

University of Nairobi

School of computing and informatics

Predicting recidivism among inmates population using Artificial Intelligent (AI) techniques: A case study of Kenya prisons department

SUBMITTED BY: JUDY W. GIKARU

P58/76338/2012

SUPERVISOR: DR. C. CHEPKEN

A RESEARCH SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT OF MSC COMPUTER SCIENCE

Declaration

This project is my original work and to the best of my knowledge this research work has not been submitted for any other award in any University

Judy W. Gikaru: (P58/76338/2012) Date: _____

This project report has been submitted in partial fulfillment of the requirement of the Master of Science Degree in Computer Science of the University of Nairobi with my approval as the University supervisor

Dr. C. Chepken:	Date:
School of Computing and Informatics	

Abstract

Currently in the Kenya prisons department there is no defined way of checking the rate of recidivism among the prison inmates population. The officers rely only on manual tallying of prisoners during admission which is not efficient. With the increase use of computerized systems in the department there is need to implement those that can help in rehabilitation and reformation. In this research Artificial intelligent techniques that is decision tree, neural networks and bayesnets are used to check on the rate of recidivism in the inmate's population. This is illustrated by the development of the Recidivism Prediction System (RPS) prototype, using the WEKA tool and the python GUI application, which play a major role in risk assessment of the inmates by checking their rate of recidivism. Currently congestion in the prisons institutions is a major challenge to the management, since the resources provided doesn't match up the need on the ground. Using the RPS prototype the department management can be able to visualize various patterns on recidivism from predicted result and most importantly show the prisoners likely rate of recidivism. Assisting the users in the decision making process, as rehabilitation and reformation is not just about incarnation but also include Community Service Order and parole.

The RPS prototype is important to the users as it can be used to predict recidivism rate and plan on various programs on rehabilitation and reformation to introduce or not. As from the prototype results the prediction outcome vary from one instance to another, where those with value above TWICE are of higher recidivism risk compared to those with ONCE and below. The prediction results is also compared with other attributes and displayed for better understanding.

Acknowledgement

To the Almighty for this great gift of life so as to accomplish this far I have come.

To my loved ones; family and friends, for their great support and encouragement throughout my academic years

To my supervisor Dr. Chepken, for his support, guidance, time, and positive criticism during my research process and the panellist for their positive criticisms thus have led to success of this project.

To my classmates, who shared ideas and provided assistance during this project, I say Thank you

Table of contents

Declaration	. i
Abstract	ii
Acknowledgementi	ii
ist of Abbreviations	1
ist of figures and Tables	2
CHAPTER ONE	3
1.0 Introduction	3
1.1 Background	3
1.2 Problem statement	5
1.3 General Objectives	5
1.3.1 Specific objectives	6
1.4 Research questions	6
1.5 Justification	6
1.6 Significance of the research to Kenya Prison Department	7
1.7 Limitation and assumptions	7
1.8 Project scope	7
CHAPTER TWO	8
2. Literature review	8
2.1 Introduction	8
2.2 Importance of risk assessment	9
2.3 Overview of data mining techniques and related work1	0
2.3.1 Nearest neighbor1	0
2.3.2 Clustering1	0
2.3.3 Rule induction1	1
2.3.4 Bayesian Methods1	2
2.3.5 Neural networks1	2
The advantages of using ANN includes:1	3
2.3.6 Decision trees1	3
2.4 Data mining tools for prediction1	4
2.4.1 WEKA1	5
2.4.2 KNIME	5

2.4.3 Rapid Miner	
2.4.4 Orange	15
2.5 Summary of Literature Review Findings	16
CHAPTER THREE	19
3. Methodology	19
3.1 Introduction	19
Overview of CRISP-DM methodology	19
3.2 Research analysis and design	20
3.2.1 Requirements Analysis	20
3.2.2 Data collection and analysis	22
3.2.3 Data preparation	25
CHAPTER FOUR	27
4.0 Prototype development	27
4.1 Introduction	27
4.2 Prototype development Process	27
4.2.1 WEKA Tool	27
BayesNets	27
J48	29
Multilayerperceptron	
4.2.2 The Graphical User Interface application	34
Functionalities	35
CHAPTER FIVE	
5. Results	
5.1 Introduction	
5.2 System Evaluation	
5.3 System testing	
5.3.1 User acceptance testing	40
CHAPTER SIX	42
6. Conclusion, Recommendation and Future Works	42
6.1 Conclusion	42
6.2 Recommendation	43
6.3 Future works	

References	45
Appendices	48
Appendix A – Interview Questions	48
Appendix B– Sample code	49

List of Abbreviations

ANN- Artificial Neural Networks **API** – Application Programming Interface CART – Classification & Regression Tree **CART** – Classification and Regression Tree CRISP-DM- Cross Industry Standard Process for Data Mining **CT**- Classification Tree **DA**- Discriminate Analysis **DM** - Data Mining **DM**- Data mining FSOM – Fuzzy Self Organizing Map **ICT-** Information Communication Technology LR- Logistic Regression **NN-** Neural Networks **ORMS** – Offenders Record Management System **PMML** – Predictive Model Markup Language **RPM**- Recidivism Prediction Model **SEMMA-** Sample Explore Modify Model and Assess **SKU** – Store Keeping Units **SOM** – Self Organizing Map **CSV-** Comma Separated Values **ARFF-** Attribute-Related File Format

List of figures and Tables

Figure 1: Data mining steps

Figure 2: CRISP – DM

- Figure 3: The architecture design of the prototype
- Figure 4: The raw data from the ORMS

Figure 5: The processed data to be loaded to WEKA

Figure 6: Prediction results of BayesNet

Figure 7: Second part of prediction results of BayesNet

Figure 8: Prediction results using decision tree

Figure 9: Second part of the J48 prediction results

- Figure 10: Graphical representation of MLP
- Figure 11: Prediction of the results of Multilayerperceptron
- Figure 12: Results of the Multilayerperceptron
- Figure 13: Report generated by the GUI application
- Figure 14: Graph on previous conviction prediction and occupation
- Figure 15: Report on male convicts on rate of recidivism

Figure 16: Graphical representation on age and previous conviction prediction

Tables

Table 1: Tabulations results from the WEKA algorithms

CHAPTER ONE

1.0 Introduction

1.1 Background

The Kenya prison department is a correctional service which is mandated by the constitution the responsibility of safe custody of both convicted and un-convicted prisoners. It has a total of 108 penal institutions countrywide and a total of 109,629 convicted prisoners this is as at 2014 which is a 41.6 percentage increase from 77,405 in 2013. For the previously convicted population in 2014 was 24,927 a 8.8 percentage increase from 22,910 in 2013 (KNBS, 2014). The population increase of the prisoners has resulted in congestion in most penal institutions mostly due to the fact that the infrastructure growth does not match that of the population among other factors. Therefore, there is the need for a system to help manage the population of inmates in the penal institutions to complement the existing methods.

The Recidivism Prediction System is to help the Kenya Prison department in its operations to study the cases of a person being released and the chances of being convicted again. For example the system could aid in the adoption of a policy based on the prevention of recidivism, adequate release planning and referrals to community based services among others. The risk levels of a person's chance of committing another crime after release will be helpful to the department in decision making on scenarios of labor allocation, Compulsory Supervision Order and parole among others.

Details of the prisoner like age, gender, offence committed, area of residence, education background and employment among others are fetched from the ORMS (Offenders Record Management System) which is maintained by Kenya Prison Department and used as variables in predictions on a prisoner's history of arrest.

The results will help the department meet its core functions effectively, and ensure public safety and effective rehabilitation of the offenders. With the rise in the number of the convicted persons in the penal institutions, there is need to increase the budget allocation among other resources for the persons to be effectively rehabilitated. By prediction of recidivism and its risk level of the inmates the department can segment those who need incarnation and those that can be sent on community supervision order among others depending on their level of risk to the society.

Recidivism is the act of a person repeating an undesirable behavior after they have either experienced negative consequences of that behavior, or have been treated or trained to extinguish that behavior. It is also used to refer to the percentage of former prisoners who are rearrested for a similar offense (Hensil J., 2008).

Recidivism is one of the most fundamental concepts in criminal justice. It refers to a person's relapse into criminal behavior, often after receiving sanctions or undergoing intervention for a previous crime. Recidivism is the most common outcome (dependent) variable in all of criminal justice research and the rate determines the success or failure of a correctional system (O'Connor, 2013).

A research by (Gray, Birks, Allard, Ogilvie, Stewart and Lewis,2008) states that risk assessment procedures occupy a central role in the Criminal Justice System decision making process and typically involve a prediction about the likelihood that an individual will re-offend.

Use of data mining techniques like decision trees and neural networks has proved to have the potential of improving prediction accuracy of risk assessment compared with the traditional statistical techniques like the regression model, because with model efficiency prediction results will be of great significance to the public safety and offender rehabilitation.

A study by (Howard, 2000) states that Canadian criminal justice system relies heavily on prediction of risk though inherently error prone, due to the fact that there are no 'laws 'of behavior that can be applied to a set of circumstances to determine the behavioral outcome that will follow. Criminal behavior in particular is motivated and supported by an unquantifiable number of factors; therefore to assess an individual's as 'high risk' is not to say that he/she will definitely recidivate. Despite its shortcomings, risk assessment can to a certain extent, differentiate offenders who pose a significant risk for re-offending in the future from those who are likely to refrain from committing future offenses.

The Recidivism Prediction System prototype for this research is developed using the WEKA software package for its full functionality as it includes API, Database system support, visualization, PMML support, statistical capabilities among others. More so WEKA is highly

robust for a variety of users irrespective of their knowledge level in data mining and the fact that it's readily available as its open source. Together with a Front end application for better and easier visualization of the predicted results to help the management in decision making.

1.2 Problem statement

Currently Kenya Prison Department is the correctional service provider in Kenya with a number of mandates among them being containment and safe custody of inmates, rehabilitation and reformation of prisoners, facilitation and administration of justice among others. As from 2010 to 2014 the inmate's number in Kenya prisons varied between 56,051 and 109,629 and for the recidivism during the same duration range between 12,949 and 30,547 (KNBS, 2014). Therefore there is the need to have a number of ways to check recidivism. One of them will be a system to check recidivism among the inmate population which would be more accurate and efficient. By predicting the level of risk of an offender re-offending to help in determining whether an offender can be sent on various programs like parole and community service order among others, thus helping in dealing with the congestion in various prisons institutions countrywide.

The system will solely provide the Kenya Prisons Department management with more insightful information to aid in decision-making process of the day-to-day running of the department operations.

This is especially with the convicted prisoners who sole responsibility lays with the Kenya prison department until they have completed their sentence.

Currently there is no existing system in place to predict recidivism in the prison department. What exists is the use of manual and some features from the ORMS which are not specific, nor are they efficient and effective.

1.3 General Objectives

The purpose of this project was to develop a RPS (Recidivism Prediction System) prototype using Artificial Intelligent (AI) techniques to check recidivism among the inmate's population with an aim to help the prison department management in decision making.

1.3.1 Specific objectives

- 1. To identify and analyze the variables used in predicting recidivism in the prison inmates population
- 2. To identify a data mining technique suitable to predict recidivism in the prison inmates population
- 3. To develop a prototype application using an identified data mining technique
- 4. To test and validate the prototype

1.4 Research questions

- 1. Which is the suitable technique to use to predict recidivism in the Kenya prison population?
- 2. What variables in the provided dataset that most determine the probability of recidivism in the Kenya prison population?
- 3. How can data mining techniques be used in recidivism prediction?

1.5 Justification

Kenya Prison Department, being a Government agency, is guided by the current Kenya Vision 2030 project which puts much emphasis on technology development by using Information Technology. This is to make work easier and manageable as there is a tremendous increase in data volume. On security one of the goals includes installation of effective ICT infrastructure in all security agencies which can be achieved by a crime prevention strategy, by use of ICT (Government of Kenya, 2007).

The System will assist the department run its operations effectively considering the increase in population and the resources allocated which may not be enough and most importantly be able to utilize other modes of rehabilitation apart from confinement of prisoners. As a result the prisoner is rehabilitated and reformed to be able to re-integrate back to the society.

To the society the system will be helpful as there will be a reduction in resources used to cater for the prisoners while confined as the population is bound to decrease.

1.6 Significance of the research to Kenya Prison Department

It will provide knowledge to help the department management in decision making this is especially in adoption of various policies like on Compulsory Supervision Orders (Cap. 90, Rev 2009), parole, and pardon (Power of Mercy Act 2012 part III section 47 1 (a)) among others.

Also provide a foundation for studying the prisoner's criminal careers and may provide insight into effective reentry programs.

1.7 Limitation and assumptions

The Recidivism Prediction System takes into consideration all offenders even the life and death sentenced with the assumption that at some point there are those that appeal and are released or sentence reduce.

The system will not take into consideration of the pretrial detainees/remands prisoners as despite them being confined in the prisons their release is determined by the courts and there is the likely of the person not being sentenced as the case is ongoing.

1.8 Project scope

This project was based on selected number of prisons within the Kenya Prison department they include Nairobi medium prison, Nairobi west prison and Langata women prison. Why the stated prison considering that there around 108 prisons country wide, due to their proximity and data availability and the time given to conduct the research is limited.

The system is intended for the management team in the department; the commissioner General of Prisons, directors and the officers in charge heading the prisons countrywide.

CHAPTER TWO

2. Literature review

2.1 Introduction

In this section various techniques used in data mining for prediction are discussed and previous work which has been done on the subject.

Data mining technology has been used in various fields like business, games, science & engineering, medical among others with the goal to extract information from a data set and transform it into an understandable structure for further use. The technology has shown to be a powerful and effective methodology to help business users facilitate intelligent decision support. In particular it enables criminal investigators to explore criminal acts quickly and efficiently (Li, Kuo and Tsai, 2010).

Data mining process is best thought of as a set of nested loops rather than a straight line. The steps do have natural order, but it is not necessary or even describes to completely finish with one before moving on to the next. The tasks involved in data mining include: classification, estimation, prediction, affinity grouping and clustering (Berry, Linoff, 2010 pg 44).



Fig 1: Data mining steps (Zaiane, 1999)

Various studies have been undertaken on recidivism especially on the risk assessment at different angle; that is female recidivism, male sexual offender's recidivism, juvenile's recidivism among others using various methods like anamnestic, clinical, and actuarial. Anamnestic (recollection) methods use historical data to determine the future actions of an individual. Clinical methods involve the human judgment of professionals such as probation officers and psychologists to make risk assessments. Actuarial methods use quantitative analyses of individual characteristics to determine risk. Both clinical and actuarial methods are commonly used today, but studies have shown that the actuarial risk prediction consistently outperforms the results of clinical risk prediction this is as stated by Gettredson &Moriarity, (2006) cited by Harris, menus, Obradovic, Izenman, Gruwald, Lockwood, Jupin and Chisholm, (2012).

The actuarial methods are more efficient as research findings consistently indicate that decisionmaking based on actuarial risk assessment tools is more accurate, valid and reliable than clinical decision-making this is by Ægisdottir, White, Spengler et al., 2006; Dawes et al., 1989; Gambrill & Shlonsky, 2000; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson, 2005 as cited by Gray, Birks, Allard, Ogilvie, Stewart and Lewis, 2008).

2.2 Importance of risk assessment

The risk assessment on the likelihood of re-offending in the justice system is highlighted by broad range of processes that require assessment and given its role in improving public safety and offender rehabilitation. The processes that require risk assessment includes; bail, sentencing, prisoner classification, parole, the case management and supervision of community based orders and the provision of effective treatment (Silver & Miller, 2002; Gottfredson & Moriarty, 2006). This is because any improvement in the ability to accurately assess risk would improve the efficiency of criminal justice decision making. Risk assessment provides a useful tool for the attainment of public safety by enabling the identification of offenders who pose an elevated risk of recidivism who require greater supervision. Consistent with the principles of best-practice for offenders receiving intensive interventions and low-risk offenders receiving either none or minimal interventions (Andrews et al., 2006; Gray et al., 2008).

2.3 Overview of data mining techniques and related work

Data mining is the process of identifying interesting patterns from large database. It is best described as an iterative and exploratory process achieved through either automated or manual methods. The two primary roles of data mining are prediction, which involves the use of variables to predict unknown future events or values of a given outcome and description involving the identification of patterns that describe the data in a meaningful manner (Gray et al., 2008).

Data mining involves using a range of techniques which are stated in various approaches like statistical, mathematical algorithms, database oriented and machine learning among others to examine potential relationships in data sets and are often used to form predictive models of either continuous or categorical variables.

According to Gray et al., (2008), some of the more common data mining methods include neural networks, decision trees, support vector machines and algorithms for mining association rules. The algorithms can be classified according to the various distinction like; methods used to discover predictive relationships for categorical variables (i.e.: classification methods), methods used to discover predictive relationships for numeric variables and methods of association rule discovery.

2.3.1 Nearest neighbor

This is among the oldest technique used in data mining. It has similarity with clustering as its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record.

It is among the easiest to use and understand because they work in a way similar to the way that people think, by detecting closely matching examples (Berson, Smith, & Threarling, 2000).

2.3.2 Clustering

Clustering is the methods which like records are grouped together. This is done to give the end user a high level view of what is going on in the database. Mostly applied in the business area of marketing where it's believed to give one a bird eye view of the business happenings. The main difference between the two techniques that is clustering and nearest neighbor being one is called unsupervised learning technique and the other supervised respectively. Where the unsupervised learning techniques has no particular reason for the creation of models the way there is for supervised that are trying to perform prediction (Berson et al., 2000).

2.3.3 Rule induction

It is one of the major forms of data mining and perhaps most common of knowledge discovery in unsupervised learning systems as when applied to a database its helpful in that it can allow possible patterns which are systematically pulled from data and added accuracy and significance. The retrieval of all possible interesting patterns in the database is a strength in the sense that it leaves no stone unturned but also a weakness as users can easily become overwhelmed with such a large number of rules that it's difficult to look through all of them.

Mostly is used on databases with either fields of high cardinality or many columns of binary fields like from the retail shops that is supermarket basket data from store scanners that contains individual product names and quantities and may contain tens of thousands of different items with different items with different packaging that create hundreds of thousands of SKU identifiers (Berson et al., 2000).

According to Li et al., (2010), the framework of intelligent decision support model based on a fuzzy self organizing map network to detect and analyze crime trend patterns from temporarily crime activity data. It also incorporates rule extraction algorithm to uncover hidden casual effect knowledge and reveal the shift around effect. It is intended to identify crime trend pattern for different criminal activities, conduct temporal rule extraction to uncover their shift around effect and provide a reference for experts when analyzing the different types of crimes. The FSOM model is used to discover crime pattern which combine the features of SOM networks and fuzzy logic in dealing with clustering, visualization and linguistic information processing. The rule extraction algorithm is used to find the hidden casual effects between different temporal linguistic crime data that can help police management understand more clearly the criminal acts. Thus providing actionable information for the police management to make better use of its duty deployment and help criminal experts to develop and implement more effective law enforcement policies and crime control programs.

2.3.4 Bayesian Methods

Bayesian approaches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can probably achieve the optimal result. Bayesian method is based on the probability theory. One limitation that the Bayesian approaches cannot cross is the need of the probability estimation from the training dataset. It is noticeable that in some situations, such as the decision is clearly based on certain criteria, or the dataset has high degree of randomality, the Bayesian approaches will not be a good choice.

According to Blattenberger, Fowles and Krantz, (2010) where they use various Bayesian statistical methods the Bayesian model averaging, extreme bounds analysis and classification & regression tree. This is to explore criminological, sociological and economic factors to predict parolees' returns to prison by comparing their results to provide useful public policy guides. The results from the extreme bounds analysis and Bayesian model analysis may differ from those of the classification and regression tree in that they are based on traditional Bayesian linear specifications within the context of a normal gamma conjugate framework. The Bayesian CART model does not necessarily lead to terminal tree nodes that have high degree of homogeneity. Using extreme bounds analysis one is able to determine the variables associated with a higher risk of recidivism by showing variation of variables how they affect the results that is recidivism from economic to the number of incarnations prior to conviction despite lack of clear policy prescription from the number of prior incarnations and age of the parolee. But there is the short run solution that could reduce the total cost of crime that is development of policies aimed at enhancing the opportunities for parolees to gain employment.

2.3.5 Neural networks

A neural network is a form of statistical method that may be used to construct dynamic models of interactions among variables for the purposes of regression and classification (Paik, 2000). Neural networks are generally composed of a collection of elementary processing units interconnected by weighted connections or "relationships" of a particular strength (Gray et al.,2008). Neural networks can be used for both regression of a numeric dependent variable and classification of a categorical dependent variable.

Research by Palocsay, Wang, & Brookshire, (2000) uses neural networks models to predict criminal recidivism by splitting an offender population into two groups: non-recidivists and eventual recidivists. The results suggested that the NN models obtained significantly higher predictive accuracy in offender's classification as recidivists and non recidivists compared to logistic regression models. As prediction accuracy heavily depends on the scope of network topology, such as the number of hidden layers and nodes in each layer, the training methodologies used and node activation functions (Gray et al., 2008).

The Artificial Neural Networks (ANN) compared with other computational function, process information in parallel rather than as with conventional computing where each task is broken down into discreet subtasks and processed sequentially. By use of a cost function it's able to process complex and non linear information as it's a mathematical computation system.

The advantages of using ANN includes:

- It can be applied to incomplete, fragmented data sets.
- It can understand and analyse incomplete, nonlinear data, the sort of data produced by human behaviour ,data that linear processors (conventional computers) cannot.
- They are arguably fairer, as they recognise numerous pathways towards an end goal, and do not focus on traditional stereotypes.
- They learn from existing data, they allow for "local" validation and prediction studies that would be costly and less effective using traditional methods.

ANN has been used extensively in prediction of behavior for example the Research from the USA has looked into predicting juvenile recidivism. Traditional methods of identifying the factors that separate repeat and non-repeat offenders had accounted for 20% of the variance in recidivism. While the ANN was trained using part of a data set (120) and tested on the remaining 46. The predictability rate rose to 74% for the test population. This represents a significant increase in the ability to predict human behaviour (Booth, 2007).

2.3.6 Decision trees

Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favored technique for building understandable models. Auditors can use them to assess, for example, whether the organization

is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit (Silltow, 2006).

Rosenfield & lewis (2005) application of a CART approach to violence risk assessment using a sample of 204 stalking offenders. The model prediction accuracy was found to be high compared to logistic regression models and relative simplicity of its application in clinical practice compared to logistic regression models (Gray et al, 2008).

Example of an application of the decision tree is the random forest modeling used by Richard Berk working with NIJ - funded researchers Geoffrey Barnesand Jordan Hyatt (2013) to build the risk prediction tool for Philadelphia's Adult Probation and Parole Department. Which can be described as hundreds of individual decision trees, where data are organized using a technique called "classification and regression trees." The computer then runs an algorithm that selects predictors at random and repeats and repeats this process to build several hundred trees which then allow the randomly selected predictors to average themselves into a single outcome. In the case of the Philadelphia tool, this outcome was assignment to one of three risk categories (high, Moderate or low) for probation-super vision purposes.

The random forest model prediction tool, allows agencies to base their personnel and policy decisions on a scientifically proven method. A tool like the one developed in Philadelphia provides an opportunity to advance the capabilities of the criminal justice system to protect communities, particularly for jurisdictions with large probation populations that must be managed with fewer dollars. This has helped probation officials manage cases more efficiently, and allowed concentration of resources where most needed (Ritter, 2013).

2.4 Data mining tools for prediction

There are various tools available that have been developed for various usage example we have Waikato Environment for Knowledge Analysis (WEKA), Rapidminer, Knostanz Information Miner (KNIME), Clementine among others. They provide a set of methods and algorithms that help in better utilization of data information available to users; that is data analysis, cluster analysis, genetic algorithms, nearest neighbor, data visualization, regression analysis, decision trees, predictive analytics, text mining among others (Wahbeh, Al-Radaideh, Al-Kabi, and Al-Shawakfa 2008).

2.4.1 WEKA

It contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interface for easy access to this functionality. It supports several standard data mining tasks like data processing, clustering, classification, regression, visualization and feature selection. WEKA capabilities include; API, database system support, visualization capabilities, PMML support and statistical analysis capabilities (Witten, frank,& Hall, 2011).

2.4.2 KNIME

KNIME is an open source data analytics, reporting and integration platform, as it integrates various components for machine learning and data mining through its modular data pipelining concept. Mostly has been used in pharmaceutical research, customer data analysis, business intelligence and financial data analysis (Tiwaria, Abhishek, Sekhar, and Arvind K.T.,2007).

Its capabilities includes; API, database system support, visualization, statistical analysis capabilities among others (Kavoc, 2012).

2.4.3 Rapid Miner

Comparing it with the above tools rapid miner has full API support, which makes it possible to access a wide variety of functionality and support. It capabilities are same like for WEKA and KNIME but the variation comes in on users using it as using rapid miner an advanced user will be able to achieve more functions compared to less advanced user (Kavoc, 2012).

2.4.4 Orange

It is similar with the other data mining tools mentioned above on functions that can be performed. Though for one to achieve full functionality additional add-ons, widgets have to be obtained and added to the program as it's a library of objects and routines written in C++. Thus may have some effect on the software's functionality and performance. It has no additional functionality that seems relevant for the end user, as it's quite basic in its performance and operations (Kavoc, 2012).

2.5 Summary of Literature Review Findings

The Neural networks and Decision tree techniques have great potential to assist in improving the predictive accuracy of decision-making processes and instruments aimed at assessing and predicting the risk of recidivism in criminal justice settings. From various researches conducted the techniques display high level of predictive accuracy over traditional statistical methods. As with their efficiency thus more improved and efficient criminal justice decision making and they are more intuitively appealing to professionals in criminal justice practice (Gray et al., 2008).

Research by Yang, Liu and Coid, (2010) which compares the traditional models, verses the data mining models accuracy measures on various scenarios. This includes overall accuracy a combination of sensitivity and specificity. The traditional methods LR and DA are more robust and controllable though limiting with number of categories involved while CT models are flexible, comparable and not restricted to large data sets with inter-correlated variables involving small effects though less plausible in risk assessment practice. When developing a model they can be manipulated technically to achieve a rather high predictive accuracy, thus resulting to poor performance in other external samples or very low accuracy in prediction of the outcome category that is relatively small.

For Neural networks it is favorable for scenarios where there are many parameters (variables) as it has the greatest flexibility to reflect complex relationships between inputs and outputs of the data. Though may be restricted by various issues like parameters change, sample size, misclassification error which may result to poor performance on an external sample. A change in parameter often causes a change in model performance in terms of predictive accuracy, having in mind that those parameters are interrelated. NN is preferred where there are large variables or target population with better homogeneity.

Neural networks and decision trees are the methods widely adopted mostly due to their prevalence in the field of data mining and proven ability to form models across a wide range of application areas. More so with advancement of data mining the two methods have proved to be most versatile and accurate techniques available. Also compared with other techniques, they are well established for adoption in criminological data (Gray et al, 2008).

Decision tree models produce a tree of decisions based on the values of the independent variables which is used to assess the predicted outcome. With its transparency helps analyst determine the exact structure of the model and how independent variables are used to arrive at its prediction. The decision trees internal workings are binary compared to the networks which are continuous. As each point of the decision tree model decision process is a discrete decision tree point. The tree slices the independent variable space into regions of different predictions.

Considering the various data mining tools as discussed earlier in the literature review section that is Rapid miner, KNIME, WEKA, Orange and jHep which most are freely available for use, and no single machine learning scheme is appropriate to all data mining problems as stated by Kavoc, 2012. Thus a tool like WEKA through its workbench provides a collection of state of the art machine learning algorithms and data preprocessing tools. It includes virtually all algorithms in data mining thus its diverse functionality characteristic, so one can quickly try out existing methods on new datasets in flexible ways. It also provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically and visualizing the input data and the results of learning (Witten, Frank, & Hall, 2011 pg 404-406).

Considering WEKA full range of API and PMML capabilities it allow importation of files from a variety of database formats thus if the RPM is to be implemented countrywide; in all prison stations the database types which may vary will not be a problem. More so the robust nature of the WEKA software on provision of various interfaces that is explorer, knowledge flow and experimental. The explorer interface is easy to navigate data and results, knowledge interface allow the user to connect various functions together in order to perform data mining functions and experimental interface allows one to compare results of more than one dataset. Therefore, it can be used by a variety of users in various setups with different levels of skills in data mining (Kavoc, 2012).

Therefore, for this project WEKA was appropriate as it could be used on existing dataset of the prisoners and analyze its output to learn more about the prisoners recidivism and also use learned models to generate predictions on new instances example to predict the prisoners likely to re-

offend or apply several different learners and compare their performances to choose one for prediction. The risk factors variables for the RPS include: age, sex, socioeconomic status and unemployment. More so due to the fact that WEKA can be fed data using a file and output to a file too, thus applicable for a small scope meant for checking viability of its implementation; the development of a prototype on RPS. Due to WEKA limitation on visualization properties we have incorporated a GUI application where the prediction results are displayed using graphs and summarized into report to assist the users in decision making.

The GUI application was developed using python programming language as it's a widely general purpose high level programming language, and supports multiple programming paradigms that is object oriented, functional or procedural styles. Therefore it is used to display the prediction results from the WEKA tool into a format that the end users can easily understand to allow easy and insightful decision making process.

CHAPTER THREE

3. Methodology

3.1 Introduction

This chapter presents the research process with details on key aspects on research methodology such as design, data, procedure and analysis which are important for a successful research activity.

It also states why specific methodology and tools were used to come up with the conclusion in line with the research area. As in data mining there are various methodologies and no standard one for applying. Thus several vendors have created their own proprietary methodologies where the approaches are strongly correlated with the design of their own software packages and solutions. The popular methodologies include Sample Explore Modify Model and Assess (SEMMA) and Cross Industry Standard Process for Data Mining (CRISP-DM). SEMMA may contain essentials elements of data mining project that is statistical, modeling and data manipulation but it lacks some fundamental parts of any information systems project like analysis, design and implementation phase. While CRISP-DM comprises of six (6) phases which are not rigid and they include; business understanding, data understanding, data preparation, modeling, evaluation and deployment much emphasis is on data which must be divided into training and validation sets. But it is limiting as techniques are selected according to data available only and not on organization goals and requirements, though it's a good approach to the general process, therefore considered for the development of the RPS for this project (Rohanizadeha, Moghadama, 2009).

Overview of CRISP-DM methodology

The methodology describes the activities as shown in the Figure 2, that are done to develop a data mining project. Every activity is composed of tasks. For every task, generated outputs and needed inputs are detailed. CRISP-DM comes up to resolve the problems that existed in data mining project developments.

The main objectives include ;ensurealing quality of data mining projects results, reducing skills required for data mining, capturing experience for reuse, general purpose (i.e., widely stable

across varying applications), robust (i.e., insensitive to changes in the environment), tool and technique independent and tools supportable(Presutti, 1999).

CRISP-DM is the most commonly used methodology for developing data mining projects.

Though it has the limitation that it just defines what to do and not how to do. Another inconviniences is that CRISP-DM does not include project management activities such as quality management or change management.



Fig 2 CRISP-DM (Rahim, 2014)

3.2 Research analysis and design

Using the CRISP-DM methodology in the research enabled a better understanding of the data from the Offenders Record Management System (ORMS) by analyzing it using Ms Excel and WEKA tool for the pattern and prediction on occurrence of re-offending of an already convicted prisoner. It involved:

3.2.1 Requirements Analysis

Understanding Recidivism

As stated by various researches like Howell, 2003 and Omboto, 2010 a number of factors like education, vocational training, counseling, farming skills and financial support are sought to

affect the recidivism in prisons from a social perspective. And according to Haseltine and day, 2011, prisoners with higher level of education found it difficult to stay in prison and tried their best to move out of prison, as education is enlightening and equips the prisoners with positive attitude and outlook of life which enables them to overcome crime and other high risk behaviors (Hoffman, 2004 and Chappel, 2002). With this we were able to narrow down to the most likely attributes that can be used to determine a prisoner's likely hood of re-offending.

With the range of factors which are thought to affect recidivism from different research work which has been done it was a guide for the attributes (variables) to be used for the RPS. Though no clear cut line on how it can be prevented or reduced the RPS can be used with the existing measures to help in minimizing the congestion in the prisons institutions and ensuring that the prisoners are rehabilitated.

Considering the prisons department mandates which include; containment and safe custody of inmates, rehabilitation and reformation of prisoners, facilitation of administration of justice among others. Prediction of recidivism in the department would be much helpful in measuring whether the various activities on rehabilitation that are in place are helpful in meeting the mandates especially rehabilitation and reformation of prisoners. And more so give a guide line on the various policies to be implemented for efficient and effective service delivery.

The data mining goals being to:

- Extract recidivism patterns by analyzing of the dataset from the ORMS of the three stations
- Prediction of recidivism based on the existing data and anticipation of recidivism rate using data mining techniques

This is with an aim of helping in the current state of congestion in the various prison institutions.

Architecture design

This represent a conceptual design of the recidivism prediction system (RPS) based on the various subsystems that were interlinked. By showing how the various processes of the systems are interacting from data inputs, data cleaning, artificial intelligent using WEKA, input and output file, decision support systems and the decision makers.



Fig 3: The architecture design of the prototype

3.2.2 Data collection and analysis

As stated from the previous phase the already known factors that affect recidivism act as a guide to the attributes (variables) to be used in the RPS. They include date of sentence, age, religion, region, occupation, education, marital status and previous conviction as the target dataset.

The data collected of the ORMS database from the three stations that are Langata women, Nairobi west and Nairobi medium existed in SQL format therefore had to be extracted to meet the intended need.

To manage the collection of data from the three stations, it involved acquiring permission letter from the prisons Headquarter. This was during the initial stage of the research when writing the proposal.

Due to the state of the databases at the station level which had a lot of incomplete fields which would have posed as an error during prediction, there was need to extract a target dataset file from the SQL files collected from the three prison stations.

Thus the use of Comma Separated Values (CSV) and Attribute-Related File Format (ARFF) file which can be feed into the WEKA tool.

The methods used for data collection in this research work included: visiting of site (secondary data) and interview conducted during the testing period to check the viability of the prototype to the users.

The data collection process

At this stage for the researcher to collect the initial data, it involved visiting the site and interacting with the Offenders Record Management System (ORMS) of the station in question. Being that all of the three stations were using the same database My SQL the process involved was the same. This is illustrated by the following steps:

Steps:

- a) Using Mysql admin window to assess the database, which showed the databases operational and for this case the interest was on the Inmates database
- b) Import the inmates database sql file
- c) Save in a portable memory (flush disk) for later use, as the sizes of the files was manageable; with a size between 500mb and 1000mb.

The **Figure 4** shows a sample of the collected data from one of the station extracted into a Ms excel format

All Commit Commit <th></th> <th>Home</th> <th>Insert</th> <th>Page Layour</th> <th>t For</th> <th>rmulas D</th> <th>ata Re</th> <th>eview VI</th> <th>ew Load</th> <th>i Test 🛛 🕅</th> <th>tro Pro 8</th> <th>PDF T</th> <th>eam</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th>۲</th> <th>_ 0</th>		Home	Insert	Page Layour	t For	rmulas D	ata Re	eview VI	ew Load	i Test 🛛 🕅	tro Pro 8	PDF T	eam								۲	_ 0
Protect Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Product Pr	ľ	K Cin		Times New Ro	m y 12	• A A	= =	<u></u> = ≫ ~]	≣≉Wrap	Text	General		- 1	4		Fn I	*	Σ AutoS	um * A	A		
At 1 Start Grant Dispiration Dispiration Dispiration Dispiration Dispiration Dispiration A1 - A Using the WEKA tool, to preprocess the target dataset (immates.csv) being that it is a case servitive tool to check the uniformity of the data in the inmates file which is to be an analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be analysis of the data in the inmates file which is to be data interm	Pas	ste	1	BIU	1	Q . A .	EE	= (: ÷	Meron	e & Center *	\$ - 2/2	1 5.0	Cond	itional Form	at Cell	Insert De	ete Format	S cm.	Sort &	Find &		
Clease Fart C Algement Lumber C Open Cols Eding A1 - A Using the WEKA tool, to persoccess the target dataset (immates.cw) being that it is a case service tool to check the uniformity of the data in the inmates file which is to be M N O P O R S - V V Y Z A A C A A C A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A	~	- V For	mat Painter	1					J =				Form	atting ~ as Tab	le • Styles •	7	* *	C/ Clear	* Filter∗	Select *		
A1 A Using the WEIA tool, to preprocess the target dataset (immates.cov) being that it is a case sensitive tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to check the uniformity of the data in the inmates file which is to be analytic tool to chec	2	Clipboar	d la		Fort	5		Align	ment	- Ga	Nu	nber	6	Styles		G	elis		Editing			
k N O P O R S T U V W/V X Y Z AA AE AD AE AF AA Immail introl intro introl introl <th>1</th> <th>A1</th> <th></th> <th>(fs</th> <th>Using</th> <th>g the WEKA</th> <th>tool, to p</th> <th>reprocess</th> <th>the target</th> <th>dataset (in</th> <th>mates.csv)</th> <th>being that</th> <th>t it is a ca</th> <th>se sensitive</th> <th>tool to che</th> <th>ck the unif</th> <th>ionnity of</th> <th>the data in</th> <th>the in mate</th> <th>s file wh</th> <th>ich is to be</th> <th>e 🕴</th>	1	A1		(fs	Using	g the WEKA	tool, to p	reprocess	the target	dataset (in	mates.csv)	being that	t it is a ca	se sensitive	tool to che	ck the unif	ionnity of	the data in	the in mate	s file wh	ich is to be	e 🕴
Image Image <th< th=""><th></th><th>k</th><th>N</th><th>0</th><th>P</th><th>Q</th><th>R</th><th>S</th><th>-</th><th>U</th><th>V</th><th>W</th><th>×</th><th>Y</th><th>Z</th><th>A.4</th><th>Æ</th><th>AC</th><th>AC.</th><th>AE</th><th>AF</th><th>AC</th></th<>		k	N	0	P	Q	R	S	-	U	V	W	×	Y	Z	A.4	Æ	AC	AC.	AE	AF	AC
17 NA VA 10 SELEVERS TEPATINL 2(F) 6006S VA NA CE (jungge 321 0.1% NA CT 18 NA VA 4 VERMU207RS TEPATENIL 57F 480S VA NA CE va A486 1984 VA NA CT 388 4861 VA NA CT 78 4861 VA NA CE VA NA CE VA NA CE VA NA CE VA NA VA NA CE VA NA	1	emai	idna	tribe p	CCUD	ane	educ	DC	ht	wi	hman	remm	sent	station	district	mageuri	division	location	sublocatif	bod	end	chiel
It It A V/A 4 LE-ALEFRIL 2+LEI SYLAS V/A It/A	17	N/A	N/A	10 SI	ELFEM	F52rRS	TEP4	V NIL	ECF	60KGS	N/4		N/A		62	./oimages	289	321	0 `	VA.	N/A -	GITH
15 NA VA 6 CAMPLOYNER TEPATENIL 2FF 40kGS VA NA 2, formages 200 238 611 VA NA VA 11 NAA VA 6 CELESUT22VTG TEPATENIL 2EF 50KGS VA NA VA VA NA VA NA VA VA NA VA VA VA NA VA VA NA VA NA VA VA NA VA VA VA NA VA VA VA VA <t< td=""><td>15</td><td>NA</td><td>N/A</td><td>4 _</td><td>VEMPU</td><td>0.29YRS</td><td>LIE-AI</td><td>E FIIL</td><td>: 4FEET</td><td>5/KGS</td><td>N/4</td><td></td><td>N/A</td><td></td><td>41</td><td>./oimages</td><td>3772</td><td>4386</td><td>10384 \</td><td>A)</td><td>NVA</td><td>GIOL</td></t<>	15	NA	N/A	4 _	VEMPU	0.29YRS	LIE-AI	E FIIL	: 4FEET	5/KGS	N/4		N/A		41	./oimages	3772	4386	10384 \	A)	NVA	GIOL
CC NAA VAA 16 ELEFENT 22/T13 TETATENIL 2FF 50kGS VA NAA 15C. formages 2005 2008 6178 VA NAA VA CL NAA VAA 10 SELFENT 22/T13 TETATENIL 2CF 50kGS VA NAA 15C. formages 1008 1198 100 VA NAA VAL CL NAA VAA 6 SELFENT 26/KRS TETATENIL 2CF 50kGS VA NAA 42C. formages 1008 1198 1002 VA NAA VAL VA NAA VAL	19	N/A	N/A	6	EMPL	(20YRS	_ TEPAT	ENIL	EPET	49K.GS	14		N/A		2	./oimegae	710	798	4541	A/A	NA	EDhoà
11 NA YA YA NA YA NA YA YA NA YA YA NA YA YA <td< td=""><td>20</td><td>N/A</td><td>N/A</td><td>6 SI</td><td>ELF E VI</td><td>F 32YEG</td><td>_ TERAT</td><td>ENIL</td><td>EEF</td><td>59K.GS</td><td>14</td><td></td><td>N/A</td><td></td><td>36</td><td>./oimcges</td><td>2092</td><td>2388</td><td>6178</td><td>YA .</td><td>N/A</td><td>N DF</td></td<>	20	N/A	N/A	6 SI	ELF E VI	F 32YEG	_ TERAT	ENIL	EEF	59K.GS	14		N/A		36	./oimcges	2092	2388	6178	YA .	N/A	N DF
122 NAA VA NAA 22.5 SMRGS VA NAA 22.5 VAA NAA VAA VAA <thvaa< th=""> <thvaa< th=""> <thvaa< th=""></thvaa<></thvaa<></thvaa<>	-21	N/A	N/A	10 3	сп с м	F261/R3	_TERAT	TNL	10	56KGS	144		N/4		190	./aimages	0951	4596	0 '	VA .	N/A	NAH I
121 NA VA NA NA VA NA NA NA VA NA NA <t< td=""><td>22</td><td>N/A</td><td>N/A</td><td>6 E</td><td>VPLO73</td><td>E3ZrRS</td><td>_ TEPAT</td><td>ENIL</td><td>52FT</td><td>58KGS</td><td>1,4</td><td></td><td>N/A</td><td></td><td>25</td><td>. /oimages</td><td>1088</td><td>1196</td><td>1905 `</td><td>MA.</td><td>N/A</td><td>AUR.</td></t<>	22	N/A	N/A	6 E	VPLO73	E3ZrRS	_ TEPAT	ENIL	52FT	58KGS	1,4		N/A		25	. /oimages	1088	1196	1905 `	MA.	N/A	AUR.
24 NA YA 1 VERDL02YRS TEATE ⁶ ETF 6000S YA NA E /imegas 258 2800 7673 YA NA Spir 15 N/A YA 6 6 DELFEVID 34/TR3 TEPATENIL 2 F 556.63 YA NYA 25 , imegas 2501 310 703 YA NYA A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A	23	NVA	N/A	6 SI	ELFEM	F28YRS	_ TERAT	THIL	EFF	60KGS	N/A		N/A		64	./oimages	1336	1471	2489	A)	NVA	WWA
15 NAA YAA 102 SELFEVFIAHTS TETATENUL EUF SSR0S YA NYA SSL/process 401 447 10027 YA NYA SSL 20 NYA YAA 21 CLAMPL(24/RB3 TETATENUL 24TELT 50K6S YA NYA YAA <	24	N/A	N/A	1.	EMPL	(22YRS	_ TEPAT	£"	EIF	60KGS	N/4		N/A		E	./oimegas	2583	2380	7673 \	VA .	N/A	PAU_
EC IVA V/A ET ATTERATINE C 4TECT Step 3 V/A N/A V/A	25	N/A	N/A	6 GI	ELFEM	F34MTG	_ TERAT	THIL	ECF	55KQS	1/4		N/A		33	./oimages	401	447	10027	(A)	N/A	SOP-
IZ INA VA 15 SELFE VR 32/RS TEPATINIL EFF B0RGS V/A N/A I, //imrugas 2446 2300 0710 V/A N/A VIL2 2F NAA V/A 4 SELFEVR 32/RS SEMI FINIL SEF R0RGS V/A N/A IF //imrugas 3407 9084 V/A N/A SILE X/A N/A V/A N/A N/	33	N/A	N/A	21 U	CEMPL	(24)R3	TT =4	S NIL	I 4TEET	55KGS	1/4		N/A		53	./bimeqes	561	519	759 \	VA .	N/A	ABLA
12F NAA V/A 4 SELEFVE/3785 SEMI Field FFF RNL5S V/A NVA LF /simegas 3377 3107 9084 V/A NVA N	27	N/A	N/A	15 SI	ELF E VI	F32YRS	_ TEPAT	ENIL	EEF	80KGS	1/4		N/A		1	./bimeyes	2449	2330	10710	VA I	NA	VUP.
12 NA YA 6 SELFEWF30YRS - LE-AFLENUL 2FEEL N/A N/A N/A Comegas 3105 3033 8288 V/A N/A V/A 3C N/A V/A 4 SELFEWF30YRS - LE-AFLENUL 2FEF 67K6S V/A N/A EF / formegas 3003 8268 V/A N/A V/A 3C N/A V/A 4 SELFEWF307RS - TE-AFLINIL 2 FEF 67K6S V/A N/A EF / formegas 3000 3003 8268 V/A N/A V/A 3E N/A V/A 6 UNEMPL(22YRS -TE-AFLINIL 2 FEF 60K6S V/A N/A FE / formegas 736 1102 3003 V/A N/A V/A 3E N/A V/A Comegas 732 1102 3003 V/A N/A V/A 3E N/A V/A Comegas 736 1102 3003 V/A N/A V/A 3E N/A V/A Comegas 736 1003 N	78	N/A	N/A	4 5	FLEEM	F.32YBS	SEMI T	FIJII	= 8F	808.65	1/4		N/4		1E	(himages	3377	3907	9084	VA.	N/A	3DA-
EC N/A V/A 4.5 ELF EVF 80X BS TE PATEINIL EFF 67K6S V/A N/A EF / /omegac 3200 3713 9601 V/A N/A M/A M/J C1 N/A V/A 0 UNEMPL/307KB TE PATEINIL 1 4TEET 79K6S V/A N/A N/A C. /omegac 2000 3713 9601 V/A N/A N/A S1 N/A V/A 0 UNEMPL/207KB TE PATEINIL 51F 60K6S V/A N/A N/A 1 //imegac 2445 2330 0373 V/A N/A V/A S2 N/A V/A N/A N/A N/A N/A N/A 1 //imegac 2445 2330 0373 V/A N/A V/A S4 N/A V/A A N/A N/A N/A 1 //imegac 2445 2300 0037 V/A N/A V/A S4 N/A V/A 1 UNEMPL/20XBS TE PATEINIL 25F 68K6S V/A	25	N/A	N/A	6 SI	ELFEM	F35YRS	_ IE-AI	ENIL	= 2FEET	54KGS	1/4		N/A		33	./oimegas	31.05	3503	8268	(A)	N/A	1/4
11 N/A V/A 6_UEMPL(200RB_UEPATUNIL 2_4TECT_79K96 V/A N/A V/A 0_Ummcgas 2050 2719 9100 V/A N/A 2/400 12 N/A V/A 6_UEMPL(200RB_UEPATUNIL 25T 600635 V/A N/A 1_/ummcgas 2445 2300 0373 V/A N/A N/A 12 N/A V/A 6_UEPATUNIL 25F 600635 V/A N/A F/A 6/ummcgas 2445 2300 0373 V/A N/A N/A 12 N/A V/A 6_SEFEPATUNIL 21F 680635 V/A N/A F/A 681 788 1035 \ N/A N/A A/A 126 N/A V/A 1_UEMPL(200RB_UEPATUNIL 21F 680635 V/A N/A 82.//ummcgas 888 808 909 V/A N/A 72/V 126 N/A V/A 1_UEMPL(240RB_UEPATUNIL 21F 584635 V/A N/A 82.//ummcgas 717 30	30	N/A	N/A	4 SI	ELFEM	F46YRS	TER/	(NIL	EEF	67KGS	N/A		N/A		E7	./oimagoo	3200	3713	9501	MA.	N/A	VUSE
SE NA VA 6_NEMPL(22YRS_IDEPATENT EESE 6069S V/A NVA 1_/uimegas 2446 2330 10373 VA NVA SE N/A V/A 6_SETEFVEMINE ESEF 6069S V/A NVA F4_(imingas 746 1102 3033 VA NVA V/A VA V/A V/A NVA V/A V/A NVA F4_(imingas 746 1102 3033 V/A NVA V/A VA V/A V/A NVA V/A NVA F4_(imingas 746 1102 3033 V/A NVA V/A VA V/A 1_NEMPL(22YRS_IDE TEPATENT EVE 62KGS V/A NVA E_1/imingas 768 1035 \ N/A V/A V/A <td< td=""><td>-01</td><td>N/A</td><td>N/A</td><td>6</td><td>NEMPL</td><td>(00/R3</td><td>_TERAT</td><td>TNL</td><td>I 4TEET</td><td>79KGS</td><td>N/A</td><td></td><td>N/A</td><td></td><td></td><td>./bimages</td><td>2050</td><td>2719</td><td>9100 °</td><td>VA.</td><td>N/A</td><td>240</td></td<>	-01	N/A	N/A	6	NEMPL	(00/R3	_TERAT	TNL	I 4TEET	79KGS	N/A		N/A			./bimages	2050	2719	9100 °	VA.	N/A	240
15 NAA VAA 6 SEE FEME AVRS TEPATTHUL FFF 75K 5S V/A N/A FA (imegas 746 102 3303 V/A N/A V/A 12 N/A V/A V/A N/A N/A N/A 102 3303 V/A N/A V/A 12 N/A V/A V/A V/A 12. (strenges 102 12/9 1909 V/A N/A V/A 12 N/A V/A 1 NEME/C2VRS TEPATENIL 52F 68K0S V/A N/A 2. (strenges 886 080 4970 V/A N/A QL 12 N/A V/A 1 SEE FA/TENIL 52F 68K0S V/A N/A 2. (strenges 886 080 4970 V/A N/A 52K 12 N/A V/A 1 SEE FA/TENIL 52F 68K0S V/A N/A 2. (strenges 717 306 2851 V/A N/A 52K 12 N/A V/A 10 SEE FA/TENIL 57F 48K6S V/A N/A 124. (str	32	N/A	N/A	6 J	EMPL	(22YRS	_ TEPAT	ENIL	ESF	60KGS	N/A		N/A		1	. /bimeyes	2449	2330	10373 1	VA.	NA	WWA
12 N/A V/A / L_NEMPL/22/NS SEMIL EIR LF BRGS L_E.R.F.GHA N/A 12: /gmmages 112: 1/39 1909 V/A N/A V/A EE N/A V/A L_NEMPL/22/NS SEMIL EIRL 2F 62KGS V/A N/A 2, /gimages 68: 768 1035 \ N/A GLE EE N/A V/A 11 SELFEMIS3/NIS TEPATENIL 2FF 62KGS V/A N/A 2, /gimages 88: 809 4970 V/A N/A GLE EV N/A V/A 1 SELFEMIS3/NIS TEPATENIL 2FF 68KGS V/A N/A 2, /gimages 88: 909 4970 V/A N/A GLE F N/A V/A 1 SELFEMIS3/NIS TEPATENIL 2FF 58KGS V/A N/A 209 209 2000 V/A N/A 2/A V/A V/A 10 SELFEMIS3/YEARS TEPATENIL 2FF 58KGS V/A N/A 2./gimage	11	N/A	N/A	6 5	FIFFM	F34YBS	TEPAT	ŦЫI	= FF	75KGS	N/A		N/A		F4	(oimages	792	1002	3303	(A)	N/A	3./A
EE NA V/A 1_VEMPL(22YRS_TEPATENIL E2FF 62KGS V/A N/A E./oimegad 668 768 1035 \ N/A GLE SE N/A V/A 1_UEMPL(22YRS_TEPATENIL E2FF 62KGS V/A N/A E./oimegad 688 768 1035 \ N/A GLE SE N/A V/A 1_UEMPL(24YRS_TEPATENIL E2FF 68KGS V/A N/A E./oimegad 209 20002 V/A N/A 72K SE N/A V/A 1_UEMPL(24YRS_TEPATENIL E2FF 58KGS V/A N/A 2./oimegad 217 300 2002 V/A N/A 72K VA V/A 1UEMPL(24YRS_TEPATENIL E2FF 58KGS V/A N/A 2./oimegad 217 300 2051 V/A N/A 72K VA V/A 1UEMPL(24YRS_TEPATENIL E2FF 58KGS V/A N/A 12./oimegad 216 1005 V/A N/A 74N N/A 74N	34	NA	N/A	1.	VEMPU	(23YRS -	SEMI _	ENIL	: LF	68KGS	-E ER-	GHA	N/A		12	./oimages	112E	1239	1909 1	./A	NVA	VAN
15 NA YA 11 DELFENTISYNE TEPATENIL ESF 68KGS X/A N/A E. /oimcgas 88E 383 4970 Y/A N/A 52K 17 N/A Y/A Y/A Y/A Y/A N/A Y/A N/A Y/A N/A Y/A N/A Y/A N/A Z/A N/A Z/A Y/A N/A Z/A N/A Z/A N/A Z/A N/A Z/A Y/A N/A Z/A Y/A N/A Z/A Y/A N/A Z/A	35	N/A	N/A	1	EMPL	(22YRS	LITERAT	ENIL	E 2FT	62KGS	14		N/A		2	./oimagoo	683	768	1035 \		N/A -	SILEL
S2 N/A V/A 1_UNEMPL(14/RS_UTEPATENIL 5 CF54KGSV/A N/A 2 / uimeges 2 / ui	36	N/A	N/A	11.3	ELFEM	ES3MTICE	LITERAT	THIL	E SFT	68K.GS	EXIMMAT	"HA	N/4		8	./oimages	88E	389	4970 °	VA.	N/A	EXK
SE N/A V/A 6_UNEMPL(24)/RS LTE-ATINIL 57F 58KGS V/A N/A CE Commages 717 306 2851 V/A N/A GEX 11 N/A V/A 10_EXPLOY-23/NS LE-ATINIL 27F 58KGS V/A N/A 1224 //mmages 375 V/A N/A GEX 12 N/A V/A 10_EXPLOY-23/NS LE-ATINIL 24FEL1 50KGS V/A N/A 1224 //mmages 395b V/A N/A GEX 41 N/A V/A 5 SELFEMB/37YEARS_TEPATENIL 24FE 36KGS V/A N/A N/A GEX 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400 400	37	N/A	N/A	1	EMPL	(*4YRS	_ TEPAT	ENIL	ECF	54KGS	N/A		N/A		2	./bimeges	2094	2390	10062	VA .	NA	CAVE
11 11 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12<	38	N/A	N/A	6	LEMPL	(24rRS	TEP/	VINE -	5 7FT	58KGS	N/A		N/A		35	. /oimages	717	306	2851 `	VA.	N/A	(IEA)
LC N/A V/A 5 SELF EXF37 YEARS _ TEPATENIL 5 SF 45K GS V/A N 26 ./oimaged 216 1305 V/A N/A 2AN 41 N/A V/A 15 UNEMTL032/TB3 TEPATENIL 411 FEETSEKGS V/A N/A 1./oimaged 2375 3300 8889 V/A N/A CUN 42 N/A V/A 11 UNEMTL032/TB3 TEPATENIL 211 FEETSEKGS V/A N/A 1./oimaged 3300 8889 V/A N/A CUN 42 N/A V/A 11 UNEMTL032/TB3 TEPATENIL 211 02KGS V/A N/A C./oimaged 3910 4557 1005 V/A N/A CAT 42 N/A V/A 6 UNEMTL0337RS TEPATINIL 211 65KGS V/A N/A KA V/A N/A V/A N/A V/A V/A N/A	16	NA	N/A	10 ±	VPLOY:	23YRS	_ 1E-441	EMIL	: 4HEET	50KGS	1/4		N/A		124	./oimages	3555		,	A)	NVA	(44)
41 N/A V/A 15 UNEMPL(32/TR3 TETATENIL 411 FEETS2KGS V/A N/A 1/oimcges 3275 3300 8889 V/A N/A CIN 45 N/A V/A 11 UNEXTINE 211 EEES2KGS V/A N/A 1/oimcges 3275 3300 8889 V/A QIN 45 N/A V/A 11 UNEXTINE 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 211 2133 2264 214 211 211 2131 2264 214 211	4[N/A	N/A	5 SI	ELFEM	F37YEARS	_TEPAT	ENIL	ESF	45K.GS	1/4		N		23	./oimegas	216	1305	,	(A)	N/A	CANI
KE N/A 11 NEMPL(29/RB3 TEPATINIL 2EF 02KG8 V/A N/A C.,/sincges 091C 4577 10095 V/A N/A Call K5 N/A V/A 6 NEMPL(39/RB3 TEPATINIL 2EF 65KGS V/A N/A 86./sincges 391C 4507 10095 V/A N/A V/A V/A V/A V/A N/A V/A <	-1	N/A	N/A	15 🗉	EMPL	(32YDB	LITERAT	THIL	411 FEET	52KGS	N/A		N/4		1	./oimcges	3275	3300	8889	A.	N/A	CUNT
KVA V/A 6 UNEMPL/33/RS LTEPATONCE 5 (F) 65KGS V/A N/A K4 6 (Jointages) 3546 4110 9657 V/A V/A KVA V/A V/A 6 (Jointages) 3546 4110 9657 V/A V/A<	42	NA	N/A	11 .	EMPL	(29/R3	_TERAT	TINL	0.21	02KGS	14		N/A		C	./oimages	091 C	4557	10095	VA.	NA	(A)U
222 NVA VVA 6. NEMPLU37 YEARS TERATNIL ELEFT 75KGS BIKUNYO KIMAN NVA É (joinciges 112) 1233 5264 V/A NVA TRAV 26 NVA 11 NEMPLU38 SEMI TENII ELEFT 75KGS BIKUNYO KIMAN NVA 167 Jointenate Jointenate Jointenate Jointenate	45	N/A	N/A	6 .	EMPL	(33rRS	TE 74	ONCE	ECF	65KGS	NjA -		N/A		33	. /oimages	3546	4110	9657	NA.	N/A	VA
2E N/A V/A 11 NEMPI (38/YES SEMI TENII F1FT 74/GS TUNG, MV/ANGA N/A 1157 tournestady 2007 3391 7752 X/A N/A court	2.2	NVA	N/A	6 _	EMPL	(37 YEARS	TE =.4	5 MIL	ECF	75kGiS	BIKENMO	KIMAN	N/A		ŝ	Jaimeass	112	1233	5264	(A)	NA	Beh
	4Ē	N/A	WA _	11	FMPL	(38/BS	SEMI T	FNI	F 1F	74KGS	TUNG., I	/WANGA	N/4		157	Come rues	7202	3391	7752	14	N/A	GUT.

Figure 4: The raw data from the ORMS

Steps for extracting the data:

At this stage the data saved in the flush disk from the station was extracted to a format that could be input to the WEKA tool. The steps undertaken for extraction are as follows:

- i. Using ODBC application interface was able to transfer the SQL data from the three stations from MYSQL platform to MS Access
- ii. Then from the MS Access database exported the inmates table to MS Excel
- iii. Using Ms Excel cleaned the data using the filter option; this involves removing blank spaces and non-uniformed data among others.

		₹.				final data - 1	Microsoft Excel			
9	Home Insert	Fage Layout	t Formulas Dat	ta Review	View Load	Test Nitro Pro 8 I	PDF Team			
Faste	∦a Cut R⊇ Copy Ø Format Painter	Calibri B I U	• Ш • А́ ѧ́ 	= <u>-</u> _ (» = = = [#	** Herge	General & Center + \$ + %	▼ 0.00 0.00	Conditional Format C	cl Insert Delete Format	Σ Auto
	Clipbeard 19		Font 🕞		Al gnment	rs Numb	er 🦷	Styles	Cells	
	F64	• (* fx	DRIVER							-
	А	В	С	D	E	F	G	Н	1	N
1	DOS 🗸	Gende -	Maritalstatı -	Religio.	Region 🗸	Occupation -	Age	Education	Previous convicti	-)
59	16 Sep 12	Male	Married	Christian	Eastern	CASUAI LABOURER	21	STD 7	ONCE	
60	5-Jan-13	Male	Married	C <mark>hristia</mark> n	Western	CASUAL LABOURER	26	SID 7	ONCE	
61	6-Feb-12	Male	Married	Christian	Western	CARPENTER	35	STD 8	ONCE	
62	23 Aug 12	Male	Married	Christian	Central	CARPENTER	45	FORM 2	TWICE	
63	15-Oct-12	Male	Married	Christian	Central	WAITER	20	FORM 3	ONCE	
64	25-May-11	Male	Married	Christian	Western	DRIVER	32	FORM 4	TWICE	
65	6-May-14	Male	Married	Muslim	Nyanza	SHOE SHINE	23	STD 7	ONCE	
66	5 Jan 11	Male	Married	Muslim	Nairobi	CARWASH	34	FORM 4	ONCE	
67	23-Aug-12	Male	Single	Christian	Riftvalley	BUSINESSMAN	19	FORM 4	ONCE	
68	6-May-14	Male	Single	Christian	Nyanza	SECURITY GUARD	56	FORM 4	ONCF	
69	1/-Jun-10	Male	Single	Christian	Nyanza	CASUAL LABOURER	24	STD 8	ONCE	
70	29-Jul-12	Male Sheet2	Single	Muslim	Central	FLORIST	48	FORM 1	ONCE	
Ready		Constant 1 Certin								

The Figure 5 shows a sample of the cleaned data displayed using Ms excel

Figure 5: The processed data to be loaded to WEKA

iv. Picked the nine (9) attributes the date of sentence, age, religion, region, occupation, education, marital status and previous conviction as the target dataset to a separate workbook

v. Then converted the Excel file containing the target dataset to a CSV format for easy use in the WEKA tool, when saving it.

The interview

The interview mode of data collection was used to check the prototypes viability to the end users need on recidivism at the testing stage. This is after the RPS prototype was developed.

Involved two types of interviews the personal interview and telephone interview; this is because of the time available for research and the availability of the end users due to their tight schedules at work.

The questions for the interview included:

- 1. The level of automation of prisoners records in the department, whether it's efficient enough to enable service delivery.
- 2. Whether there is any advantage in automation of prison activities; example the use of the RPS?
- 3. Considering the rate of recidivism in prisons, would the RPS be of help in the day to day running of the department.
- 4. Whether he/she could advocate for the RPS implementation in the department

3.2.3 Data preparation

The collected data from various stations is diverse and due to the fact that the ORMS is still in its initial state of implementation in the department thus there were missing values, inconsistence data and not useful data. Thus data preprocessing was inevitable as it's a process that consists of data cleaning, data integration and data transformation, with intent to reduce some noises, incomplete and inconsistent data.

Using the WEKA tool, to preprocess the target dataset (inmates.csv) being that it is a case sensitive tool to check the uniformity of the data in the inmates file which is to be used in the system for prediction.

This is to enhance the quality of the output from the system.

The preprocessing includes the following tasks:

i. **Data cleaning:** fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies (there are many modest proposals for filling missing values).

Different preprocessing techniques were used to get clean data, these include:

- Removing outliers, some of the data in the inmate's (inmates.csv) datasets represent outliers and cannot be included in the analysis algorithms and techniques, so these data records were deleted from the, set.
- Filling missing data,
- ii. **Data integration:** using multiple databases, data cubes, or files (since our data are collected from various stations, the data are integrated to build uniform datasets).
- Data transformation: normalization and aggregation
 There was no much normalization involved as all attributes were a determining factor for the end result on recidivism rate.
- iv. **Data reduction:** reducing the volume but producing the same or similar analytical results (Omitting entire records because they have more than three missing values so the filling will cause noisy).
- v. **Data discreetisation:** part of data reduction, replacing numerical attributes with nominal ones

For easy interaction the value for the number of convicted times was changed from numeric to alphabetic for easy interaction

Therefore out of the 2000 instances collected from the three (3) stations, after preprocessing process there was 624 instances that could be used in the WEKA tool.

CHAPTER FOUR

4.0 Prototype development

4.1 Introduction

In this section the process of developing the prototype using both WEKA tool for prediction and the Python GUI to assist the end user in accessing the relevant data through better visualization is detailed.

The WEKA tool use the data inmates file to predict on the rate of a person who had been earlier convicted being convicted again, using a number of algorithms like the; BayesNets, J48 and multilayerperceptron. The result from all algorithms is compared to see that with a high level of accuracy among others. Providing a platform to compare practically the algorithms (techniques in data mining) those with the highest level of accuracy, thus helping in the identification of the optimal results to assist the users in the decision making.

The output from the WEKA tool is then input to the Python GUI application for better visualization into reports and graphs. This is to give the end users a better view of predicted results.

4.2 Prototype development Process

In this section it includes detailed illustration on how during development of the prototype the researcher interacted with both the WEKA tool and the Python GUI application to the accomplishment of the third objective stated earlier.

4.2.1 WEKA Tool

This involves the comparison of results from various models built using different algorithms (techniques) with an essence of identifying that with the highest prediction rate on recidivism.

BayesNets

A Bayesian classifier is a program which predicts a class value given a set of attributes.

Using the Bayes rule where C is a class value and the attributes are A_1, A_2, \ldots, A_n

$$P(C|A_1, A_2...A_n) = \frac{(\prod_{i=1}^n P(A_i \mid C)) P(C)}{P(A_1, A_2...A_n)}$$

For each known class value,

- 1. Calculate probabilities for each attribute, conditional on the class value.
- 2. Use the product rule to obtain a joint conditional probability for the attributes.
- 3. Use Bayes rule to derive conditional probabilities for the class variable.

Once this has been done for all class values, output the class with the highest probability.

The Figure 6 shows the results from the BayesNets algorithm run using the percentage split test option. It comprise of four columns the instance, actual value, predicted and error prediction. Whereby like for instance 1 to 4 the predicted class is 2 whose value is ONCE and that of instance 5 predicted classes is 1 but value is TWICE with a probability that instance 5 actually belongs to class 1 is estimated at 0.693.

reprocess	Classify	Cluster	Associate	Select attribute	s Visualize				
Classifier									
Choose	Dayes	Net -D -	Q weka.clas	sifiers.bayes.net	.search.local.	K2P 1 -5 DA	YES -E w	eka.classifiers.bay	es.net.estimate
est options				Classifier out	put				
	ining set					· · · · · · · · · · · · · · · · · · ·	- -		
C. C. marks		1		1nat#	actual	predicted.	error	prediction	
Supplie	d test set		Set	1	1:TWICE	2 : ONCE	+	0.514	
Crose-	alidation	Folds	10	2	2 : ONCE	2 : ONCE		0.646	
Percen	tage split	%	66		3: TURTCE	2 : ONCE	+	0.06	
		1000		4	2 : ONCE	2: ONCE		0.604	
	More opt	lone		5	2 : ONCE	1:TWICE	+	0.693	1
				1 6	1:IWICE	1:IWICE		0.518	
Nom) Previe	ous convict	Ion	-	7	2: ONCE	1:IWICE	+	0.444	1
				8	2: ONCE	2: ONCE		0.361	
Start	•] [5	stop	9	2:ONCE	3:THRICE	1	0.555	-
cault list fr	abt click fo	or option	13)	10	2 : ONCE	1 : TWICE	+	0.525	
cource as to	ight cick to		(3)	, 11	2: ONCE	2:ONCE		0.713	
0:49:49 - E	payes.Daye	sNet		12	2:ONCE	1:TWICE	+	0.532	
num sa r	аусальаус	anter		13	1:IWICE	1:IWICE		0.732	
				14	2: ONCE	2:ONCE		0.667	
				15	2:ONCE	2:ONCE		0.672	
				16	1 : TWICE	1 : TWTCE		0.76	
				17	1 : TWICE	1 : TWICE		0.75	
				18	4 : FORTH	2: ONCE	+	0.565	
				19	1:IWICE	1:IWICE		0.525	
				20	2:ONCE	2:ONCE		0.774	
				21	2:ONCE	2:ONCE		0.865	
				22	1:TWICE	2:ONCE	+	0.091	
				2.3	2: ONCE	2 : ONCE		0.610	
				24	1:TWICE	2 : ONCE	+	0.802	
				25	1: TWICE	2: ONCE	+	0.597	
				26	2: ONCE	2:ONCE		0.711	
				27	SIFIFIH	1:TWICE	- T	0.855	
				28	2:ONCE	2:ONCE		0.641	
				29	1 : TWICE	1 : TWICE		0.605	
				30	1 : TWICE	2 : ONCE	+	0.722	
				31	2:ONCE	2:ONCE		0.927	
				32	2:ONCE	2:ONCE		0.936	
				33	2:ONCE	1:IWICE	+	0.568	
				34	2:ONCE	2: ONCE		0.739	
				35	2 : ONCE	2 : ONCE		0.005	
				36	2:ONCE	2:ONCE		0.425	
				4					5.

Fig 6: Prediction result of BayesNet

While the Figure 7 shows the confusion matrix which shows the class proper placing and the percentage level of accuracy (correctly classified instances) of the same test option.

reprocess Classify Cluster Associate	Select att	ribute	s Visu	alize							
Classifier											
Choose BayesNet -D -Q weka.clas	sifiers.bay	es.net	.searc	h.local	.K2F	9 1 -S	BAYE	ES -E v	veka.classifiers.t	ayes.net.es	timate.SimpleEs
lest options	Classifi	er out	put								
 Use training set Supplied test set Cross-validation Folds 10 Percentage split More options 	Corre Incor Kappa Mean Root Relat	abso mean ive	7 Cla :ly C atist olute a squ abso	ssif lass ic err ared lute	ied Ir ified or error error	Ins	nces tanc	es	134 78 0.21 0.11 0.25 85.38	75 42 35 73 %	63.2075 36.7925
	Root	rela	tive	squ	ared e	rro	r		98.53	75 💲	
Nom) Previous conviction	Mean Tota	rel.	of c reg	ases ion	(0.95 size (1e	vel) 5 le	vel)	91.98 31.19 212	11 8 1 8	
Start Stop	J Tota.	- null	WEL 1	01 I.	is callo	-ca			212		
Result list (right-click for options)	1	etai	led	Accu	racy E	у с	lass				
0:19:55 bayes.BayesNet				TP	Rate 0.415 0.822 0.067 0 0 0 0	F	P Ra 0.1 0.5 0.0 0 0 0	.82 .84 .05	Precision 0.431 0.712 0.5 0 0 0 0 0	Recall 0.415 0.822 0.067 0 0 0 0	F-Measure 0.423 0.763 0.118 0 0 0 0 0 0
	Weigh	nted	Avg.	2	0.632		0.4	19	0.596	0.632	0.6
	0	Confu	sion	Mat	rix ==	-					
	a 222 22 1 0 0 1	b 30 111 10 4 0 1 0	C 0 1 0 0 0	a 1 1 0 0 0 0	e 0 0 0 0 0 0 0	± 0 0 0 0 0 0 0 0	0000000000	n 0 0 0 0 0 0	<pre>< Class1: a = TWIC b = ONCE c = THRI d = FORT f = FIFT f = SIXT g = TEN h = SEVEN</pre>	tied as E CE H H H N	•

Fig 7: Second part of prediction results of BayesNet

J48

At this section the researcher illustrate the use of the J48 algorithm (decision tree) whose accuracy level may not vary much with the previous results but has much difference on the error prediction. This is shown in the figure 8 below where the prediction result is different in that most of them are predicted for class 2 whose value is ONCE and the probability is constant compared to that of Bayesnet which valid with some as high as 0.972.

eprocess Classify Cluster Associate	Select attribut	es Visualize				
Classifier						
Choose 348 -C 0.25 -M 2						
est options	Classifier ou	tput				
O Use training set	586	2: ONCE	2:ONCE		0.623	
	587	2:ONCE	2:ONCE		0.623	
Supplied test set Set	588	2:ONCE	2:ONCE		0.623	
Cross-validation Folds 10	589	2:ONCE	2:ONCE		0.623	
Dercontage galit 9/ 66	590	2:ONCE	2:ONCE		0.623	
Percentage split % 00	591	1:TWICE	2:ONCE	+	0.623	
More options	592	2:ONCE	2:ONCE		0.623	
	593	2:ONCE	2:ONCE		0.623	
om) Provious conviction	594	2:ONCE	2:ONCE		0.623	
only Previous conviction	595	2:ONCE	2:ONCE		0.623	
Start Stop	596	2:ONCE	2:ONCE		0.623	
	597	2:ONCE	2:ONCE		0.623	
esult list (right-click for options)	598	5:FIFTH	2:ONCE	+	0.623	
:28:17 - bayes.BayesNet	599	2:ONCE	2:ONCE		0.623	
:28:58 - bayes.BayesNet	600	2:ONCE	2:ONCE		0.623	
:29:05 - bayes.BayesNet	601	2:ONCE	2:ONCE		0.623	
:42:29 - trees.348	602	2:ONCE	2:ONCE		0.623	
:47:51 - trees. J48	603	1:TWICE	2:ONCE	+	0.623	
	604	1:TWICE	2:ONCE	+	0.623	
	605	1:TWICE	2:ONCE	+	0.623	
	606	2:ONCE	2:ONCE		0.623	
	607	2:ONCE	2:ONCE		0.623	
	608	2:ONCE	2:ONCE		0.623	
	609	2:ONCE	2:ONCE		0.623	
	610	2:ONCE	2:ONCE		0.623	
	611	2:ONCE	2:ONCE		0.623	
	612	2:ONCE	2:ONCE		0.623	
	613	1:TWICE	2:ONCE	+	0.623	
	614	1:TWICE	2:ONCE	+	0.623	
	615	2:ONCE	2:ONCE		0.623	
	616	2:ONCE	2:ONCE		0.623	
	•			111		•
-	100			_		

Fig 8: Prediction results using Decision tree (J48)

The Figure 9 shows the decision tree confusion matrix and the correctness of instances that have been well classified.

lassifier												
Choose 348 -C 0.25 -M 2												
est options	Classifie	er outp	ut									
🕑 Use training set	Corre	ctly	Clas	ssif	ied In	nst	ances	3	135		63.6792	8
Supplied test set Set	Incor	rect	ly C	lass:	ified	In	stand	ces	77		36.3208	8
Cross-validation Folds 10	Kappa	sta	tist:	ic					0			
	Mean	abso	lute	erre	or				0.13	2		
Percentage split % 66	Root	mean	squa	ared	erro	r			0.25	72		
More options	Relat	ive	absol	lute	erro	r			98.76	48 %		
	Root	rela	tive	squ	ared e	err	or		99.99	19 💲		
Iom) Previous conviction	Cover	age	of ca	ases	(0.9	5 1	evel))	95.75	47 %		
	Mean	rel.	reg:	ion	size	(0.	95 le	evel)	37.5	90		
Start Stop	Total	. Numi	ber (of In	nstan	ces	1		212			
esult list (right-click for options)	D	etai	led 1	Accu	racy I	Ву	Class	3 ===	-3			
):49:49 - bayes.BayesNet					200							
:49:55 - bayes.BayesNet				TP	Rate		FP Ra	ate	Precision	Recall	F-Measure	ROC Are
:26:31 - trees.J48				1			0		0	0	0	0.5
::26:37 - trees.J48					1		1		0.637	1	0.778	0.5
							0		0	0	0	0.5
							0		0	0	0	0.5
							0		0	0	0	0.5
							0		0	0	0	0.5
					5		0		0	0	0	0 5
	Weigh	nted i	Avg.		0.637		0.0	637	0.406	0.637	0.495	0.5
	0	Onfu	sion	Mat	rix =							
	a	b	G	d	e	f	a	h	< classi	fied as		
	0	53	0	0	0	0	0	0 1	a = TWIC	E		
	0	135	0	0	0	0	0	0 1	b = ONCE			
	0	15	0	0	0	0	0	0 1	c = THRI	CE		
	0	6	0	0	0	0	0	0 1	d = FORT	н		
	0	1	0	0	0	0	0	0 1	e = FIFT	н		1
	0	1	0	0	0	0	0	0	f = SIXT	н		
	0	0	0	0	0	0	0	0 1	g = TEN			
	0	1	0	0	0	0	0	0	h = SEVE	N		
	1							111				•

Fig 9: Second part of the J48 prediction results

Multilayerperceptron

In this section the researcher illustrates the result from the multilayerperceptron, which is the most common neural network model, also known as supervised network as it requires a desired output in order to learn. Its goal is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.

A graphical representation of an MLP is shown below:



(Mu-sigma,2014)

D ! -	10.	C		- FMI D
F1g	10:	Graphical	representation	OI MLP

contactor Classify Churter Acception Co	loct attribute	Migualiza				
leprocess classify cluster Associate Se	iect attribute	s visualize				
Choose MultilayerPerceptron -L 0.3	-M 0.2 -N 5	DO -V O -S O -E 2	:0-Ha			
est options	Classifier out	put				
lice training set	103	2:ONCE	2:ONCE		1	1
	104	2:ONCE	2:ONCE		1	
Supplied test set Set	105	2:ONCE	2:ONCE		0.994	
Cross-validation Folds 10	106	1:TWICE	2:ONCE	+	1	
	107	2:ONCE	2:ONCE		1	
Percentage split % 66	108	2:ONCE	2:ONCE		1	
More options	109	2:ONCE	2:ONCE		1	
	110	1:TWICE	2:ONCE	+	1	
	111	2:ONCE	2:ONCE		1	
lom) Previous conviction 🔹	112	6:SIXTH	2:ONCE	+	1	
	113	2:ONCE	2:ONCE		1	
Start	114	1:TWICE	1:TWICE		0.951	
esult list (right-click for options)	115	2:ONCE	2:ONCE		1	
4:43:14 - bayes.BayesNet	116	2:ONCE	2:ONCE		1	
4:43:22 - bayes.BayesNet	117	2:ONCE	2:ONCE		1	
1:44:38 - trees. J48	118	3:THRICE	2:ONCE	+	1	
1:44:59 - trees.348	119	1:TWICE	2: ONCE	1	0.996	
47:51 - functions.MultilayerPerceptron	120	2:ONCE	2: ONCE		1	
5: 14:44 - functions. MultilayerPerceptron	121	2:ONCE	2:ONCE		1	
	122	2:ONCE	2:ONCE		1	
	123	1 : TWICE	2 : ONCE	+	1	
	124	1:TWICE	2:ONCE	+	1	
	125	2:ONCE	2:ONCE		1	
	126	2:ONCE	2:ONCE		1	
	127	2:ONCE	2:ONCE		1	
	128	2:ONCE	2:ONCE		1	
	129	2:ONCE	2:ONCE		1	
	130	3: THRICE	2:ONCE	+	1	
	131	3:THRICE	2:ONCE	+	1	
	132	1:TWICE	2 ; ONCE	+	1	(
	100	1 . THIT OF	0.0105			

Fig 11: Prediction of the results of multilayerperceptron

The figure 11 shows the multilayer-perceptron prediction results the predicted value and the value it's estimated at.

Compared from the rest of the algorithms i.e. the BayesNets and J48 the Multilayer-perceptron has a high accuracy level and the probability of the predicted class being in the said predicted class is high, showing that the Neural Network is a better option as a data mining technique.

As it learn using an algorithm called back-propagation, where the input data is repeatedly presented to the neural network with each presentation the output of the NN is compared to the desired output and an error is computed.

As shown in the Figure 12.

Proposes Cluster Associate Select attributes Visualize Closeffer Choose MultilayerPerceptron -1.0.3 +M 0.2 +M 500 +V 0 -5 0 -2 20 +H a Test options Carectly Classified Instances 138 65.0943 % Supplied test set Set Correctly Classified Instances 74 34.9057 % Percentage splt % 66 0.0879 Mean absolute error 0.0808 Nore options Correctly Classified Instances 74 34.9057 % More options More options 0.0879 Mean absolute error 0.2947 Result (right-tick for options) Test mean rel. region size (0.95 level) 65.566 % Mean rel. region size (0.95 level) 12.7385 % Total Number of Instances 212 Detailed Accuracy By Class ==-1 10.0577 0.0157 0.0157 0.0157 0.077 0.676 1.0770 0.676 1.0727 0.676 1.0727 0.676 1.0727 0.676 1.0727 0.676 1.0727 0.676 1.0727 0.676 1.000 0 0 0 0 0.0728 0.0752 0.051 0.0.651 0.528 0.6681 Measure Nor d e f g h < classified as 3 50 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	O Weka Explorer												- 0 ×
Classifier Choose MultilayerPerceptron 1.0.3 +M.0.2 -M 500 + V 0 - S 0 - E 20 + H 0 Test options Use training set Use training set Coss-validation Folds 10 Percentage spit % 666 Root mean squared error More options Norm Previous conviction Status Status Coss - Coss - Status Coss -	Preprocess Classify Cluster Associate	Select att	ributes	Visu	Jalize								
Choose MultilayerPerceptron -1.0.3 -M 0.2 -N 500 -4 0 -5 0 - E 20 -H a Test options Classifier output Use training set Ocreatly Classified Instances 138 65.0943 % Suppled test set Set Ocreatly Classified Instances 74 34.9057 % More options Ocot mean squared error 0.0867 More options More options Ocot mean squared error 0.2947 Relative absolute error 65.797 % Got relative squared error 114.5524 % Coverage of cases (0.95 level) 12.7358 % Total Number of Instances 212 man rel. region size (0.95 level) 12.7358 % Total Number of Instances 212 more options TP Rate FP Rate Precision Recall F-Measure ROC Area 0.057 0.006 0.757 0.676 10:6949-byses.MayeNet 112:203-functions.MultilayerPerceptron 0 0 0 0 11:22:03 - trees.M8 1 0.946 0.649 1 0.757 0.676 0.661 1:22:03 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0	Classifier												
Test options Classifier output Output biast set Set Suppled test set Set Orrectly Classified Instances 138 65.0943 % Correctly Classified Instances 0.0579 More options 0.088 More options 0.0519 More options 0.0519 More options 0.0519 More options 0.0510 Nom Previous conviction 0.051 Start Stop Result list (right-tack fro options) 0.057 Dife9:97 - bayes.BayesNet 1126:37 - trees.148 T128:37 - trees.149 11 T128:37 - trees.149 1 T128:37 - trees.149 1 T128:37 - trees.140 1	Choose MultilayerPerceptron -L	. 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a											
Outer training set Supplied test set Set Supplied test set Set Set Cross-validation Folds O @ Percentage split % 66 Mare applications Nom) Previous conviction Start Start Stop Start Stop Result ist (jn)t-dck for options) In Correctly Classified Instances 10:69:79 Mare applications Result ist (jn)t-dck for options) Incorrectly Class 10:69:79 Dayses.BayesNett 11:20:37: rese.148 TP Rate FP Rate Precision Recall F-Measure ROC Arel 11:20:37: rese.148 O 11:20:37: rese.149 O </th <th>Test options</th> <th>Classifie</th> <th>er outp</th> <th>ut</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>	Test options	Classifie	er outp	ut									
Suppled test set Set Cross-validation Folds 0 Procendings split % 6 Percendings split % 6 More options 0.057 Noon) Previous conviction 0.058 Realt list (nght-click for options) 10.057 11:28:37 - trees.348 11.052 (0.95 level) 11:28:37 - trees.348 11.000 for clions. MultilayerPerceptron 11:28:37 - trees.348 11.000 for clions. MultilayerPerceptron 11:28:37 - trees.348 0.051 0.000 for clions. MultilayerPerceptron 0.000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	🔘 Use training set	Corre	ctly	Cla	ssif:	ied I	nst	ances	3	138		65.0943	8
Cross-validation Folds ID	Supplied test set Set	Incor	rect	ly C	lass	ified	In	stand	ces	74		34.9057	8
• Percentage split % 66 Mean absolute error 0.088 • More options Root mean squared error 0.2947 • More options • More options • 66 • More options • • • • • • • • • • • • • • •	Cross-validation Folds 10	Kappa	sta	tist:	ic					0.0	579		
Nore options Nore options More options Relative absolute error 65.7897 % Root relative squared error 114.5824 % Coverage of cases (0.95 level) 65.566 % Mean rel. region size (0.95 level) 12.7358 % Total Number of Instances 212 === Detailed Accuracy By Class === 10:49:49 - bayes.BayesNet 11:26:37 - trees.J48 1 0.948 0.649 1 0.787 0.676 11:26:37 - trees.J48 0 0 0 0 0 0 11:26:37 - trees.J48 0.051 0.0651 0.605 0.0611 0.651 0.729 0.787 0.676 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Perceptage cplit 96 66	Mean	abso	lute	erro	or				0.0	88		
More options Relative absolute error 65.7697 % Noor relative squared error 11.45824 % Coverage of cases (0.95 level) 12.7358 % Total Number of Instances 212 === Detailed Accuracy By Class === 10:49:49 - bayes.BayesNet 11:26:37 - trees.J48 0.057 0.006 0.75 0.057 0.105 0.671 11:26:37 - trees.J48 1 11:28:09 - hunchons.MultilayerPerceptron 0 12:44:00 - functions.MultilayerPerceptron 0 14:44:00 - functions.MultilayerPerceptron 0 14:44:00 - functions.MultilayerPerceptron 0 15:50 or 0.000 - 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Percentage spire 78 00	Root	mean	squi	ared	erro	r			0.2	947		
Root relative squared error 114.5824 % Coverage of cases (0.95 level) 65.566 % Mean rel. region size (0.95 level) 12.7358 % Total Number of Instances 212 These.348 0.057 0.066 0.75 0.057 0.105 0.671 T1263:0 - Kurctons.MultilayerPerceptron 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 <t< th=""><th>More options</th><th>Relat</th><th>ive</th><th>abso</th><th>lute</th><th>erro</th><th>r</th><th></th><th></th><th>65.7</th><th>397 💲</th><th></th><th></th></t<>	More options	Relat	ive	abso	lute	erro	r			65.7	397 💲		
(Nom) Previous conviction 65.566 % Start Stop Result ist (right-dick for options) TP Rate FP cation 10:49:49 - bayes.BayesNet 12:7358 % 10:49:59 - bayes.BayesNet 10:57 - trees.J48 11:26:37 - trees.J48 1 0.948 0.649 1 0.787 0.671 11:28:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 Verighted Avg. 0.651 0.651 0.651 0.528 0.691 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		Root	rela	tive	squa	ared	err	or		114.5	824 %		
Wear rel. region size (0.95 level) 12.7358 % Start Stop Resultist (right-ddx for options) Intervention of Instances 212 Intervention of Instances 212 = Detailed Accuracy By Class = D:9:5: 0.057 0.105 0.057 0.057 0.057 0.057 0.676 11:26:37 trees.348 1 0.946 0.649 1 0.787 0.676 11:26:37 trees.348 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	(Nom) Previous conviction	Cover	age	of c	ases	(0.9	5 1	evel)		65.5	56 %		
Start Stop Result list (right-click for options) Detailed Accuracy By Class 10:49:49 - bayes.BayesNet 0.057 0.006 0.75 0.057 0.105 0.671 11:26:31 - trees.J48 1 0.948 0.649 1 0.727 0.676 11:26:37 - trees.J48 1 0.948 0.649 1 0.727 0.676 11:28:09 - functions.MultibayerPerceptron 0 0 0 0 0 0.752 14:44:00 - functions.MultibayerPerceptron 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		Mean	rel.	reg	ion :	size	(0.	95 le	evel)	12.7:	358 %		
Result list (right-click for options) === Detailed Accuracy By Class === 10:49:49 - bayes.BayesNet 0.057 0.006 0.75 0.057 0.105 0.671 10:39:49 - bayes.BayesNet 1 0.948 0.649 1 0.787 0.676 11:26:37 - trees.J48 1 0.948 0.649 1 0.787 0.676 11:28:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 14:49:00 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Start Stop	Total	Num	ber (of In	nstan	ces			212			
10:49:49 - bayes.BayesNet 10:49:55 - bayes.BayesNet 11:26:31 - trees.J48 11:26:37 - trees.J48 11:26:09 - functions.MultilayerPerceptron 14:44:00 - functions.MultilayerPerceptron 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 -	Result list (right-click for options)	=== I)etai	led I	Accu	racy 1	Ву	Class	3 ====	= X			
101-9925-bayes.payesNet 0.0817 P-redsuite ROC AFE 112:63:37 - trees.J48 0.057 0.0057 0.0057 0.0057 0.0057 0.0057 0.0105 0.671 11:28:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0.722 13:44:00 functions.MultilayerPerceptron 0 0 0 0 0 0 0.752 0 0 0 0 0 0 0.01729 0.775 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.0105 0.01	10:49:49 - bayes.BayesNet				TD	Data		ED D.		Duccicics	Decell	E Managemen	DOC Ana
11126:13 - trees348 1 0.937 0.103 0.103 0.101 1126:13 - trees348 1 0.948 0.649 1 0.787 0.676 1128:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0.787 0.676 1128:19 - functions.MultilayerPerceptron 0 0 0 0 0 0 0.787 0.676 1128:19 - functions.MultilayerPerceptron 0 0 0 0 0 0 0.787 0.676 14:14:200 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0 0.782 0 0 0 0 0 0 0 0 0.782 0 0 0 0 0 0 0 0 0.796 0 0 0 0 0 0 0 0 0.782 0 10 0 0 0 0 0 0 0.651 0.551 0.561 0.551 0.551 0.551 0	10:49:55 - bayes.BayesNet				IF	Rate		EF Re	ne	Precision	Recall	r-measure	ROC APE
11125:37 - Grees. 340 1 0.787 0.679 1 0.787 0.676 1125:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0 0 0.729 11:25:09 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 <th>11:26:31 - trees.J48</th> <td></td> <td></td> <td></td> <td>1</td> <td></td> <td></td> <td>0.0</td> <td>000</td> <td>0.75</td> <td>0.057</td> <td>0.105</td> <td>0.671</td>	11:26:31 - trees.J48				1			0.0	000	0.75	0.057	0.105	0.671
11120-00 - functions.MultilayerPerceptron 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 </td <th>11:20:37 - trees. J+o</th> <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td>0.3</td> <td>940</td> <td>0.649</td> <td>1</td> <td>0.767</td> <td>0.070</td>	11:20:37 - trees. J+o					-		0.3	940	0.649	1	0.767	0.070
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	14:44:00 - functions MultilaverPerceptron					0		0		0	0	0	0.729
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	r in the faileastandiaidych creepa sh							0		0	0	0	0.752
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0						0		0		0	0	0	0 706
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0						0		0		0	0	0	0.750
Weighted Avg. 0.651 0.605 0.601 0.651 0.528 0.684 Confusion Matrix a b c d e f g h <						0		0		0	0	0	0.015
Confusion Matrix a b c d e f g h < classified as 3 50 0 0 0 0 0 0 0 0 1 b = ONCE 0 135 0 0 0 0 0 0 0 0 1 b = ONCE 0 15 0 0 0 0 0 0 0 0 1 d = FORTH 1 0 0 0 0 0 0 0 0 0 0 1 d = FORTH 1 0 0 0 0 0 0 0 0 0 0 1 f = SIXTH 0 1 0 0 0 0 0 0 0 0 1 g = TEN 0 1 0 0 0 0 0 0 0 0 0 1 h = SEVEN <		Weigh	nted.	Avg.	(0.651		0.4	505	0.601	0.651	0.528	0.684
a b c d e f g h < classified as 3 50 0 0 0 0 1 a TWICE 0 135 0 0 0 0 1 b ONCE 0 155 0 0 0 0 1 b ENCE 0 6 0 0 0 0 1 d FRICE 0 6 0 0 0 0 1 d FRICE 0 6 0 0 0 0 1 d FRICE 0 1 0 0 0 0 1 fritt fritt 0 1 0 0 0 0 0 fritt fritt 0 1 0 0 0 0 fritt fritt 0 1 0 0 0 0 fritt fritt 0 1 0 0 0		(Confu	sion	Mat	rix =							
3 50 0 0 0 0 1 a = TWICE 0 135 0 0 0 0 1 b = ONCE 0 15 0 0 0 0 1 c = THRICE 0 6 0 0 0 0 1 d = FORTH 1 0 0 0 0 0 1 f = SIXTH 0 1 0 0 0 0 1 f = SIXTH 0 1 0 0 0 0 1 h = SEVEN		a	b	с	d	e	f	a	h	< class:	ified as		
0 135 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 <th></th> <th>3</th> <th>50</th> <th>0</th> <th>0</th> <th>0</th> <th>0</th> <th>ō</th> <th>0 </th> <th>a = TWI</th> <th>Œ</th> <th></th> <th></th>		3	50	0	0	0	0	ō	0	a = TWI	Œ		
0 15 0 0 0 0 1 c = THRICE 0 6 0 0 0 0 0 1 d = FORTH 1 0 0 0 0 0 0 1 e = FIFTH 0 1 0 0 0 0 0 1 f = SIXTH 0 0 0 0 0 0 0 1 f = SIXTH 0 1 0 0 0 0 0 1 h = SEVEN		0	135	0	0	0	0	0	0 1	b = ONCI	2		
0 6 0 0 0 0 0 1 d FORTH 1 0 0 0 0 0 0 1 e FIFTH 0 1 0 0 0 0 0 1 f SIXTH 0 0 0 0 0 0 0 1 f SIXTH 0 1 0 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 1 h SIXTH 0 1 0 0 0 0 1 h		0	15	0	0	0	0	0	0 1	c = THR	ICE		
1 0 0 0 0 0 1 e FIFTH 0 1 0 0 0 0 0 1 f 0 0 0 0 0 0 0 1 f 0 1 0 0 0 0 0 1 f 0 1 0 0 0 0 0 1 h 0 1 0 0 0 0 0 1 h		0	6	0	0	0	0	0	0 1	d = FOR	СН		
0 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 <th></th> <th>1</th> <th>0</th> <th>0</th> <th>0</th> <th>0</th> <th>0</th> <th>0</th> <th>0 1</th> <th>e = FIF</th> <th>СН</th> <th></th> <th></th>		1	0	0	0	0	0	0	0 1	e = FIF	СН		
0 0 0 0 0 0 1 g = TEN 0 1 0 0 0 0 1 h = SEVEN		0	1	0	0	0	0	0	0 1	f = SIX	СН		
0 1 0 0 0 1 h Image: Contract of the server of the		0	0	0	0	0	0	0	0 1	g = TEN			
Status OK		0	1	0	0	0	0	0	0	h = SEVI	EN		
Status OK													
Status OK		•						III					•
	Status OK											Log	x0

Fig 12: Results of the Multilayerperceptron

The figure 12 shows the confusion matrix of the multilayer-perceptron, the correctness of instances that have been well classified.

4.2.2 The Graphical User Interface application

At this section there is the detailed illustration of how the researcher developed the Python GUI application for the project. This is to help in better and clear visualization of the predicted results by use of reports and graphs.

Its developed using the python software which is a widely used general purpose high level programming language and features includes a dynamic type system, automatic memory management and a large comprehensive standard library. More on the need of the python GUI being used is in the summary part of the literature review part of this report.

Dos	Gender	Marital Status	Religion	Region	Occupation	Age	Education	Predicted Previous	Choose a graph to draw
6-Feb-12	Male	Single	Christian	Nyanza	UNEMPLOYED	21	FORM4	1 =	plot_convictions_against_education
29-Apr-14	Male	Married	Christian	Central	MASONRY	40	STD8	1	plot_convictions_against_occupatio
22-Apr-14	Male	Single	Muslim	Western	HAWKER	25	NIL	1	
6-May-14	Male	Married	Muslim	Central	BUSINESSMAN	34	POSTGRADUATE	1	
07-May-14	Male	Divorced	Christian	Western	CARPENTER	32	FORME	1	
14-Jun-10	Male	Married	Christian	Western	SECURITYGUARD	40	FORME	1	
29-Jul-12	Male	Married	Christian	Central	DRIVER	35	STDB	1	
29-Jul-12	Male	Married	Christian	Riftvalley	SECURITYAIRPORT	29	UNIVERSITY	1	
17-Mar-14	Male	Married	Muslim	Eastern	STEELFEXER	51	FORM4	1	
19-Feb-14	Male	Married	Christian	Nyanza	DRIVER.	30	FORM4	1	
24-Nov-12	Male	Married	Christian	Nairobi	PLUMBER	50	FORM4	1	
09-Apr-14	Male	Married	Muslim	Riftvalley	MASON	25	STD8	1	
13-Jan-14	Male	Married	Muslim	Nyanza	TAILOR	50	STD4	1	
16-Sep-12	Male	Married	Muslim	Riftvalley	BUSINESSMAN	76	STD8	1	
23-Aug-12	Male	Married	Christian	Riftvalley	тоит	30	STD7	1	
05-Jun-14	Male	Single	Christian	Western	CASUALLAEOURER	18	FORM2	1	
2/8/2008	Female	Single	Christian	Eastern	SELFEMPLOYED	36	LITERATE	1	
15-0ct-12	Male	Married	Christian	Riftvallev	CHARCCALSELLER	23	STDB	1 *	
Gender	_	coupation	Marital Statu	is [Relgion	Reg	ian (Education	
				Generate R	eport		1000 - 100 		

Fig 13: Report generated by the GUI application

The Figure 13 shows the actual GUI application interface where we have the report part and the graphs, where the nine attributes are shown of the instances. In this case the figure 13 is a report of all the attributes and some instances from the predicted result. From the report the user can opt to fetch specific data, for instance the female or male, occupation, marital status, religion and education.

Functionalities

- 1. Input of predicted results from WEKA
 - Predict on recidivism rate
- 2. Decision support function
 - Visualization of prediction on graphs to the users



Fig 14: Graph on previous conviction prediction and occupation

The Figure 14 shows a graph from the application which shows the rate of recidivism against various occupations, where the driver and hawker are the likely persons to be reconvicted.

	Dos	Gender	Marital Status	Religion	Region	Occupation	Age	Education	Predicted Previous Conviction	Previous Conviction	
1	6-Feb-12	Male	Single	Christian	Nyanza	UNEMPLOYED	21	FORM4	1	1	
	22-Apr-14	Male	Single	Muslim	Western	HAWKER	25	NIL	1	7	
	23-Aug-12	Male	Single	Chr <mark>ist</mark> ian	Riftvalley	WATCHMAN	25	STD8	1	1	
	14-Jun-10	Male	Married	Christian	Western	SECURITY GUARD	40	FORM3	1	1	
	29- <mark>.</mark> ul-12	Male	Married	<u>Christian</u>	Central	DRJVER	35	STD8	1	2	
	29-Jul-12	Male	Married	Christian (Riftvalley	SECURITYAJRPORT	29	UNIVERSITY	1	2	
	17-Mar-14	Male	Married	Muslim	Eastern	STEELFIXER	61	FORM4	1	2	
	19-Feb-14	Male	Married	Christian	Nyanza	DRJVER	30	FORM4	1	1	
	09-Apr-14	Male	Married	Muslim	Riftvalley	MASON	25	STD8	1	1	
	13-Jan-14	Male	Married	Muslim	Nyanza	TAILOR	60	STD4	1	1	
k	22-Apr-14	Male	Married	<u>Christian</u>	Riftvalley	CONDUCTOR	31	FORM4	1	2	
2	16-Sep-12	Male	Married	Muslim	Riftvalley	BUSINESSMAN	76	STD8	1	3	
3	23-Aug-12	Male	Married	Christian	Riftvalley	TOUT	30	STD7	1	2	
4	05-Jun-14	Male	Single	Christian	Western	CASUALLABOURER	18	FORM2	1	1	
5	15-0ct-12	Male	Married	Christian	Central	WAITER	20	FORM3	1	1	
.6	15-0ct-12	Male	Married	Chr <mark>isti</mark> an	<mark>Riftvalle</mark> y	CHARCOALSELLER	23	STD8	1	1	
7	13-May-14	Male	Married	Muslim	Eastern	NIL	27	UN <mark>EVER</mark> SETY	1	3	
9	16.Sen.17	Male	Married	Muelim	Factore	DP1/ER	18	FORM	1	1	-531
Ge	nder			ton		Marital Status		E	Relgon	Region	Education

• Report generation

Fig 15: Report of male convicts on rate of recidivism

The Figure 15 shows a report sample of the male convicted persons and other attributes.

CHAPTER FIVE

5. Results

5.1 Introduction

In this section, the results obtained from the developed prototype are described. The purpose is to establish if the prototype met the functional requirements of the system and if the results can be relied on to make a decision on various management issues on prisoner's rehabilitation. This is on the viable rehabilitation programs to be used on a prisoner be it incarnation, parole, community service among others.

	Test options	Training set	Percentage split
	Algorithms	Correctly classi	fied instances
1	BayesNet	76%	63%
2	J48	62%	63%
3	Multilayerperceptron	62%	65%

Table 1: Tabulation results from the WEKA algorithms

The table 1 shows the variation on various algorithms accuracy level done using different methods; the training data and the percentage split, whereby the results from the ANN (multilayerperceptron) are more reliable since it has a higher accuracy level compared to the other techniques used i.e. the BayesNets and J48.

Where the prediction value is TWICE and ONCE, thus for those with a higher value from twice and above are considered of high risk so they can be proposed to be treated in special way to avoid their chance of being convicted again after release, or introduction of various programs that will help cub the chances of the convicted prisoners being convicted again.



Fig 16: Graphical representation on Age and previous conviction prediction

Figure 16 shows a graph on result of the predicted values visualized that can assist the user in various decisions as far as recidivism is concerned, where the rate of recidivism is compared with the age. From the graph there is the age group which is more prone to recidivism than other, the age between 23 and 32.

The RPS assist in strategic recidivism analysis as it is concerned with long term problems and planning for long term projects, by allowing examination of long term increase or decrease in recidivism.

Also include administrative analysis focus by providing summary data, statistics and general trend information to the prison management.





From the graph in figure 17 the user of the system can be able to tell the level of education of those with a high risk of recidivism.

As assessing recidivism through analysis helps in prevention efforts, because prevention will cost less, compared to the cost incurred when there is high population in the prison institutions especially due to high rate of recidivism.

5.2 System Evaluation

Considering the results from the algorithms, of all the instances there is a prediction of a prisoner being convicted again. As observed there are those whose chances are once or twice, depending on other attributes of that specific instance.

5.3 System testing

In this section the objective was to verify that the system had the functionalities required to monitor recidivism in the prisoner's population. How well the two applications interface and give an end result which can be used by the prison management in decision making.

This is from the reports and the graphs generated or displayed when running. Some of the reports and graphs include:

- ▶ Figure 15 report of male convicts on rate of recidivism
- > Figure 16 Graph representation age and previous conviction prediction
- > Figure 17 Graph representation level of education and previous conviction prediction

5.3.1 User acceptance testing

This formed the final stage of testing the developed RPS prototype. The officers working in the three (3) stations; langata women, Nairobi west and Nairobi medium prisons at the data entry point of prisoners' records and release of prisoners were given access to use the prototype. The main objective being to check if the user expectations were met by the prototype developed.

In order to ensure proper testing of the prototype, the researcher interviewed a number of officers from the three stations; Nairobi west prison, Nairobi medium and Langata women. The officer's interviewed were senior, middle and junior officers in the institution. Those interviewed were eight officers at least two from the three stations that data had been corrected from and one officer based at the prison headquarters.

Out of the eight officers interviewed six of them were positive towards the use of the prototype as a tool to help in the rehabilitation and reformation in the department. This is because it provides the knowledge to the users on determining the recidivism rate of a new prisoner who has just been brought from court by comparing his or her details provided with the prototype existing predictions.

Summary on the interview results

From the interview conducted, the officers from the various stations in the KPS agreed that recidivism truly exist in the department. The various modes/programs for rehabilitation and reformation of the prisoners include:

- Vocational training
- Professional courses
- Formal learning
- Counseling

- Chaplaincy
- Sports and recreation
- Offender development
- Case management
- Volunteer and placement

Though it was noted that the level of automation in the department is generally poor, but if systems like RPS were implemented they would be of great help in the listed programs on rehabilitation.

The benefits of the RPS to the department from the interview result:

- a) The system would be of much help to the department if used together with the existing measures due to the sensitivity of the issue; the convicted person, as a person whose chance of reconviction is once can be considered for other rehabilitation programs like parole or community service after serving his sentence for a while among other factors.
- b) Allow development of other programs that would be of help to control the rate of recidivism in the department

Challenges encountered from the interview result

a) Being a new technology in the department enough training is needed to show how well the RPS is relevant to the needs of the department on recidivism

Thus it meets the intended goal of a recidivism pattern and the prediction rate which were the goals during the initial phase on recidivism understanding.

CHAPTER SIX

6. Conclusion, Recommendation and Future Works

6.1 Conclusion

The prison department has a large volume of data especially on prisoners that if it was well stored and data mined it would be of much assistance to the prison department management, as illustrated by the development of the operational prototype on the RPS.

The big data within the department has not adequately been used for analysis and predicting of future trends to aid in decision making process. There is no prediction in place if any they only rely on numbers especially on recidivism and projected future numbers which is not realistic.

The effective knowledge discovery techniques and tools of data mining in the modern world are important in the building of intelligent analysis and prediction systems from the big data in various industries. Data mining and prediction tools like WEKA used in this research and the prototype building have proved to be very efficient in prediction from the big data available in the department on recidivism.

The objectives set earlier at the introduction of the project, have been realized as follows:

a) To identify and analyze the variables to be used to predict recidivism in the prison inmates population.

From the existing knowledge on recidivism illustrated in the literature review during the research the objective was achieved. By the identification of the risk factors which are factors that if prisoner posses will have a higher rate to be reconvicted. This includes occupation, age, level of education and marital status.

b) To identify a data mining technique suitable to predict recidivism in the prison inmates population.

This is also realized from the existing knowledge in the literature review, where various techniques have been used in the criminal justice system to check recidivism among other areas.

The most common techniques being; Bayesian, neural networks, rule induction and decision tree whose accuracy levels vary depending on the area applied.

c) To develop a prototype application using an identified data mining technique

Using the WEKA tool which comprises of a number of algorithms (techniques) it's used to predict the rate of recidivism. The results from the algorithms (Bayesnets, J48 and multiperceptron) are compared for that with best accuracy level. Whose results are displayed using the python developed application in the form of reports and graphs in the 5th chapter on results.

d) To test and validate the prototype

After the prototype development the end users from the prison department get to interact with it during the testing phase, and from their response the system is found viable to the needs on the ground. As illustrated in the system testing section in the 5^{th} chapter.

e) To display existing recidivism patterns using the prototype application

Using the python application there is a better visualization of the prediction results from the WEKA tool by use of graphs and reports, which are crucial in the decision making process.

The prototype is therefore a useful piece of invention that prison department management can use to predict recidivism rate and plan on various programs to introduce or not. The only limitation of the system is that it can only help the prison management in decision making but not replace the management.

6.2 Recommendation

The efficiency of the prototype depends largely on availability of accurate data from all the prisons institutions in the country. My recommendation to the prison department is to implement proper and full automation of the prisoner's records using the Offender Records Management System (ORMS). To enhance the functioning of prediction systems built from the data.

This will aid in advancement of the system to enable it to include more specific cases on recidivism at a larger scale.

This prototype has been built using python and data feed using file in CSV and AARF format while WEKA is implemented using Java platform. Thus predicted results could not seamlessly accessed by the python GUI and had to be uploaded manually, reducing the flexibility of scenarios that the user can try within the prototype outside WEKA in case such data has not been uploaded.

Therefore it's recommended that the system in future be built in java to aid a seamless integration of WEKA with the system.

6.3 recommendations for future work

This project confined the research to only three (3) prisons within the prison department and it can be expanded to other prisons within the country. More so it can also be implemented in other justice administration bodies like the Police, Probation Department and Judiciary.

References

1. Andrews, D. A., Bonta, J., & Wormith, J. S. (2006) 'The Recent Past and Near Future of Risk and/or Need Assessment', Crime & Delinquency, 52(1), pp.7-27.

2. Barnes .H, Keller .M .(2009). "Predicting Recidivism in Adolescent Males Using the Minnesota Multiphasic Personality Inventory" – A and the Trauma Symptom Checklist for Children (Doctoral dissertation, Pacific University). Available at: <u>http://commons.pacificu.edu/spp/104</u>

3. Berry, M. J. A. and Linoff, G. S. (2010) *Data mining techniques for marketing, sales & customer relationship management.* 2nd.New jersey: Wiley publishing, inc

4. Booth. T. 2007." Neural Networks and Artificial Intelligence; Predicting human behaviour", Activ8 intelligence, pp, 4-7. Available at: <u>http://www.a8i.co.uk/uploads/whitepapers/nn_and_ai_research2.pdf</u> [Accessed 24 January 2014].

5. Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical Risk Assessment: Old Problems and New Applications. Crime & Delinquency, 52(1), 178-200.

6. Government of Kenya. *Power Of Mercy Act* (2011) Nairobi: National Council for Law Reporting

7. Government of Kenya. *The Prisons Act Cap. 90* (rev.2009) Nairobi: National Council for Law Reporting

Gray. B, Birks .D, Allard .T, Ogilvie .J, Stewart . A and Lewis .A.(2008). "Exploring the Benefits of Data Mining on Juvenile Justice Data" Available at: http://www98.griffith.edu.au/dspace/bitstream/handle/10072/21293/53136_1.pdf?sequence = 1 [Accessed 24 January 2014].

9. Harris .P, Mennis .J, Obradovic. Z, Izenman .A, Grunwald .H, Lockwood B, Jupin .J, and Chisholm .L. (2012). "Investigating the Simultaneous Effects of Individual, Program and Neighborhood Attributes On Juvenile Recidivism Using GIS and Spatial Data Mining", Available at: <u>https://www.ncjrs.gov/pdffiles1/nij/grants/237986.pdf</u> [24 January 2014].

10. Henslin, J. 2008 "Recidivism Using Neural Networks. Socio-Economic Planning Sciences, 34, 271-284.

11. Howard.J.(2000)."OffenderRiskAssessment",Availableat:http://www.johnhoward.ab.ca/pub/pdf/C21.pdf[Accessed 14 January 2014].

12. Kavoc S. (2012) "Suitability analysis of data mining tools and methods" Degree thesis. Available at: <u>http://is.muni.cz/th/255695/fi_b/suitability_analysis_of_data_mining_tools.pdf</u> [Accessed on 10 March 2014].

13. Kenya National Bureau of Statistics (2014) Economic Survey 2014. Nairobi

14. Li .S, Kuo .S and Tsai .F. (2010)" An intelligent decision-support model using FSOM and rule extraction for crime prevention" Expert Systems with Applications. Available at: http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/HK2%20-%202012%20-

%202013/Tieu%20luan/2010%20An%20intelligent%20decision-

support%20model%20using%20FSOM%20and%20rule%20extraction%20for%20crime%20
prevention.PDF [Accessed on 24 January 2014]

15. O'Connor, T. (2013). "Recidivism prediction" Developmental Prevention of Crime and Terrorism. Available at: <u>http://www.drtomoconnor.com/3440/3440lect06.htm</u> [Accessed on 14 January 2014]

16. Palocsay, S. W., Wang, P., & Brookshire, R. G. (2000). Predicting Criminal SocialProblems:ADown-To-EarthApproach."Availableat:http://www.drtomoconnor.com/3440/3440lect06.htm/

17. Rahim .A, (2014). "Best practices for business intelligence and predictive analytics". Available at: <u>http://www.informationbuilders.com/new/newsletter/13-04/3ali</u> [Accessed 14 march 2014]

18. Ritter .N. (2013). "Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise", Available at: <u>https://www.ncjrs.gov/pdffiles1/nij/240696.pdf</u> [Accessed on 14 January 2014]

19. Rohanizadeha S. S, Moghadama M. B (2009). "A Proposed Data Mining Methodology and its Application to Industrial Procedures" Journal of Industrial Engineering 4, pp.37-50

20. Silver, E., & Miller, L. L. (2002). A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control. Crime & Delinquency, 48(1), 138-161.

21. Tiwaria, Abhishek, Sekhar, and Arvind K.T. (2007). "Workflow based framework for life science informatics". *Computational Biology and Chemistry* **31** (5-6): 305–319.

22. Wahbeh, A. H, Al-Radaideh, Q. A., Al-Kabi, M. N. and Al-Shawakfa E. M (2008)"A Comparison Study between Data Mining Tools over some Classification Methods". Available at: <u>http://www.thesai.org/downloads/SpecialIssueNo3/Paper%204-</u> <u>A%20Comparison%20Study%20between%20Data%20Mining%20Tools%20over%20some</u> %20Classification%20Methods.pdf [accessed on 10 March 2014].

23. Witten, I.H, Frank, E & Hall, M.A (2011) *Data mining practical machine learning tools and techniques*. 3rd. Burlington: Morgan Kaufmann

24. Yang .M, Liu .Yand Coid. J. (2012). "Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism". Available

at: <u>www.justice.gov.uk/downloads/publications/research-and-analysis/moj-research/neural-</u> <u>networks-research.pdf</u> [Accessed 20 January 2014].

25. Mu-sigma (2014), Neural network "A neural network is a powerful data modeling tool that is able to capture and represent complex input/output relationships." <u>http://www.mu-sigma.com/analytics/thought_leadership/cafe-cerebral-neural-network.html</u>

26. Howell, J. (2003). *Preventing and Reducing Juvenile Delinquency: A Comprehensive Framework*. Thousand Oaks: Sage Publications.

27. Omboto. J. (2010). *Challenges Facing the Control of Drugs and Substances Use and Abuse in Prisons in Kenya: The case of Kamiti Prison*. Unpublished MA Project, University of Nairobi.

28. Heseltine, K., Sarre, R. & Day, A. (2011). *Correctional offender treatment programs: The 2009 national picture in Australia*. Canberra: Australian Institute of Criminology.

29. Hoffman, J. (2004) *Youth Violence, Resilience, and Rehabilitation*. New York: LFB Scholarly Publishing LLC

Appendices

Appendix A – Interview Questions	
1. Full Name (optional):	
2. Name of Prison	
3. Position held	
Question 1.	□ Yes
Does recidivism exist in Kenya Prison service?	□ No
Question 2:	
What are the various modes of rehabilitation and reformation of the prisoners	
within Kenya prison Service?	
a)	
b)	
c)	
d)	
Question 3:	
What is the level of automation within the Kenya Prisons service on prisoner's	Der Poor
records?	Average
	🖂 Good
Question 4:	Yes
Would the RPS system be of help in rehabilitation and reformation of prisoners?	No No
Question 5:	
In your opinion what should be taken into consideration of the final system for	
better performance.	
	1

Appendix B- Sample code

...

```
Plot graphs
....
import os
import re
from csv import DictReader
from collections import OrderedDict
from pylab import *
PATH = os.path.join(os.path.dirname(___file___), 'data_.csv')
class DataRow(object):
  ....
  Represents a single row(instance) of data
  ....
  def __init__(self):
    self.dos = None
    self.DOS = 0
    self.gender = None
    self.GENDER = 1
    self.marital_status = None
    self.MARITAL_STATUS = 2
    self.religion = None
    self.RELIGION = 3
    self.region = None
    self.REGION = 4
    self.occupation = None
    self.OCCUPATION = 5
    self.age = None
    self.AGE = 6
    self.education = None
    self.EDUCATION = 7
    self.predicted_previous_conviction = None
    self.PREDICTED_PREVIOUS_CONVICTION = 8
    self.previous_conviction = None
```

```
self.PREVIOUS_CONVICTION = 9
  def __getitem__(self, key):
    return getattr(self, key)
  def __setitem__(self, key, value):
    return setattr(self, key, value)
  def is_equal(self, **kwargs):
    ...
    If the passed kwargs match the datarow values
    ...
    for key, value in kwargs.items():
       slugified key = slugify(key)
      if not hasattr(self, slugified_key):
         return False
      if not getattr(self, slugified_key) == value:
         return False
    return True
class Data(object):
  def __init__(self, f_path):
    self.f = open(f path, 'rb')
    self.reader = DictReader(self.f)
    self.fieldnames = self.reader.fieldnames
    self.x = 'predicted_previous_conviction'
    self.set_data()
  def yield_rows(self, return_object=False):
    ...
    Returns the rows in the file
    ...
    self.f.seek(0)
    for row in self.reader:
       if return object:
         obj = DataRow()
         [setattr(obj, slugify(key), row[key])
         for key in row]
         yield obj
      else:
         yield row
  def set data(self):
    ....
    Sets the data
    ....
    enum = enumerate(self.yield_rows(True))
```

```
self.data = []
  for i, data in enum:
    self.data.append(data)
  self.data = self.data[1:]
def map_age_count(self, data):
  ...
  Returns the count of ages
  ...
  field = 'age'
  # Get the min and max ages
  min age = 0
  max age = 0
  ages = []
  for row in data:
    ages.append(int(row[field]))
  min_age = min(ages)
  max_age = max(ages)
  # Get the age map
  age_map = OrderedDict()
  for age in range(min_age, max_age + 1):
    age map[age] = 0
  for row in data:
    try:
      age_map[int(row[field])] += int(row[self.x])
    except TypeError as e:
      print e
  return age_map
def plot_convictions_against_age(self, data):
  ....
  A graph of convictions against age
  ....
  graph = self.map_age_count(data)
  # Use one figure
  figure(0, figsize=(15, 10))
  hold(True)
  # Set grid
  grid(True)
  # Set the grid parameters
  yticks(arange(min(graph.values()), max(graph.values()), 2))
  xticks(arange(min(graph.keys()), max(graph.keys()), 4))
  # Set the labels
  xlabel('Age [Years]')
  ylabel('Predicted Convictions [count]')
```

```
# Title
  title('Predicted Previous Convictions vs Age')
  # Plot the graph
  plot(graph.keys(), graph.values(), 'x-', color='#000000', lw=2)
  # bar graph
  bar(
    graph.keys(),
    graph.values()
  )
  # Show the graph
  show()
def map occupation count(self, data):
  ....
  A mapping of occupations and convictions of each occupation
  ...
  name = 'occupation'
  occupations = set()
  occup_map = OrderedDict()
  for row in data:
    if row[name] not in occupations:
      occup_map[row[name]] = 0
    occupations.add(row[name])
  # Get the occupation mapping
  for row in data:
    occup_map[row[name]] += int(row[self.x])
  # Return the map
  return occup_map
def plot_convictions_against_occupation(self, data):
  ш
  A graph of convictions against occupations
  ...
  try:
    graph = self.map_occupation_count(data)
    # Use figure 0
    figure(0, figsize=(15, 10))
    # plot the bars
    bar(arange(0, len(graph)), graph.values(), align='center', width=.8)
    # Set the labels
    xticks(arange(0, len(graph)), graph.keys(), rotation=85, fontsize=9)
    yticks(arange(0, max(graph.values()), 2))
    # grid
    grid(False)
    # title
```

```
52
```

```
title('Predicted Previous Convictions vs Occupation')
    # labels
    xlabel('Occupation')
    ylabel('Convictions [count]')
    # show
    show()
  except Exception as e:
    print e
def map_education_count(self, data):
  ...
  Education mapping
  ...
  name = 'education'
  educations = set()
  edu_map = OrderedDict()
  for row in data:
    if row[name] not in educations:
      edu map[row[name]] = 0
    educations.add(row[name])
  # Get the occupation mapping
  for row in data:
    edu map[row[name]] += int(row[self.x])
  # Return the map
  return edu_map
def plot_convictions_against_education(self, data):
  A graph of convictions against education
  ...
  try:
    graph = self.map_education_count(data)
    # Use figure 0
    figure(0, figsize=(15, 10))
    # plot the bars
    bar(arange(0, len(graph)), graph.values(), align='center', width=.8)
    # Set the labels
    xticks(arange(0, len(graph)), graph.keys(), rotation=50, fontsize=8)
    yticks(arange(0, max(graph.values()), 4))
    # grid
    grid(False)
    # title
    title('Predicted Previous Convictions vs Education')
```

```
# labels
xlabel('Education')
ylabel('Convictions [count]')
# show
show()
except Exception as e:
print e

def slugify(string_value):
""
Slugifies a string
""
return string_value.lower().replace(' ', '_')
```

```
def wordify(slug):
```

```
...
```

Undos slufigy

```
return re.sub('_+', ' ', slug).title()
```