A Decision Support System to Analyze Criminal Modus Operandi

by B.M.U.K.K. Basnayake

149205R

Dissertation submitted to the Faculty of Information Technology,

University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of

the Master of Science in Information Technology

June 2017

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any other university or any other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged and a list of references is given.

Name of Student	Signature of Student
B.M.U.K.K. Basnayake	…………………………
	Date:
Name of Supervisor	Signature of Supervisor
S. C. Premaratne	…………………………
	Date:

Acknowledgement

I would like to express my gratitude to my supervisor, Mr. S. C. Premaratne, Senior Lecturer at University of Moratuwa, Sri Lanka, for providing intellectual support, expertise and guidance and whose vast knowledge, skills in many areas, understanding and patience immensely contributed the research.

I would like to thank all the lecturers of University of Moratuwa, Sri Lanka, specially Prof. Asoka Karunananda, for the knowledge and assistance they provided at all levels of the research project.

Special thanks goes out to S.P.A. Seram, Director of the Criminal Records Division, Department of Police, Sri Lanka and his staff members, specially Mr. Suresh, for assisting me in gathering information and giving me their valuable comments.

I would also like to show my appreciation to all my friends and batch mates of the M.Sc. in IT degree program for their valuable support, assistance and feedback. I must acknowledge my loving family, particularly my father, for their encouragement and support.

Abstract

Crime analysis is the method of analyzing crime activities. The process of finding the relationship between criminals and crime is still under-developed, but there is an increase in the complexity of police information and intelligence. “Criminal intelligence” is information with added value used by law enforcement to deal with crimes. Criminal intelligence helps to direct and prioritize resources when it comes to preventing, minimizing and detecting crimes using the modus operandi of criminals. Strategically and tactically intelligence makes police decision making more accurate, efficient and justifiable. Up to date and useful criminal intelligence is necessary for preventing, minimizing and investigating very organized and dire crime.

After doing field work and gathering data and information, it was evident that the crime analysts in the Crime Records Office (CRO) are extracting and analyzing criminal records to solve crimes using a very primitive manual system to store, extract and analyze criminal information with limited resources and trained personnel. This makes analyzing large volumes of criminal data very tedious. Existing crime analysis tools do not meet the requirements of the CRO in Sri Lanka.

The solution to this problem is a user friendly system, connected to the Sri Lanka police Criminal Database, that is used to enter and store details of apprehended criminals, with descriptions of their modus operandi, and compare those details with the modus operandi, uncovered from crime scenes, to predict a list of suspects possibly connected to those crimes and help crime analysts make decisions on the next course of action based on the analysis. The analysis results include details such as the suspect names, the suspect’s present place of residence and the Supervisor in charge of tracking the suspect.

This research proves that the new approach for criminal records analysis through data mining is fruitful. The k-means algorithm used for criminal records analysis has proven its efficiency and accuracy in clustering criminals. The criminal records analysis system is built to help crime analysts in the CRO and minimize the paperwork for police officers when recording criminal information. It supports the crime analysts in decision making but cannot replace them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prolegomena

This thesis introduces a crime analysis tool for the Sri Lanka Police that provides the necessary criminal intelligence to solve cases faster and bring criminals to justice. Developing a crime analysis tool to identify criminals based on their modus operandi is a challenging field of research [1].

Crime analysis is the process of analyzing crime activities. The value of information is identified through analysis. Once information is collected, it will be "evaluated" according to the reliability of its source and the relevance and validity of its content before being filed, cross referenced and organized, which is known as "collation". Analysis will consider the information in context, draw conclusions as to what it means and produce reports, briefings and other documentation depicting that meaning. The results of this process will be distributed or "disseminated" to those who need to know it.

The art of connecting crime with criminals is primitive and undeveloped [2] but there is an increase in the sophistication of police information and intelligence. "Criminal intelligence" is information with additional value used by law enforcement to deal with crimes. Criminal intelligence helps direct and prioritize resources in the prevention, reduction and detection of crimes through the identification and analysis of modus operandi. Strategically and tactically intelligence makes police decision making more accurate and justifiable. Timely and usable criminal intelligence is essential for the prevention, reduction and investigation of severe and organized crime [3].

1.2 Objectives of the Research

- Create a useful crime analysis tool that can be of great service to the society.
- Help Crime Analysts to easily and efficiently analyze the modus operandi (MO) and other information of criminals without having to go through each and every criminal record manually.
- Help Crime Analysts to easily enter and store the details of arrested criminals and descriptions of their MOs without having to manually fill forms.
- Reduce the space required to store stacks of crime records in the CRO by saving and maintaining the records in the database.
- Reduce the time and efforts put in by Crime Analysts and make their jobs easier.
- Help the Sri Lanka Police to easily identify possible suspects connected with a crime.
- Prevent criminals from getting away with crime.

1.3 Background and Motivation

With the rapid technological development in the world the Crime Records Division (CRD) of the Sri Lanka Police is turning towards technology to automate their manual procedures such as entering and storing details of arrested criminals with descriptions of their modus operandi and analyzing those details with the modus operandi of Crime Scenes to provide valuable suggestions on the most probable suspects. This helps to narrow down the suspect list but since it is done manually, it consumes a lot of effort and time for Crime Analysts.

Currently there are a handful of expensive software tools that analyze the modus operandi of criminals but there are none that cater to the requirements of crime analysts in the Crime Records Office (CRO), the limited budget and the limitations of VPNs at the CRD. As a government organization, the Sri Lanka Police needs to do more with less. Managing under tighter policing budgets as government revenues shrink, they must continue to improve public safety and respond to stronger calls for transparency and accountability in all policing actions. With budget cuts pressuring law enforcement to fight crime using fewer resources, the use of analytics to both respond to crime is now more critical than ever [4].

Even though there is a decrease in the crime rate in Sri Lanka the last few years, the severity of crimes have increased significantly and many of the convicted criminals have successfully eluded capture by law enforcement agencies, and are, to date, still at large. This is a serious concern for the society. Crime has become a career, an industry, an art, with established customs and practices and a recognized technique. On its perpetration an amazing amount of skill and finesse is utilized.

According to the Criminal Records of the CRD, out of 369000 criminals arrested in Sri Lanka, 40000 are repeat offenders (Island Registered Criminals – IRC). Most crimes are committed by Habitual Criminals and they are rarely versatile in their crimes. They tend to commit the same kind of crime over and over again; they do not often venture upon forms of crime with which they are unfamiliar. Professional criminals have characteristically individual and unique ways of committing crimes known as the Modus Operandi. Major W.L. Atcherley, Chief Constable of the West Riding of the Yorkshire Constabulary, devised the Modus Operandi System.

Success of solving a crime is too often the result of chance or accident rather than of a premeditated plan. The finger-print system has added greatly to the effectiveness of police service. Its scope, however, is limited. It is of use only where the measurements of finger-prints of a criminal are already on file. For the most part, they merely establish the identity of a man under arrest and connect him with a previous criminal record. There are very few experienced Crime Analysts and they are too often armed with poor tools or with no tools at all [2].

Therefore this study focuses on a system which aims at reducing the effort and time to identify possible suspects while helping the Sri Lanka police make the best use of the people and information at hand to solve cases faster.

1.4 Problem Statement

After conducting a comprehensive literature review and gathering data and information it was realized that the crime analysts in the Crime Records Office (CRO) are using a very primitive manual system to store, retrieve and analyze criminal records to solve crimes with limited resources and trained personnel. This causes the difficulty of analyzing large volumes of data involved in criminal and terrorist activities. The existing crime analysis tools do not cater to the requirements of the CRO in Sri Lanka.

MOs of offenders include all phases of the crime, specially the methods used, from planning to the use of tools, weapons and other resources and the completion of the crime which involves getting rid of or hiding stolen goods, weapons and covering tracks. The MOs of crimes can be compared to and linked to previous crimes and criminals to narrow down the list of suspects. Therefore this process needs to be made more accurate and efficient.

With the increase in the severity of crimes in Sri Lanka and the fact that many convicts have managed to escape justice by hiding in the society it is crucial to automate and improve the efficiency of the existing manual system of crime analysis in bringing these criminals to justice. This is a very serious issue that needs to be addressed and technology can play a crucial role in solving this problem.

1.5 Hypothesis

Data mining methods, such as k-means clustering, can be utilized to group and develop criminal profiles to predict a list of suspects possibly connected to a crime.

1.6 Proposed Solution

The solution is to introduce a user friendly web system which can be directly connected to the Criminal Database in the Sri Lanka police to enter and store details of arrested criminals with descriptions of their modus operandi and analyze those details with the modus operandi of crime scenes to provide valuable suggestions on the most probable suspects connected to those crimes. This system should be able to display a list of suspects close to the crime scene, which includes details such as their names, current residential addresses and the Supervisor in charge of tracking the suspects, based on descriptions of their MO and other information. This is a very effective way to help the Sri Lanka Police capture criminals who escaped justice.

1.7 Structure of the Thesis

The rest of the thesis is organized as follows. The first chapter is a summary and overview of the research and the second chapter is a critical literature review which identifies the research problems. The third chapter summarizes the technologies adapted in criminal records analysis and the fourth chapter describes the approach to develop the criminal records analysis tool. Chapter 5 presents the design and analysis. Chapter 6 and 7 describe the implementation and evaluation respectively. The final chapter concludes the research with a discussion on the results, limitations and further work.

Chapter 2

Developments and Challenges in Crime Records Analysis

2.1 Introduction

This Chapter provides a critical review of the literature in relation to developments and challenges in crime records analysis. Finally this chapter defines the research problem and identifies the technologies that can be used to address the problem.

Defining the problem and choice of technology or algorithm, must be justified with reason. From this chapter you will get not only the problem and the technology, but also major researchers, their projects, achievements, software/hardware and methodology used.

2.2 The Modus Operandi System

There are many researches done on detection of criminals by analyzing the Modus Operandi of criminals such as the research done by Raymond B. Fosdick in 1916 based on the Modus Operandi System devised by Major W.L. Atcherley, Chief Constable of the West Riding of the Yorkshire Constabulary in 1913 [2]. This system allows detecting criminals by an intensive study of their methods of operation. It is a cooperative arrangement, by which habitual or traveling criminals can be traced from community to community by a comparison of their methods of work. However since the procedures of this system are done manually, it consumes a lot of effort and time for Crime Analysts and requires clearing-houses to collect, sort, store and analyze facts relating to crime methods which is a very tedious task. This system has been used by law enforcement agencies around the world and has proven to be very effective in facilitating the apprehension of criminals. However with the rapid technological development in the world the manual system needs to be automated to improve the efficiency of crime analysis.

2.3 CODER

Fox and France proposed the CODER (Composite Document Expert/Extended/Effective Retrieval) Architecture for composite document analysis, representation and retrieval [5]. It is an Artificial Intelligence (AI) based system. The architecture is used to analyze a large amount of documents and efficiently retrieve information from those documents. However this architecture will require the development of a more refined lexicon, specification of heuristics with regard to retrieval techniques for certain types of queries and intensive integration of user models into an interactive system for information needs. Similar work includes an expert system for document retrieval [6]. By May 1987, roughly 13 megabytes of data, including over 6500 messages by different authors in different formats, were collected through the CODER architecture. Therefore it can be concluded that the various goals of this research are met. A feasibility study needs to be conducted on CODERs ability to analyze and retrieve documents.

2.4 Textual Analysis of Modus Operandi

Rogerson has done research which studies the usability of applying textual analysis to descriptions of criminal modus operandi for the prevention of high volume crime [7]. The research successfully shows that MO fields have a potential source of intelligence relevant to crime prevention and research and this information can be extracted using novel computer-aided textual analysis techniques. The two theoretical frameworks introduced in the thesis presents a connection between data patterns, previous literature findings, theories of causes of crimes and preventative methods. However, the analysis process is not fully objective or automated. In [8], the study identified hunting process scripts from a sample of 361 serial sex offends perpetrated by 72 serial sex offenders, using multiple correspondence analysis and hierarchical cluster analysis. The research concluded that criminological frameworks are a pre-requisite to interpret intelligence even though the research questioned the strict categories and hierarchies imposed by the frameworks which do not completely mirror the flexibilities of real-life crime commission. Nevertheless it will be useful to apply the techniques used in the research for other types of crimes other than theft.

2.5 Clustering Techniques

There are many surveys done on crime data analysis such as the survey done by Bharathi and Shilpa on crime data analysis through data mining using clustering techniques [1]. This paper presents a detailed study on clustering techniques and its role on crime applications based on many researches conducted on clustering. It helps to better predict and categorize crimes. However, the survey focuses only on three clustering algorithms. This survey discusses the importance of clustering and similarity measures in detail. Nevertheless it is better to conduct a survey on clustering algorithms, other than the three clustering algorithms described in this paper, which can be used in crime data analysis.

2.6 Crime Data Mining

Chen et al. has reviewed crime data mining techniques and presented four case studies done in the COPLINK project [9]. This research is based on a number of crime characteristics and analysis techniques. A major advantage of the review is that it helps identify suitable data mining techniques to analyze large volumes of data involved in criminal and terrorist activities. However the neural network-based entity extractor did not perform as well for addresses and personal properties (47 – 60%) [10]. Therefore through the results of the case studies it can be concluded that crime data mining increases the effectiveness and efficiency of criminal and intelligence analysis. More visual and intuitive data mining techniques can be developed for identifying crime patterns and visualizing networks.

2.7 Data Mining in Criminal Career Analysis

Bruin et al. describes a tool that extracts criminal career information based on four important factors: crime, nature, frequency and severity [11]. The research is based on a more individually oriented approach suggested by Blumstein [12]. By visual clustering of criminal careers, the research enables the identification of classes of criminals. However the research only focuses on career criminals and does not take into account one time offenders who have a high probability of becoming career criminals. The enormous cluster of one time offenders gave an ambiguous situational picture of the distance space which caused a large group of un-clustered criminals. Xianga et al. [13] proposes the use of a hyperbolic tree view and a hierarchical list view to visualize criminal relationships. The FLINTS and FinCEN project identifies relationships between offenses and offenders and fraudulent transactions [14]. Oatly et al. [15] connected robberies in the OVER project. The tool analyzed criminal records and provided users with a realistic 2-D clustering of criminal careers. This research has demonstrated the applicability of data mining in criminal career analysis. However the system did not achieve optimal runtime and the clustering algorithm utilizes too much computational time resulting in performance delays. Further research on clustering the group of un-clustered criminals, analysis of one time offenders, using progressive multi-dimensional scaling approach [16] to optimize the performance of the tool and incorporating the tool in a data mining framework for automatic police analysis, should be done.

2.8 Future Challenges

So far discussion shows that there are many tools and frameworks for criminal data analysis. The above researches have used various technologies and methods for this purpose. Nevertheless there are many limitations of these solutions. This indicates the need of further research on these tools and frameworks. It is needed to integrate more features, functionalities and technologies required into these systems for the real world adoption. Further research on criminal record analysis needs to be done to figure out which tools are best for criminal records analysis. There is no complete solution for criminal records analysis as it is still a developing field.

Table 2.1 summarizes the achievements and the benefits of the key research discussed in this chapter.

Research	Key Benefits	Limitations	Remarks
Bharathi [1]	Helps to better predict and categorize crimes. This survey discusses the importance of clustering and similarity measures in detail.	The survey focuses only on three clustering algorithms.	A survey should be conducted on other clustering algorithms that can be used in criminal records analysis.
Bruin [11]	Enables the identification of classes of criminals and demonstrated the applicability of data mining in criminal career analysis.	Research only focuses on career criminals. Cluster of one time offenders maybe the cause of the creation of a group of un-clustered criminals and creates an ambiguous situational picture of the distance space. The approach caused performance delays.	Further research on clustering the group of un-clustered criminals, analysis of one time offenders, using progressive multi-dimensional scaling approach [16] to optimize the performance of the tool and incorporating the tool in a data mining framework for automatic police analysis, should be done
Chen [9]	Helps identify data mining techniques to analyze large volumes of criminal and terrorist data.	The neural network-based entity extractor did not perform well for addresses and personal properties.	More visual and intuitive data mining techniques can be developed for identifying crime patterns and visualizing networks.
Fosdick [2]	Habitual or traveling criminals can be traced by a comparison of their MO. This system has been used by law enforcement agencies around the world and has proven to be very effective in apprehending criminals.	The manual system consumes a lot of effort and time for crime analysis and requires clearing-houses to collect, sort, store and manage criminal records.	The manual system needs to be automated to improve the efficiency of crime analysis.

Fox [17]	The architecture analyses and efficiently retrieves information from a large amount of documents.	This architecture requires a refined lexicon, specification of heuristics with regard to retrieval techniques of queries and integration of user models into an interactive system.	A feasibility study needs to be conducted on CODERs ability to analyze and retrieve documents.
Rogerson [7]	The research successfully shows that MO fields have a potential source of intelligence relevant to crime prevention and research and this information can be extracted using novel computer-aided textual analysis techniques. The frameworks present a connection between data patterns, previous literature findings, theories of causes of crimes and preventative methods.	The analysis process is not fully objective or automated. The categories and hierarchies of the frameworks do not completely mirror the flexibilities of real-life crime commission.	It will be useful to apply the research techniques for other types of crimes other than theft.

Table 2.1: Summary of Literature Review

2.9 Summary

This chapter presented a critical review of literature on the research problem by giving an overview of the research presented by others in this thesis and discussing the findings of the literature review, which includes the research problems and descriptions of the solutions. The next chapter will describe in detail the technology used to solve the research problem.

Chapter 3

Technology Adopted in Crime Records Analysis

3.1 Introduction

The previous chapter includes findings of a comprehensive literature review on criminal records analysis. This chapter presents the technology adapted in criminal records analysis which describes in detail the different methods used and a brief review of the technologies applicable to solve the problem of predicting the likely perpetrators of crime based on previous data.

3.2 Predicting Suspects through Modus Operandi

As discussed in the literature review per the findings of Fosdick on the modus operandi system devised by Major W.L. Atcherley [2], predicting suspects based on their modus operandi proved to help solve crimes faster. This is very true according the statistics obtained from the CRD which shows that out of 369000 criminals arrested in Sri Lanka 40000 are habitual criminals.

Out of the large pool of one time offenders, most crimes are committed by career criminals as they are not afraid to commit the same crimes over and over again. These criminals are comfortable in committing crimes in a manner they’re familiar with, which is like their own signature of committing crimes. This signature or modus operandi maybe very unique to the criminal or a few criminals may have the same modus operandi. Either way, these criminals can be suspected of committing a crime by identifying the modus operandi after analyzing a crime scene.

The current manual system of going through each and every criminal record manually is a tedious task. Therefore automating this process improves the accuracy and efficiency of apprehending criminals in connection with crimes and helps solve more crimes much faster. The suggestions of crime analysts with respect to crimes have proven very important in solving many crimes in Sri Lanka. However the limited budget and lack of resources has hindered the progress of solving crimes.

As per the grave crime abstract for the year 2016 it can be seen that out of 36768 grave crimes committed there were 120 mistakes committed by the Police, in 5625 cases the perpetrator is unknown and only 3296 crimes have been solved so far. There are 17241 unsolved cases. These statistics show that the current manual system of criminal records analysis is very slow and analysis results maybe prone to human error. The fact that 5625 criminals are on the loose is disturbing news and the fact that there are 17241 unsolved cases is a cause for concern. However statistics show that there is a significant decrease in grave crimes since 2013. This number can be further reduced by improving the current manual criminal records analysis system using existing technologies.

3.3 Current Methods and Technologies used in Criminal Records Analysis

Many researches have used various technologies and methods for criminal records analysis. Research done by Oatley et al. [15] shows that spatial statistics can be used to predict criminals. Crime matching techniques used are logic programming and ontologies, and naïve Bayes augmented with spatio-temporal features. Bayesian networks were used for crime prediction.

CrimeNet Explorer [18] incorporates several advanced techniques such as concept space approach, hierarchical clustering, social network analysis methods, and multidimensional scaling. The COPLINK [13] project uses hyperbolic tree view and a hierarchical list view to visualize criminal relationships. All these methods and technologies have proven valuable in crime analysis. Likewise the field of crime analysis can benefit from many such technologies and methods.

3.4 Technologies and Methods Adopted

To solve the current problem of automating the criminal records analysis system of the CRO, a system comprising of a web GUI and database connected to an analysis tool is implemented. The web GUIs are developed using Visual Studio, the database is developed using SQL Server Management Studio and RapidMiner is connected to the database as the analysis tool.

As the data mining method k-means clustering is used with the density performance operator to evaluate the accuracy of the clustering method used. The clustering results are visually represented through a scatter plot graph. Querying is used in the search functionality of the system to enter certain details of the criminal to search and retrieve the criminal records. The reporting extension of RapidMiner is used to generate reports of the analysis results in html and pdf formats.

Web GUIs are used because it will be easier for police officers to access the criminal database online through any device, from any place at any time. Visual Studio is the best and most widely used software to develop interfaces to connect to the SQL database. Since the criminal records database of the MIS is an SQL database, the MOS database is also an SQL database. RapidMiner is a sophisticated open source analysis software used specifically for the purpose of data mining. It consists of a rich library of data mining algorithms.

K-means clustering algorithm is used to cluster the criminal data because the data being clustered are of the mixed data types of numeric, nominal and date values. Compared to the other algorithms and the requirements of the criminal records analysis system, k-means clustering algorithm accurately clusters the criminals into groups.

3.5 Summary

This chapter describes in detail the technologies and methods used to solve the research problem and provide a brief review of the technologies and methods used in criminal records analysis. It also explains the inspiration behind the research and why the mentioned technologies and methods were used with justifications. The next chapter describes the approach to develop the crime analysis tool using the technologies and methods described in this chapter.

Chapter 4

Approach to Develop the Crime Records Analysis Tool

4.1 Introduction

The previous chapter describes in detail the technology and different methods used to solve the research problem. This chapter explains in detail the approach to develop the crime analysis tool using the technology described in the previous chapter to predict suspects based on their modus operandi and to classify the criminals based on the number of offenses committed and their modus operandi.

4.2 Hypothesis

The automated criminal modus operandi analysis system can be easily used to enter, store, retrieve, manage and analyze criminal information in a database to predict suspects based on their modus operandi and to provide decision support to the Sri Lanka police to solve crimes more accurately and efficiently, minimizing the human error. Computer aided queries can be used to retrieve criminal data and suspects can be predicted in connection with a crime based on their modus operandi. Criminals can be clustered using the k-means clustering algorithm based on two important factors: the modus operandi and number of crimes committed by each criminal. This information can be used to sort the list of suspects predicted in connection with a crime thus narrowing down the list of suspects.

4.3 Input Data

Criminal data or information is input through a web form, connected to the central database, by Police officers when a criminal is apprehended for a crime. These details include the details of the criminal, details of the criminal’s family and associates, modus operandi of the criminal, observations made at the crime scene, details of the police station, details of the crime, details of the court case/s of the criminal and the details of the Supervisor in charge of the criminal. The data types are mixed and include numerical, nominal and date attributes.

4.4 Process Model

It is important to understand the overall approach prior to extracting information from the data. Simply knowing data analysis algorithms is insufficient for successful data mining [19]. Therefore this research is conducted according to the KDD (Knowledge Discovery in Databases) process introduced by Fayyad et al. [20]. This process includes the following steps as displayed in Figure 5.1,

1. Developing and understanding the application domain. This includes field work which involves gathering the required data and getting a clear understanding of the requirements of the CRO by visiting and interviewing the police officers at the CRD. It also involves conducting a literature review on researches done on crimes and criminal modus operandi to help gather more information on the domain.
2. Creating a target data set. This includes querying the gathered data to select the desired subset.
3. Data cleaning and preprocessing. This consists of removing irrelevant data and information, dealing with duplicate data values such as the similarity of the names of criminals.
4. Data reduction and projection.
5. Choosing the data mining task. This involves matching the goals defined in step 1 with a specific data mining method, such as classification, regression, clustering, etc.
6. Choosing the data mining algorithm. This involves choosing methods to cluster the criminals based on the criminal data available.
7. Data mining. This step helps predict information using the mining model and predict patterns in a particular representational form, such as decision trees etc.
8. Interpreting mined patterns. Here visualization of the retrieved information can be done through histograms, tables etc.
9. Consolidating discovered knowledge. The final step is generating reports of the information discovered in html and pdf formats. This helps comparison of the current analysis with previously conducted analyses.

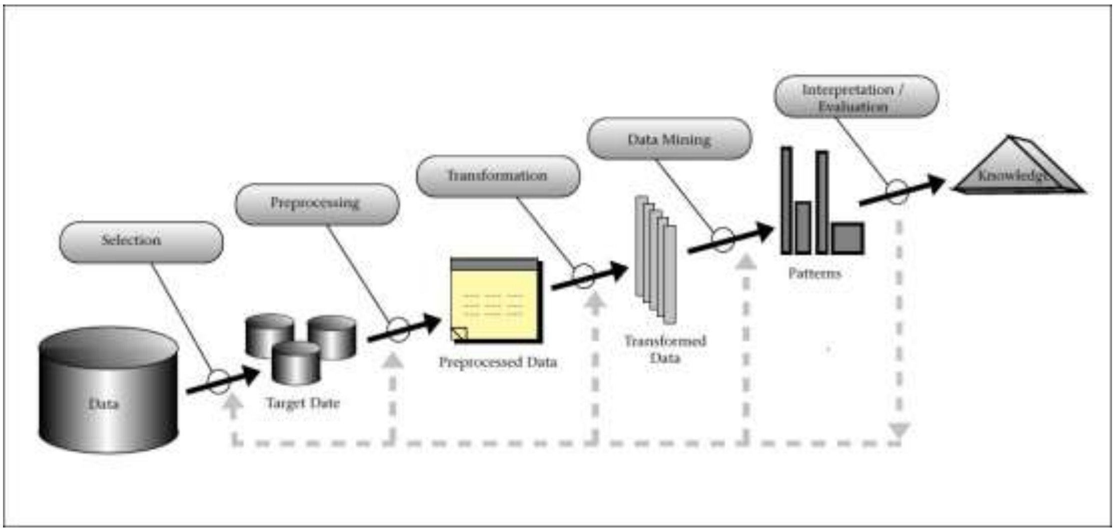


Figure 4.1: Overview of the steps in the KDD process

4.5 Results of the Analysis

After gathering evidence from a crime scene and identifying the modus operandi used to commit the crime, the crime analyst can enter the details of the modus operandi and other details, uncovered from the crime scene, into the system and get a list of suspects sorted by the number of crimes committed matching the modus operandi. The analysts can view the details of each and every suspect listed in the search query and also view the analysis results in the form of clusters, histograms and decision trees.

4.6 System Features

The system comprises of a simple user friendly user interface through which criminal data is entered and a central database to which the UIs are connected to store the criminal information. The database is stored and maintained within secure servers. Analysis is done using the k-means clustering algorithm to cluster criminals based on their common traits.

4.7 System Users

The system can be accessed by policemen, of every police department in Sri Lanka, through the web portal to enter and update details of apprehended criminals. They can also easily retrieve the details of criminals. However the crime analysis is conducted, through the system, by the Crime Records Officers in the CRO to provide valuable suggestions on the most probable suspects in connection with a criminal case to solve crimes accurately and efficiently. The results of the analysis are distributed in a need to know basis to the relevant departments and officers in charge of ongoing investigations. The Police officers at the CRD are responsible for the maintenance of the system and act as the system administrators. They create user profiles for the relevant police officers.

4.8 Summary

This chapter discussed in detail the approach adopted to develop the crime records analysis tool for decision support including the hypothesis, input data, process adopted, output of the analysis, system features and system users. It is important to accurately and efficiently analyze the large volume of data collected by the system. In most cases it is difficult for the police to obtain precise information on criminals. Clustering is the most suitable method to group criminals according to their common traits. The results of the analysis can help compare results from previous analyses to come up with predictions and assumptions on crimes and criminals. The next chapter will discuss the design of the solution in detail.

Chapter 5

System and Database Design

5.1 Introduction

The previous chapter describes the approach adopted to solve the identified problem. This chapter describes the system and database design to analyze crime records and describes each component in the system in detail providing a clear picture of the overall system.

5.2 Top Level Architecture

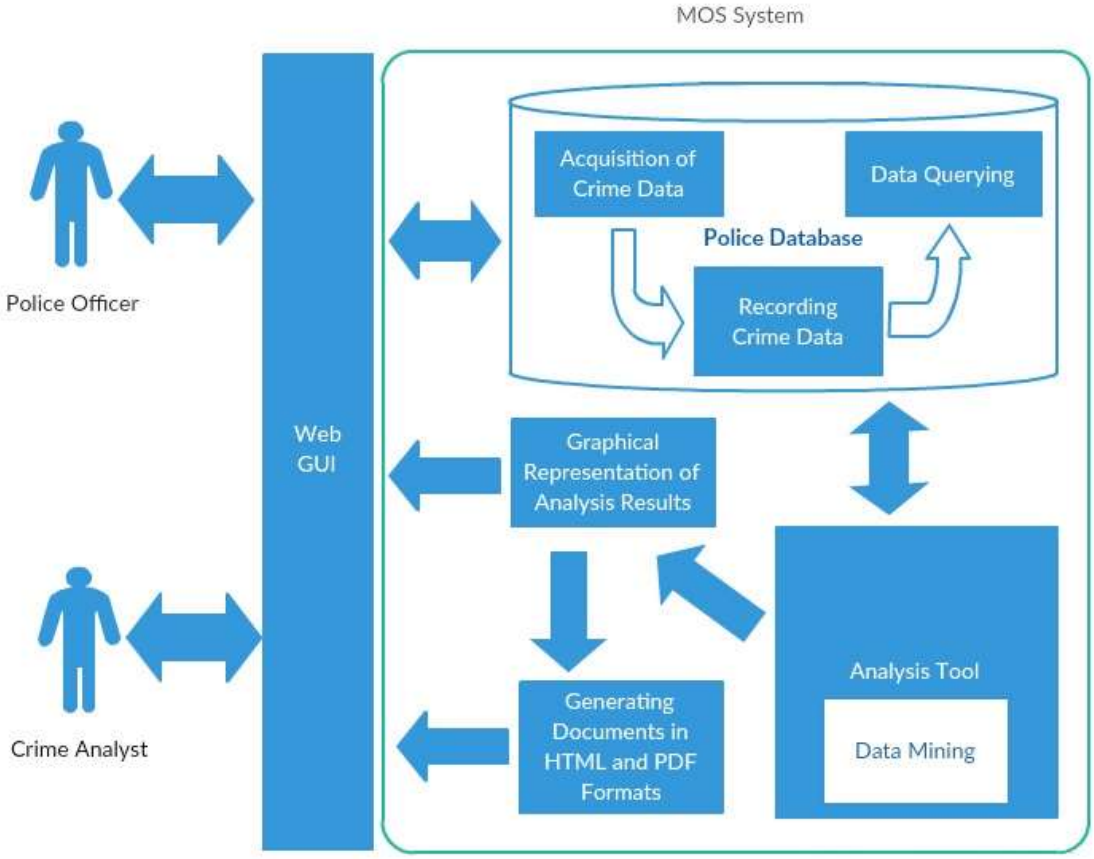


Figure 5.1: Top level architecture of the MOS (Modus Operandi System)

As represented in Figure 5.1 above, the criminal records analysis system can be divided into three main modules listed as follows,

- The web graphical user interface
- The Police database
- The criminal records analysis tool

5.2.1 Web Graphical User Interface

Police officers can login to the system through the web portal and enter and manage the details of apprehended criminals through the simple user-friendly web form. However only authorized police personnel have access to the system. The web portal can be accessed from every Police department in Sri Lanka and the web form follows a similar format to the paper form used to manually fill in the details of criminals. These details include the details of the criminal, details of the criminal's family and associates, modus operandi of the criminal, observations made at the crime scene, details of the police station, details of the crime, details of the court case/s of the criminal and the details of the Supervisor in charge of the criminal. This information can be easily accessed and viewed by police officers through the search functionality.

The form consists mostly of drop down menus to select values from, radio buttons and date fields to make it easier for the Police officers to enter the details, as a solution to the problem of the language barrier. It is the responsibility of the Police officers at the CRD to maintain the system and they act as the system administrators. They can create user profiles for the authorized police personnel. The police user profile contains details such as the name, badge number and rank of the policeman.

After gathering the required evidence and details of the modus operandi from the crime scene, this information can be entered into the system and the system will display a list of suspects close to the crime scene, which includes details such as their names, current residential addresses and the Supervisor in charge of tracking the suspects, based on descriptions of their MO and other information. The modus operandi is entered into the system based on a classification code as in the following Tables.

Code	Meaning
ABD	Abduction/kidnapping
AMF	Arson and mischief
BGL	Breaking into houses
HOM	Homicide
AHM	Attempted murder / abetting suicide
RAP	Rape
ROB	Robbery
THF	Theft
CBT	Conning
CAT	Stealing livestock
GHT	Grievous bodily harm
HNF	Stabbing
RIT	Riot
UNN	Unnatural offence
EXT	Extortion
BTF	Bicycle theft
COY	Counterfeit currency
OST	Offences against the state
OW	Illegal weapons
SEL	Child abuse and sexual exploitation of children
DD	Illegal drugs
OPO	Obstruction of police officers

Table 5.1: Classification codes of crimes

Code	Meaning
BGL HO1	By gaining the trust of the home owners
BGL HO2	Through windows and ventilators
BGL HO3	By breaking down the door
BGL HO4	Stealthily
BGL HO5	While the inhabitants of the house are asleep
BGL HO6	Entering through the roof
BGL HO7	By poisoning the inhabitants or their pets
BGL HO8	When nobody's home
BGL HO9	To prevent being chased
BGL HO10	Random house break ins by travelling burglars
BGL HO11	Break in to steal computers
BGL HO12	Forcibly entering a place to steal from the safe
BGL HO13	Abetting

Table 5.2: Further classification of a crime (Break in)

5.2.2 Police Database

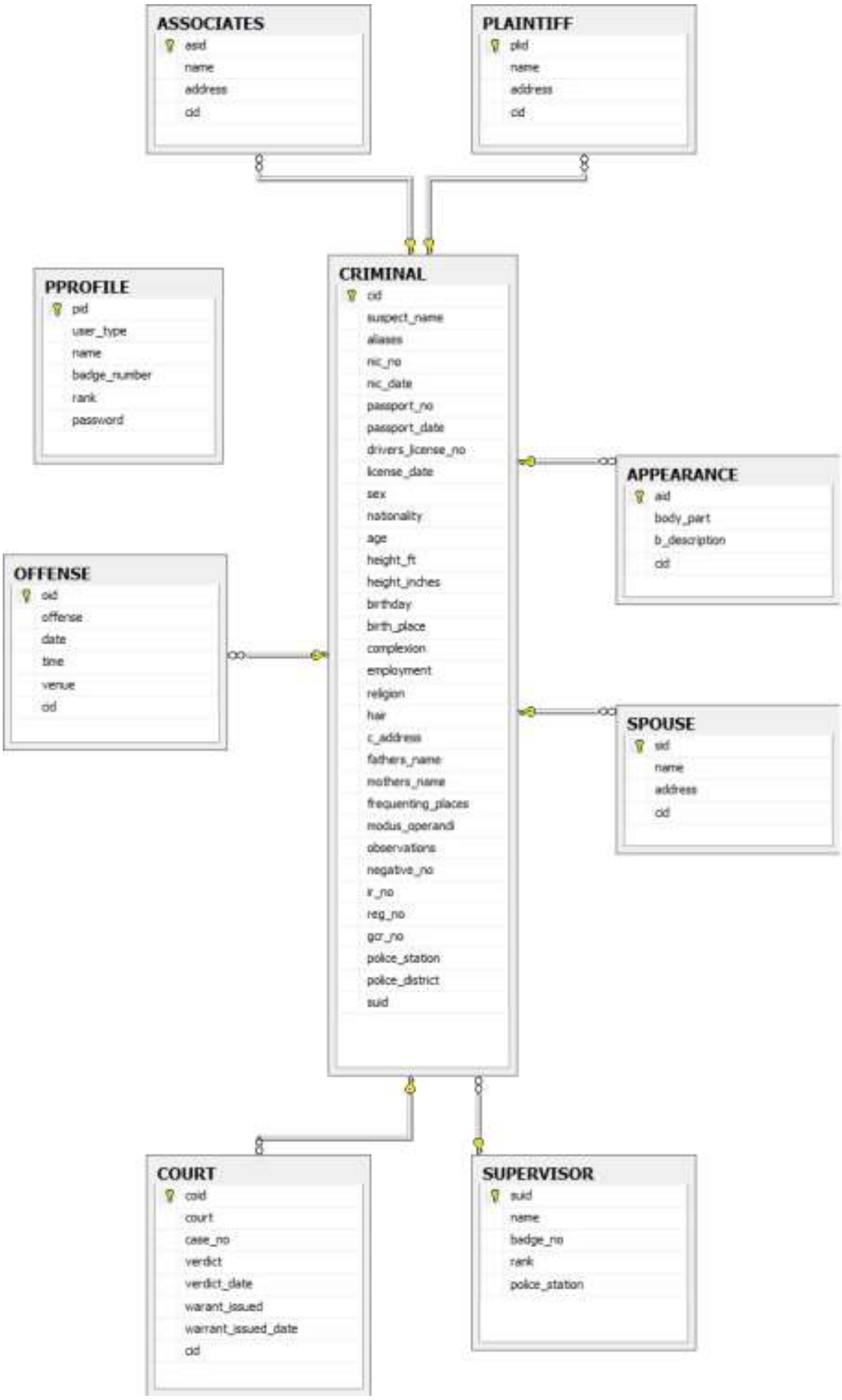


Figure 5.2: Design of the database diagram

The police database is used to store a large volume of criminal data of mixed data types which include numerical, nominal and date attributes. The data acquired are stored in tables as displayed in Figure 5.2 and retrieved through data querying. The database is maintained within a secure data server, therefore security is not a concern.

As in Figure 5.2 the database consists of nine tables; criminal, associates, plaintiff, profile, appearance, offense, spouse, court and supervisor. The criminal table stores details of the criminal, modus operandi and the police station. The offence table includes the details of the crime committed by the perpetrator. The supervisor table includes the details of the supervisor in charge of the criminal. The court table includes the details of court cases filed against the criminal. The associates and spouse table contains the information on people the criminal frequently associates with and details of the criminal's spouse. The appearance table describes the appearance of the criminal which is useful for policemen to track down the criminal. The plaintiff table contains the information of the person who accused the criminal of committing the crime. This could be an eyewitness, the police or the victim. The profile table is a separate table to store the details of the policeman who entered the relevant details of the criminal.

5.2.3 Analysis Tool

After gathering evidence and details related to the modus operandi by analyzing a crime scene the crime analysts of the CRO enters the discovered information into the system and extract the details of the most probable suspects in connection with the crime. These suspects are listed in a table sorted by the number of crimes committed matching the modus operandi with details such as their name, address and supervisor in charge. The analysts can view the details of each and every listed suspect and also view the analysis results in the form of clusters, histograms and decision trees. Criminals can be clustered using the k-means clustering algorithm based on two important factors: the modus operandi and number of crimes committed by each criminal. This information can be used to narrow down the list of suspects.

The results of the analysis are distributed in a need to know basis to the relevant departments and officers in charge of ongoing investigations and can help compare with results from previous analyses to come up with predictions and assumptions on crimes and criminals.

All this can be achieved through the analysis tool connected to the database which includes in-built algorithms and functionalities to accurately and efficiently analyze the criminal data records.

5.3 Summary

This chapter explained the design of the criminal records analysis system for decision support in detail. The top level architecture of the system gives a clear overall view of the system and the database design diagram gives a clear overall view of the database design. The next chapter will explain the implementation of the criminal records analysis system in detail.

Chapter 6

Implementation of the Criminal Modus Operandi Analysis System

6.1 Introduction

The previous chapter presented the design of the proposed solution and this chapter specifies how each component of the design is being implemented. Moreover, this chapter will provide details about the hardware, software, operating system, algorithms and important code segments utilized in the implementation of the crime records analysis system for decision support.

6.2 Software and Hardware

The system was implemented using an Intel core i7 personal computer with 4GB of RAM and a capacity of 1 TB in the Windows 7 environment which was sufficient for the task. For the implementation of the web GUIs practically any software that can be connected to the SQL database can be used. The most preferred software is Visual Studio. The SQL database is used because the current criminal records in the MIS (Arrested Monitoring Information System) of the CRD are stored in an SQL database maintained within secure servers provided by the SLT (Sri Lanka Telecom).

Due to legal issues the CRD cannot provide actual data on criminals. Therefore for data generation the Redgate SQL Data Generator was used. However, in the practical scenario the necessary information can be migrated from the MIS database to the MOS database and fill in the missing information, such as the modus operandi details, through the web portal. Nevertheless it is better to maintain a separate database for the criminal records analysis system as the MIS database is used for a completely different purpose and contains information irrelevant to the MOS system.

RapidMiner tool is used as the analysis tool because it is a well-accepted open source software that can be easily connected to the SQL database and consists of a rich library of data mining algorithms.

6.3 Graphical User Interfaces

There are five main graphical user interfaces:

- The login page
- The home page
- The profile page
- The results page
- The create profile page.

The login page is used to login to the system by providing a username and password. Only authorized police officers with a profile in the system can login. The profiles are created by the system administrators.

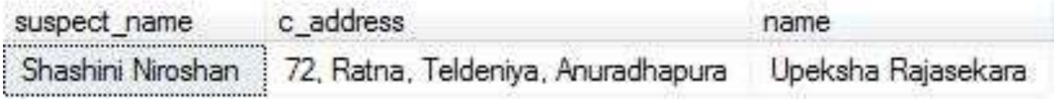
The home page is the web form through which the details of apprehended criminals are entered into the database. As described in the design chapter, the form consists mostly of drop down menus to select values from, radio buttons, text fields and date fields. Users can submit criminal data, reset and search for criminal records through the web form. Police officers can edit their profile details in the profile page.

Crime analysts of the CRO can enter the known information about the criminal through the web form or information gathered from a crime scene, which includes the modus operandi details, and the results of the search query are displayed in the results page as a table with the names of the list of suspects, address and the name of the supervisor in charge of the criminal. The list of suspects are sorted by the number of crimes committed matching the modus operandi.

E.g.

```
SELECT c.suspect_name, c.c_address, s.name
FROM CRIMINAL c, OFFENSE o, SUPERVISOR s
WHERE c.modus_operandi = 'TFT HO79h' AND o.offense = 'Theft' AND o.venue
= 'Pavement' AND c.police_district = 'Anuradhapura' AND c.cid = o.cid AND c.suid
= s.suid
```

The results of the query are displayed in a table as in table 6.1,



suspect_name	c_address	name
Shashini Niroshan	72, Ratna, Teldeniya, Anuradhapura	Upeksha Rajasekara

Table 6.1: Search query results

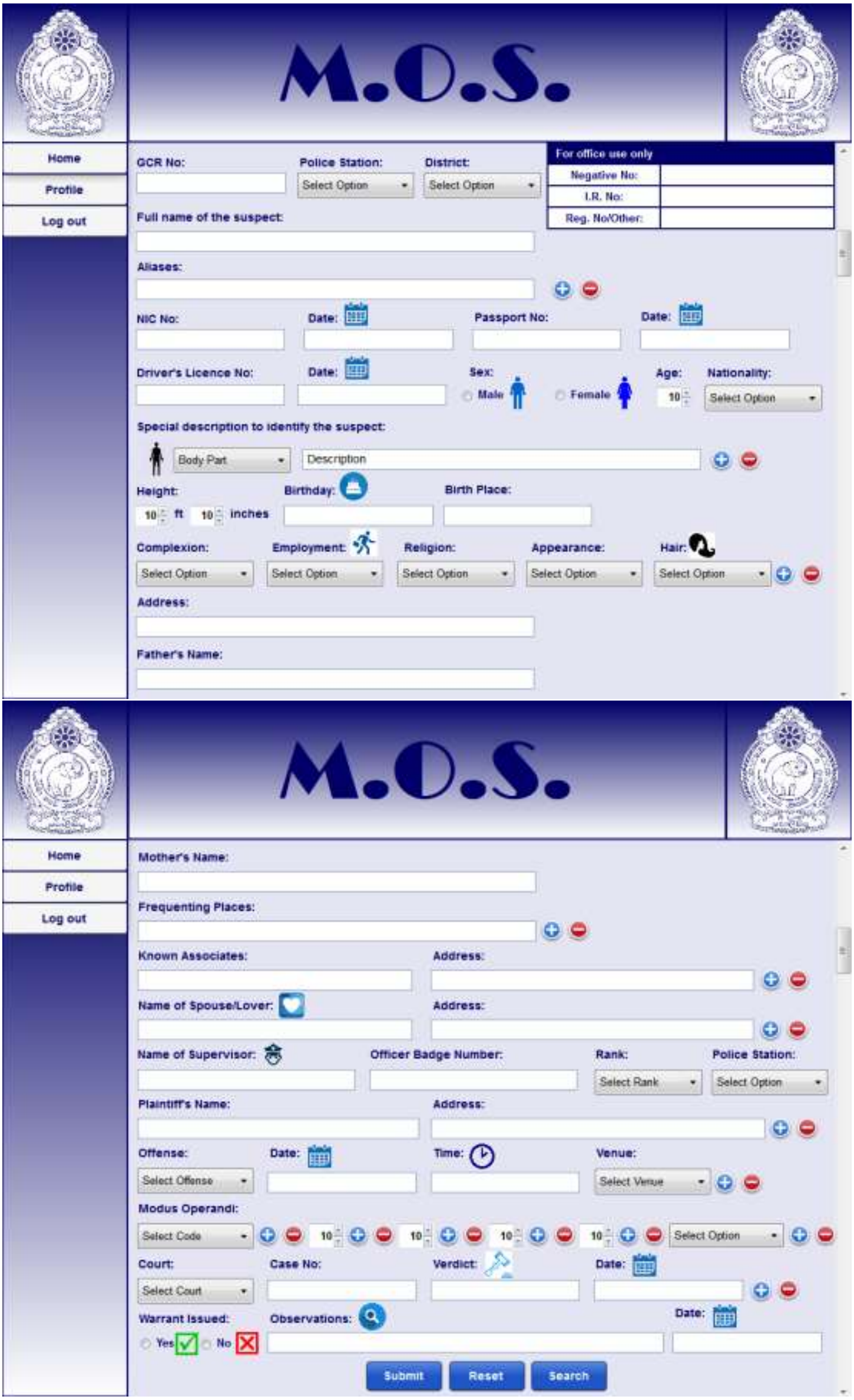


Figure 6.1: Home page of the MOS

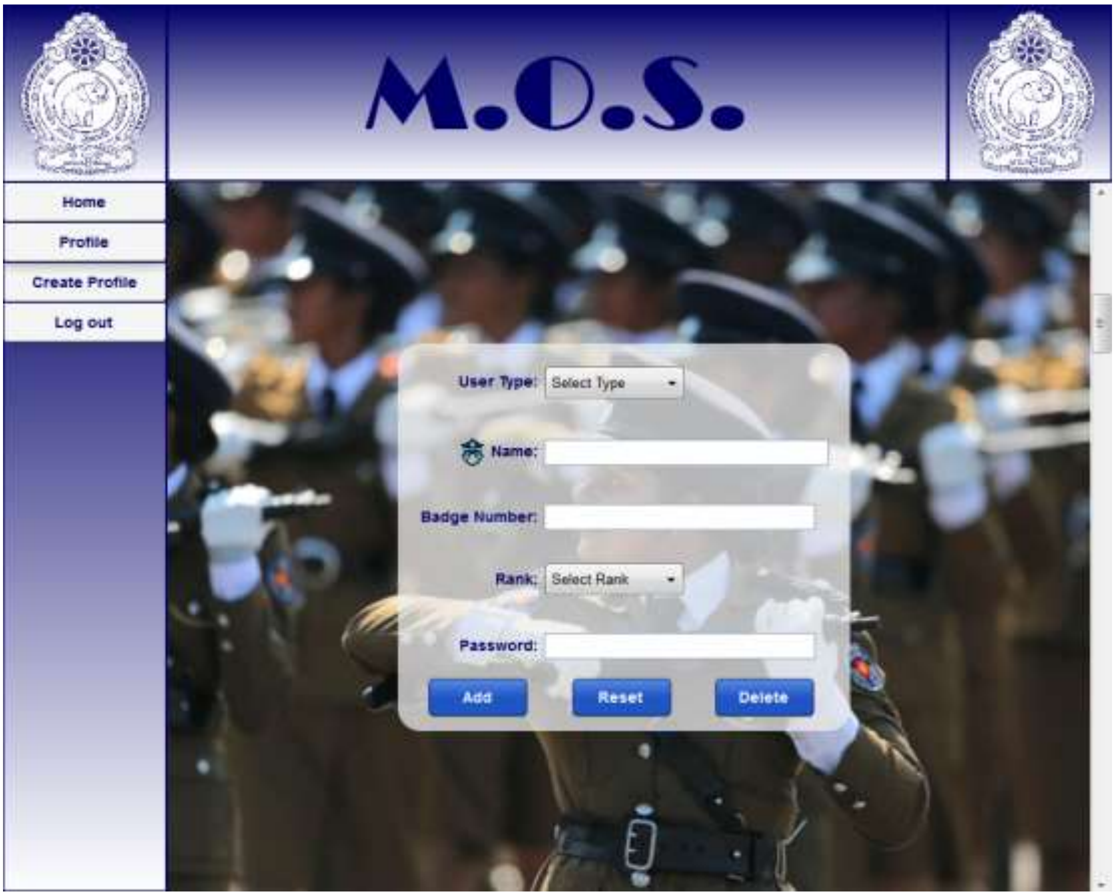


Figure 6.2: Create profile page of the MOS

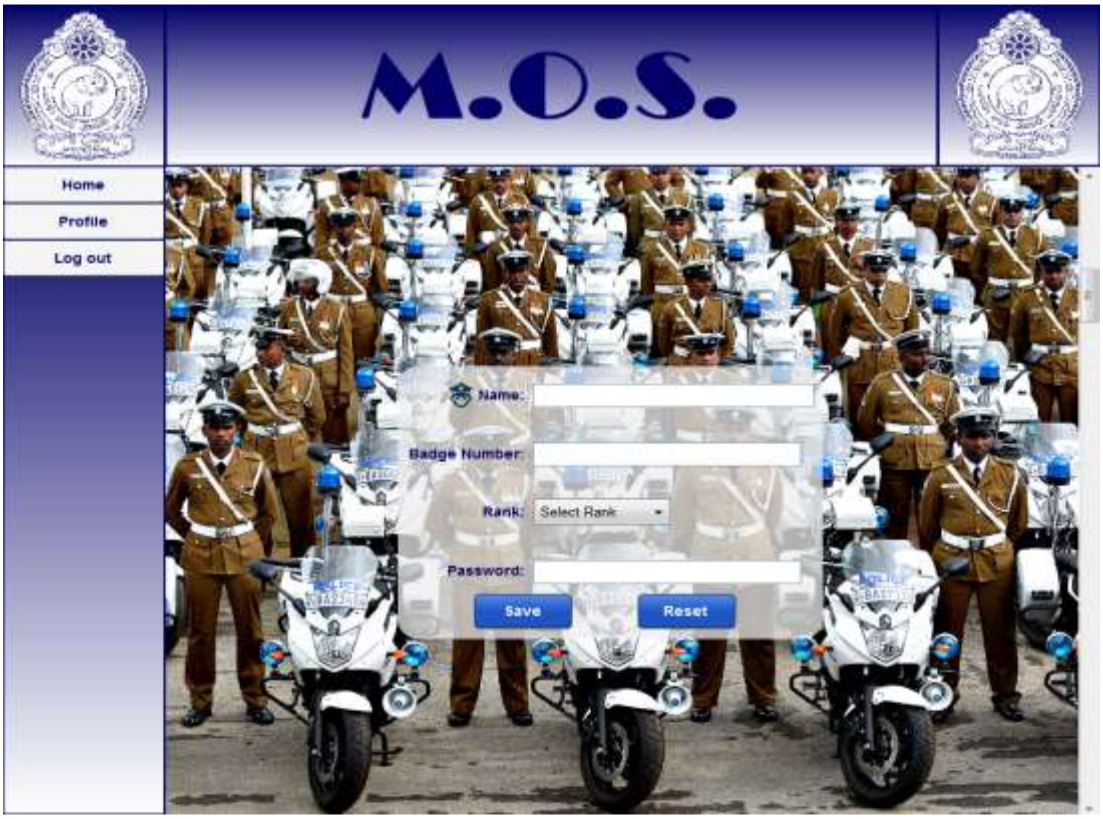
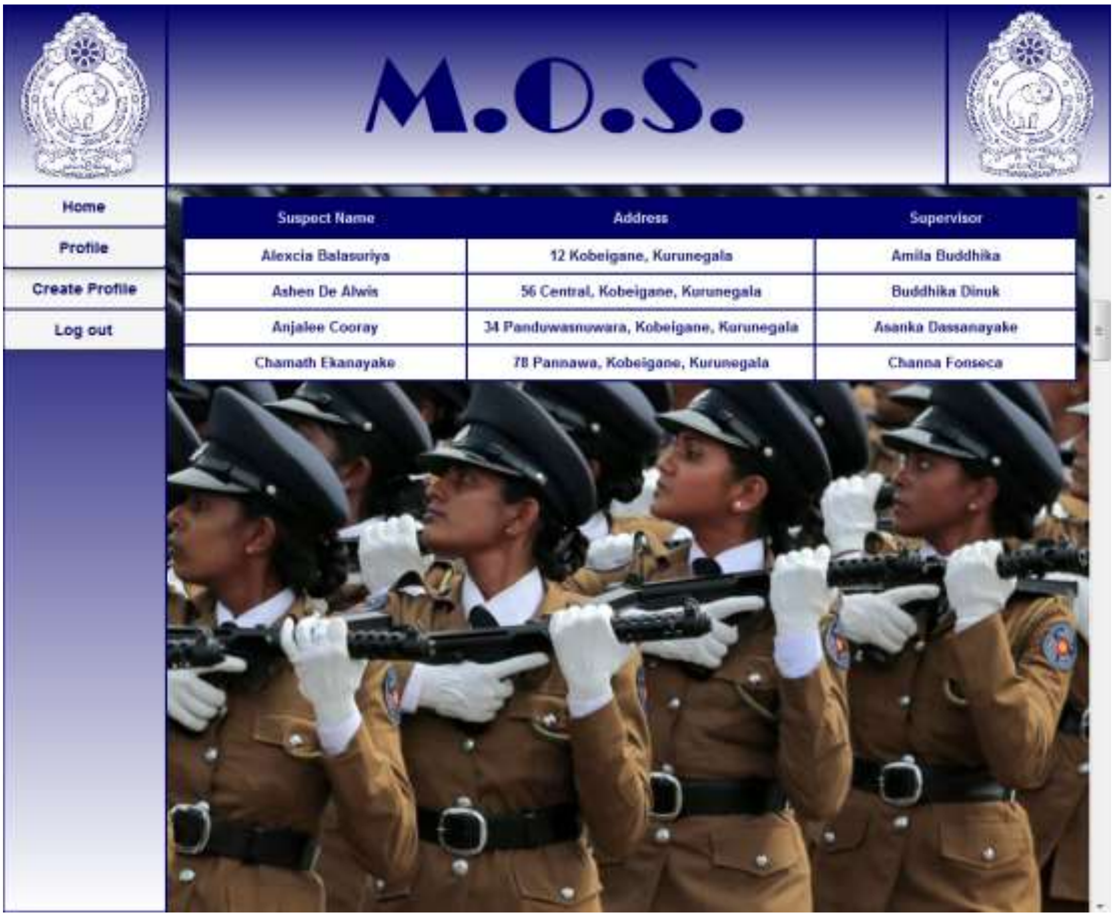


Figure 6.3: Profile page of the MOS

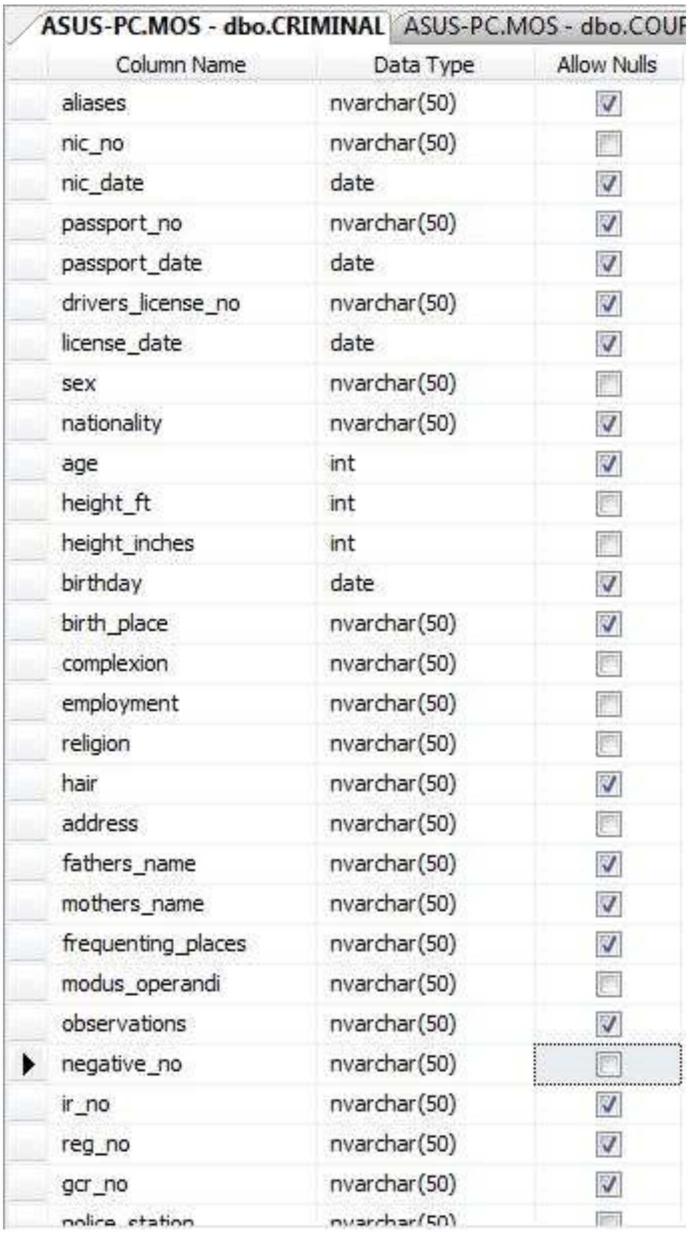


Suspect Name	Address	Supervisor
Alexcia Balasuriya	12 Kobeigane, Kurunegala	Amila Buddhika
Ashen De Alwis	56 Central, Kobeigane, Kurunegala	Buddhika Dinuk
Anjalee Cooray	34 Panduwasnuwara, Kobeigane, Kurunegala	Asanka Dassanayake
Chamath Ekanayake	78 Pannawa, Kobeigane, Kurunegala	Channa Fonseca

Figure 6.4: Results page of the MOS

6.4 Creating the Database

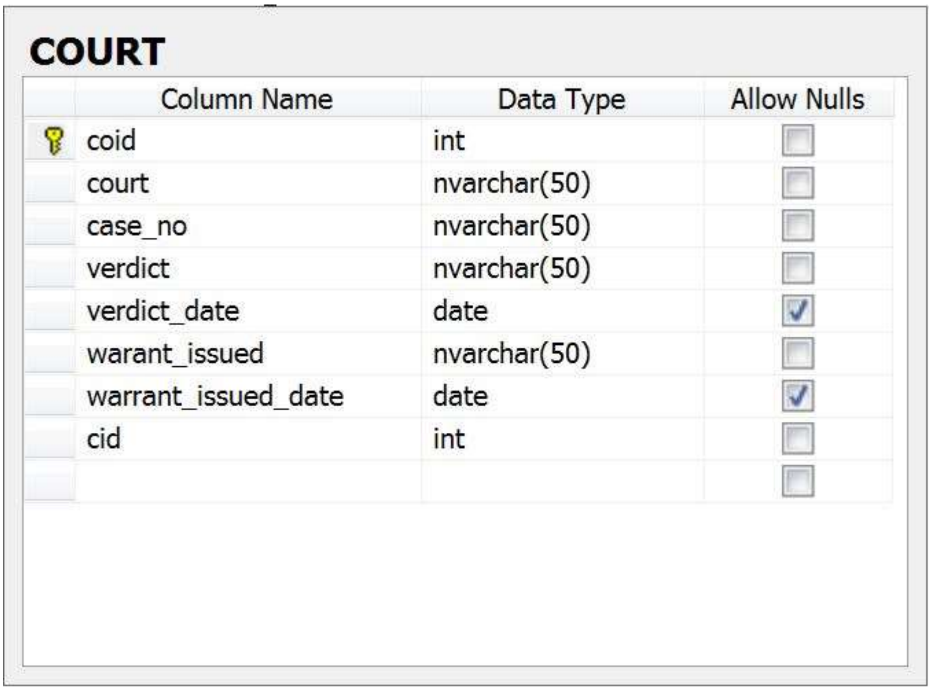
The database is designed as in Figure 5.2 in the design chapter and Figure 6.1 below using SQL Server Management Studio 2008 R2. The criminal table has been vertically partitioned into seven new tables to improve the overall performance of the database. This will reduce the time to retrieve data from a table, especially since all columns have to be selected from the table sometimes.



ASUS-PC.MOS - dbo.CRIMINAL ASUS-PC.MOS - dbo.COUF

	Column Name	Data Type	Allow Nulls
	aliases	nvarchar(50)	☑
	nic_no	nvarchar(50)	□
	nic_date	date	☑
	passport_no	nvarchar(50)	☑
	passport_date	date	☑
	drivers_license_no	nvarchar(50)	☑
	license_date	date	☑
	sex	nvarchar(50)	□
	nationality	nvarchar(50)	☑
	age	int	☑
	height_ft	int	□
	height_inches	int	□
	birthday	date	☑
	birth_place	nvarchar(50)	☑
	complexion	nvarchar(50)	□
	employment	nvarchar(50)	□
	religion	nvarchar(50)	□
	hair	nvarchar(50)	☑
	address	nvarchar(50)	□
	fathers_name	nvarchar(50)	☑
	mothers_name	nvarchar(50)	☑
	frequenting_places	nvarchar(50)	☑
	modus_operandi	nvarchar(50)	□
	observations	nvarchar(50)	☑
▶	negative_no	nvarchar(50)	□
	ir_no	nvarchar(50)	☑
	reg_no	nvarchar(50)	☑
	gcr_no	nvarchar(50)	☑
	police_station	nvarchar(50)	□

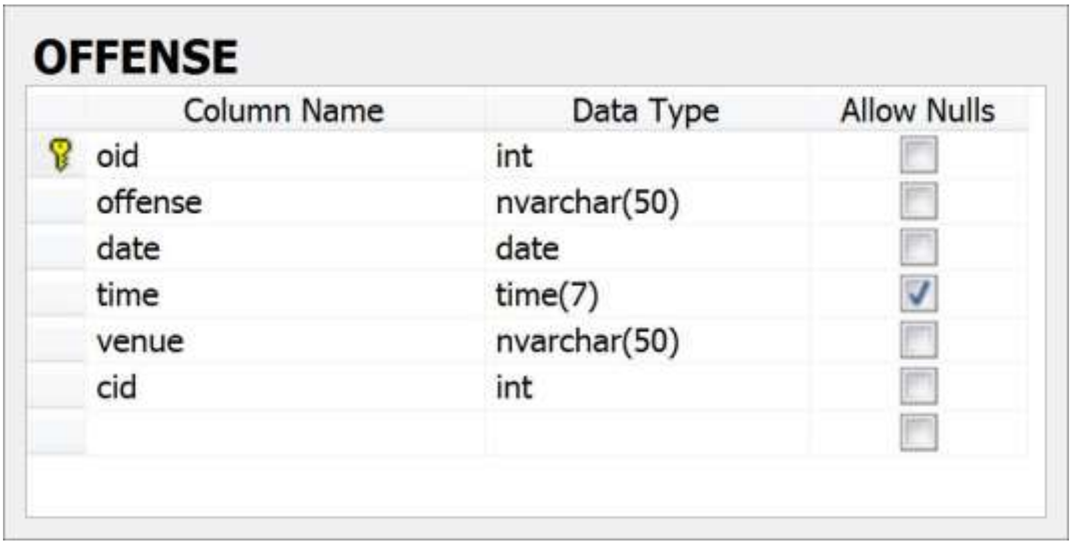
Figure 6.5: Design view of the criminal table



COURT

	Column Name	Data Type	Allow Nulls
🔑	coid	int	☐
	court	nvarchar(50)	☐
	case_no	nvarchar(50)	☐
	verdict	nvarchar(50)	☐
	verdict_date	date	☑
	warant_issued	nvarchar(50)	☐
	warrant_issued_date	date	☑
	cid	int	☐
			☐

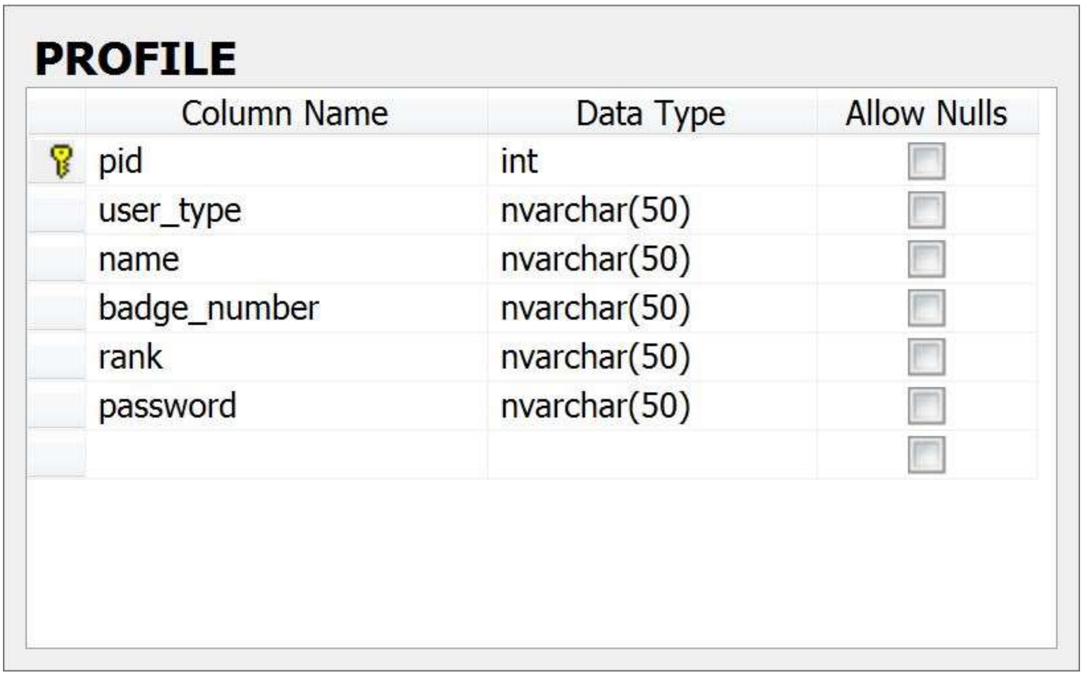
Figure 6.6: Design view of the court table



OFFENSE

	Column Name	Data Type	Allow Nulls
🔑	oid	int	☐
	offense	nvarchar(50)	☐
	date	date	☐
	time	time(7)	☑
	venue	nvarchar(50)	☐
	cid	int	☐
			☐

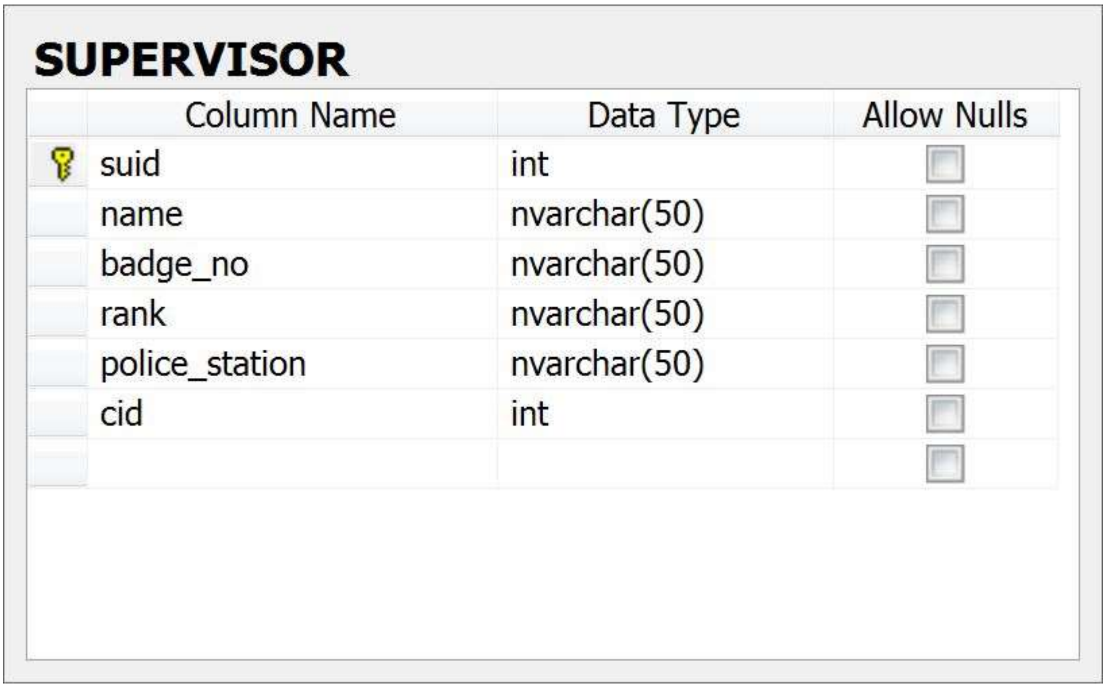
Figure 6.7: Design view of the offense table



PROFILE

	Column Name	Data Type	Allow Nulls
🔑	pid	int	□
	user_type	nvarchar(50)	□
	name	nvarchar(50)	□
	badge_number	nvarchar(50)	□
	rank	nvarchar(50)	□
	password	nvarchar(50)	□
			□

Figure 6.8: Design view of the profile table



SUPERVISOR

	Column Name	Data Type	Allow Nulls
🔑	suid	int	□
	name	nvarchar(50)	□
	badge_no	nvarchar(50)	□
	rank	nvarchar(50)	□
	police_station	nvarchar(50)	□
	cid	int	□
			□

Figure 6.9: Design view of the supervisor table

6.5 Data Generation and Pre-processing

Hundreds of data collected through onsite fieldwork and during the literature review were organized in Microsoft Office Excel sheets. Microsoft SQL constraints, such as unique and check constraints, were written and executed to accurately generate data through the SQL Data Generator.

```
E.g. ALTER TABLE CRIMINAL
ADD CONSTRAINT check_unique UNIQUE (nic_no);
```

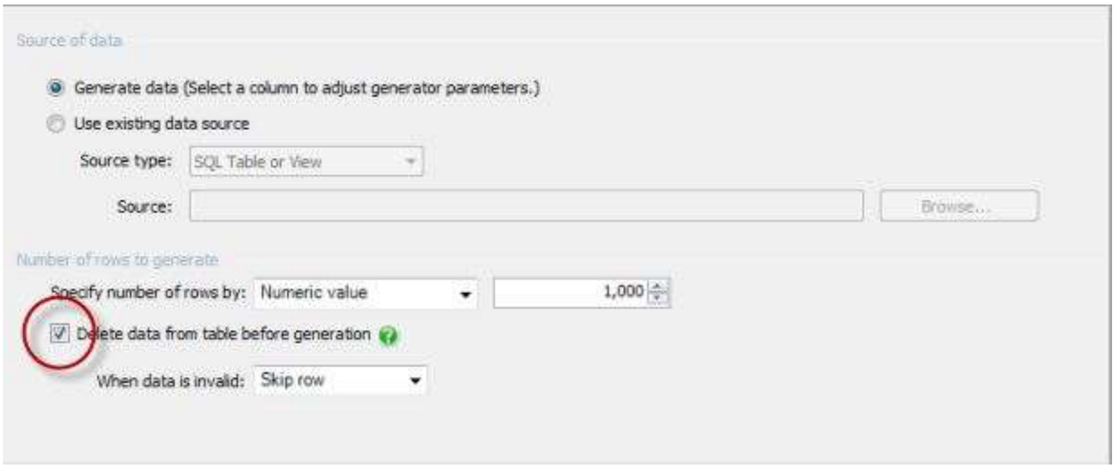
```
ALTER TABLE CRIMINAL
ADD CONSTRAINT chk_colombo CHECK (((police_district = 'Colombo') AND
(police_station = 'Angulana' OR police_station = 'Athurugiriya' OR police_station =
'Avissawella' OR police_station = 'Bambalapitiya' OR police_station = 'Bluemendhal'
OR police_station = 'BMICH' OR police_station = 'Boralesgamuwa' OR
police_station = 'Borella' OR police_station = 'Cinnamon Garden' OR police_station =
'Colombo Habour' OR police_station = 'Dam Street' OR police_station = 'Dehiwala'
OR police_station = 'Dematagoda' OR police_station = 'Foreshore' OR police_station
= 'Fort' OR police_station = 'Gothatuwa' OR police_station = 'Grandpass' OR
police_station = 'Hanwella' OR police_station = 'Homagama' OR police_station = 'Ja-
Ela' OR police_station = 'Kadawatha' OR police_station = 'Kahathuduwa' OR
police_station = 'Kandana' OR police_station = 'Kelaniya' OR police_station =
'Keselwatta' OR police_station = 'Kiribathgoda' OR police_station = 'Kirulapone' OR
police_station = 'Kohuwala' OR police_station = 'Kollupitiya' OR police_station =
'Kosgama' OR police_station = 'Kotahena' OR police_station = 'Kottawa'
OR police_station = 'Maharagama' OR police_station = 'Maligawatta' OR
police_station = 'Maradana' OR police_station = 'Mattakkuliya'
OR police_station = 'Mirihana' OR police_station = 'Modara' OR police_station =
'Moragahahena' OR police_station = 'Moratumulla' OR police_station = 'Moratuwa'
OR police_station = 'Mt. Lavinia' OR police_station = 'Mulleriyawa' OR
police_station = 'Narahenpita' OR police_station = 'Nawagamuwa' OR police_station
= 'Padukka' OR police_station = 'Peliyagoda' OR police_station = 'Pettah' OR
police_station = 'Piliyandala' OR police_station = 'Ragama' OR police_station =
'Slave Island' OR police_station = 'Thalangama' OR police_station = 'Wattala'));
```

```
ALTER TABLE CRIMINAL
ADD CONSTRAINT chk_address CHECK (((police_station = 'Akkaraipattu') AND
(c_address LIKE '%Akkaraipattu%' OR c_address LIKE '%Hingurana%' OR
c_address LIKE '%Panamkadu%' OR c_address LIKE '%Thambiluwil%')) OR
((police_station = 'Ampara') AND (c_address LIKE '%Deegavapiya%' OR c_address
LIKE '%Mayadunna%' OR c_address LIKE '%Paragahakele%')) OR ((police_station
= 'Bakkiella') AND (c_address LIKE '%Bakkiella%')) OR ((police_station = 'Central
Camp') AND (c_address LIKE '%Annamalei%' OR c_address LIKE
'%Navithanveli%' OR c_address LIKE '%Rajagama%')) OR ((police_station =
'Chavalakade') AND (c_address LIKE '%Chavalakade%' OR c_address LIKE
'%Periyakallar%' OR c_address LIKE '%Periyaneelawanai%')) OR ((police_station =
'Damana') AND (c_address LIKE '%Damana%' OR c_address LIKE '%Madana%'
OR c_address LIKE '%Thottama%' OR c_address LIKE '%Wadinagala%' OR
c_address LIKE '%Waripathancenai%' OR c_address LIKE '%Bakmitiyawa%'))
OR ((police_station = 'Dehiattakandiya') AND (c_address LIKE '%Dehiattakandiya%'
OR c_address LIKE '%Bakmeedeniya%' OR c_address LIKE '%Kadirapura%' OR
c_address LIKE '%Nawamadagama%' OR c_address LIKE '%Rideela%' OR
c_address LIKE '%Sandamadulla%' OR c_address LIKE '%Sandunpura%' OR
c_address LIKE '%Seruditiya%' OR c_address LIKE '%Thalagama%')) OR
((police_station = 'Inginiyagala') AND (c_address LIKE '%Galahitiyagoda%' OR
c_address LIKE '%Namaloya%' OR c_address LIKE '%Padagoda%')) OR
((police_station = 'Kalmunai') AND (c_address LIKE '%Kalmunai%' OR c_address
LIKE '%Malkampiddy%' OR c_address LIKE '%Maruthamunai%' OR c_address
LIKE '%Nintavur%' OR c_address LIKE '%Oluvil%' OR c_address LIKE
'%Palmunai%')) OR ((police_station = 'Mahaoya') AND (c_address LIKE
'%Mahaoya%' OR c_address LIKE '%Malwatte Wewa%' OR c_address LIKE
'%Uhapana%')) OR ((police_station = 'Padiyathalawa') AND (c_address LIKE
'%Dorakumbura%' OR c_address LIKE '%Kehelula%')) OR ((police_station =
'Panama') AND (c_address LIKE '%Cahdayanlalawa%' OR c_address LIKE
'%Mawanagama%' OR c_address LIKE '%Pallegama%' OR c_address LIKE
'%Parahagama%' OR c_address LIKE '%Tempitiya%')) OR ((police_station =
'Pothuvil') AND (c_address LIKE '%Pothuvil%' OR c_address LIKE '%Hulannuge%'
OR c_address LIKE '%Komari%' OR c_address LIKE '%Lahugala%'));
```

As data generation is a one-time thing for the purpose of research, SQL Data Generator was connected to the database and configured the data generation to automatically populate the selected tables in the database. The tool generated a random set of data for each different data type in the DB that is a realistic match for the context it’s going into. For example regular expressions had to be used to generate data for address fields as follows,

([1-9][0-9][0-9]?)(Dharmarama Mawatha |Galle Road |Katukurunda |||||)(Payagala |Kalutara South |Beruwala)(Jaffna |Kalutara |Kandy)

After the necessary configurations, 4000 records of realistic data were generated within minutes. However data preparation took 60% of the overall effort for the project.



Source of data

● Generate data (Select a column to adjust generator parameters.)

○ Use existing data source

Source type: SQL Table or View

Source:

Number of rows to generate

Specify number of rows by: Numeric value 1,000

☑ Delete data from table before generation

When data is invalid: Skip row

Figure 6.10: Generating data for numerical values

6.6 The Analysis Tool

The open source RapidMiner tool, which is a statistical and data mining package written in Java with flexible data mining support options, was used for analysis because it conducted an accurate and efficient data analysis. Before connecting the RapidMiner to the SQL database it was ensured that the SQL server browser services were running and the latest JTDS drivers were installed for Windows.

Here the Data to Similarity operator was used to measure the similarity of each criminal with the other criminals of the given data set. All the attributes related to the criminal and the crime is taken into consideration. The data was read from the SQL database using the "Read Database" process and the "Select Attribute" operator was used to select the attributes required for testing. Then the "Select Subprocess" is used to hold the cross validation of the different algorithms as in figure 6.11.

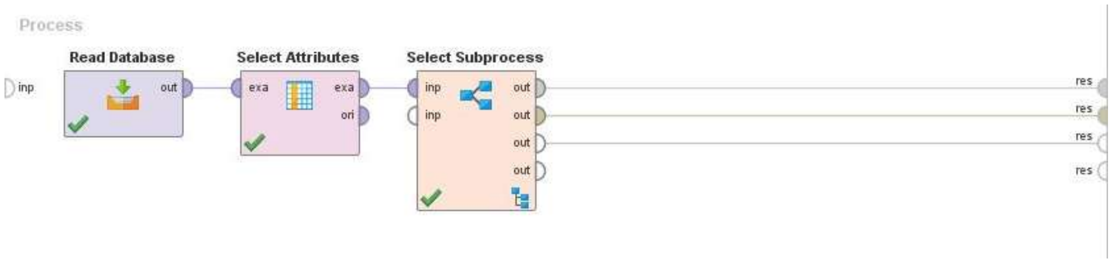


Figure 6.11: Reading and selecting attributes from the database

Inside the select subprocess, 4 "Cross Validation" operators were added for each selected algorithm and the "Set Role" operator is used to set the right roles for the selected attributes as in figure 6.12.

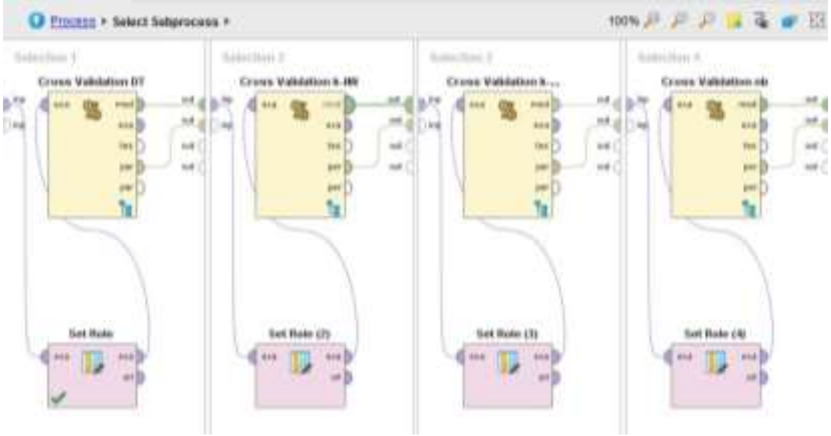


Figure 6.12: Cross validation operators

The relevant algorithms are within the Cross Validation operators as in figure 6.13 which displays the k-means clustering algorithm. “Apply Model” operator applies an already learnt or trained model on the ExampleSet and a cluster density performance operator was used to evaluate the accuracy of the clustering method used. It delivered a list of performance criteria values based on cluster densities.

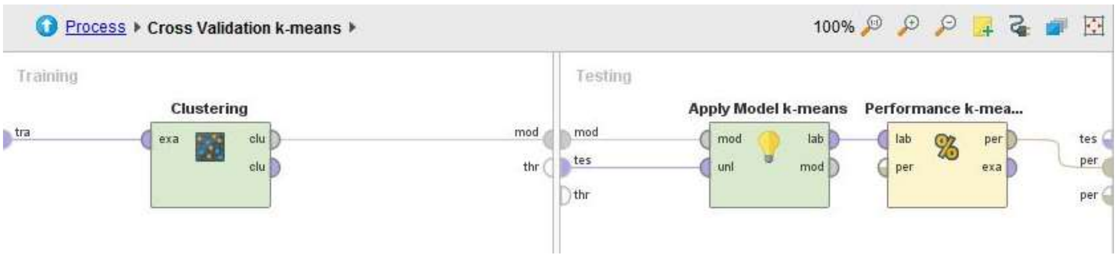
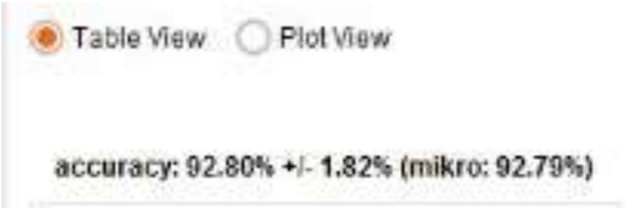


Figure 6.13: k-means algorithm

Analysis is done using the k-means clustering algorithm to cluster criminals based on their common traits. Each criminal is assigned to precisely one of a set of clusters with the nearest mean. Initially the number of clusters was defined as 5 and a set of 5 instances were selected as cluster centers. The algorithm considered the instances and assigned them to the nearest cluster. Then the cluster centroids were calculated and the entire process was repeated.

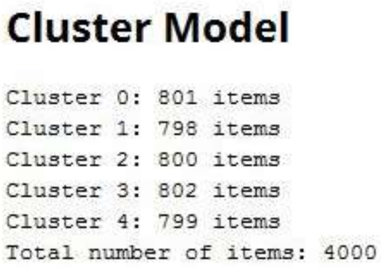
Accuracy of the k-means algorithm was displayed as in figure 6.14.



accuracy: 92.80% +/- 1.82% (mikro: 92.79%)

Figure 6.14: Accuracy of the k-means algorithm

The cluster model is displayed as in figure 6.15.

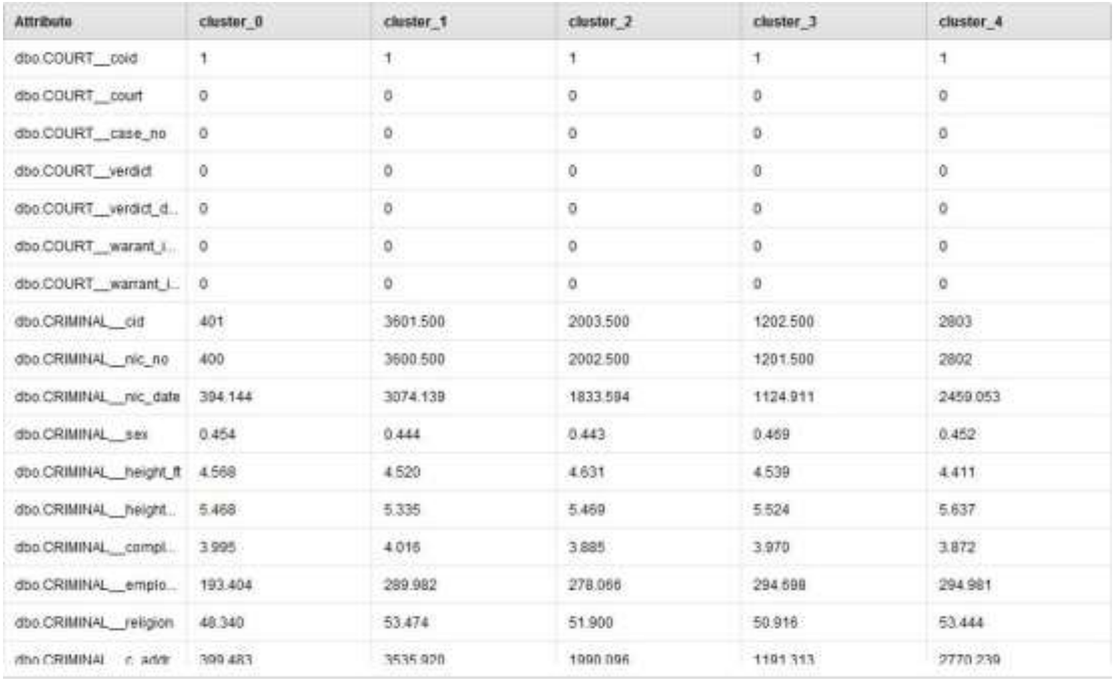


Cluster Model

```
Cluster 0: 801 items
Cluster 1: 798 items
Cluster 2: 800 items
Cluster 3: 802 items
Cluster 4: 799 items
Total number of items: 4000
```

Figure 6.15: The cluster model

The centroid table is displayed as in table 6.1.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
dbo.COURT__coid	1	1	1	1	1
dbo.COURT__court	0	0	0	0	0
dbo.COURT__case_no	0	0	0	0	0
dbo.COURT__verdict	0	0	0	0	0
dbo.COURT__verdict_d...	0	0	0	0	0
dbo.COURT__warant_i...	0	0	0	0	0
dbo.COURT__warrant_i...	0	0	0	0	0
dbo.CRIMINAL__cid	401	3601.500	2003.500	1202.500	2803
dbo.CRIMINAL__nic_no	400	3600.500	2002.500	1201.500	2802
dbo.CRIMINAL__nic_date	394.144	3074.139	1833.594	1124.911	2459.053
dbo.CRIMINAL__sex	0.454	0.444	0.443	0.469	0.452
dbo.CRIMINAL__height_ft	4.568	4.520	4.631	4.539	4.411
dbo.CRIMINAL__height...	5.468	5.335	5.469	5.524	5.637
dbo.CRIMINAL__compl...	3.995	4.016	3.885	3.970	3.872
dbo.CRIMINAL__emplo...	193.404	289.982	278.066	294.698	294.981
dbo.CRIMINAL__religion	48.340	53.474	51.900	50.916	53.444
dbo.CRIMINAL__c_addr	399.483	3535.920	1990.096	1191.313	2770.239

Table 6.2: The centroid table

The centroid plot view is displayed as in figure 6.16.

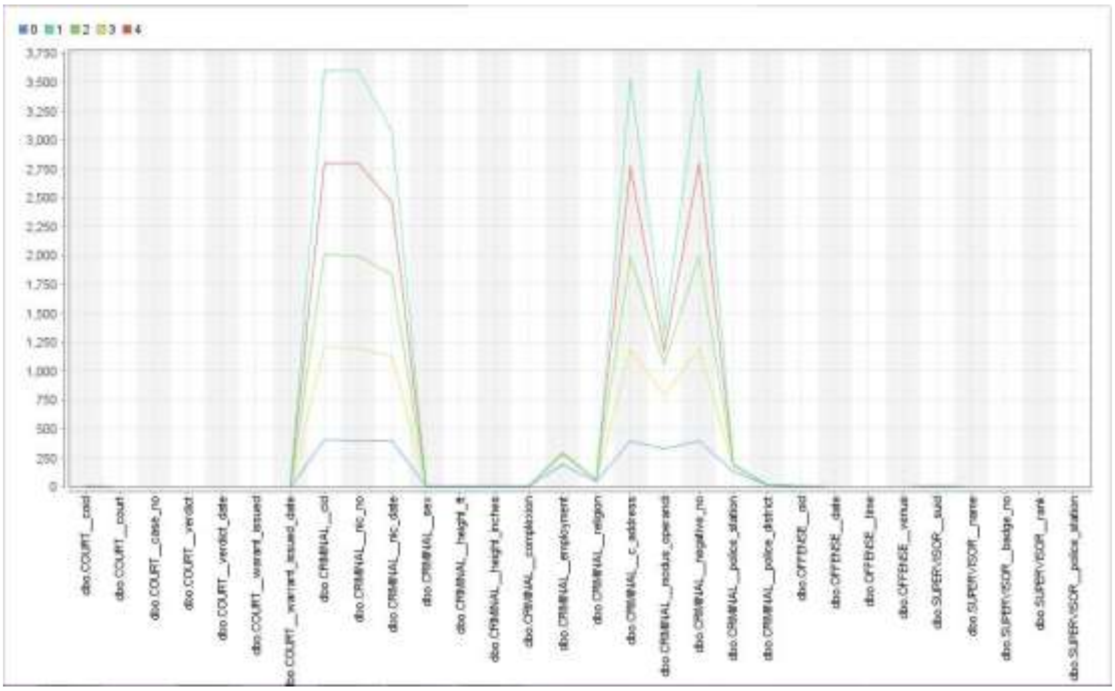
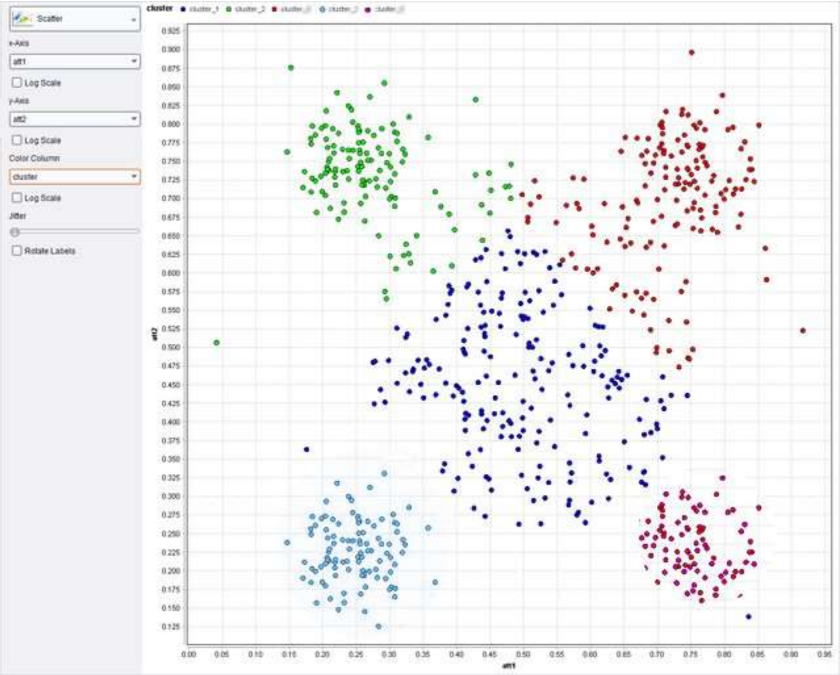


Figure 6.16: The centroid plot view

k-means clustering, can be utilized to group and develop criminal profiles to predict a list of suspects possibly connected to a crime. k-means clustering algorithm is used to cluster the criminal data because the data being clustered are of the mixed data types of numeric, nominal and date values. Therefore the measure type was selected as Mixed Measures and the centroid distance is calculated using Mixed Euclidean Distance. Compared to the other algorithms and the requirements of the criminal records analysis system, k-means clustering algorithm accurately clusters the criminals into groups.

The tool can also be used to find patterns in the crimes committed by offenders. For example the criminals can be categorized based on the number of offenses committed and the severity of crimes. It can also be used to identify the crime trends, the number of crimes committed within a given period and identifies areas with high crime. This helps identify the probability a crime might occur in a given area in the future. It graphically represented the analysis results as clusters as in figure 6.17. The analysis results can also be displayed as decision trees, histograms, bar and pie charts as in the following figures. The reporting extension of RapidMiner helped generate reports of the analysis results in html and pdf formats. To use the system, users must be connected to a network. The cost of implementation is minimal, which is an advantage for the Sri Lanka Police.



Scatter

x-Axis

att1

□ Log Scale

y-Axis

att2

□ Log Scale

Color Column

cluster

□ Log Scale

Jitter

□ Rotate Labels

Figure 6.17: Clusters of criminals based on their similarities

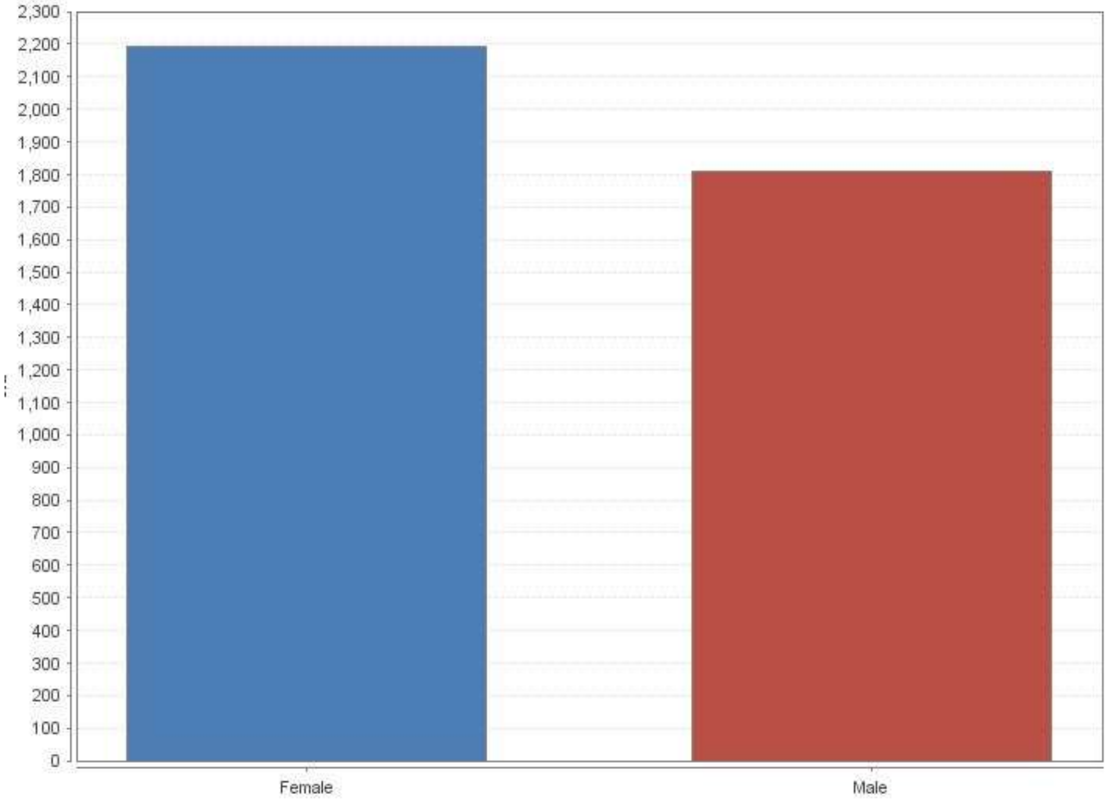


Figure 6.18: Comparison of incarcerated criminals based on their gender

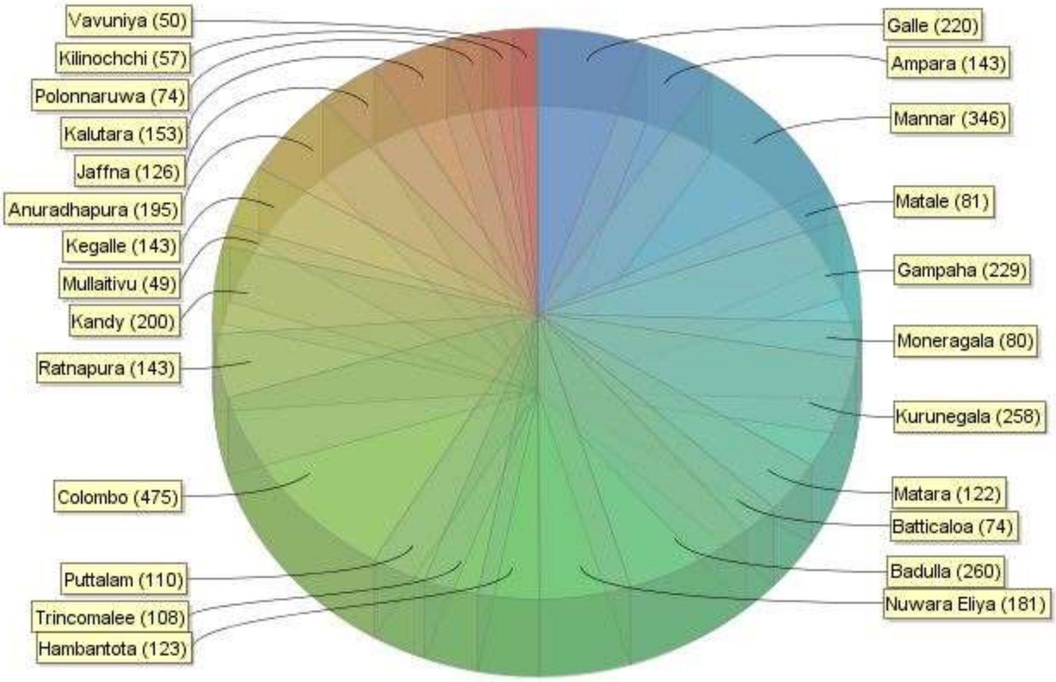


Figure 6.19: Comparison of crimes in different districts

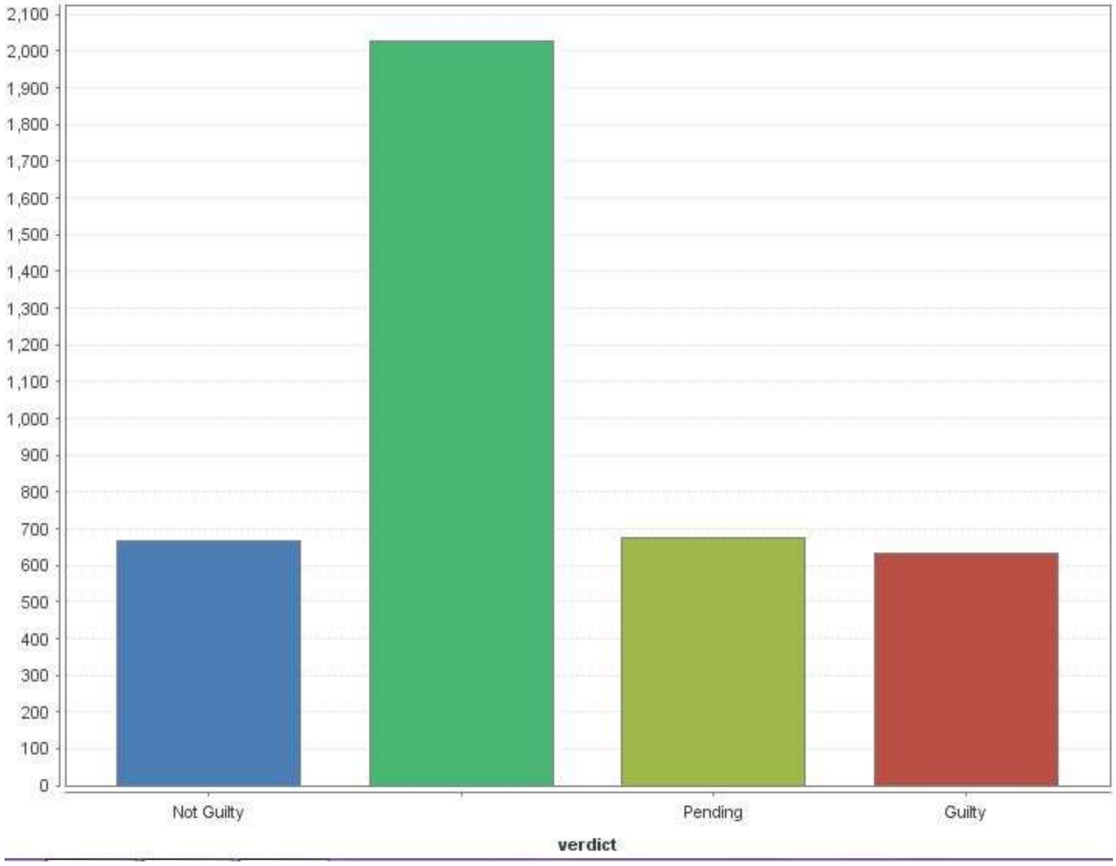


Figure 6.20: Statistics on verdicts

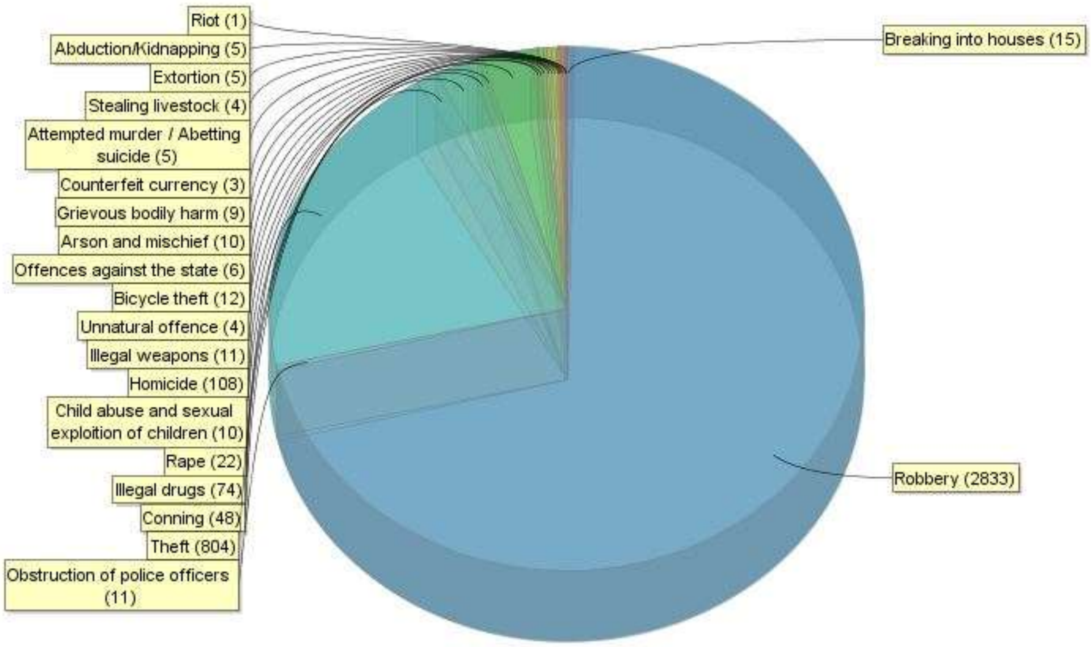


Figure 6.21: Statistics on crimes

6.7 Summary

This chapter provided the implementation of each component of the system which includes details about the software, hardware, operating system, algorithm and some code segments. The system can not only be used to extract a list of suspects in connection with a crime but it can also be used to generate criminal profiles and make predictions and assumptions of future crime by comparing the analysis obtained from the system with previous analyses. The next chapter is an evaluation of all that has been implemented so far.

Chapter 7
Evaluation

7.1 Introduction

The previous chapter explains in detail how each of the components of the system has been implemented. This chapter discusses the evaluation of the approach by assessing whether the objectives of the research have been achieved. In particular the software prototype is evaluated to analyze whether the hypothesis can be substantiated. Here the evaluation strategy is discussed in detail.

7.2 Evaluation of the Research Objectives

The objectives of this research are listed as follows,

- Create a useful crime analysis tool that can be of great service to the society.
- Help crime analysts to easily and efficiently analyze the MO and other information of criminals without having to go through each and every criminal record manually.
- Help crime analysts to easily enter and store the details of arrested criminals and descriptions of their MOs without having to manually fill forms.
- Reduce the space required to store stacks of crime records in the CRO by saving and maintaining the records in the database.
- Reduce the time and effort put in by crime analysts and make their jobs easier.
- Help the Sri Lanka Police to easily identify possible suspects connected with a crime.
- Prevent criminals from getting away with crime.

The prototype of the system has proven to be useful in accurately and efficiently analyzing criminal records. It not only reduces the time and effort taken to go through each and every record and analyze the data. It also reduces the space and costs required to manually store and maintain a physical file storage unit by storing all the criminal records in the police database maintained within a secure server. The secure server provides the necessary security for the sensitive confidential information.

By reducing the time, cost and effort to analyze criminal data helps crime analysts to provide valuable suggestions on suspects, who most probably might have committed a crime, to accurately and efficiently solve cases is a great service to the society. The simple user-friendly web GUIs allow police officers to easily enter or update the details criminals without having to do a lot of paperwork thus ensuring that criminals do not get away with crime. Therefore it can be stated that the objectives of the research have been met.

7.3 Evaluating the Results of the Analysis

RapidMiner Studio provides the means to accurately and appropriately estimate model performance. Where other tools tend to too closely tie modeling and model validation, RapidMiner Studio follows a stringent modular approach which prevents information used in pre-processing steps from leaking from model training into the application of the model. This unique approach is the only guarantee that no over fitting is introduced and no overestimation of prediction performances can occur. For this purpose the RapidMiner tool has many performance criteria for numerical and nominal or categorical targets. It also uses performance estimation for cluster models based on distance calculations and density calculations.

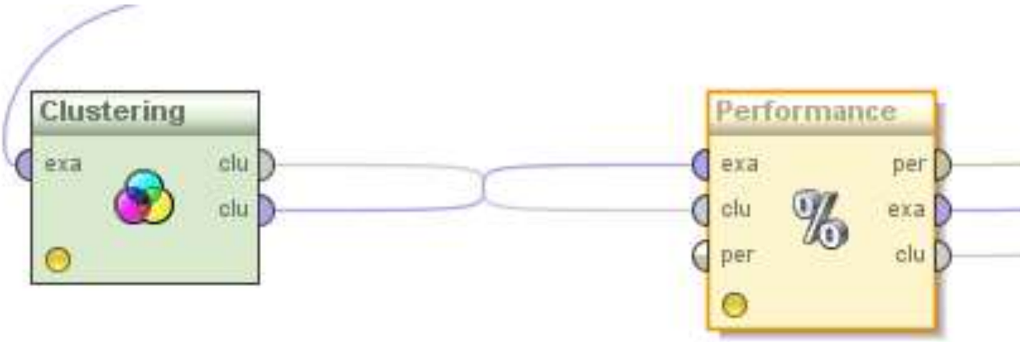
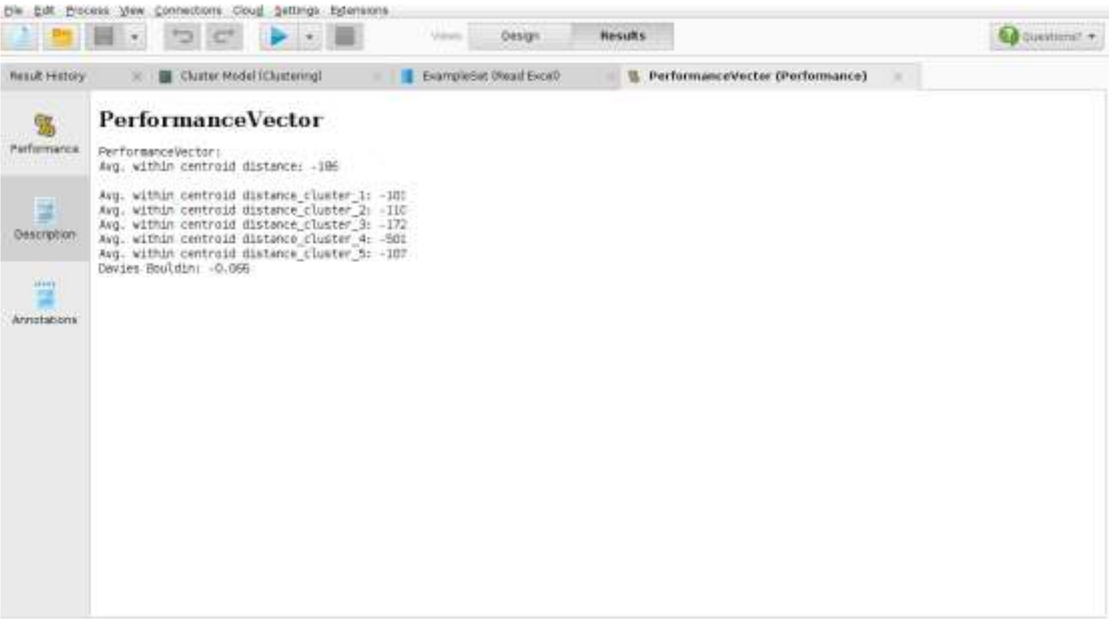


Figure 7.1: Adding performance estimation for cluster analysis



File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Result History
Cluster Model (Clustering)
ExampleSet (Read Excel)
PerformanceVector (Performance)
Performance
Description
Annotations

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -106

Avg. within centroid distance_cluster_1: -101
Avg. within centroid distance_cluster_2: -110
Avg. within centroid distance_cluster_3: -172
Avg. within centroid distance_cluster_4: -501
Avg. within centroid distance_cluster_5: -107
Davies Bouldin: -0.066
```

Figure 7.2: Results of the performance estimation for cluster models based on distance calculations

The results of the performance estimation in Figure 7.2 proves that the k-means clustering algorithm is the most suitable to cluster the criminals sharing similar traits.

Table 7.1 displays the sum of squared errors within clusters depending on the number of clusters. Depending on the values obtained it can be assumed that the sum of squared errors is consistent from five clusters onwards. Therefore the number of clusters to group criminals in using the k-means algorithm is taken as five.

Number of Clusters	Sum of Squared Error
2	936.893
3	925.907
4	807.166
5	693.118
6	691.859
7	631.819
8	508.97
9	502.344
10	469.702

Table 7.1: The sum of squared errors within clusters based on the number of clusters

Number of Clusters	Kappa	Root Mean Squared Error	Root Relative Squared Error
2	1	0	0
3	0.5950	0.1815	64.17%
4	0.5950	0.1815	64.17%
5	0.3232	0.1898	74.15%
6	0.481	0.1672	82.35%
7	0.3477	0.1486	87.45%
8	0.2590	0.1346	85.14%
9	0.3778	0.1179	85.25%
10	0.2481	0.1068	84.34%

Table 7.2: The calculations of kappa, root mean squared error and root relative squared error based on the number of clusters

Furthermore, Table 7.2 shows the calculations of kappa, root mean squared error and root relative squared error based on the number of clusters where the values become consistent after five clusters.

7.4 Summary

This chapter explains how the research has achieved its objectives as well as the evaluation of the k-means clustering algorithm using performance estimation for cluster models based on distance calculations, the sum of squared errors within clusters, kappa, root mean squared error and root relative squared error based on the number of clusters. The next chapter will explain the conclusion of the research and future work.

Chapter 8
Conclusion and Further Work

8.1 Introduction

The previous chapters explain the problem and the solution. This chapter provides the overall conclusion of the project, future work, limitations, problems encountered and how they were solved.

8.2 Overall Conclusion

The research shows that the new approach for criminal records analysis through data mining has performed well. As explained in the previous chapter the k-means algorithm used for the criminal records analysis shows improvements in efficiency and accuracy of clustering criminals in the evaluation test results done using the methods: performance estimation for cluster models based on distance calculations, the sum of squared errors within clusters, kappa, root mean squared error and root relative squared error based on the number of clusters.

As such overall conclusion supports the hypothesis that k-means clustering algorithm can offer high accuracy and efficiency in criminal records analysis. k-means algorithm partition criminals into clusters based on the mean value. The results of the performance estimation in proves that the k-means clustering algorithm is the most suitable to cluster the criminals sharing similar traits. The sum of squared error values is consistent from five clusters onwards. Therefore the number of clusters to group criminals in using the k-means algorithm is taken as five. The calculations of kappa, root mean squared error and root relative squared error based on the number of clusters also proves this. As explained in the previous chapter it can be stated that the objectives of the research have been met.

8.3 Limitations

The criminal records analysis system is developed to aid the crime analysts in the CRO and reduce the paperwork of police officers when recording criminal information. It supports the crime analysts in decision making but cannot replace the crime analysts. There is also the problem of human errors when entering data. Field validation is provided in the interfaces and check constraints on the database tables can minimize the human errors during data entry, but it can do so much. Also the web form consists mostly of drop down menus to select values from, radio buttons and date fields to make it easier for the Police officers to enter the details. Plus gathering accurate information about criminals is difficult, therefore inaccurate details maybe entered.

8.4 Further Work

Future work includes adding features such as displaying the GPS locations of criminals on a map of Sri Lanka. The addresses of suspects listed in the system in connection with a crime can be pinpointed in a map of Sri Lanka using the GPS technology. This will help the police to allocate resources from the nearest police station to where the suspect is residing to track them down. It also clearly shows the suspects closest to the crime scene.

Another such feature that can be added is the option to predict crime patterns and trends. To do so more tables need to be added to the database to hold the information of crimes. This data can be analyzed to identify past and present trends and patterns to predict future trends and patterns in crimes. Also adding an option to analyze terrorist activities is also very useful as more and more terrorist organizations are born in the world. It also helps identify, prevent and control future terrorist situations that might occur in Sri Lanka.

8.5 Summary

This chapter explains in detail the conclusion arrived at after completing the research, the future work, limitations, problems encountered and how they were solved. This study aims to help crime analysts to identify suspects in connection with crimes based on their modus operandi and make decisions on the appropriate actions needed to be taken regarding the analysis results.

.

References:

[1] A. S. Bharathi and R. Shilpa, “R.:” A survey on crime data analysis of data mining using clustering techniques”,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 2, no. 8, 2014.

[2] R. B. Fosdick, “The modus operandi system in the detection of criminals,” *J. Am. Inst. Crim. Law Criminol.*, vol. 6, no. 4, pp. 560–570, 1915.

[3] U. N. O. O. D. A. C. Vienna, *Police Information and Intelligence Systems*. Vienna International Centre, 2006.

[4] IBM, “2016-04-06 Crime Prevention Software Info with Case Studies.pdf.” IBM.

[5] E. A. Fox and R. K. France, “Architecture of an expert system for composite document analysis, representation, and retrieval,” *Int. J. Approx. Reason.*, vol. 1, no. 2, pp. 151–175, 1987.

[6] M.-K. Yip, “An Expert System for Document Retrieval,” MIT, Department of Electrical Engineering and Computer Science, 1981.

[7] M. Rogerson, “The Utility of Applying Textual Analysis to Descriptions of Offender Modus Operandi for the Prevention of High Volume Crime,” University of Huddersfield, 2016.

[8] E. Beauregard, J. Proulx, K. Rossmo, B. Leclerc, and J.-F. Allaire, “Script Analysis of the Hunting Process of Serial Sex Offenders,” *Crim. Justice Behav.*, vol. 34, no. 8, pp. 1069–1084, Jun. 2007.

[9] H. Chen *et al.*, “Crime data mining: an overview and case studies,” in *Proceedings of the 2003 annual national conference on Digital government research*, 2003, pp. 1–5.

[10] M. Chau, J. J. Xu, and H. Chen, “Extracting meaningful entities from police narrative reports,” in *Proceedings of the 2002 annual national conference on Digital government research*, 2002, pp. 1–5.

[11] J. S. De Bruin, T. K. Cocx, W. A. Kosters, J. F. Laros, and J. N. Kok, “Data mining approaches to criminal career analysis,” in *Sixth International Conference on Data Mining (ICDM’06)*, 2006, pp. 171–177.

[12] A. Blumstein, National Research Council (U.S.), and Panel on Research on Criminal Careers, *Criminal careers and “career criminals.” Volume I Volume I*. 1986.

[13] Y. Xiang, M. Chau, H. Atabakhsh, and H. Chen, “Visualizing criminal relationships: comparison of a hyperbolic tree and a hierarchical list,” *Decis. Support Syst.*, vol. 41, no. 1, pp. 69–83, Nov. 2005.

[14] H. G. Goldberg and R. W. Wong, “Restructuring transactional data for link analysis in the FinCEN AI system,” in *AAAI Fall Symposium*, 1998, pp. 38–46.

[15] D. Addison and J. MacIntyre, *Intelligent Computing Techniques: A Review*. Goldaming: Springer London, 2005.

[16] M. Williams and T. Munzner, “Steerable, progressive multidimensional scaling,” in *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, 2004, pp. 57–64.

[17] E. A. Fox, “Development of the CODER System: A Test-bed for Artificial Intelligence Methods in Information Retrieval,” p. 53, Dec. 1986.

[18] J. J. Xu and H. Chen, “CrimeNet explorer: a framework for criminal network knowledge discovery,” *ACM Trans. Inf. Syst. TOIS*, vol. 23, no. 2, pp. 201–226, 2005.

[19] K. Cios, R. Swiniarski, W. Pedrycz, and L. Kurgan, “The knowledge discovery process,” in *Data Mining*, 2007, pp. 9–24.

[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.