 | Evaluation & Testing 73
8. Evaluation & Testing 74
8.1 Evaluation Methods 74
8.2 Overall Testing 75
8.3 Module Testing 76
8.4 Functional Testing 77
Chapter – 9 | Conclusion 80
9. Conclusion 81
9.1 Learning Points 81
9.2 Key Achievements 81
9.3 Future Work 82
Chapter – 10 | References 83
10. References 84

List of Figures

Figure 2.1 - Crime analysis method 9
Figure 2.2 - Crime cluster partition 10
Figure 2.3 - Homicide crime pattern (C0) 12
Figure 2.4 - Homicide crime pattern (C1) 13
Figure 2.5 - Homicide crime pattern (C2) 13
Figure 2.6 - Homicide crime pattern (C3) 14
Figure 2.7 - Homicide crime pattern (C4) 14
Figure 2.8 - Crime matching process engine 15
Figure 3.1 - Use case diagram 36
Figure 4.1 - Work breakdown chart 39
Figure 5.1 - High-level dataflow diagram 42
Figure 5.2 - 0th level dataflow diagram 43
Figure 5.3 - 1st level dataflow diagram 43
Figure 5.4 - 2nd level dataflow diagram 44
Figure 6.1 - Sequence diagram 47
Figure 6.2 - K-Means clustering 49
Figure 6.3 - High-level architectural diagram 53
Figure 6.4 - Calculating neighbors using the method of the original code 56
Figure 6.5 - Removing the outliers using the method of the original code 58
Figure 7.1 - Colombo city map 62
Figure 7.2 - Kandy city map 62
Figure 7.3 - Crime clustering in Colombo city using whole data 63
Figure 7.4 - Crime clustering in Colombo city for selected crime types 64
Figure 7.5 - Crime clustering in Colombo city for selected crime location 65
Figure 7.6 - Crime clustering in Colombo city for gender wise 66
Figure 7.7 - Crime clustering for Kandy city 67
Figure 7.8 - Error measurements in map 69
Figure 7.9 - Prediction measurements in map 70
Figure 7.10 - Stats automation of Weka 71
Figure 7.11 - Stats results of Weka 71
Figure 7.12 - Plots automation of Weka 72
Figure 7.13 - Plots results of Weka 72

List of Tables

Table 2.1 - Homicide crimes 12
Table 3.1 - End-user survey questions 30
Table 3.2 - System expert survey questions 31
Table 3.3 - Public survey questions 31
Table 7.1 - Cluster count for whole data in Colombo city 63
Table 7.2 - Cluster count for selected crime type in Colombo city 65
Table 7.3 - Cluster count for a selected location in Colombo city 66
Table 7.4 - Cluster count for gender-wise in Colombo city 67
Table 7.5 - Cluster pattern for Kandy city map 67
Table 7.6 - Weka Actual, Predicted values 68
Table 7.7 - Error measurements in the table 70
Table 7.8 - Prediction measurements in the table 70
Table 8.1 - Evaluation 75
Table 8.2 - Module Testing 77
Table 8.3 - Functional Testing 78

List of Abbreviations

MO	Modus Operandi
AREST	Armed Robbery Eidetic Suspect Typing
ICAP	Integrated Criminal Apprehension Program
VIP	Very Important Places
FIR	First Information Report
SOM	Self-Organizing Maps
Weka	Waikato Environment for Knowledge Analysis
DNA	Deoxyribo Nucleic Acid
UCI	University of California Irvine
FSOM	Fuzzy self-organizing map
STAC	Spatial-Temporal Analysis of Crime
GAM	Geographical Analysis Machine
DSIM	Dynamic Systems Development Method

Chapter – 1 | Introduction

1. Introduction

1.1. Project Overview

The volume of crime data is expanding alongside the occurrence and multifaceted nature of violations. Data mining is a powerful tool if extensive training on data mining is provided to the data analyst it can help them to explore large databases fast and efficiently. Criminal justice and law enforcement specialist are having high ability to solve crimes at a higher rate. Computer data analysts have helped the law enforcement officers to solve their crimes fast and efficiently with the help of technological advantages. Here the author will be developing a platform that will relate computer science and criminal justice which can help to solve the crimes faster when the platform is developed. The author will be using clustering based models to help to identify crime patterns in a more unique way.

Crime detection problems can be figured out using a data mining tool. Crimes can affect society and it can be a nuisance to everybody. Researches that can solve problems at a faster rate will be highly appreciated. 50% of the crimes are committed by about 10% of the criminals. Clustering algorithm for data mining can be used to detect the crime patterns and help to solve the crime quickly. K-Means clustering with a few enhancements can help in the process of crime patterns identification. These techniques were applied to real crime data which were obtained from a Sheriff's office and then the author also validated our result. To increase the predictive accuracy and to receive knowledge discovery from the crime records the author uses a semi-supervised learning technique. A weighting scheme was developed for various attributes to help in a deal with out of the box limitations from clustering tools and technique. Data mining framework that works with the geo-spatial plot of crime can be implemented that helps to improve the productivity of the detectives and other law enforcement officers. This can also aid in counter-terrorism for homeland security.

Definition of crime mapping, crime analysis and crime prevention:

- With the help of crime mapping, the author is able to manipulate and process spatially referenced crimes that are in order to deliver the output in a visually display informative format. The location of hotspots can be provided by this and also help to report the high-level of crime.
- The set of processes that are applied to relevant information about crime patterns are crime analysis. The result of the analysis can be used by administrative and police to prevent and suppress criminal activities and for investigations.
- The risks of criminal events and anti-social behavior can be reduced in their causes with the help of crime prevention.

1.2. Problem definition

The manual recordings of the crime incidents cannot be the ways forward when the police authorities have goals with regards to improve crime management. The responsibility of the different police stations is to handle various crime incidences in line with the boundary under their jurisdiction. But there are problems with their recording systems and the aggregation conducted in all the police station of crime incidents. The recordings which are done in the police station are manually written in blotters which make it difficult for decision making for all the police stations.

The below following issues were identified during a crime investigation.

- How to find high-density crime areas
- How to find the crime type over those areas
- How to identify the crime pattern over an area
- Investigation of the crime takes a longer duration due to the complexity of issues
- Identifying crime similarities over an area
- Increase in the size of crime information that has to be stored and analyzed

1.3. Project Objectives

The main objective of this project is to develop a methodology for crime clustering and spatial analysis for the police department.

Specific objectives;

- The proposed system helps to identify the crime clusters in relation to criminal activity. The crime pattern can be understood by the crime analyst with the help of this.
- The investigation can be expedited and analyzing the growing volumes of the crime data can be accurately carried out with the help of this proposed system.
- The proposed system helps to identify and analyze common crime patterns which can help to reduce further occurrences of similar incidents.
- The proposed system is utilized alongside the geo-spatial plot. The crime analyst may pick a period range and the type of crimes from certain topography and show the outcome graphically. So the futuristic crime rates will be forecasted with proper visualizations.

1.4. Project Scope

From Kandy and Colombo districts, the author has collected around 5000 records from crime departments. Since the data from other districts were not available to the expected level, the author was not able to collect data related to those districts. Social criminal activities can only be included in these data sets. With this kind of an initiative, it helps to focus more on finding out criminal activities. The author was not able to find out the person who did the crime because of security concerns. Normally it is a difficult task to trace out the individual name. The author has to note that there are no critical or political crimes. The amount of the datasets depends completely on theses cluster based predictions. So the author has to make sure these cluster based predictions are in order to derive the amount of dataset. The accuracy of the solution can be figured out and decided with the amount of data. The data in Colombo and Kandy districts were collected for one year, so the prediction is based on this period. A wide range of period provides an accurate prediction for the solution.

1.5. Case Study

Colombo is the capital city in Sri Lanka. In the past years, Colombo is growing and developing at a rapid rate. The current population of the metropolitan Colombo city is 5.6 million. Colombo is the financial hub throughout the country and more business transactions are happening in Colombo. There are global firms which are investing in Colombo to expand a structure and develop the infrastructure. As a result of the environmental contrasts between Colombo and other areas, urban growth has been well influenced. The Colombo city is segmented into 15 segments starting from Colombo 1 to Colombo 15. Kandy is the second major city in Sri Lanka which is located in the Central Province. The city is the hill station area which attracts many tourists. Kandy stands next to Colombo in term of city development. Since Kandy has many tourists visiting every year there are more chances for crime occurrences. So based upon this the author has selected Colombo and Kandy for my crime data collection.

Chapter – 2 | Project Background

2. Project Background

2.1. Background Analysis

The author will be comparing and contrasting the terminologies that are used in criminal justice and police departments with respective to data mining systems. The suspect is identified as the person who has relatively committed the crime. The suspect is actually not convicted unless he is proved guilty. The target of the crime is the victim. The person who is reporting the crime in most cases is the victim and also the crime may have some external witnesses. Manslaughter or killing someone can be called as homicides [40]. There are also other categories like infanticide who kill infants. In order, for our modeling author will not go in deep into criminal justice but the author will confine himself to some other kinds of crimes. The geographical group of crime can be referred to as clusters, which means a lot of crimes in a given geographical region. Using a geo-spatial plot of the crime such clusters can be visually represented which are overlaid on the map of police jurisdiction. The hot-spots 'of crime can be located with the help of a densely populated group of crime. In case the author talks of clustering from a data mining point of view and the author will be able to refer to some similar kinds of crime according to the given geography of interest. These kind of clusters are very helpful in identifying the pattern of the crime spree. The DC sniper, a serial rapist or a serial killer are well-known examples of crime patterns. A single suspect or group of suspects can be involved in these crimes.

2.2. Background Study

The relevant information provided by the analytical process in relation to crime patterns is called crime analysis. Respectively the specific trend correlations assist personally in planning the deployment of resources and also help to prevent and suppress criminal activities. With the help of the crime analysis, this task is mainly completed. It is advised to analyses crime due to below following reasons [40], [7].

1. Analyze the methodical crime to inform law enforcers about general and specific crime trends in a timely manner.
2. Analyze the relevant crime to take advantage of the plenty of information existing in the justice system and public domain.

These methods also help to find out the patterns without the help of prior knowledge. The main objective of crime analysis includes:

The hidden patterns of crimes are found by the improved analysis and the rates are rapidly changing. Without the help of prior knowledge, these methods also aid to find out the pattern. The main objectives of crime analysis are defined as below.

1. Extraction of crime patterns by analysis of available crime and criminal data
2. Prediction of crime based on the spatial distribution of existing data and anticipation of crime rate using different data mining techniques
3. Detection of Crime

The crime dataset helps to analyze crime analysis by applying K-Means clustering algorithms which uses the rapid minor tool. The procedure is given below:

The data mining tool implemented easily by the open source, therefore, the analysis can be done easily and quickly. Crime dataset analyzed by using the K-Means clustering guidelines. This is used as a rapid minor tool. The procedure is defined as below in Figure 2.1.

1. Initially, the researcher has to consider crime dataset.

2. The datasets are filtered according to requirement and this helps to create a new dataset which has relevant attribute according to analysis.

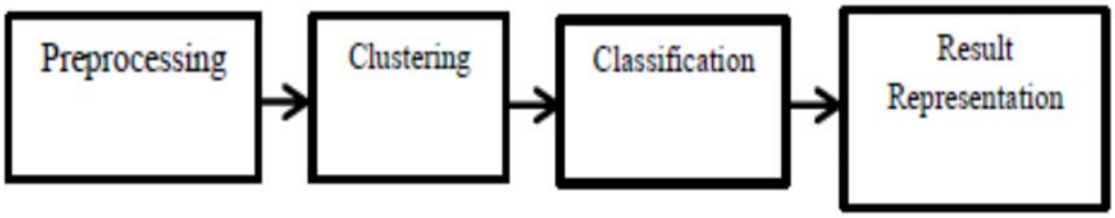
3. Initially, the rapid miner tool should be opened and then read the excel file of the crime dataset and then apply “Replace Missing value operator" on it and then execute the operation.

4. Perform “Normalize operator” on the resultant dataset and execute an operation.

5. The K-Means clustering on resultant dataset should be performed which were formed after normalization and execute an operation.

6. The results from the view must be plotted between crimes and then get the required cluster.

7. The relevant analysis can be done on the cluster formed.



Source: Analysis and Prediction of crimes by clustering and Classification, (2015) [7]

Figure 2.1 - Crime analysis method

The researcher here introduces a new framework for clustering and prediction of cluster members to analyze crimes. The researcher uses the rapid minor for the purpose of the implementation and the dataset of crime which is used for this was recorded by police in England and Wales from 1990 to 2011.

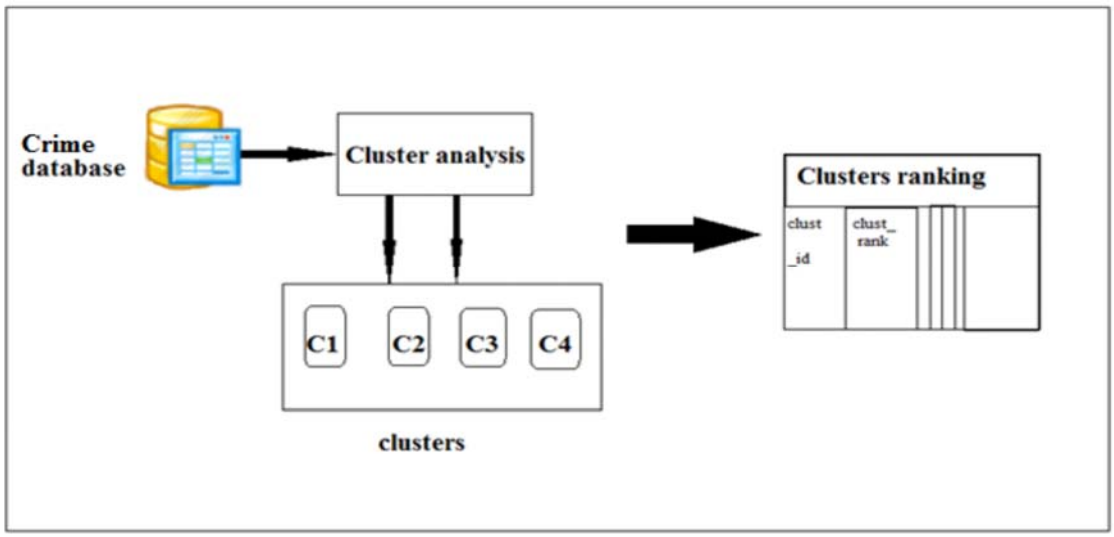
Crime profiling based on categories of areas are mentioned in this research [6]. The result yielded will show six clusters and each shows crime in six types of areas. The categorized areas considered for profiling are: [7]

In this research, the researcher clearly mentioned crime profiling based on categories of areas.

- Slums
- Residential areas
- Commercial areas
- Very Important Places (VIP) zones
- Travel points
- Markets

In accordance with the First Information Report (FIR) records from Delhi police the primary database is created. Then random sampling is done and the database is refined on the basis of area category. Then data is fed to the Waikato Environment for Knowledge Analysis (Weka)

and then the clustering is done. This is done in order to identify the association between crime type and area category. This helps to find in which area category which type of crime occurs the most and least [40], [7]. The diagram is shown in below Figure 2.2.



Source: Analysis and Prediction of crimes by clustering and Classification, (2015) [7]

Figure 2.2 - Crime cluster partition

This is a process of considering the challenges of leveraging classic hierarchical and partitioning clustering techniques which helps in intelligent crime analysis. Relatively the different approach was introduced which had utilized SOM neural network for crime data clustering. The commitment of clustering and analyzing the process on the binary nature of criminal behavior requires some elegance. Generally, because of the binary encoding, the commonly used continuous types of variables become useless this happens because of the popular Euclidian distance measure. It can lead to misleading results in the clustering process using these valuables. The relevant distant function which is specific for achieving the similarity between binary data objects should be leveraged. There is a comparative dissimilarity between two objects which are calculated here based on the corresponding distance. There is a specific equation calculated between two objects.

The Spatial-temporal analysis [6], [9] can be involved to analyze the cluster in a location wise which can be within the cluster or across the clusters. The Spatial-temporal analysis was done over the specified clusters by the researchers. To get the crime location the population of the city can help. Crime analysis becomes more accurate with the help of this method. There are multiple crime clusters within the country, the centroids of the clusters are found. The centroids converge to one point as a result of this. Then the specific distance to each and every city from the cluster centroid point was found and the needful forecasting analysis was done.

In the crime data analysis, the Modus operandi (MO) had played a vital role. According to the criminology theory, the professional criminal particularly tends to have his/her own MO which cannot be changed by him/her freely. The MO is used to relentlessly describe criminals' characteristic pattern and style of work. The probability of the crime incidents committed by the same criminal or criminals will rise if two or more crime incidents show similar MO. Generally, in this theory, there are clustering techniques which use data points into clusters. Then the same cluster in more similar points is grouped into different clusters which are often used to link crime incidents and identify the particular criminal in a similar way. In order to solve the problem of linking criminal incidents, lots of research methods were introduced and conducted.

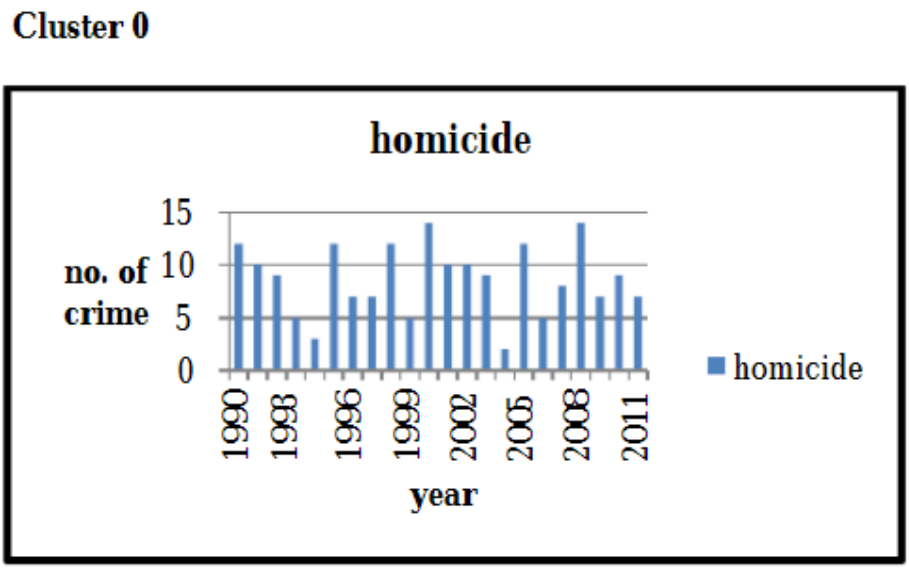
A Literature proposal based on similarity approach is particularly aimed at the outlier analysis. Some the actual systems which were developed to solve the criminal incident association problem are the Armed Robbery Eidetic Suspect Typing (AREST), The Integrated Criminal Apprehension Program (ICAP), COPLINK project and RECAP [5]. There is a large volume of outputs which were produced by the MO analysis but the really valuable clues always hide behind other useless ones. Even if crime data analysis spends more time on the set of results, it is actually not guaranteed to get a valuable clue. It is a difficult problem to deal with the results effectively. The clustering results quality can be determined by the similarity measure. The crime transforming was done by one of the researchers. In England and Wales, the crime dataset is used for crime analysis recording is an offense. This is obtained from England and Wales by offense and police force area from 1990 to 2011-12 [1]. The sample crime dataset is shown in Table 2.1 [40].

Table 2.1 - Homicide crimes

Year	Homicide	Attempted Murder	Child destruction	Causing death by careless driving
1990	10	19	0	7
1990	6	10	0	5
1990	6	8	0	9
1990	6	2	0	15
1990	10	5	0	1

Source: Crime Analysis using K-Means clustering, (2013) [40]

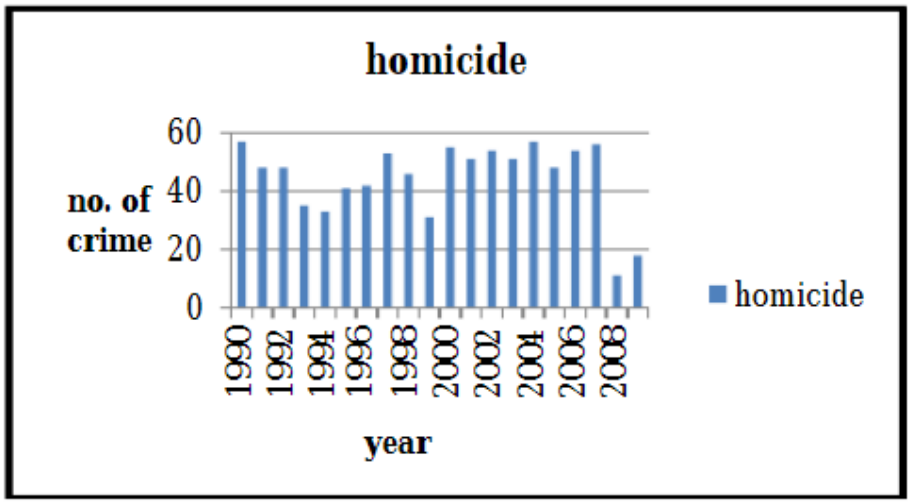
This included transferring crime rate changes from one year to next and how to make these changes in the future. Here researcher considered the offense of murder and the year it was created by the group and the analysis variation in the plot [4]. The cluster analysis is shown in below Figures such as Figure 2.3, Figure 2.4, Figure 2.5, Figure2.6 and Figure2.7.



Source: Crime Analysis using K-Means clustering, (2013) [40]

Figure 2.3 - Homicide crime pattern (C0)

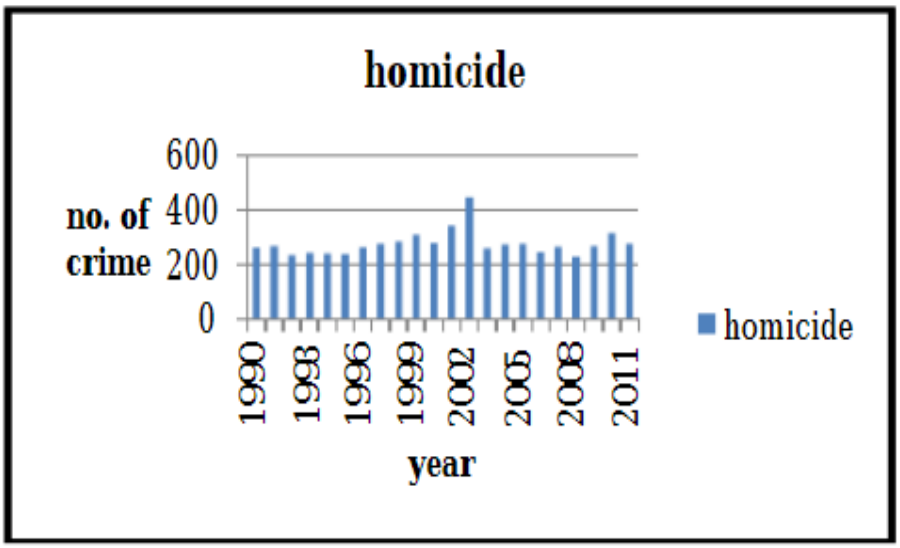
Cluster 1



Source: Crime Analysis using K-Means clustering, (2013) [40]

Figure 2.4 - Homicide crime pattern (C1)

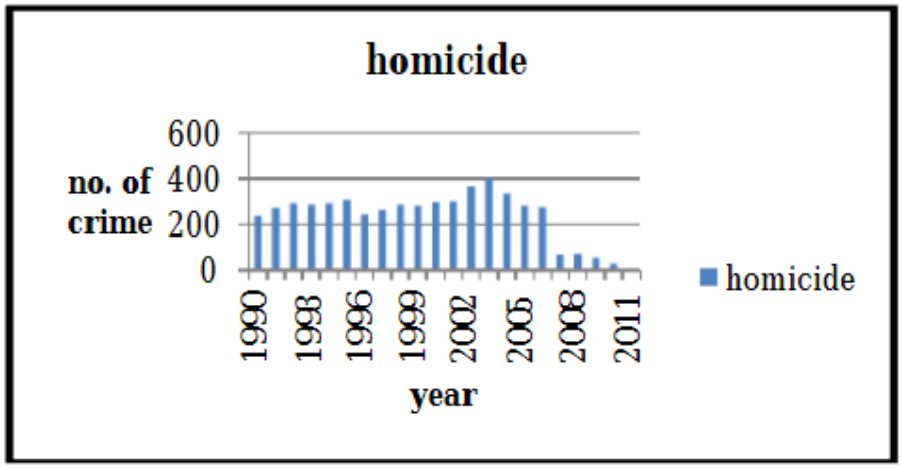
Cluster 2



Source: Crime Analysis using K-Means clustering, (2013) [40]

Figure 2.5 - Homicide crime pattern (C2)

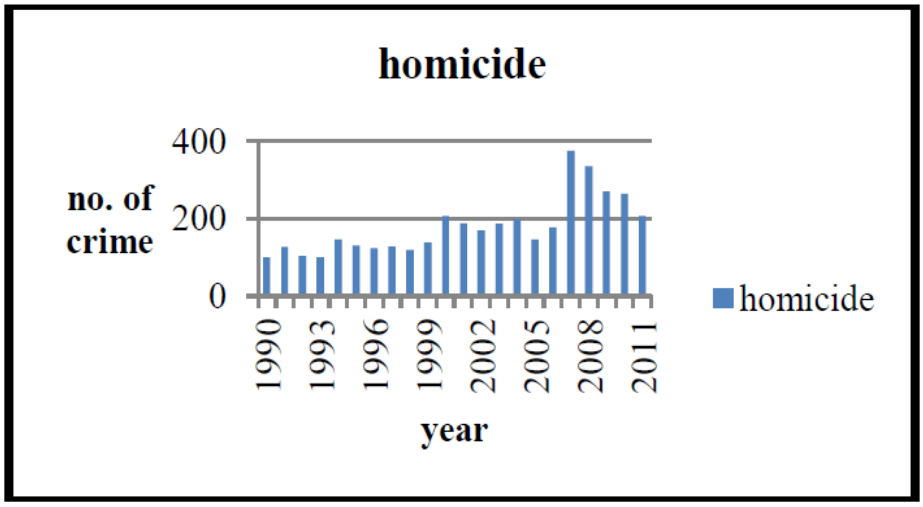
Cluster 3



Source: Crime Analysis using K-Means clustering, (2013) [40]

Figure 2.6 - Homicide crime pattern (C3)

Cluster 4



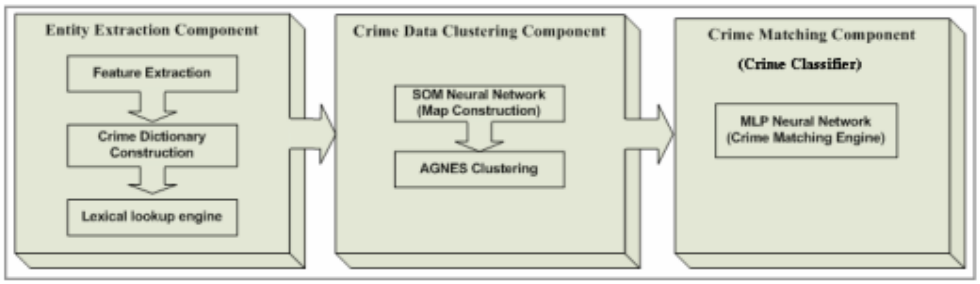
Source: Crime Analysis using K-Means clustering, (2013) [40]

Figure 2.7 - Homicide crime pattern (C4)

According to the researchers, there were details about the crime analysis. Criminals choose their specific areas for illegal activities over their crime period. For crime matching, there was a utilization for self-organized MAP (SOM) neural network for crime matching purpose [44] and the relevant K-Means clustering algorithm [1] is used. The main techniques are mentioned below.

- Crime data clustering
- Neural Network as an engine for crime matching process

The main techniques are mentioned below in Figure 2.8.



Source: Detecting and investigating crimes by means of data mining: a general crime matching framework, (2011) [44]

Figure 2.8 - Crime matching process engine

The method and challenges of leveraging classic hierarchical and partitioning clustering techniques in intelligent crime analysis have been discussed in this section [40]. A proposed approach for crime data clustering is represented which utilizes SOM neural network. Generally, the specific binary nature of crime behavioral variables may be regarded as a challenge, because committing clustering and analysis process on these types of variables requires some elegance.

The popular Euclidian [45] distance measure is commonly used in this approach for different types of variables. In a clustering process, the behaving binary quantities which are similar to continuous quantities can lead to misleading results. The functions should be leveraged in some other distances which are specific for the similarity in between binary data objects. The dissimilarity between two objects as their corresponding distance is calculated by these functions.

Based on the process of analyzing the cluster location the Spatial-Temporal analysis is involved [16], this can be particularly within the cluster or across the cluster. Over the clusters, the researchers have done the Spatial-Temporal analysis. Based on the population of the city the researchers get the crime location. Within the country, there are multiple crime clusters and then the centroids converge to one point. As a result of this, the distance to each and every city from the cluster centroid point is found and further forecasting analysis is done.

Spatial and hot spot crime analysis [40], [42] was done by one of the researchers in England. The researcher selected a particular area and did the research. This means Merseyside County is in England from December 2010 to August 2012 (21 months), several crimes have been studied in this dataset including theft and antisocial behavior. By this analysis, many interesting findings have been blamed on crime in Merseyside. It included: Severe crime offenses levels, Hotspot with coarse crime levels, coherent changes over the entire Merseyside. The criminal trends some hot spots were contraindicated, and one strong contact between crime hotspot locations and boroughs or Zip code slots. We are statistically and trusted by this type of integrated analysis of crime systems can help law enforcement. The agent predicts criminal proceedings, excludes resources and encourages them to social awareness to reduce overall crime rates.

2.3. Existing Crime Analysis Approach

There are some existing crime analysis approaches are available.

2.3.1 Statistical Approach

Sometimes an extensive range of data sources and a complex structure can provide imperfect information and this barrier for crime exploration and prevention. The warehouse of crime profiles of recent and historical along with the socio-economic factors like population, poverty, unemployment is taken into consideration along with the multiple

Statistical analysis like correlation and regression, make the process soother. The relationship crime pattern and to build a process administration tactic are revealed by the work [42].

The crime was forecasted on the basis of criminality and modus operandi of the sequence criminals by the police. The preferred scale for inspection is miniscule for the forecast models are designed depending on the distinctive attributes of the problem domain and data, and this why the crime is not forecasted. The crime must be identified in regions as small as feasible for all the tactical purposes, this must be done by the police at the patrol district level or similar [43].

The main argument is to figure out if it is possible to forecast accurately the selected crime in advance, in a small geographical area. On the accuracy level comparison between the two approaches namely model-based forecasting and present practices of police are made, these are the two main arguments which are discussed frequently. In the short term crime forecasting, the unilabiate methods are combined along with the multivariate methods to help in this. With the help of the forecasting process, the hotspot and the criminality tracing of certain places can be attained. Certain explorations are made and the result must deal with the essential recognition of the problem that specifically happens when trying to forecast outcomes. Investigating some of the pragmatic matters involve in anticipating city-level crime rates by means of a general panel dataset in this light [10].

The combinations of socio-economic and socio-demographic issues were utilized for the task of crime mining. A MATLAB program is used to generate synthetic data using the multivariate Gaussians, this is done for the experimental purposes. The conversion to real data is carried out based on the base map. Then ten datasets are produced and the average output is considered. The framework starts with the normalizing approach by using min-max technique then weighted, directed multi-graph for each dataset is created, and this is done in a detailed manner. To find a similar co-distribution the correlation analysis is used. In order to find a correlation, a person's correlation coefficient or Jaccard's index is involved [11].

The agent-based representations of offense based on the geographical location and also aids in finding the individual victims characterization are permitted by the model. Generally, the population Reconstruction Model was appropriately established with forecasting the idea of utilizing the mixture comprising the census of the small zone and the sample of anonymized data to afford the synthetic list pertaining to the city or region population in the country [12].

2.3.2 Expert Knowledge Approach

In the investigation model developed by incorporating the data mining techniques, the Deoxyribo Nucleic Acid (DNA) and fingerprint from the crime scenes were used. Based on the skill set of training they underwent and the capability to examine the crime scenes, the crime scene investigations are classified into three levels. With the information about time, day, date happened along with the complete and also forensic samples, only along with these the activity of investigators for each crime scene is recorded. The cross-industry standard procedure for data mining is utilized in this investigation model. After the preparation of the data, it is let into the modeling with the domain expert knowledge to bring out the effective results of the investigation. The collected data of fingerprint from the crime scene into insufficient, eliminated, matched and outstanding is classified by the experts. The classification is done even for the DNA. The two flag fields that were created helps in the matching purpose [13].

2.3.3 Collective Data Mining Techniques

After providing the offender information or the data mining approach is applied, this paves the way to understand the behavioral pattern of the criminals. The preview of the relationship is provided by the rules mined from the profiles. The associated rule mining is a successful technique. The concept hierarchies are constructed in such a manner which in turn generates rules. It is significant that the incorporation of dissimilar statistics such as the Pearson's sample coefficient correlation and multiple correlations, and an incomplete correlation coefficient, all these improves viable benefits when considering the view of obscured circumstances and the final report [14].

From the early 20th century, it is known that anticipating has been a substantial part of criminal justice. For the necessary policing, with the help of data mining the decision support system is provided, and the achievement is gained using the soft forensic proof like modus operandi, temporal and geographical crime features. Forensic computing extracting practical lessons relating to the purpose of computer science and to set of conclusions that include the need for multidisciplinary contribution to guide, all these are included by the key things of the system. In the investigation of crimes like geographical information systems will display, link analysis algorithms clustering, and association along with the more complex process as per requirements, all these are processed by the use of various data mining techniques. Generally, the contemplation focuses on the intelligence that is based on conceptual knowledge and reasonable logistics of different kinds of statistical models that are specifically related with the calculation of a wide quantity of data that goes through mimicking illegal incidents and their particular correlations [17]. Here the data mining techniques classification and the clustering are appropriately applied over a felonious dataset in order to determine the hotspot and also in guiding in the calculation of the crimes and felons. As a result, the classification is carried out in three ways which are based on crime place, types and time to produce the relevant crime hotspot or for the crime cold spot depending on the incidents. The use of the weighted attributes along with the threshold ways to generate the cluster is done by the clustering [19].

Generally, the crimes are devised into eight categories of crime type which are based on the increasing harm to the public. Traffic violation, sex crime, theft, fraud, arson, drug offenses, violent crime and cybercrime are some of the crime types. Entity extraction, clustering techniques, classifications, association rule mining, deviation detection, sequential pattern mining, string comparator and social network analysis are the various data mining techniques. The result between the crime types and data mining methods like classification, clustering, association and trend visualization clearly depicts the intensity of the crime types are produced by the relevant crime data mining frameworks. The usage of single or combined data mining techniques can be made based on the investigators [33].

2.3.4 Clustering Approach

The useful information is provided by data fusion and data mining. Data fusion is used to combine data from multiple resources and the data mining is used to discover the pattern and also equivalently meant for automation to get the relationship among various attributes that have an impact on the crime. The nearest neighborhood clustering is performed when the work is carried out the spatial data [18]. Acknowledging the tactical operation support that is for the law enforcement agencies on a weekly basis, a multivariate prediction model for the hotspots was introduced. Prediction based on space-time event, a point pattern based on the transition density model was used, and it particularly depends on observed passed offense preferences. Geo-Space Feature density, Spatial Transition density and temporal transition density are the models of the components. According to the time scale of the Richmond, [41] the team had worked. The crime area considering the counts of criminal activities, demographic features, consumer expenditure features and distance features are some of the places where the team had worked on the Richmond time scale [15].

The tactical police resources were used by the model to identify the high crime spots. The process of crime analysis is derived from predicting the crime future. This is done along with space and time combining people's behavior and it is termed as spatial choice analysis. The approach is normally depending on the clustered results which are from the offense that has happened and the activities with respect to time and location. Space adjusted and key featured adjusted model are the two proposed models where the decision is derived. Spatial analysis [8] that performs point density analysis, cluster analysis, cluster allocation and

bounding are employed to perform. Consisting of gamma test and model implementation for crime prediction purpose, the sequential cluster modeling is performed [16].

Along with the police strength of both civil and armed, the real-time data is considered by the crime prediction model. To get rid of missing data on careful analysis special care is taken. The quality data that is used for further clustering process are ensured by the combination of required algorithms. DB scan with a hybrid of K-Means is used to cluster for clusters stating crime is ready. This is rising and also generally increasing as in flux. Semi-supervised learning was utilized for knowledge discovery and this will eventually boost the analytical precision. The ultimate aim of the work is to set the tool development for Sri Lankan scenario which in turn will help the law enforcement department to tackle the crime investigation in a better way [20][38]. With the conventional weights, the city level crime data is worked. K-Means which is weighted works on the basis logical difference between the clusters inspected for the variations and between smallest and largest weight. The city level crime data is worked along with the conventional K-means and they are proposed by the weighted K-Means which are along with Radial Basis Function classification in both the methods. When the fine tuning is done with the help of weights, the accuracy is higher. According to the logical difference between the clusters and the weights, the weighted K-Means works. Along with the Radial basis, function accuracy is higher in this proposed technique [21].

Along with the data mining technique, the behavior of violent criminals is taken into consideration in the simulation model. Violent Psychopaths, Persons with Antisocial Personality disorder and Persons with Intermittent Explosive Disorder are the three types of criminals. In crime data, inconsistent in data, mapping the criminal behavior to the crime, the pertinent challenges are increasing here. Clustering, classification and outlier detection is applied by the simulation model. The process of cleaning and retaining the required attributes are performed. The grouping is performed by the clustering by tracing common properties in Classification and analysis done based on specificity. The properties that are used in the model are based on criminal behavior literature. Informatics output [41] is derived from the insertion of probabilistic and timing factors, and they are high in accuracy [15].

2.3.5 Machine Learning Approach

Generally, the lack of software impact based on the police need has paved the way to the over project. They are more often aimed at various records of the past to address the issues. Using the MS Access the work was started and they were analyzed using structured query language. The developing system normally affords the mapping and visualization tools which are for the existing data along with predicting perspective. The last stage of the work focuses on criminal profiling that generally helps to solve the crimes based on victimization and the possible offenders of the same. To specifically matched crimes with the criminal Kohonen, the neural network is used and to overview. Fuzzy self-organizing map (FSOM) network [23] was proposed by "LI" team. They have worked on relevant temporal criminal activity data which are from the non-western real world. The main objective of the work was to find the crime trend patterns and rule generation to reveal hidden causal –effect knowledge. In Taiwan, the data from the national police agency is obtained and in the pre-processing phase, data standardization and fuzzification are applied. On the process data, the FSOM works on it, they work by setting the parameter for a number of clustering confidence and support for the rule extraction. The crime trends, offenses, locations and side around effects are revealed by the output. The information about the typical, gradual increase, sharp increase and winter time crime patterns are provided by the analysis. The above-said crime patterns were true and they were proved by statistical tests supports the [24].

The Arabic documents were worked by Alruily in order to categorize into a type of crime. There were various crimes that were taken into consideration some of them were the theft, fraud, drug and alcohol smuggling, magic and sorcery, sex and violent crime. Generally, normalization starts the process that heads to resolve the difference that exists for the same word. After this, the information extraction is performed which is combined with the stemming process. The affixes are eliminated by the stemming and the works are based on unique word concept along which that removes the stop word. In order to cluster the documents, clustering and visualization are invoked. They are done with the help of Self Organizing Map technique. With the initialization of the weight randomly and neighborhood ratio, the process begins. Until the convergence criterion is met, the input pattern is set, Euclidean distance is calculated, winner neuron is found and updating occurs [25].

The deriving patterns based on different aspects of crime, this is done when the work is concentrated and applied to this. The aspects here refer to the data sources that are versatile. The objectives that disclose the trend is a pertinent thing, this is done help of exploring the data at various stages based on granularity and also aspect criteria. A method of hierarchical clustering is deployed which is the growing self-organizing map. This process is involved during the initialization phase and also growing and smoothing phase. The granularity is called the further works based on the concept of hierarchy. This is mainly based on multi-modal data and the model is constructed. The identification on the basis of horizontal and vertical pattern identification with global concept hierarchy, this was mainly identified by the architecture. The different level of granularity and abstraction are provided by the same [36].

The work of classification technique on the University of California Irvine (UCI) Communities and crime dataset was clearly proposed by Iqbal. Along with 12 attributes, the work was carried out. Using the Decision tree and Naïve Bayesian algorithm with the help of Weka tool the prediction was done. After removing the missing values attribute and then multiple linear regressions on the UCI Communities and crime dataset, then the model was clearly built after the above steps and process. Along with the wine quality dataset, a spam-based dataset with various techniques to bring out the performance the work was performed. The prediction rate on the crime prediction over crime data is observed as 83% [26].

Keyvanpour [1], [3] according to the range of crime analysis suggested a framework for detection and investigation. According to the proposed methods, the crime variable and matching are the two pillars. The crime spatial-Temporal crime variables, crime natural specification and offender profiles are taken into consideration for crime variables are irrespective of the crime and the previously mentioned variables. Starting with the entity extracting, clustering and finally shows the matching and they are the component in crime matching holds. Starting with the feature extraction turns the entity extracting starts, they create the crime dictionary along with lexical lookup engine, the accuracy is guaranteed. To the selected entities where the features map is, retrieved using the self-organizing neural network and k-Means is applied to the same, here the SOM natural network clustering is performed [1].

Chapter – 3 | Requirement Analysis

3. Requirement Analysis

3.1. Introduction

The analysis of the requirement gathered for the proposed solutions along with the different techniques used for the solutions are discussed in this chapter.

Before implementing a software project, the main priority is to check whether it meets the needs and conditions of the end-user. The requirement must be correctly defined and the project background must be analyzed in order to identify the needs of the user. The fundamental requirement and the relevant research done in the crime clustering can be identified in this chapter.

The literature review which was carried out on existing solutions and systems was used as the main source for gathering requirements for crime clustering.

Both the functional and non-functional requirements were gathered in an above-mentioned manner. These requirements were gathered, then they were confirmed and proven further with the help of an online survey which was provided with multiple sets of questions. This was mainly targeted at gathering non-functional requirements. These gathered requirements were recorded into functional and non-functional requirements based upon their priority considering the available resources.

The personal experiences of the author and his knowledge played a pivotal role in the requirement gathering process.

Further, the author includes the use case diagram for the proposed system. This will help end-users to understand the system properly.

3.2. Requirement Gathering Techniques

The requirement gathering is an essential part of the software developing process which is also mentioned by the author. In order to succeed in the project, the author has to get a clear idea of what a project will deliver. Even though this is a commonly taught, many times people fail to give more attention to it.

Various requirements from stakeholders are brought in using the requirement gathering techniques. The author has identified various requirement gathering techniques to gather requirement from stakeholders. The below mentioned are the recommended techniques.

1. Interviewing
2. Brainstorming
3. Documenting use cases
4. Prototyping
5. Analyzing documents
6. [Bonus] Business Process Modeling
7. Questionnaires
8. Following people around

3.3. Brainstorming

To generate new effective ideas brainstorming is very essential. Brainstorming has two effective phases. Creating ideas and validating ideas. Unfortunately, brainstorming has some drawbacks and limitation to the author's project. Those are given below

- Entering the criminal record in paper
- Maintaining the records in excel sheet
- Identifying the similar records manually

3.4. Document Analysis

One of the important gathering technique is document analysis. When performing an execution document AS-IS or when analyzing the current system's documentation, the interval analysis for the purpose of immigration plans can be evaluated.

In the present condition, chunks of information are mostly buried in documents that guide us in placing question as a part of the validation of the requirement completeness. As a result of the invalid source specified the author did not select this method to use and gather requirement phase and document analysis. Additionally, the below following features of this method can't be in cooperating with this project.

- It is not suitable to evaluate user opinions, needs or satisfaction with services.
- Some documents may be sensitive and not publicly available.

3.5. Document Use Cases

The use cases are newer and the most agile format for capturing software requirements. They are generally contrasted to large monolithic documents which attempt but fail completely in conveying all the possible requirements before construction of a new system.

There are some mismatched features with the proposed systems,

- Use cases are not well suited to easily capturing non-functional requirements of a system.
- Use cases templates do not automatically ensure clarity. Clarity depends on the skill of the writer(s).
- Use case developers often find it difficult to determine the level of user interface (UI) dependency to incorporate in the use case. While use case theory suggests that UI not be reflected in use cases, many find it awkward to abstract out this aspect of design as it makes the use cases difficult to visualize.

The author was able to identify the most appropriate methods to carry out requirements with the help of this analysis. These requirements are an interview, questionnaire survey, observation and literature survey methods. Interviewing the stakeholder is the most common technique used in requirement gathering and analysis. These interviews will provide us details not previously envisaged and been within the mentioned scope of the project and sometimes the requirement may be contradictory. Generally, the stakeholders will carry on with their expectation or will have visualized the requirements.

The requirements were gathered from a variety of stakeholders in this system. The varieties of stakeholders include policemen, criminal lawyers, jailors, crime branch officers and the public. The crime area and the patterns are understood by the crime branch officers. The key users of the proposed system are the crime branch officers. Therefore, the crime branch officers are valuable resources to gather fundamental and non-functional requirements.

Normally the policemen use the report which was given by the crime branch officers, as a result, in this case, they will be only able to understand the drawback of the current process. The criminal lawyers work for crime investigations for their clients in court, as a result of this, they have more workload and in order to reduce the workload they are expecting this kind of tools. Therefore, the author is coming to an assumption that they are interested in this system, so based upon this the author has gathered requirement from theses category people.

3.6. Interview

The author has interviewed policemen and the crime branch officers who are familiar with the crimes, this has been an interview phase conducted by the author. The main scope for conducting the interview is to make sure if the interview provides an opportunity to explore and clarify the topic in more details. Within the time constraint, the author was able to collect minimal data. This time constraint is generally because the people who are working in the police station or the crime branch are busy. As a result of the busy schedule, the crime branch officers and the policemen were reluctant to give their valuable time and sit for a long answering interview.

3.7. Survey Method

To gather requirement from a large number of stakeholders this method is very effective. Electronic and hand copies of questionnaires were shared with the policemen, criminal lawyers, jailors, crime branch officers and public from various regions. The author was able to gather a large amount of requirement from different range of age, various regions and a variety of designations.

An online survey contains three types. One is for policemen, criminal lawyers, jailors and crime branch officers. These have 9 questions on crime types and crime pattern over the mentioned area are associated. The criminal activities and other general questionnaire were targeted to the public who were affected by criminal activities. This questionnaire contains 7 questions and the other one contains 5 questions for system experts who are already developed the system like this. Around 30 despondences have made their feedback on crime types and crime pattern over the area associated. This was carried out over 30 days counting from 15.4.2018 to 15.7.2018. There were seven on criminal activities according to public view and eight on experts view. This survey provides insights into the mind of an experienced set of researches, so this survey is very valuable. Generally, the view on criminal activity is different from the public to crime branch officers or policemen. The crime scene in public perception doesn't give insight idea about the criminal who did the crime and what is behind that. The crime branch officers and the policemen have to think deeply on the crime and identify the criminal activity properly and find the real reason behind it. If there is any crime related to it, they have to investigate in all the ways which are linking to other crimes. Many new aspects have been revealed which were not explored before with the help of this survey.

3.7.1 Observations

This is also one of the valuable and helpful methods that make it easy to understand the current process and easily identify the drawbacks in the existing approaches. Furthermore, this method is very useful in getting an understanding of the context more accurately, this also provides cross-checking information differences between what people do and what they say. It also helps to access the quality of the relationship. The criminal activities and indirect crime matching process in the crime branch office were observed by the author in this case.

3.7.2 Online Survey

There are three types of question in this survey.

- End-user related – This survey helps to gather information and requirements from of crimes from the end user's view. Please find the link below and the questions are mentioned in Table 3.1.
  https://docs.google.com/forms/d/e/1FAIpQLScX_SONPpHEPmhTANBoHLJxFlhdQBowIX2kU21hGeW8YEi4dg/viewform
- Tools and technology related – This particular survey guides us to gather what are the tools that can be applied in this system and what are the features that can be applied in this system and what are the technologies that can be applied in this system. Please find the link below and the questions are mentioned in Table 3.2.
  https://docs.google.com/forms/d/e/1FAIpQLSdcB1pHi7gIUj3EzRCEpeEPx1M0plJBzI1JRB_btS5WNYhNdg/viewform
- Public related - This survey helps to gather information and requirements of crimes from the public view. Please find the link below and the questions are mentioned in Table 3.3.
  https://docs.google.com/forms/d/e/1FAIpQLSdoMrxSPMVzjIOwVdeYLh_tAJn8nMtrrBbXqNHA36-2n02F2g/viewform

Table 3.1 - End-user survey questions

No	Questions	Type
1	Are you using a fully automated system that can crime match or to identify crime pattern over the area?	Yes/No
2	Please select which is the easiest method, crime matching or crime pattern identification?	Multiple choices
3	Which is the harmless way crime matching or crime pattern identification over the area?	Multiple choices
4	Which method is more accurate crime matching or crime pattern identification over the area?	Multiple choices
5	Approximately how many crime entries are registered per day?	Text
6	Approximately how many policemen are there in each police station?	Text
7	Do you think this type of fully automated system will reduce the time required for crime matching or crime pattern identification over the area?	Text

8	What are the features (realistic) that you would want in crime matching or crime pattern identification over the area application if you were using it?	Text
9	Are these features helpful in the crime analysis?	Yes/No

Table 3.2 - System expert survey questions

No	Questions	Type
1	What is the best development environment in order to develop this system?	Multiple choices and text embedded
2	What is the best developing methodology that is applicable to this project?	Multiple choices and text embedded
3	Is there any new technology to achieve this task?	Multiple choices and text embedded
4	Looking at the requirement of the proposed system, do you think it is necessary to provide a solution which needs less engineering efforts than the systems that are available at the moment?	Multiple choices and text embedded
5	What are the improvements you would like to see in future versions?	Multiple choices and text embedded
6	What is the best developing methodology that is applicable to this project?	Multiple choices and text embedded
7	What are your overall comments on the project? Please include its academic impact, novelty, practical applicability, research value, etc. when forming your answer?	Multiple choices and text embedded

Table 3.3 - Public survey questions

No	Questions	Type
1	What are the criminal activities you have faced in your life?	Multiple choices and text embedded
2	Does crime happen with the one person or group?	Yes / No
3	Do you lose your property due to any criminal activity?	Yes / No
4	Have you heard of any similar crime in any other area?	Yes / No
5	Do you think that this crime is related to the crime which you faced?	Multiple choices and text embedded

3.7.3 Limitation of the Online Survey

Wide area and variety of researchers were covered in this survey. There are a number of 24 participants who may not be representing the whole research community of this area.

This online survey was sent to policemen, criminal lawyers, jailors, crime branch officers and the public who were initially affected by these criminal activities, but it may not be represented of the whole policemen, criminal lawyers, jailors, crime branch officers and the public. The policemen, criminal lawyers, jailors, crime branch officers and the public who were affected by the criminal activities were not in a position to answer the questionnaire because they were inexperienced and not exposed to the real industry.

Question regarding motivation and suggestions factor of this project were not focused well in the survey. Most of the participants were not able to answer the questions, the feedback provided by the domain experts in data mining area helps to equalize this gap.

3.8. Requirement Analysis

The necessary investigations were carried out with one of the crime branch officers of Colombo district, by an interview. According to the investigation, it was stated that the proposed system will be a helpful tool to solve the current difficulty in crime pattern identification. In addition to this, the end-user has suggested developing an automotive system to analyze a bunch of crimes. Because the crime pattern identification is very vital. The result generally depends on the crime details over the area during the time, therefore it could be automated.

Using the information which was gained from the witness, the crime branch officers and policemen are trying to analyze and identify the criminals. This kind of analysis is vulnerable to errors. This is because the truth of the information which is provided by the witness cannot be trusted. With a computer-aided crime, the analyzing system can be proposed to eliminate those errors and this will help them with their decision.

From the information based on the crime or with the past experience, the crime branch officers and policemen find out the appropriate crime which is similar to the other crime. This

is the exact reason for making it an automated one. Therefore, it is recommended to have an automated analyzing system for crime.

3.9. Functional Requirements

Currently, crimes records maintain in excel sheets. So the system should read these records by-line by line. Because based on these values crime pattern will be analyzed by the system. It can be excel file or CSV file.

The intended behavior of the system can be captured by the features of functional requirements and the resulted behavior can be expressed as services, tasks or function. The relevant system is required to perform five Functional requirements.

FR1: System should read the crime data from the excel file.

Currently, crimes records are maintained in excel sheets. So these records are read line by line by the system. The system will analyze the crime pattern based on this system. It can be excel file or CSV file.

FR2: The crime clusters must be created and analyzed by the system based on the excel data.

The common crime patterns are analyzed and identified to reduce further occurrences of similar incidence, so a string array should be created by the system based on the crime data and matching the crimes based on the array.

FR3: Based on the geographical image the system should plot and also plot according to the cluster.

Alongside the geo-spatial plot, the proposed system should be utilized. Based on certain crimes in order to show the outcome graphically, the crime analyst may pick a period range and the different type of crimes. This will help to forecast the futuristic crime in a proper visualized way.

FR4: The system should be able to do the cluster analysis using the Weka tool:

The clusters can be analyzed by the system. The can ideally be within the cluster or across it. This can be done using Weka tool. Then finally, the analysis result is displayed in the UI by the system.

FR5: The system should be able to configure the cluster number using user input:

The end-user must be allowed to select the number of clusters. The system needs to create clusters using crime matching based on the user input systems.

3.10 Non - Functional Requirement

The overall qualities or attributes of the resulting system can be defined by the non-functional requirements. The below-following properties and constraints are identified as non-functional requirements for the project, which are then categorized according to project requirements and the organizational requirements.

NFR1: Usability: Often, one of the toughest goals of software development is to create user-friendly software settings. Human-Computer interaction method should be followed. The user interface of the relevant system should be developed according to the standards which are in line with the user familiar icons and which confirm according to the reasonably usable matrix.

NFR 2: Accuracy: System should be able to provide an accurate result of the crime identification. The creation of crime clustering from the overall crime information details depends on the amount of crime information details.

NFR 3: Efficiency and Reliability: Once the analysis is completed, the system should be able to provide the results. That should be efficiency and reliability. The report should be generated by the system providing the analysis results which are based on the crime clustering.

3.11 Constraints and Dependencies

- Unrealistic time schedule: The time schedule of the project is only 8 months due to this the author have to go through all the stages from the initial stage. Suppose if there any more changes then it would be difficult to modify the changes within the limited time frame specified.
- Lack of user interaction: The system users are crime branch officers and policemen but they are extremely busy. This makes them have no time to answer for the interview or fill the questionnaire. So it is difficult to get them involved when developing the project. Therefore, the system will eventually miss out certain user requirements and they will not fulfill their requirements. They are mostly non-functional requirements.
- Lack of data set: Based on the interviews and survey data was collected. This made it not possible to collect data from all the areas of the city. As a result of this, the most critical crimes or political crimes were not collected from the crime branch due to government rules.
- Frequently change request: The release time of the project will be affected if the user changes the requirement frequently.
- Lack of quality of data: normally in the crime police stations, records are entered manually. So when the author computerized the data, there are lots of missing values in the data set. Also in the crime branch, they do not give any criminal's details.
- Lack of quality of data: The records entered manually in the crime police stations. When the data is computerized, there are lots of missing values in the data set. So in the crime branch, missing criminal details are not revealed.
- Further, in this scenario, the author has significantly used the Spatial-Temporal Analysis of Crime (STAC) algorithm to visualize the crime data in the map and this has been the limitation here. Images were used by the author to visualize the crime data. The satanic algorithm which is STAC was well used here. Therefore, real-time crimes were not captured by the author here. Apart from this, it was evident that the author was able to analyze or predict the crime based on past data.

3.12 Use Case Diagram

Here the author attached the use case diagram of the system based on the requirement gathering process.

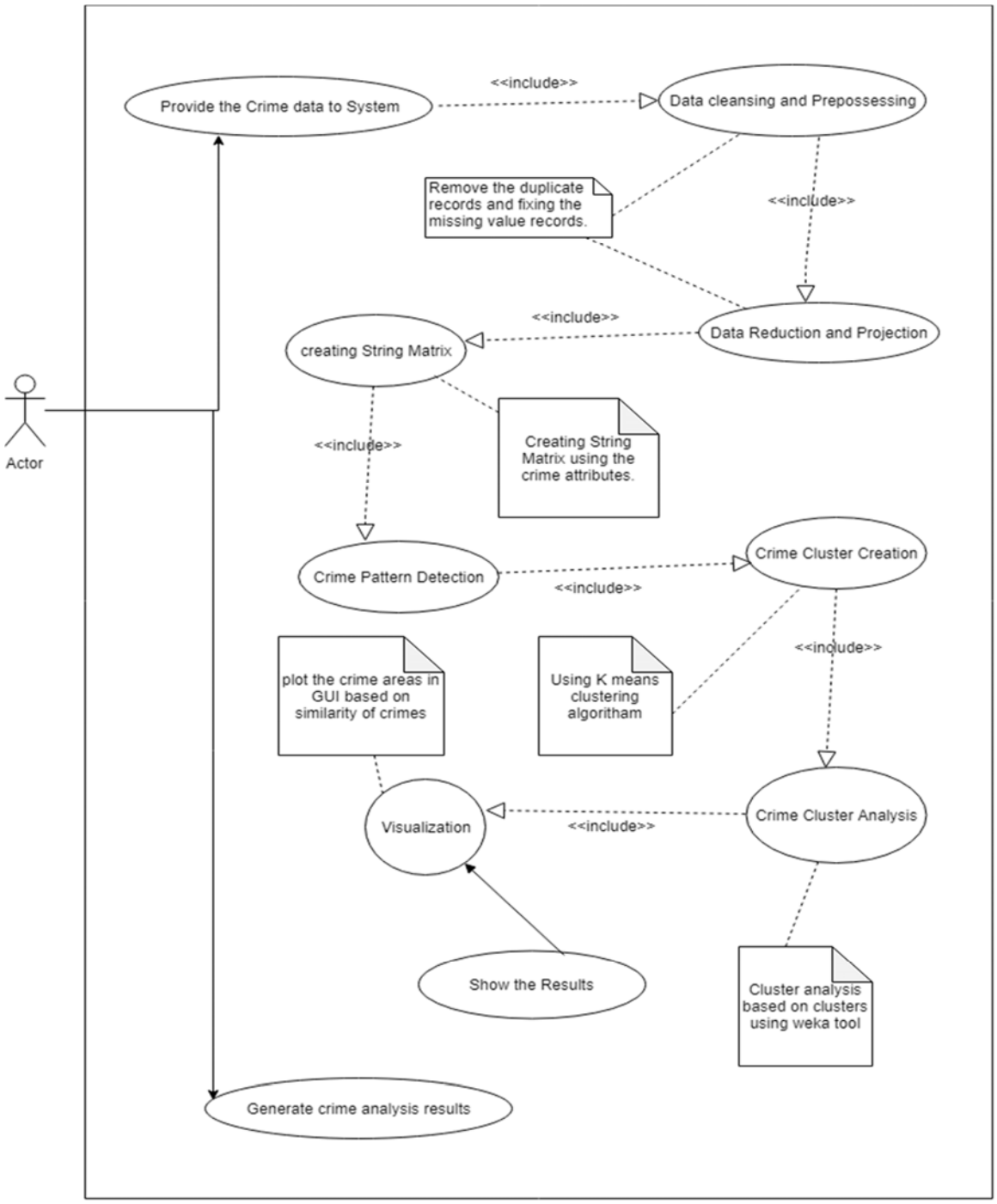


Figure 3.1 - Use case diagram

Chapter – 4 |
Project Plan

4. Project Plan

4.1 Project Management Approaches

The results are rarely predictable during the project research and development process. The author also needs to be able to control the timescale and a proper framework for project management apart from the previous method, this will eventually help the author to achieve the task successfully. The one which holds the change and uncertainty are approached by the project management, both of them are grouped under the term again and of the DSIM (Dynamic Systems Development Method). This is most widely known and used. These are the appropriate approaches when it comes to project management.

4.2 Project Planning Tools and Techniques

The author has used the Microsoft Project tool in this proposed project approach, this is mainly followed in order to plan all the tasks that are supposed to be completed as part of a project. The author was able to schedule and plan the task efficiently with the help of this tool. The author was able to monitor the achievement of project goals and to observe where remedial action needs to be taken to get a project back on course, this was mainly done during the project management phase.

This initiative was well planned and produced at the early and initial stage of the project and this is when during the submission of the project proposal document. Including the main task and subtasks, the period for each task was well estimated for each milestone. The duration of achieving certain tasks was extended due to unavoidable circumstances. Therefore, a revised plan was well prepared in order to figure out and control the tasks and preceded with the project.

4.3 Risk Management

In risk management, the author was certainly able to identify and track down some of the key risk factors during the project work. These risks were handled through effective risk management that helps the system and by adopting alternative actions. The risk management system should be well planned and executed in order to avoid and restrict risks.

4.4 Work Breakdown Chart

Project work breakdown chart is mentioned in below in Figure 4.1.

ID	Task Mode	Task Name	Duration	Start	Finish
1		**Terms Of Reference**	**42 days?**	**Sat 02-06-18**	**Tue 31-07-18**
2		Create Proposal	12 days	Sat 02-06-18	Sat 16-06-18
3		Submission of Draft Proposal	1 day	Mon 18-06-18	Mon 18-06-18
4		Evaluation of Proposal	20 days	Tue 19-06-18	Mon 16-07-18
5		Apply Modification	10 days	Tue 17-07-18	Mon 30-07-18
6		Submission of Proposal	1 day	Tue 31-07-18	Tue 31-07-18
7		**Litrature Review**	**19 days?**	**Wed 20-06-18**	**Mon 16-07-18**
8		Research on Domain	9 days	Wed 20-06-18	Sat 30-06-18
9		Analyzing on existing system	12 days	Sun 01-07-18	Mon 16-07-18
10		**Data Gathering and Analysis**	**8 days?**	**Thu 05-07-18**	**Mon 16-07-18**
11		Data gathering	8 days?	Thu 05-07-18	Mon 16-07-18
12		**System Design**	**22 days?**	**Fri 20-07-18**	**Mon 20-08-18**
13		Prepare High Level Design	9 days	Fri 20-07-18	Wed 01-08-18
14		Prepare Detail Design	13 days	Thu 02-08-18	Mon 20-08-18
15		**Implementation**	**84 days?**	**Tue 21-08-18**	**Sat 15-12-18**
16		System Development and Unit testing	85 days	Tue 21-08-18	Sat 15-12-18
17		**Testing**	**33 days?**	**Thu 15-11-18**	**Mon 31-12-18**
18		System Testing	33 days	Thu 15-11-18	Mon 31-12-18
19		**Documentation**	**130 days?**	**Thu 01-11-18**	**Wed 01-05-19**
20		Preparation of project report	85 days?	Thu 01-11-18	Wed 27-02-19
21		Submission of Draft Report	6 days?	Fri 01-03-19	Fri 08-03-19
22		Apply Modification	38 days	Sun 10-03-19	Tue 30-04-19
23		Submission of Final Report	1 day?	Wed 01-05-19	Wed 01-05-19

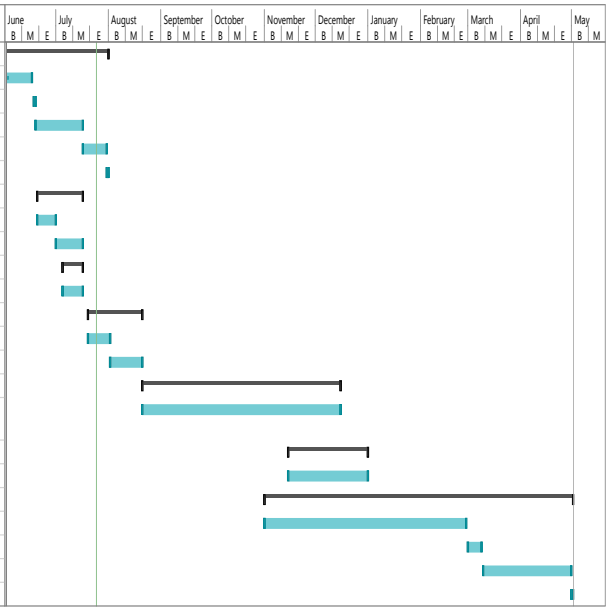


Figure 4.1 - Work breakdown chart

Chapter – 5 |
Design

5. Design

5.1 High-level Diagram

The overview of a product is provided by high-level design, so the innovation of high-level design is very important. The relevant components, interfaces and the servers which need to be developed are relatively depicted by a high-level architecture. High-level details are added to the intended project and they represent a suitable model of coding, this is done with the purpose of high-level design. The requirements analysis phase is a vital stage because this helps to identify the key functional requirements. The diagram represents all the identified functional requirements.

It is known facts that before coming to a stage author have to find out the necessary changes and analyze it, so as discussed in the project background chapter, many relevant design changes were made before coming into the stage. There are additional necessary details included in the intended project and represent a suitable model for coding before the high-level design is designed to provide an overview of the product. In the previous chapter, the author has identified the functional and non-functional requirements. The overall view of the proposed system consists of five major modules and the below picture represents the overall view of the system. By giving the relevant input process and output these modules are further described in this chapter.

In the previous chapter, the author explained that which methods were used for data collection and how data was collected. So the author found some inconsistency in the collected data. So the author needs some appropriate methods for data cleansing before creating the clusters. Because clusters are created based on the dataset and this is the input for mapping geographical location.

Below Figure 5.1 shows the high-level dataflow architecture diagram of the proposed system from the beginning of the data collection.

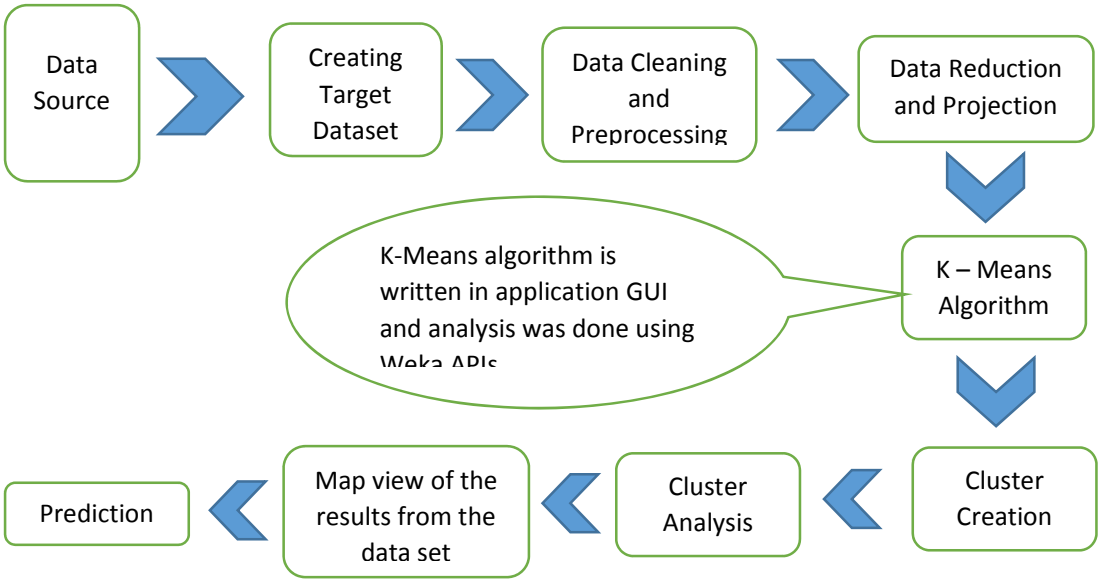


Figure 5.1 - High-level dataflow diagram

5.2 Data Flow Diagram

In this section, the author has drawn a dataflow diagram which shows the flow of the proposed system. The Figures such as Figure 5.2, Figure 5.3 and Figure 5.4 shows respectively the 0^{th}, 1^{st} and 2^{nd} level of the data flow between the proposed modules, these modules later implement in the implement chapter 6.

According to the problem domain, the author has collected the crime information in Colombo and Kandy districts to achieve the objectives. The following below five attributes in the data set, are heavily used for this research.

- Crime name
- Crime location
- Crime date and time
- Crime detail description
- Criminal activity

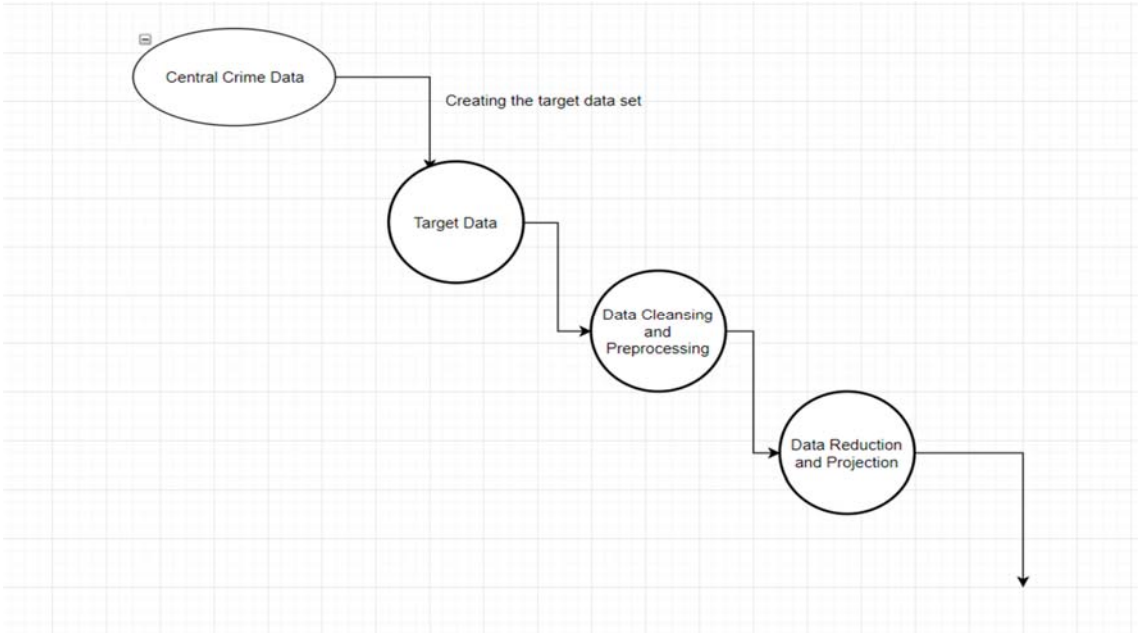


Figure 5.2 - 0th level dataflow diagram

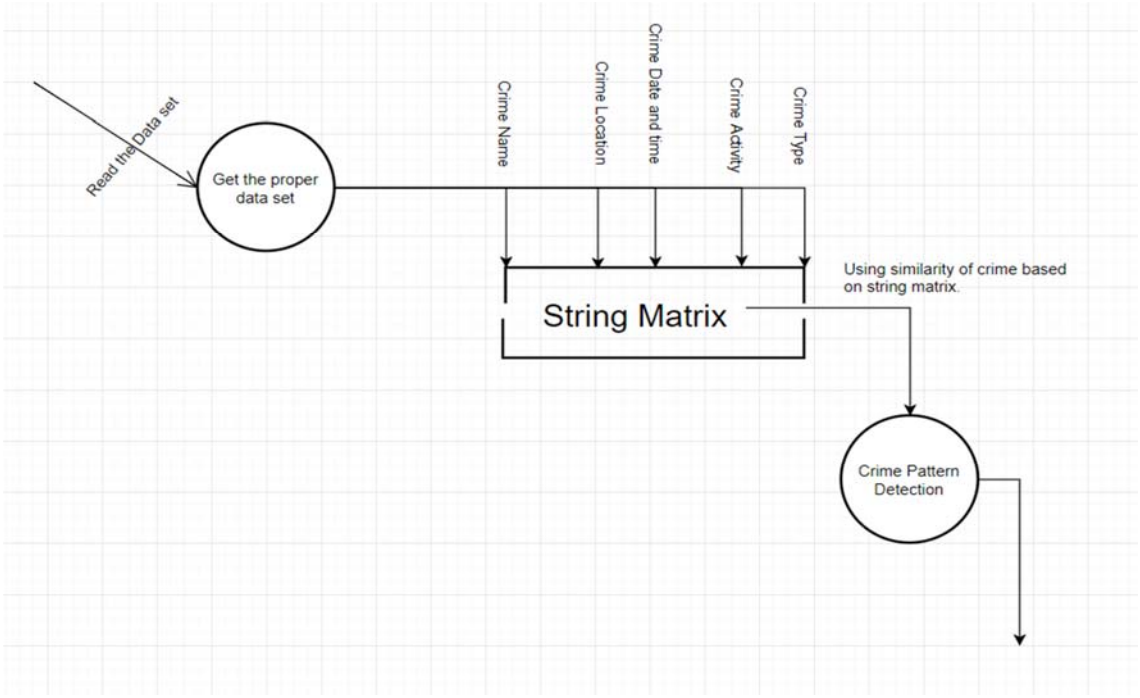


Figure 5.3 - 1st level dataflow diagram

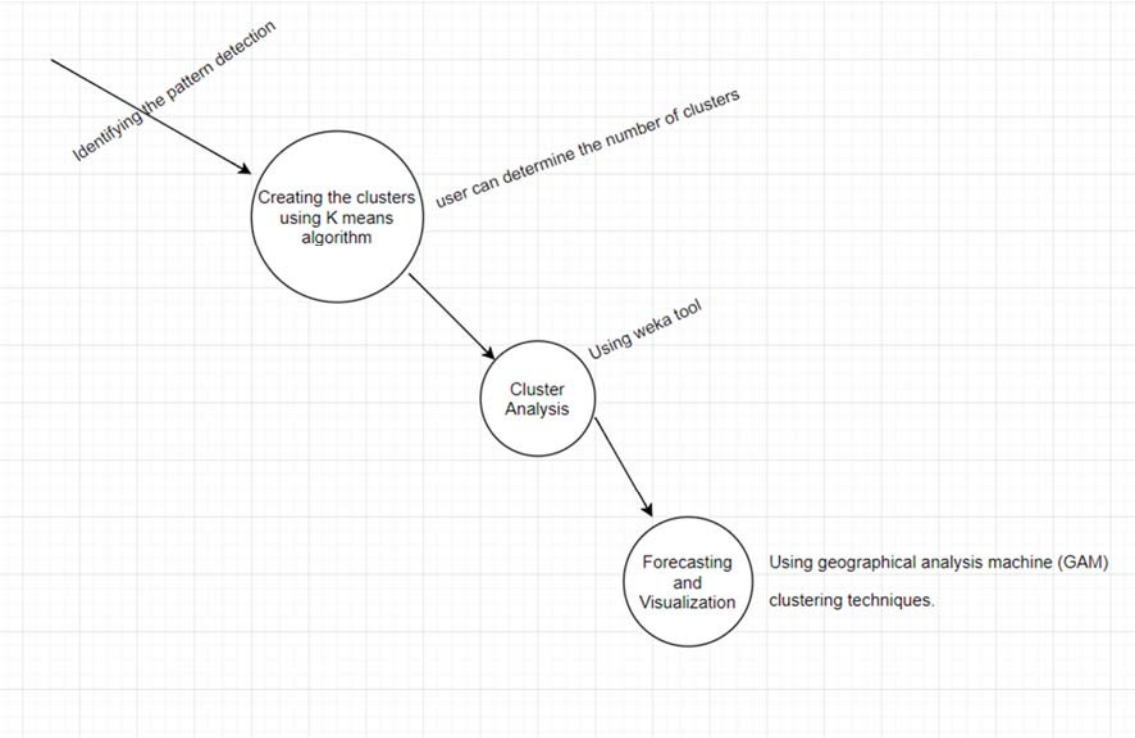


Figure 5.4 - 2nd level dataflow diagram

Further, in below chapter author discuss the implementation details of these data flow diagrams.

Chapter – 6 | Methodology & Implementation

6. Methodology & Implementation

6.1. Project Development Approach

A background study of the cluster creation using the crime data set is done here as the first approach of this project and it is used for further process. It is important to pay attention when creating the crime clustering, this is because the similarity of the crime is very important and it should be noticed well. This is mainly done to analyze further in order to recognize their similarity and differences. Data cleansing and data preprocessing are two important presses done using the Weka tool.

It is recommended to study the Weka tool and apply for data cleansing and data preprocessing, this is carried out after getting the values from the time data set. Relevant steps are taken for accuracy testing, troubleshooting or regular fine-tuning of the codes. These are very important steps required for testing the result. The missing values and the outliers from the dataset can be found out using this.

In the next stage, the corrected datasets are transferred or passed to the Graphical User Interface (GUI). After this step, the system then creates the string array which is based on the relevant values from the obtained dataset such as crime type, Location (latitude, longitude), criminal activity, etc. Then this array creates the similarity matrix of the particular dataset and then it creates clusters based on the similarity matrix. Based on the user input the number of clusters is determined.

Finally, a Graphical User Interface needed to access the created clusters, to analyze the across and within the clusters using Weka APIs and other findings and results would be presented in the GUI. Relevant technology and algorithms are clearly discussed below.

Here further author discussed the high-level architecture diagram with the main functionalities. This is helpful to understand the functionalities of the proposed system. The further author explains the detail of every process below.

Here the author attached the sequence diagram for further understanding in Figure 6.1.

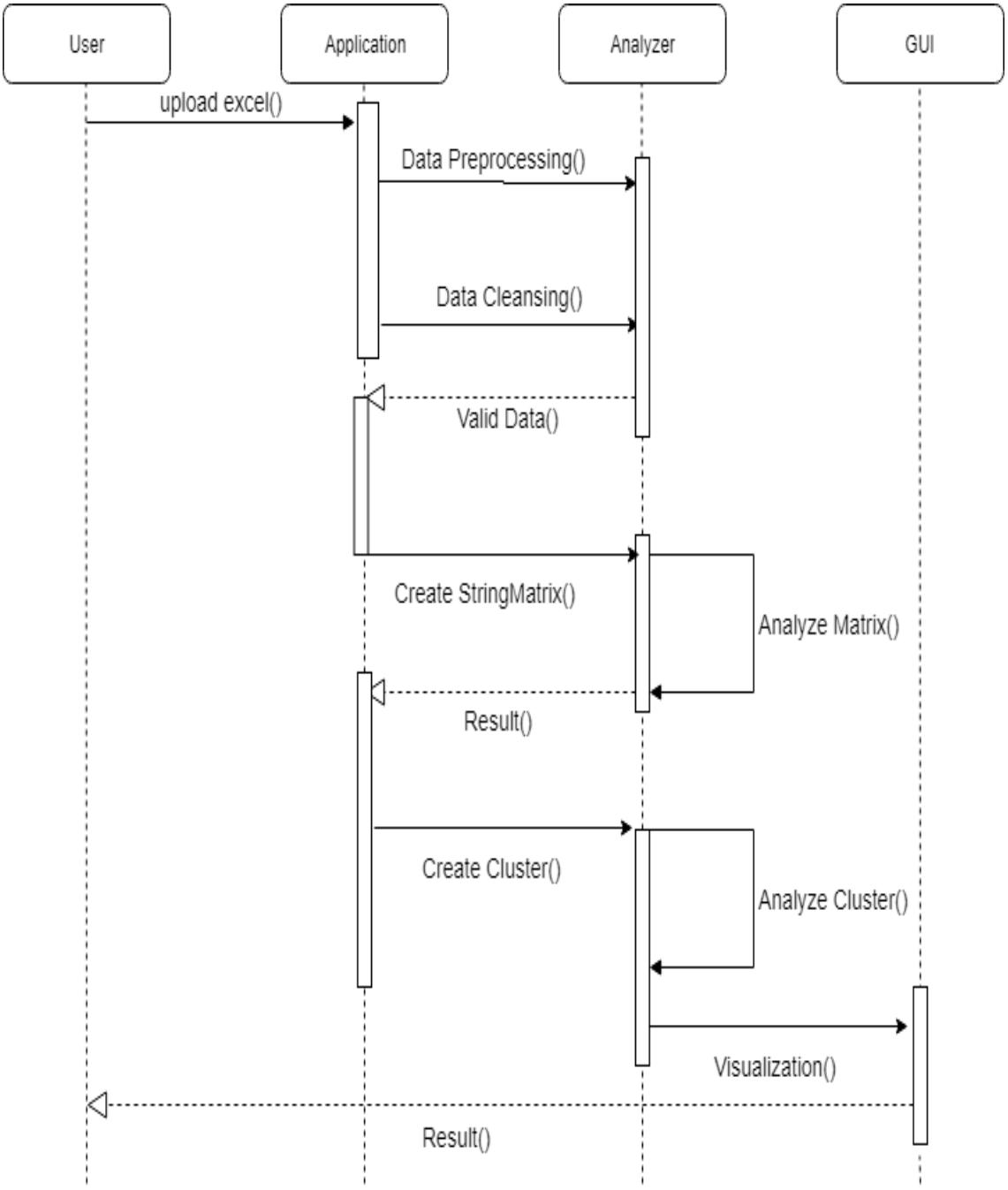


Figure 6.1 - Sequence diagram

6.2. Project Development Approach

6.2.1 Clustering

Clusters are subsets of data which are similar. The process of dividing a dataset into groups is called clustering, also called unsupervised learning. The groups are sometimes members of each group those who are similar to possible to each other and sometimes different groups who are dissimilar as possible from each other. Previously undetected relationships in a dataset can be uncovered by clustering process. There are many applications for cluster analysis. For example, in business to discover and characterize customer segments for marketing purposes and in biology, cluster analysis can be implemented. It can also help in the classification of plants and animals when their features are provided. [4]

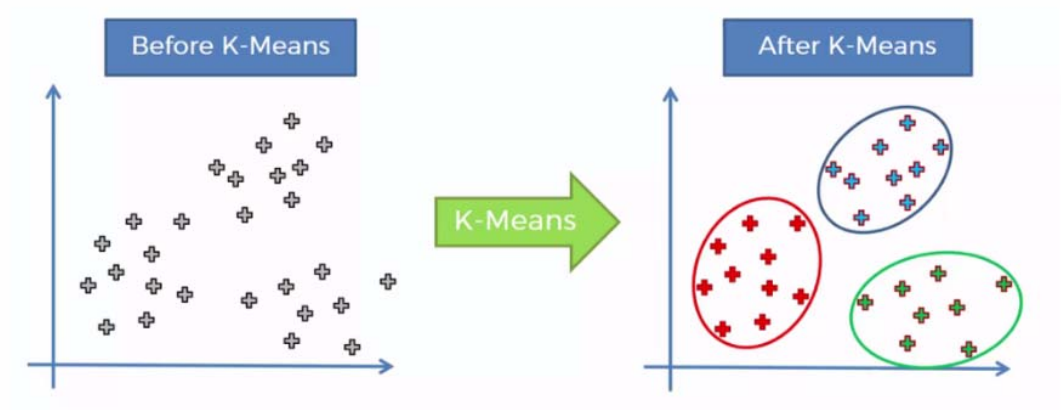
Two main groups of clustering algorithms are:

1. Hierarchical
    - Agglomerative
    - Divisive
2. Partitive
    - K-Means
    - Self-Organizing Map

6.2.2 K-Means Clustering

The most well-known clustering algorithm is K-Means. In many introductory data science and machine learning classes, this is taught well. This is an easy task to understand and implement in a program.

Further in below Figure6.2 explains that how data points are arranged using K-Means algorithm.



Source: K Means Clustering: Identifying F.R.I.E.N.D.S in the World of Strangers, (2018) [4]

Figure 6.2 - K-Means clustering

1. Initially, the number of classes or groups is selected to use and the respective center points are randomly initialized. It is good to take a quick look at the data to figure out the number of classes and then try to identify any distinct groupings. The vectors of the same length are the center points, these presents each data point vector and are the “X’s” in the graphic above.
2. Normally the distance between the specific point and each group center is classified by each data point and then classifying the point to be in the group whose center is closest to it.
3. The group center is recomputed but taking the mean of all the vectors in the group based on these classified points.
4. For a set number of iterations or until the group centers don't change much between interactions these steps are repeated. In initialize, the group centers for few times one can randomly opt for it and then select the particular run that looks like it provided the best result.

K-Means has many advantages, one of the main is, it is pretty fast and generally, the author is computing the distances between points and group centers. Eventually, it has a linear complexity of 0(n).

K-Means also has a couple of disadvantages. Initially, you have to select how many groups or classes are existing. This is generally not trivial and also it is ideal with a clustering algorithm which the author would prefer to figure those out for us because the main intention is to gain

some insight ideas from the data. Generally, K-Means starts with the random choice of cluster centers and therefore they may yield different results from clustering on the basis of different runs of the algorithm. As a result of this, the results may be not repeatable and would lack consistency. The remaining cluster methods are more consistent.

K-Medians is another clustering algorithm related to K-Means, here instead of re-computing the groups center points using for the mean, the author uses the median vector of the group. This type of methods is less sensitive to outliers, apart from this it is much slower for larger datasets as sorting is required on each iteration when computing the relevant median vector.

6.3 System Model

According to the prescribed problem domain, the author has relevantly collected the crime information from Colombo and the Kandy districts to achieve his objectives. In order for this research the following below five attributes in the data set, are heavily used and handled.

- Crime name
- Crime location
- Crime date and time
- Crime detail description
- Criminal activity

The above five attributes relatively help to categorize accordingly or guides to identify the various crime clusters by using pattern detection techniques. Text mining and the similarity of crimes based on the above attributes are used for pattern detection techniques.

This particularly detailed analysis generally talks about hotspots and the appropriate use of the K-Means clustering for crime pattern cluster creation. In this, it first identifies all the significant attributes in the provided dataset. This is which is unlike other papers, it adds weight to the mentioned attributes in the data set. Here the most important attribute ‘criminal activity’ is absolutely given the highest priority (weight) in comparison to other attributes. There is a usage of this feature of the research paper in this project. In this, the

crime analyst the author would select the number of clusters what he wants. Crime clusters are created according to the selection.

Hence, finally using the geographical analysis machine (GAM) the spatial-temporal analysis of crime is taken place. Crime clusters are created based on the above attributes using K-Means clustering with pattern detection techniques. But it has to be noted that these crimes occurred in various locations and times. So Clusters need to be analyzed with the help of GAM clustering techniques within and across the clusters and find the diversity of the particular criminal activity and forecasting suspected crime location for the crime clusters. The appropriate records are relatively grouped in light of the foreordained qualities and the weights. It has to be noted that the subsequent, bunches have the conceivable crime related designs. Then these subsequent bunches are basically plotted on the geo-spatial plot.

Algorithms used by the author for this solution

- Crime pattern detection using text mining and similarity of crime based on the above attributes.
- The creating crime clusters based on crime patterns using K-Means clustering algorithm.
- The relevant prediction model is based on an analysis of crime clusters - Spatial-Temporal Analysis of Crime (STAC) is using GAM clustering techniques.

The author already collected the past static data and based on this data the system removes the outliers and will generate the clusters using the K-Means algorithm. To visualize this clusters author uses the STAC algorithm.

The visualization is particularly obtained based on past static data. In this scenario, the author has collected the past crime data through the police area and had applied the STAC algorithm in order to visualize the data. Generally, the STAC algorithm is used to visualize the static geographical data apart from the real-time data, as it supports the static geographical data other than the real-time geographical data. When this happens based on the STAC, the author is not able to visualize the real-time happening crimes. Through this, the author is able to predict or visualize what type of crime has happened in which area and also to know the

possibility of when it has to be happening in the future. As a result of this, it has to accept that the outcome is just an alert and to an extent of assumption, it is exactly not 100% accurate.

As discussed earlier STAC is the static algorithm in order to visualize the data, and the author specifically uses area image instead of the real-time map. According to the author's point of view, the image can be used instead of a real-time map. Therefore, the author uses the past criminal records to visualize in the map. As in the image, the latitude and longitude cannot be identified by the author. As a result of this the author has actually assigned the latitude and longitude to the X axis and Y axis, when this happens the coordination points are taken into account and then the data is plotted and visualized according to the X axis and Y axis coordination.

The clustering techniques are used to analyze the clusters within the police area. This is the final step the author has taken using the GAM. The crime cannot be proved that it happens only in a particular police area, this is because the particular crime can happen across multiple police area. Particular police areas can share one or more clusters because the particular police area has multiple crimes. When particular clusters can be shared across police area, the author generally uses GAM and these can analyze the crimes across the clusters and areas. The author concludes the possibility of these types of crimes happened for those areas, based on this specific analysis.

Generally, the pattern of crime rating for the appropriate locations in the dataset for the specified types of crime in the dataset based on the regional basis could be analyzed and are generally found using the proper application such as Crime Analysis System. In this case, when the author considers for both the crime types the application is extensible, therefore in order to increase the records in the dataset for a huge amount of the data. The system provides the facility of choosing two crime types in order to analyze among the dataset, which can be provided by the system. It has to be noted that both the graphical plots and map view for analysis can be provided by the system. As a result of this, the two visualizations are introduced in the application portal of the system.

Figure 6.3 shows the high-level dataflow architecture diagram of the proposed system from the beginning of the data collection.

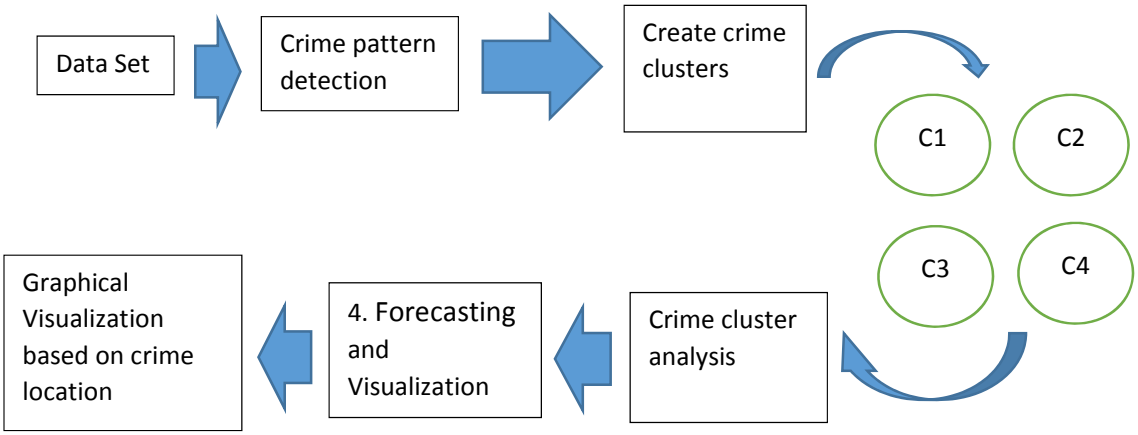


Figure 6.3 - High-level architectural diagram

6.3.1 Creating target dataset

The relevant data sources are collected for the crime analysis in this system from various police data sources. Under regional basis, two data sources are present for the experiments which are from Colombo crime branch and Kandy crime branch. The clustering algorithm requires only the target dataset, generally, the dataset consists of some 6K records. These are which relatively includes crime types and their counts for the location based on the regional basis of Colombo and Kandy. It is to be noted that this target dataset has been created using the entire dataset in the excel file as a CSV file. Also, it has been to acknowledge that the targeted dataset of crime types and its count in the particular location on a regional basis can be done using the excel pivot table.

6.3.2 Data cleaning and Data preprocessing

In this case, the author has to note that the removal of noises or outliers is carried out. After this activity or method, the necessary information is collected to model or account for noise. The missing data fields should be initiated with relevant strategies because the missing data fields are key initiatives for the design model. The accounting for time sequence information and known change should be carried out apart from this. The Data cleaning and Data preprocessing are done using the famous scripting tool called R which. Generally, the outliers are removed as well as the noise data in the dataset, by box plotting data in the set and the

removal of the outliers and noisy data using R. When this results, data preprocessing can be used in Apache Hadoop in order to accommodate the entire dataset. It reduces the space as well as time complexity when this is done through Hadoop.

6.3.3 Data Reduction and Data projection

This is done in order to find and figure out useful features that would represent relevant data by depending on the goal of the tasks. Generally, to reduce the effective number of variables under consideration or to find or figure out relatively relevant invariant representations for the data it is well recommended to use dimensionality reduction or transformation methods.

6.3.4 Data Mining Task for Crime Analysis

The K-Means, AK-mode and expectation- maximization are derived from the primary classification of the partition clustering methods. "K" partition can be constructed from the data belonging to a given dataset of "n" objects by the methods of partitioning. The process of putting similar data into groups is called data clustering. A dataset can be partitioned it to several groups by the clustering algorithm. Here the similarity within a group is larger than among the groups. The most important unsupervised learning technique can be considered as clustering. This generally deals with finding a structure in a collection of unlabeled data. Apart from other clustering techniques, the K-Means algorithm is the one used in this paper.

To partition the clusters based on their means K-Means algorithm is used. Initially, a number of objects are grouped and specified as “K” clusters. Using the mean distance between the objects, the mean value is calculated. The iterative technique which is used to improve the partitions by moving objects from one group to another is the relocation iterative technique. Till the convergence occurs, a number of iteration is done. K-Means algorithm steps are described as:

Input: Input should be Number of clusters.
Step1: Arbitrarily choose k objects from a dataset D of N objects as the initial cluster centers.
Step 2: Reassign each object which distributed to a cluster based on a cluster center which is the most similar or the nearer.
Step 3: Update the cluster means, i.e. calculate the mean value of the object for each cluster.

Output: Output should be a set of k clusters.

K-Means algorithm is a base for all other clustering algorithms to find the mean values.

6.4 System Implementation

In this case, the CSV file which the author has used for the analysis has the relevant columns such as crime type, crime location, gender, etc. after this analysis the array is relevantly created based on the appropriate attributes. According to this, the particular array can be changed or modified based on the created CSV file between the program executions.

Based on the varied police area the CSV file will differ from each other. When the CSV differs, the array created also differs between program executions based on this CSV file.

Generally, with the location of the crime, the author is able to collect the crime dataset. This particular dataset the relevantly includes the specific latitude and longitude of the particular place of crime with the crime location and crime type along with the crime date and time. Therefore, the author is able to process the data set, by following the specific way and then creates the clusters using the K-Means algorithm. After all these executions when the user runs the application, the cluster will be created using this relevant data from the array by using the K - means algorithm.

- Creating the target dataset
- Data Cleaning and preprocessing
- Data Reduction and Projection
- Cluster Creation

The K-Means algorithm works based on the distance between two points. The algorithm determines the neighbor of each point and creates the cluster according to the neighbors based on the distance. The author here explained the way of calculating the neighbors in between the points. Further, the author explained how to link between the neighbors while creating the cluster using the program and pseudo code.

The further author explained the most effective ways to calculate the distance between two points. They are Manhattan distance Euclidean distance Manhattan distance means that the sum of the horizontal and vertical distance between two points

Two most effective ways to calculate the distance between two points were well explained by the author. There are two distances they are Manhattan distance and Euclidean distance.

1. The Manhattan distance specifically means that the sum of the horizontal and vertical distance between two points.
2. The Euclidean distance means the "ordinary" straight line distance between two points in Euclidean space.

Further, the author explains the way of calculating neighbors using the original code of the method in below figure 6.4.

```
public void RemoveOutlier()
{
        for(int i=0;i<n;i++)
        {
                if(clus[i].numterms==1&&clus[i].clusId!=-1)
                {
                        clus[i].clusId=-1;

                        GlobalHeap.numberOfTerms--;
                        NumberOfCluster--;
                        int[] arrxtemp=new int[MAX];
                        int k=0;
                        for(int j=0;j<lheap[i].numberOfTerms;j++)
                        {
                                Goodness tempneighbor2=new Goodness();
                                tempneighbor2=lheap[i].tHeap.data.get(j);
                                arrxtemp[k]=tempneighbor2.neighborClusterID;
                                k++;
                        }
                        for(int l=0;l<k;l++)
                        {
                                int check;
                                check=lheap[arrxtemp[l]].removeNodewithClusId(i);
                                if(check==1)
                                        lheap[arrxtemp[l]].numberOfTerms--;
                        }
                        lheap[i].clusterID=-1;
                }
        }

}
```

Figure 6.4 - Calculating neighbors using the method of the original code

Further, the author explains the way of calculating neighbors using the pseudo code of the method.

```
FOR(start from i to n)
        IF(clus[i].numterms == 1 && clus[i].clusId != -1)
                assign clus[i].clusId to -1
                decrease GlobalHeap.numberOfTerms by 1
                decrease NumberOfCluster by 1
                create an int array "arrxtemp" with the "MAX" size
                Initialize k to 0
                FOR(start from j to each lheap's numberOfTerms)
                        create instance of Goodness
                        assign value from lheap[i].tHeap.data.get(j)
                        add element to arrxtemp array
                        increase k by 1
                END FOR
                FOR(start from l to k)
                        initialize check with lheap[arrxtemp[l]].removeNodewithClusId(i)
                        IF(check == 1)
                                increment numberOfTerms for each lheap's each
                        END IF
                END FOR
                        assign lheap[i].clusterID to -1
        END IF
END FOR
```

The method of cluster creation was explained by the author. The outliers need to be removed before the cluster creation, only after this, the cluster is accurate. The outliers are removed using the system and later the author has explained the code and the pseudo code in order to remove the outlier.

In the data mining area, cluster analysis and outlier detection are strongly coupled. By certain few outliers, the cluster structure can be easily destroyed, as a result of this, the outliers are defined by the concept of the cluster then again they are recognized as the points belonging to none of the clusters. Further, the author explains the way of removing outliers using the original code of the method in below Figure 6.5.

```
public void CalculateNeighbors()
{
        int tunion,tinter;
        int[] noterms=new int[n];
        for(int i=0;i<n;i++)
        {
            for(int j=0;j<n;j++)
            {
                tunion=tinter=0;

                for(int k=0;k<col;k++)
                {
                    if(temp[i][k].equals(temp[j][k])==true)
                    {
                        tunion=tunion+1;
                        tinter=tinter+1;
                    }
                    else
                    {
                        tunion=tunion+2;
                    }
                }
                double sim=(double)tinter/(double)tunion;
                if(sim>=theta==true)
                {
                    noterms[i]++;
                    arrneighbr[i][j]=1;
                }
                else
                {
                    arrneighbr[i][j]=0;
                }
            }
        }
}
```

Figure 6.5 - Removing the outliers using the method of the original code

Further, the author explains the way of removing outliers using the pseudo code of the method.

```
FOR(start from i to n)
    IF(clus[i].numterms == 1 && clus[i].clusId != -1)
        assign clus[i].clusId to -1
        decrease GlobalHeap.numberOfTerms by 1
        decrease NumberOfCluster by 1
        create an int array "arrxtemp" with the "MAX" size
        Initialize k to 0
        FOR(start from j to each lheap's numberOfTerms)
            create instance of Goodness
            assign value from lheap[i].tHeap.data.get(j)
            add element to arrxtemp array
            increase k by 1
        END FOR
        FOR(start from l to k)
            initialize check with lheap[arrxtemp[l]].removeNodewithClusId(i)
            IF(check == 1)
                increment numberOfTerms for each lheap's each
            END IF
        END FOR
            assign lheap[i].clusterID to -1
    END IF
END FOR
```

According to this post the cluster creation, the relevant cluster details are stored in the CSV file. Here when the cluster details are stored in the CSV file, the CSV file is then passed to the Weka tool using this application. When this is executed, the occurrence of the analysis takes place in the program using the Weka API. Hence, there is an availability of algorithms in the Weka tool in order to analyze the clusters.

The author can get more information regarding the clusters based on the analysis from these clusters. The latitude and longitude place of the crime can be obtained from each and every data entry within the cluster. Therefore, each point on the map is plotted by the system. As a result of this instance, this is called a visualization of the data with the crime type. Finally, these help to figure out which crime type of crime has happened in which place.

Chapter – 7 | Results & Discussion

7. Results & Discussion

7.1. Cluster Creations

The author relatively loads the map of Colombo and Kandy in the GUI according to the selection of Colombo or Kandy. As the below figure the system is divided into equals square. According to the end user's point of view and requirement, it is supposed to view the number of count for each square in Figure 7.1 and Figure 7.2 respectively Colombo and Kandy. The crime density for each square can be viewed by the end-user.

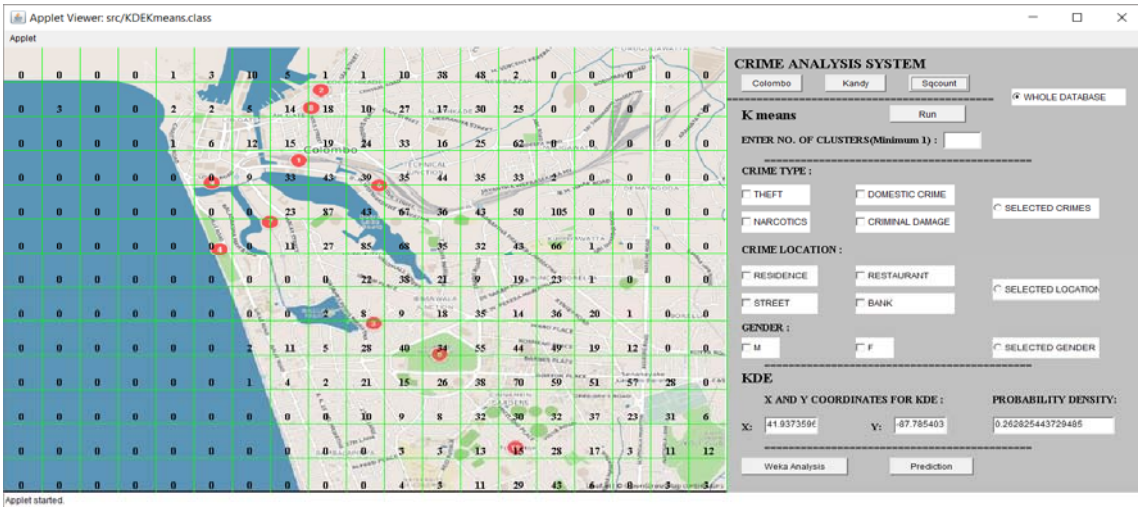


Figure 7.1 - Colombo city map

Figure 7.2 - Kandy city map

As of end-users perspective, a number of clusters need to be configured as their wish. For example, if the end-user decides crimes need to be categorized into five clusters. Finally, the end-user can analyze how many records are fallen in each cluster. These cluster analyses are written in excel file by the system. So the below Figure 7.3 explains the cluster mapping across the whole data in the Colombo police area.

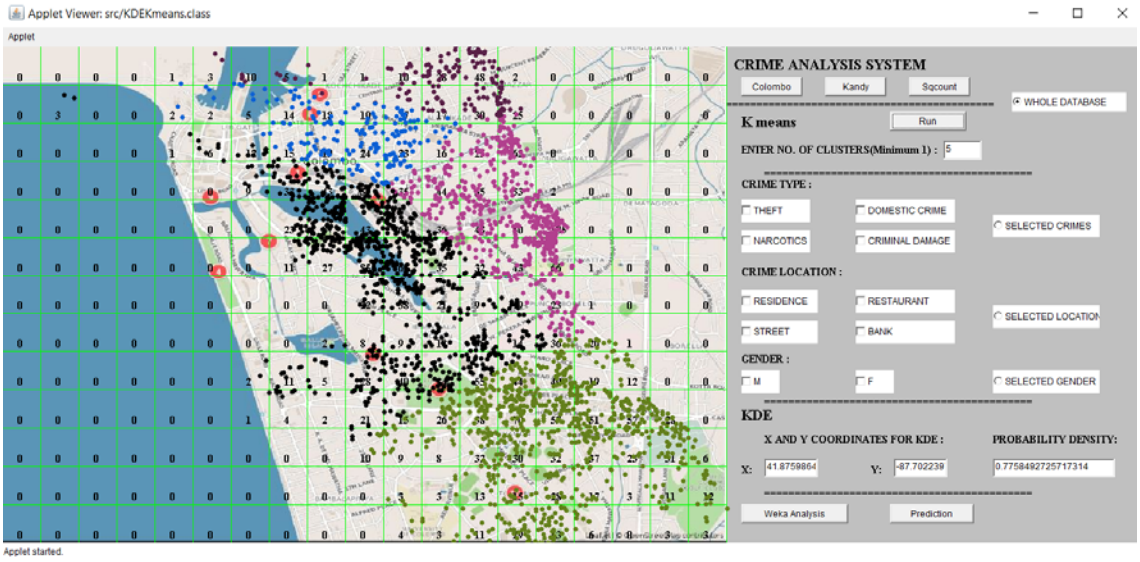


Figure 7.3 - Crime clustering in Colombo city using whole data

Further Table 7.1 is the output from the excel file for the Colombo police area. So end-user can identify the severity of the crime based on count and can identify that location which has more crimes.

Table 7.1 - Cluster count for whole data in Colombo city

Cluster Name	Crime Count
Cluster 1	450
Cluster 2	767
Cluster 3	750
Cluster 4	689
Cluster 5	337

The above detail implementation analysis is done based on the Colombo district.

Based on the crime dataset attributes, in order to create the clusters, there are various ways the end-user can select. According to the standard system, the end-user can create the relevant clusters based on the selected attributes such as crime type, crime location and gender. According to the user selection, the crime clusters are created, as a result of this further number of the crime, clusters can be configured or determined at any time by the end-user. The attribute which is going to help to create the cluster (crime type, crime location and gender) and can configure or determine the number of crime clusters can be selected by the end-user.

Based on types the crimes can be analyzed by the end-user. There are four main crime categories as per requirement gathering and limitation of the dataset. Therefore, the end-user is able to select crime categories as their wish.

1. Theft
2. Domestic Crime
3. Criminal Damage
4. Narcotics

Further below Figure 7.4 output is shown the analysis as per the crime type selection.

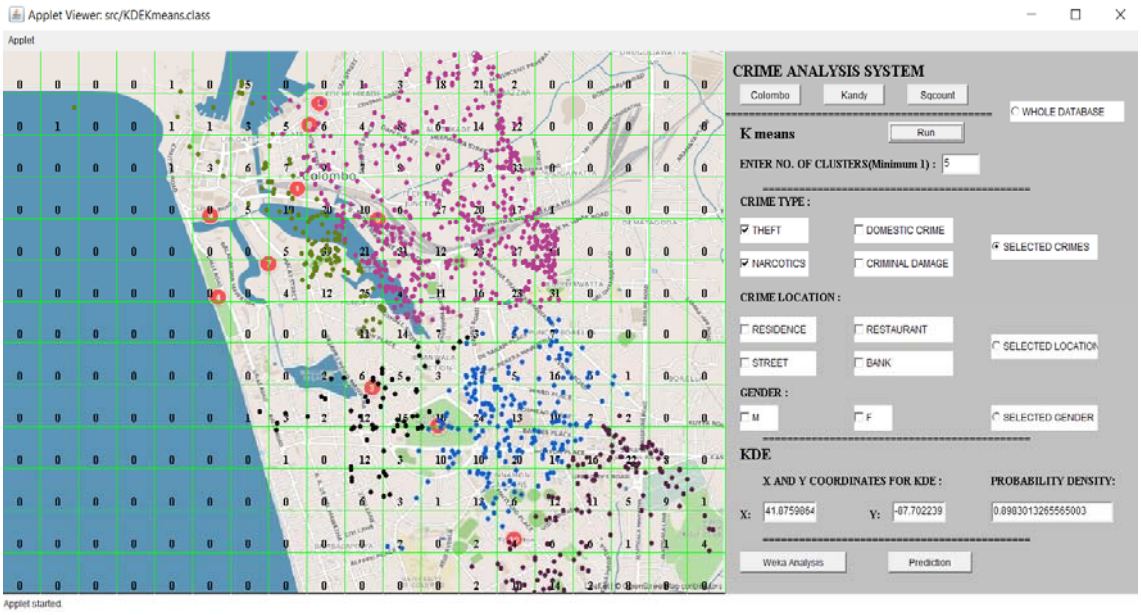


Figure 7.4 - Crime clustering in Colombo city for selected crime types

Further Table 7.2 is the output from the excel file. So end-user can identify the severity of the crime type based on count and can identify the police area which has more crimes.

Table 7.2 - Cluster count for selected crime type in Colombo city

Cluster Name	Crime Count
Cluster 1	468
Cluster 2	309
Cluster 3	173
Cluster 4	87
Cluster 5	131

Further based on locations the crimes can be analyzed by the end-user like crime types. There are four main crime locations as per requirement gathering and limitation of the dataset. Therefore, the end-user is able to select crime locations as their wish.

1. Residence
2. Restaurant
3. Bank
4. Street

Further below Figure 7.5 output is shown the analysis as per the crime location selection.

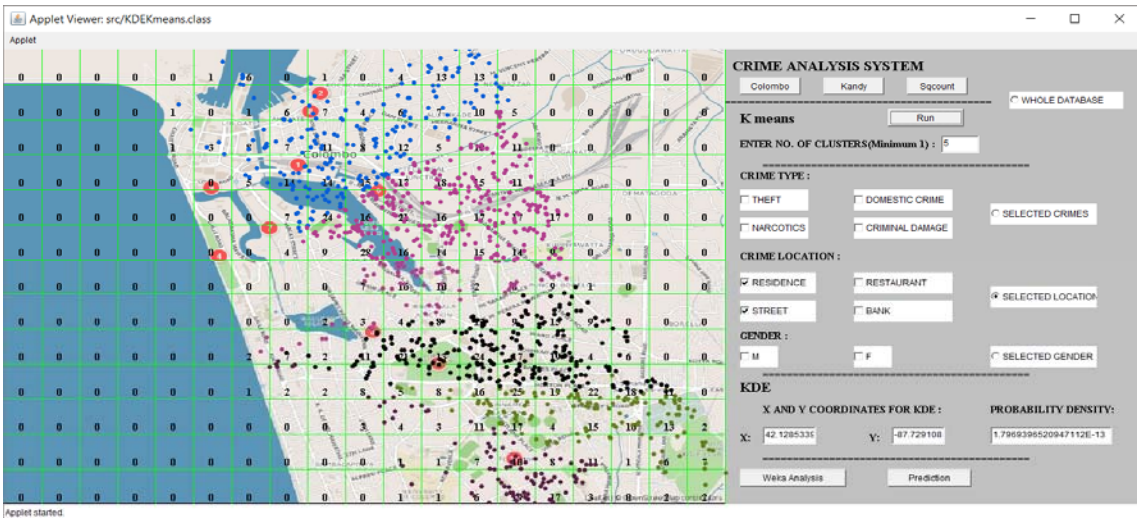


Figure 7.5 - Crime clustering in Colombo city for selected crime location

Further Table 7.3 is the output from the excel file. So end-user can identify the severity of the crime location based on count and can identify the police area which has more crimes.

Table 7.3 - Cluster count for a selected location in Colombo city

Cluster Name	Crime Count
Cluster 1	265
Cluster 2	116
Cluster 3	368
Cluster 4	213
Cluster 5	141

Further based on gender the crimes can be analyzed by the end-user like crime types. There are two major gender types as per requirement. Therefore, the end-user is able to select gender as their wish.

1. Male(M)
2. Female(F)

Further below Figure 7.6 output is shown the analysis as per the gender selection.

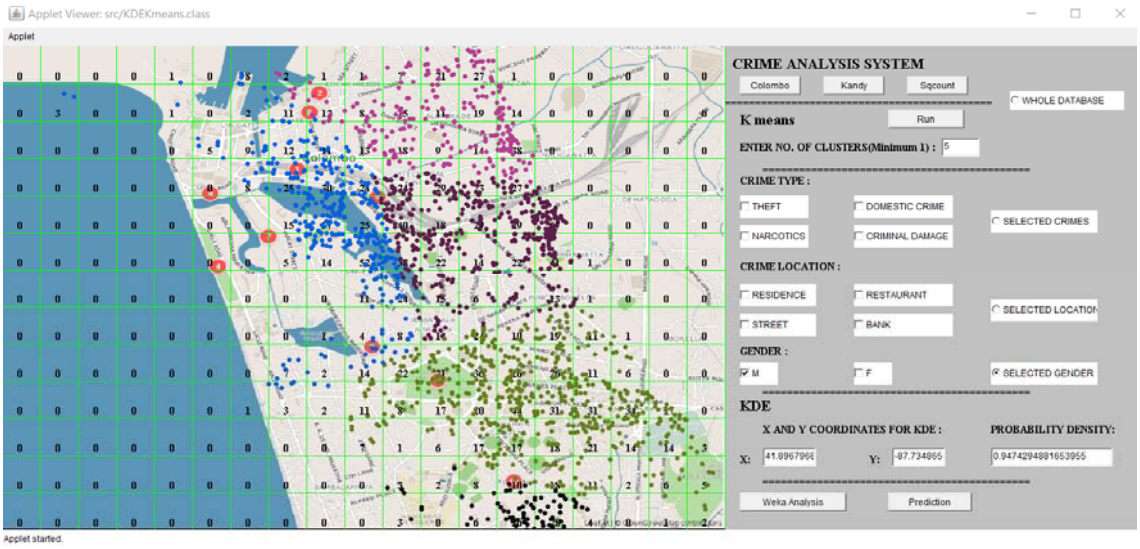


Figure 7.6 - Crime clustering in Colombo city for gender wise

Further table 7.4 is the output from the excel file. So end-user can identify the severity of the crime based on count within the cluster which is created based on gender. It helps to identify the police area which has more crimes.

Table 7.4 - Cluster count for gender-wise in Colombo city

Cluster Name	Crime Count
Cluster 1	102
Cluster 2	458
Cluster 3	266
Cluster 4	390
Cluster 5	595

Therefore, the same functionalities are implemented for Kandy city like Colombo city. This helps the user to select the district and can do the analysis. As a result of time constraint and the lack of data, this system is implemented for two main districts in Sri Lanka. The below outputs are shown for Kandy district. Just like the Colombo district, the end-user can select the crime. The below output is shown as per the whole crimes. So the below Figure 7.7 explains the cluster mapping across the whole data in the Kandy police area.

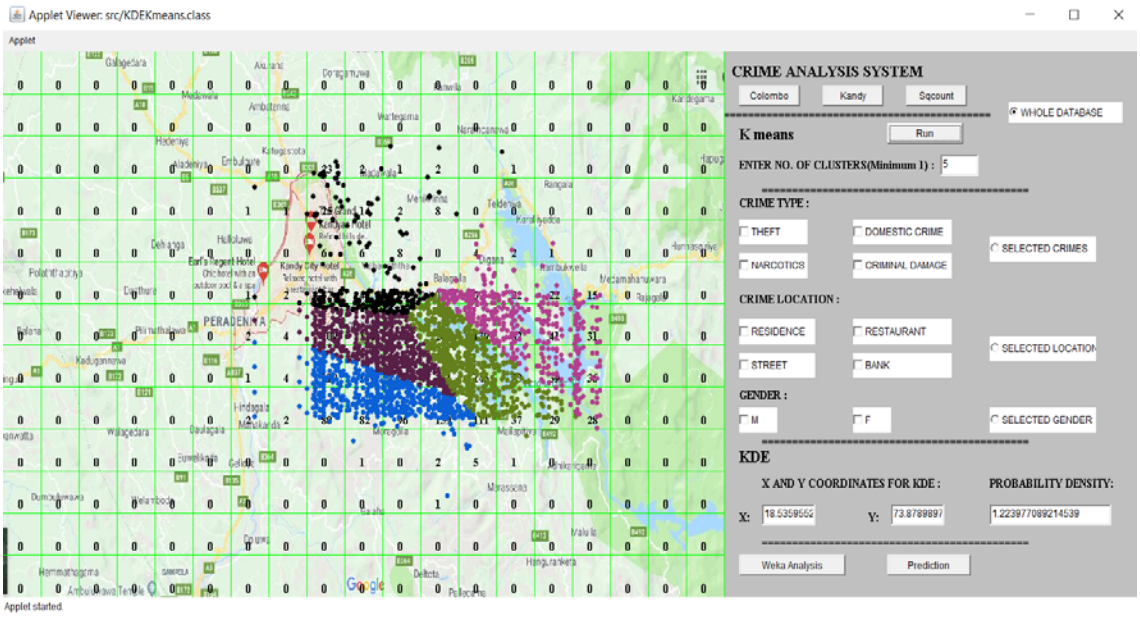


Figure 7.7 - Crime clustering for Kandy city

Further Figure 7.5 is the output from the excel file for the Kandy police area. So end-user can identify the severity of the crime based on count and can identify that location which has more crimes.

Table 7.5 - Cluster pattern for Kandy city map

Cluster Name	Crime Count
Cluster 1	622
Cluster 2	291
Cluster 3	1066
Cluster 4	321
Cluster 5	662

The further end-user can do the cluster analysis using Weka APIs. This will help to end-user to predict and forecast the crime pattern across the police area or city.

7.2. Cluster Analysis Using Weka API

Using the Weka API analysis, the researches measure the corrected classifies instances, incorrectly classify instances and the various error measurements. The precision-recall and the F measure are calculated by the author. The accuracy of the clusters is identified by the end-user by using the above parameters. The author explains how to calculate the Precision, Recall and the F Measure, before going to the system model [2].

Let's introduce Table 7.6 with the predicted and actual values.

Table 7.6 - Weka Actual, Predicted values

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Source: Accuracy, Precision, Recall or F1?, (2018) [2]

For Precision,

Precision = (True Positive / (True Positive + False Positive))

Precision = (True Positive / (Total Predicted Positive))

Immediately, you can see that Precision talks about how precise/accurate your model is out of Precision those predicted positive, how many of them are actually positive.

For Recall,

Recall = (True Positive / (True Positive + False Negative))

Recall = (True Positive / (Total Actual Positive))

Immediately, you can see that Recall talks about how many of the Actual Positives our model capture through labeling it as Positive (True Positive).

For F Measure,

F Measure = (2 * (Precision * Recall)) / (Precision + Recall)

So F Measure might be a better measure to use if the author needs to seek a balance between Precision and Recall and there is an uneven class distribution (a large number of Actual Negatives).

According to the system, the model author calculates the error measures and the prediction measurements.

Below Figure 7.8 output is shown the results with the error measurements.

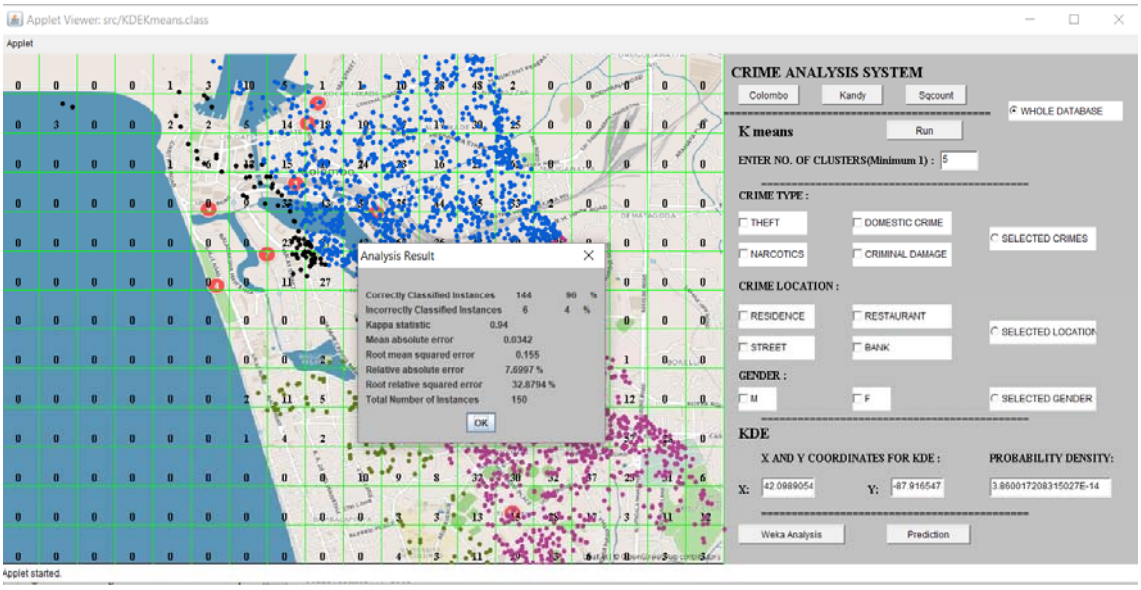


Figure 7.8 - Error measurements in map

Below Table 7.7 is shown the results with the error measurements.

Table 7.7 - Error measurements in the table

Error Measures	Readings
Correctly Classified Instances	144
Incorrectly Classified Instances	6
Kappa Statistic	0.94
Mean absolute error	0.0342
Root Mean squared error	0.155
Relative absolute error	7.6997
Root relative squared error	32.8794
Total number of instances	150

Below Figure 7.9 output is shown the results with the accurate measurements.

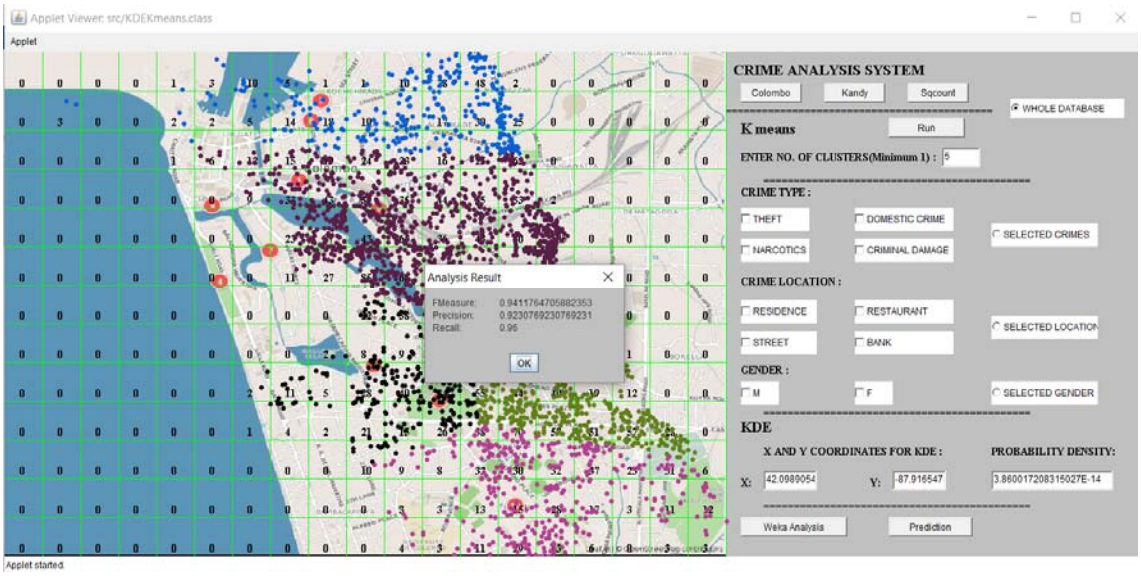


Figure 7.9 - Prediction measurements in map

Below Table 7.8 is shown the results with the error measurements.

Table 7.8 - Prediction measurements in the table

Accuracy Measures	Readings
F Measure	0.941176471
Precision	0.923076923
Recall	0.96

7.3. Cluster Analysis Using Weka Tool

Author automated the statistical analysis using Weka tool. This will be easy to analyze the statistical result. The diagram of the statistical result automation is shown below Figure 7.10

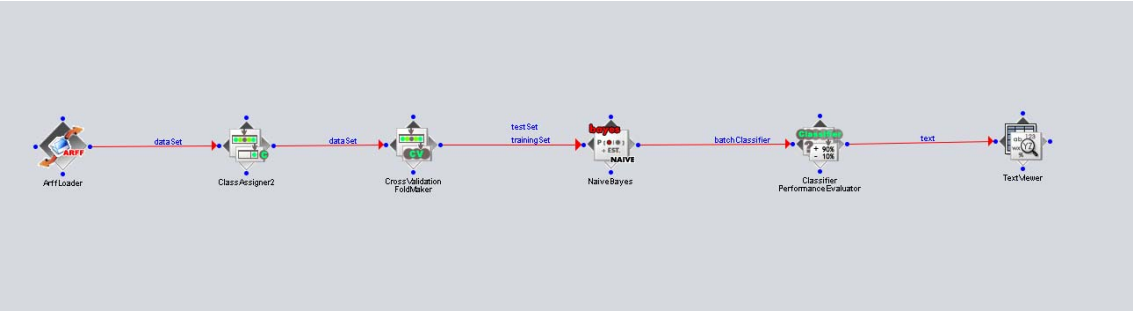


Figure 7.10 - Stats automation of Weka

The result of the above automation is shown below Figure 7.11.

Text

```
=== Evaluation result ===

Scheme: NaiveBayes
Relation: weather.symbolic


Correctly Classified Instances           4               28.5714 %
Incorrectly Classified Instances        10               71.4286 %
Kappa statistic                         -0.4583
Mean absolute error                      0.6231
Root mean squared error                  0.6376
Relative absolute error                123.4373 %
Root relative squared error            125.1823 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.167    0.625    0.167      0.167    0.167      -0.458   0.063     0.298     TRUE
                 0.375    0.833    0.375      0.375    0.375      -0.458   0.063     0.418     FALSE
Weighted Avg.    0.286    0.744    0.286      0.286    0.286      -0.458   0.063     0.367

=== Confusion Matrix ===

 a b   <-- classified as
 1 5 | a = TRUE
 5 3 | b = FALSE
```

Figure 7.11 - Stats results of Weka

So using this automation results, end-user or author calculates the error measures and the prediction measurements.

Further Author automated the plot analysis using Weka tool. So end-user can identify the measurements at every each point. The diagram of the point to point automation in the curve is shown below Figure 7.12

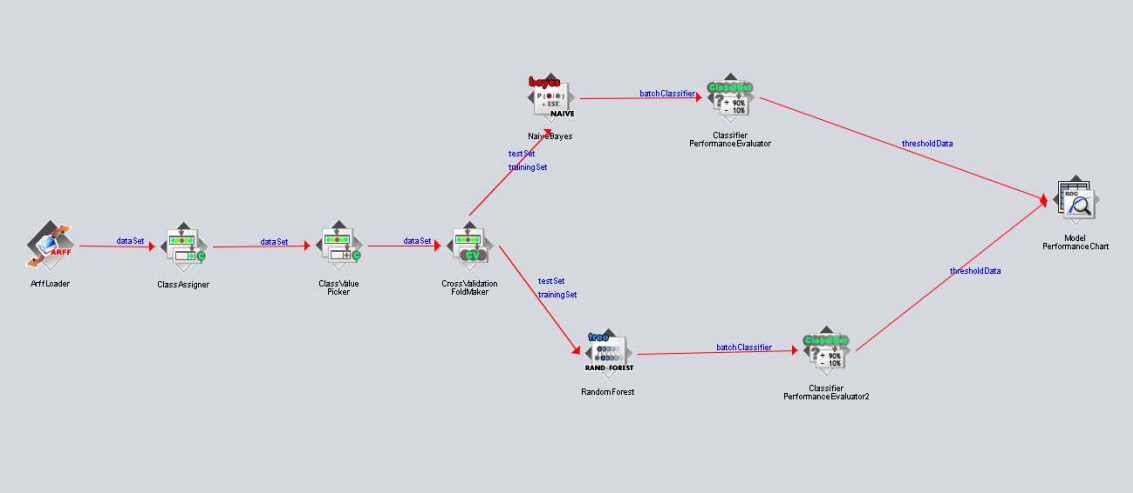


Figure 7.12 - Plots automation of Weka

The result of the above automation is shown below in Figure 7.13. So end-user or author can identify the changing crime pattern at every each point. So it will give more insight into the crime analyzing and crime forecasting in the future.

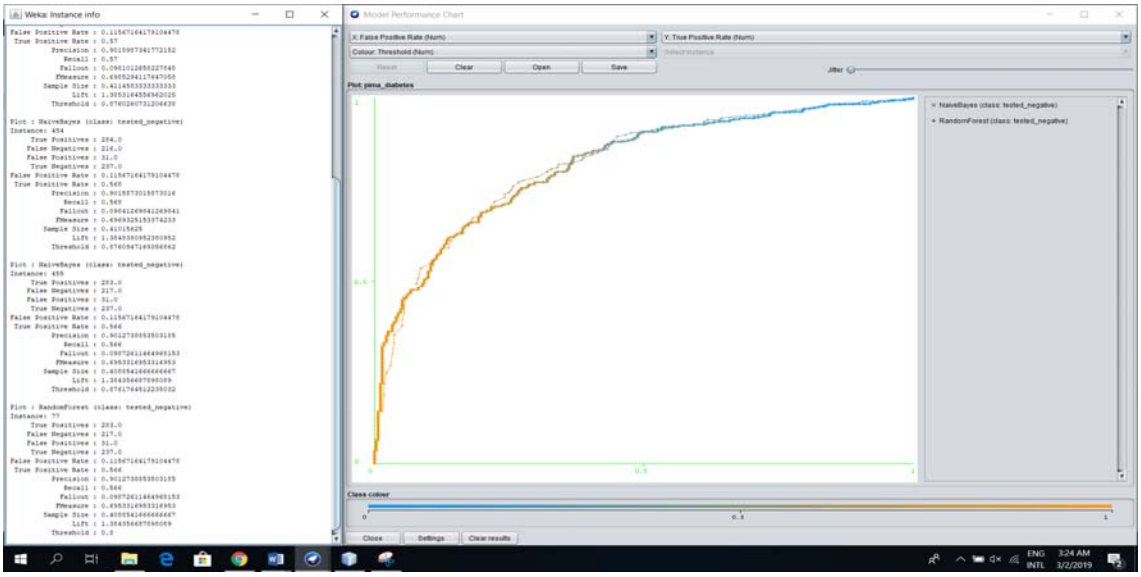


Figure 7.13 - Plots results of Weka

Chapter – 8 | Evaluation & Testing

8. Evaluation & Testing

8.1 Evaluation Methods

Many algorithms have been proposed by the task of clustering. Different techniques are in favor of different clustering purposes, therefore no clustering technique is universally applicable. Therefore, it is necessary to understand both the clustering problem and the clustering technique which is well required and to be applied to a suitable method to a given problem. In the following, the author has well described the parameters of a clustering technique which are relevant to the task of inducing a verb classification.

1. Parametric design:

Assumptions may (but need not) be made about the form of the distribution used to model the data by the cluster analysis. The parametric design should be chosen with respect to the nature of the data. It is Assumptions may (but need not) be made about the form of the distribution used to model the data by the cluster analysis. The parametric design should be chosen with respect to the nature of the data. It is often convenient to assume, for example, that the data can be modeled by a multivariate Gaussian.

2. Position, size, shape and density of the clusters:

There might be a clear desired clustering results based on the experimenter idea. This is in order to mainly position, size, shape and density of the clusters. There would be different impact on these parameters based on different clustering algorithms, this is depicted by the description of the algorithm. The design parameters are influenced by the various clustering algorithm.

3. A Number of clusters:

If the desired number is known beforehand the number of clusters can be fixed (e.g. because of a reference to a gold standard), this can also be varied in order to find the optimal cluster analysis. As Duda et al. (2000) state, 'In theory, the clustering problem can be solved by exhaustive enumeration, since the sample set is finite, so there are only a finite number of possible partitions; in practice, such an approach is unthinkable for all but the simplest problems, since there are at the order of kn k! Ways of partitioning a set of n elements into k subsets'.

4. Ambiguity:

Generally, there can be multiple senses for verbs, requiring them being assigned to multiple classes. The soft clustering algorithm defines cluster membership probabilities for the clustering objects, this is only possible by using soft clustering algorithm. The hard clustering performs yes/no decision on object membership and cannot model verb ambiguity but it has also be noted it is easier to use and interpret. The set of parameters are determined by the choice of the clustering algorithm.

8.2 Overall Testing

In the software engineering life cycle, Testing plays a major role. In this case, the Testing helps to check the software or the relevant products meet its requirement or generally, in other words, the software or if the appropriate product is ready to use by the end-user and if it is that final product that they had required. The author describes how the testing of the prototype is carried out in this chapter. It consists of the test cases and their results for every single module. Also furthermore, in this case, it has clearly mentioned about the current state of each functional requirement and also the non-functional requirements and their appropriate test results.

The below Table 8.1 explains the overall test scenario and the test result.

Table 8.1 - Evaluation

Test Scenario	Test Result
Evaluate that the data set which has sufficient data for pattern detection with proper fields.	The data set should have the proper fields which are enough to identify the data set with sufficient data for crime analysis.
Evaluate that the crime analyst can select the number of crime clusters.	The system should allow selecting the number of crime clusters and clusters should be created based on the numbers.
Evaluate that the cluster analysis happens within and across the clusters based on location and plotted on the geo-spatial plot.	The system is able to analyze within and across crime clusters and able to identify the crime location along with crime clusters and should be plotted geographical wise.

8.3 Module Testing

It has to be clearly noted that author has well prepared a testing schedule to reflect the appropriate unit, integration, system acceptance, and the significant release tests, as well as the exact time duration of each case. This schedule has particularly and well reflected the relevant personnel involved in the test effort. In the test schedule, include the following information:

- Documentation review
- Test scripts
- Data preparation
- Test execution
- Output review
- System certification
- System release

In this scenario, it has to be clearly noted that the module test significantly tests each and every single sub module in all modules and of the ones belonging to the prototype. (The persisting system setting module which is an important utility module is also tested by this), this generally checks whether if every individual module function as its appropriate design. The outputs of each module or sub-module also need to be clarified and verified it is because those outputs are using as input to the next module or sub module. It has to be noted that in each particular test case it gives a small significant description, also states the input and final result (pass/ fail).

- Creating the target dataset
- Data Cleaning and preprocessing
- Data Reduction and Projection
- Removing outliers
- Calculate neighbors
- Cluster Creation

The below Table 8.2 explains the module wise test scenario and the test result.

Table 8.2 - Module Testing

Module	Input	Actual Result	Expected Result	Result
Creating the target dataset	Pass the CSV file to the system	The relevant dataset is received according to crime.	Get the output as expected	Pass
Data Cleaning and preprocessing	Pass the CSV file to the system	The existing data are corrected and the missing values are checked.	Get the output as expected	Pass
Data Reduction and Projection	Pass the CSV file to the system	The unwanted attributes and data are removed.	Get the output as expected	Pass
Removing outliers	Pass the CSV file to the system	The outlier values are removed in order to create the cluster.	Get the output as expected	Pass
Calculate neighbors	Pass the CSV file to the system	Based on the distance between data entry points the neighbors are determined.	Get the output as expected	Pass
Cluster Creation	Pass the CSV file to the system	The cluster is created.	Get the output as expected	Pass

8.4 Functional Testing

It has to be noted that the Functional Requirement Testing is very important and well applied in order to determine how much of functional requirements are concentrated on by the system, by using following test cases it is easy to verify that system performs according to the requirements. Here it eventually helps to evaluate the current system as well as to make the future plans of the system.

The Functional test cases are carried out under input, output relevant documents against each specific requirement in a table below and it provides the current status of each functional requirement. It normally checks whether the user requirements are fulfilled by the prototype or not.

The below Table 8.3 explains the functional wise test scenario and the test result.

Table 8.3 - Functional Testing

FR	Detail	Actual Result	Expected Result	Result
The system should read the crime data from the excel file.	The records are easily read line by line because the crimes records are maintained in excel sheets. Based on the records from the excel sheet the system will analyze the crime patterns. Either this can be Excel or CSV file.	The excel file should be read by the system	Get the output as expected	Pass
The crime clusters must be created and analyzed by the system based on the excel data.	It is advisable to create a string array by the system based on the crime data and matching the crimes based on the array, in this practice the common crime patterns are analyzed and well identified to reduce further occurrences of repeated incidence.	Based on the excel data the crime clusters have to be created and should be analyzed by the system.	Get the output as expected	Pass
The system should be able to configure the cluster number using user input	The number of clusters must be selected by the end-user. Using crime matching, the system creates clusters based on user inputs.	The system must be allowed to configure the cluster number using user input	Get the output as expected	Pass

Based on the geographical image the system should plot and also plot according to the cluster.	The proposed system should be utilized based alongside the Geo-spatial plot. Period range and different types of crimes are picked in order to show the outcome graphically, this is done by the crime analyst. This will help to forecast the futuristic crime in a proper visualized way.	The system should represent the data as a visualization.	Get the output as expected	Pass
The system should be able to do the cluster analysis using the Weka tool.	The system should analyze the clusters, this can be done ideally within a cluster or across it. It is done with the help of Weka tool. The system will finally display the analysis in UI.	The clusters should be analyzed by the system	Get the output as expected	Pass

Chapter – 9 |
Conclusion

9. Conclusion

9.1 Learning Points

The completion of the project undertaken is well described here. The author reviews the achievements accomplished here, in the form of objectives and this has given him/her great pleasure in doing this task ahead. The problems that were faced during the life cycle of the project have been discussed by the author here. The chapter relevantly concludes highlighting the future enhancement to the system.

Here the students have become skilled at lots of new skills, this has become an impressive feature of this kind of project development. There has been a significant improvement in the following skills.

- In order to learn independently with minimum guidance and supervision.
- In order to work with tight deadlines.
- In order to work independently with minimum guidance and supervision.
- In order to work carefully in the presentation of system deliverables.

9.2 Key Achievements

The key achievements for the author were because most of the existing researchers could not reach with the integration of both the features which were, the visualization of crime data and the cluster analysis from the existing clusters. The implementation of a full analysis system with the prediction, forecasting and visualization was combined well and applied by the author in this case.

In order to effectively improve knowledge in different technologies especially gains depth knowledge in Weka tool and clustering techniques.

9.3 Future Work

Based on the past data of crime this system works and this generally creates the cluster, does the analysis and helps to visualize the data. In this scenario when the author considers the real-time situation, the author is actually not able to analyze the real-time crime data and the past time data using the Spatial-Temporal Analysis of Crime (STAC) algorithm which was using static data. We have to note that real-time is not static data. In this case, the author leaves this behind as research in order to figure out how to modify the STAC algorithm for the real-time data with the past data. The end-users find this more useful in order to identify the crime patterns using the real-time data together with the past data, rather than using only the past data.

Chapter – 10 |
References

10. References

[1] M. R. Keyvanpour, in *Detecting and investigating crime by means of data mining: a general crime matching framework*, Procedia Computer Science, 2011, p. 872–880.

[2] K. P. Shung, "Accuracy, Precision, Recall or F1?," 15 March 2018. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

[3] M. Keyvanpour, M. Javideh and M. R. Ebrahimi, "Detecting and investigating crimes by means of data mining: a general crime matching framework," pp. 1-9, 2011.

[4] P. Patil, "K Means Clustering : Identifying F.R.I.E.N.D.S in the World of Strangers," 20 May 2018. [Online]. Available: https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d.

[5] S. Kapoor and A. Kalra, "DATA MINING FOR CRIME DETECTION," September 2014. [Online]. Available: http://www.ijcea.com/wp-content/uploads/2014/10/11Shasha-kapoor.pdf.

[6] P. Gera and R. Vohra, "City Crime Profiling Using Cluster Analysis," 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.446.3788&rep=rep1&type=pdf.

[7] R. Kiani, S. Mahdavi and A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification," 2015. [Online]. Available: https://pdfs.semanticscholar.org/3643/74119cd633ac6396f81959700912acdf30ee.pdf.

[8] H. Zhang and M. P. Peterson, "A SPATIAL ANALYSIS OF NEIGHBOURHOOD CRIME IN OMAHA, NEBRASKA USING ALTERNATIVE MEASURES OF CRIME RATES," 2007. [Online]. Available: https://pdfs.semanticscholar.org/139a/6a864b9a30a0b346d7517f138b5059b6c089.pdf.

[9] M. Ahmadi, "Crime Mapping and Spatial Analysis," February 2003. [Online]. Available: https://pdfs.semanticscholar.org/26fc/afccd32e738286a20194dace82691c80ff0a.pdf.

[10] WilpenGorr and R. Harries, in *Introduction to crime forecasting*, 2003, p. 551–555.

[11] P. Phillips and I. Lee, in *Mining co-distribution patterns for large crime datasets, Expert Systems with Applications*, 2012, p. 11556–11563.

[12] N. Malleson, Analysis of crime patterns through the integration of an agent-based model and a population micro simulation, International Journal of Computers, Environment and Urban Systems, 2012.

[13] R. Adderley, M. Townsley and J. Bond, "Use of data mining techniques to model crime scene investigator performance," Knowledge-Based Systems, 2007, p. 170–176.

[14] T. Abraham and O. d. Vel, in *Investigative Profiling with Computer Forensic Log Data and Association Rules*, ICDM '02 Proceedings of the 2002 IEEE International Conference on Data Mining, 2002, pp. 11-18.

[15] H. Liu and D. E. Brown, in *Criminal incident prediction using a point-pattern-based density Model*, International Journal of Forecasting , 2003, p. 603–622.

[16] J. J. Corcoran, in *Predicting the geo-temporal variations of crime and disorder*, International Journal of Forecasting, 2003, pp. 623-634.

[17] G. C. Oatley, in *Decision support systems for police: Lessons from the application of data mining techniques to ''soft'' forensic evidence, Artificial Intelligence and Law*, Springer, 2006, p. 35–100..

[18] D. Brown, in *The regional crime analysis program (RECAP): A frame work for mining data to catch criminals*, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1998, pp. 2848-2853.

[19] A. Joshi and M. Suresh, "Finding similar case subset and hotspot detection in felonious data set using data mining algorithms: Weighted Clustering and Classification," International Journal of Application or Innovation in Engineering & Management, 2014, pp. 157-165.

[20] A. Malathi and S. SanthoshBaboo, in *An enhanced algorithm to predict future crime using data mining*, International Journal of Computer Applications, 2011, pp. 1-6.

[21] Anshusharma and R. Kumar, in *Analysis and Design of an Algorithm Using Data Mining Techniques for Matching and Predicting Crime*, International Journal of Computer Science and Technology, 2013, pp. 670- 674.

[22] G. C. Oatley and B. W. Ewart, in *Crimes analysis software: 'pins in maps', clustering and Bayes net prediction*, Expert Systems with Applications, 2003, p. 569–588.

[23] S.-T. Li, in *An intelligent decision-support model using FSOM and rule extraction for crime prevention*, Expert Systems with Applications, 2010, p. 7108–7119.

[24] M. Alruily, A. Ayesh and A. Al-Marghilani, in *Using Self Organizing Map to cluster Arabic crime documents*, IEEE International Multiconference on Computer Science and Information Technology, 2010, p. 357–363.

[25] Y. L. Boo and D. Alahakoon, in *Mining Multi-modal Crime Patterns at Different Levels of Granularity Using Hierarchical Clusterin*, IEEE International Conference on Computational Intelligence for Modelling Control and Automation, 2008,, p. 1268–1273.

[26] RizwanIqbal, in *An Experimental Study of Classification Algorithms for Crime Prediction*, Indian Journal of Science and Technology, 2013, pp. 4219-4225.

[27] A. M. Olligschlaeger, "ARTIFICIAL NEURAL NETWORKS AND CRIME MAPPING," [Online]. Available: https://pdfs.semanticscholar.org/4aa8/088fc73678cdf22eb7951202318492315361.pdf.

[28] A. Stec and D. Klabjan, "Forecasting Crime with Deep Learning," 5 June 2018. [Online]. Available: https://arxiv.org/pdf/1806.01486.pdf.

[29] M. A and D. S. S. Baboo, "An Enhanced Algorithm to Predict a Future Crime using Data Mining," May 2011. [Online]. Available: https://pdfs.semanticscholar.org/195a/247055cd1be24a4f27c607fc8c6a75a64f2f.pdf.

[30] R. A. Bolla, "Crime pattern detection using online social media," 2014. [Online]. Available: https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=8320&context=masters_theses.

[31] V. Jain, Y. Sharma, A. Bhatia and V. Arora, "Crime Prediction using K-means Algorithm," April 2017. [Online]. Available: http://www.grdjournals.com/uploads/article/GRDJE/V02/I05/0176/GRDJEV02I050176.pdf.

[32] S. Li, "Exploring, Clustering and Mapping Toronto's Crimes," 30 October 2017. [Online]. Available: https://towardsdatascience.com/exploring-clustering-and-mapping-torontos-crimes-96336efe490f.

[33] L. M, J. PK and R. Beatrice, "CLUSTERING ANALYSIS - US CITY CRIME 1970 DATASET," 19 March 2016. [Online]. Available: http://rstudio-pubs-static.s3.amazonaws.com/162521_c9bf11c9aa864ac198bb6d75853f622f.html.

[34] M. D. Porter, "Crime Series Identification and Clustering," 19 September 2015. [Online]. Available: https://cran.r-project.org/web/packages/crimelinkage/vignettes/crimeclustering.html.

[35] T. H. Grubesic and A. T. Murray, "Detecting Hot spots Using Cluster Analysis and GIS," [Online]. Available: https://pdfs.semanticscholar.org/c3db/1455ca4f771195525cf128efcdc866145045.pdf.

[36] C. H, C. W., J. Wang, G. Qin and Y. C. M., in *Crime data mining: a general framework and some examples*, IEEE Computer, 2004, pp. 50-56.

[37] M. W. Milo, S. C. Richards and P. Saraf, "Crime Hotspot Tracking and Geospatial Analysis in Merseyside, UK," 2012. [Online]. Available: https://pdfs.semanticscholar.org/88e9/4de27f4248b6cffa07836b28980f8d7c396a.pdf .

[38] S.Yamuna and N.SudhaBhuvaneswari, in *Data mining Techniques to Analyse and Predict Crimes*, The International Journal of Engineering And Science, 2011, p. 243 – 247.

[39] S. E. Reid, M. Valasik and G. Tita, "The Mapping and Spatial Analysis of Crime," January 2019. [Online]. Available: https://www.researchgate.net/publication/330425196_The_Mapping_and_Spatial_Analysis_of_Crime.

[40] J. Agarwal, R. Nagpal and R. Sehgal, "Crime Analysis using K-Means Clustering," December 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.1621&rep=rep1&type=pdf.

[41] ZakirHussain, in *Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model*, International Journal of computer science and Information Technology and Security, 2012, pp. 25-29.

[42] Dale Rudzkiene and Vitalija Rudzkiene, in *Multiple Regression Analysis in Crime PatternWarehouse for Decision Support, Database and Expert Systems Applications*, Springer-Verlag Berlin Heidelberg, 2002, pp. 249-258.

[43] WilpenGorr, in *Short-term forecasting of crime*, 2003, p. 579–594.

[44] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining," [Online]. Available: http://cs.brown.edu/courses/csci2950-t/crime.pdf.

[45] S. Borgatti, "Distance and Correlation," [Online]. Available: http://www.analytictech.com/mb876/handouts/distance_and_correlation.htm.

[46] A. Trevino, "Introduction to K-means Clustering," 6 December 2016. [Online]. Available: https://www.datascience.com/blog/k-means-clustering.

[47] J. Brownlee, "How To Estimate The Performance of Machine Learning Algorithms in Weka," 18 July 2016. [Online]. Available: https://machinelearningmastery.com/estimate-performance-machine-learning-algorithms-weka/.