



UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND INFORMATICS

**USE OF DATA MINING TO CHECK THE
PREVALENCE OF PROSTATE CANCER: CASE OF
NAIROBI COUNTY**

By

NGARUIYA, MARY NJERI

(P58/76265/2012)

Supervisor

CHRISTOPHER MOTURI

A project report submitted in partial fulfillment of the requirement for the award of Masters of
Science in Computer Science of the University of Nairobi

September 2014

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

Mary Njeri Ngaruiya: _____

Date: _____

(P58/76265/2012)

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

Christopher A Moturi: _____

Date: _____

Deputy Director

School of Computing and Informatics

Acknowledgement

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

To my loved ones, for their great support and encouragement throughout my academic years.

To Ms Rachel Rasjut of School of Mathematics in helping out with ANOVA.

To Martha Wambugu, for proof reading my document from proposal to the final report.

To my supervisor Mr. Christopher Moturi who has opened my eyes to the research world. His guidance, support, and positive criticism made this project a success. To the panellists, Dr. Agnes Wausi, Mr. Evans Miriti and Mr Samuel Ruhui to whom I am grateful for positive criticism that has led to success of this project.

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

Table of Contents

Declaration	i
Acknowledgement	ii
List of Figures	v
List of Tables.....	vi
List of Abbreviations	vii
ABSTRACT	viii
INTRODUCTION.....	1
1.1 Background of Study	1
1.2 Statement of the Problem	2
1.3 Research Objectives	3
1.4 Justification and Significance of the Study.....	3
1.5 Scope of the Study	4
CHAPTER 2.....	5
LITERATURE REVIEW	5
2.1 Introduction To prostate cancer	5
2.2 Prostate cancer Risk Factors.....	5
2.3 Cancer In Africa	6
2.4 Predictive Models Used In prostate cancer	8
2.5 Application of datamining in cancer.....	10
2.5.1 Data Mining In Breast Cancer	11
2.5.2 Data Mining In Cervical Cancer	11
2.5.3 Data mining in prostate cancer	12
2.6 Preferred Algorithm.....	13
CHAPTER 3.....	14
METHODOLOGY	14
3.1 Research Design	14
3.2 Overview of CRISP-DM.....	14
3.3 Overview of WEKA	15

3.4	Sources of Data and Target Population.....	16
3.5	Data Collection and Analysis	17
3.6	Proposed Model.....	18
3.7	Requirements	18
CHAPTER 4.....		19
RESULTS AND DISCUSSION		19
4.1	Data Pre-processing	19
4.2	Modeling Techniques, Tools and Modeling Infrastructure Used.....	21
4.3	Data Analysis and Results.....	21
4.3.1	Basic Classification Results using WEKA.....	21
4.3.2	R Algorithm Analysis	27
4.4	Proposed Prototype	37
CHAPTER 5.....		40
CONCLUSIONS AND RECOMMENDATIONS.....		40
5.1	Achievements	40
5.2	Contribution of the Study.....	40
5.3	Limitations of the Study.....	41
5.4	Recommendation for Future work.....	41
References.....		42
Appendices.....		45

List of Figures

Figure 1: Data mining process	1
Figure 2: Most common cancer sites in Africa by Sex and Country are shown for 2008.....	6
Figure 3: A Visual Guide to CRISP-DM Methodology.....	14
Figure 4: Header information in ARFF Source	16
Figure 5: Data in ARFF.....	16
Figure 7: MLP for Nairobi Central	22
Figure 8: Decision tree for Nairobi Central.....	22
Figure 9: Clustering Dataset A	23
Figure 10: Data Visualization for DatasetA	24
Figure 11: Clustering Dataset B	24
Figure 12: Data Visualization for DatasetB	25
Figure 13: Clustering Dataset C	26
Figure 14: Data Visualization for DatasetC	27
Figure 16: BoxPlot of Cluster A.....	29
Figure 19: Boxplot of cluster B (Nairobi East).....	32
Figure 20: Boxplot for Cluster C.....	34
Figure 21: Boxplot for Cluster D.....	35
Figure 22: How Prototype Works.....	38
Figure 23: Entity Relationship Diagram for the Database	39

List of Tables

Table 1: Part of the Raw Data	19
Table 2: Zones of Nairobi County according to City Council of Nairobi	20
Table 3: Data Generalization according to major Hospitals.....	21
Table 4 : F- Value Table for Cluster A	29
Table 5: The mean, variance and standard deviation of cluster A.....	30
Table 6: F-Value of Cluster B	31
Table 7 : the mean, variance and standard deviation of cluster B.....	31
Table 8:The F-Value table for cluster C.....	33
Table 9: The mean, variance and standard deviation for Cluster C	33
Table 10 :The F-Value table for cluster D.....	34
Table 11:The mean, variance and standard deviation for Cluster C	35

List of Abbreviations

ANN	-	Artificial Neural Network
ANOVA	-	Analysis of Variance
ARFF	-	Attribute-Relation File Format
ASCII	-	American Standard Code for Information Interchange
CLI	-	Command Language Interface
CRISP-DM	-	CRoss Industry Standard Process for Data Mining
DRE	-	Digital Rectal Examination
GUI	-	Graphical User Interface
KNCCS	-	Kenya-National-Cancer-Control-strategy
LR	-	Linear Regression
PSA	-	Prostate Specific Antigen
WEKA	-	Waikato Environment for Knowledge Analysis

ABSTRACT

Prostate cancer has been on the rise in the past years and alarming cases being found in men in their 20s. The problem is that most of the cases are diagnosed in their late stages thus the mortality rate being high. In recent years data driven analytic studies have become a common complement with new and novel research where different tools and algorithms are taking a centre stage in cancer research. In this research, the main goal is to use datamining to derive patterns which will be used in building a prognostic tool that helps in identification of the Gleason score once screened and advice on the treatment technique. In this research, we used two popular data mining tools (R Environment and WEKA) which exhibited almost same results .The dataset contained around 485 records and 7 variables. In WEKA, a 10-fold cross-validation was used in model building in comparing ANN and J48. The results showed that ANN is the most accurate predictor compared to J48 in all the instances. This study contributes to society, academics and cancer research which ultimately assist in reduction of mortality rates by use of pattern recognitions which leads in better decision making.

Keywords: Artificial Neural Network, Data Mining, GIS, prostate cancer, J48 (decision trees), R, WEKA

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Data mining is the non trivial extraction of implicit, previously unknown, and potentially useful information about data. To achieve these patterns and trends, data mining relies on sophisticated mathematical and statistical models, and substantial computing power to help user convert algorithmic behavior to user understandable rules for action (decision making) and forecasts the effects of these actions. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by healthcare administrators to improve quality of service (Kaun H and Wasan, S.K, 2006).

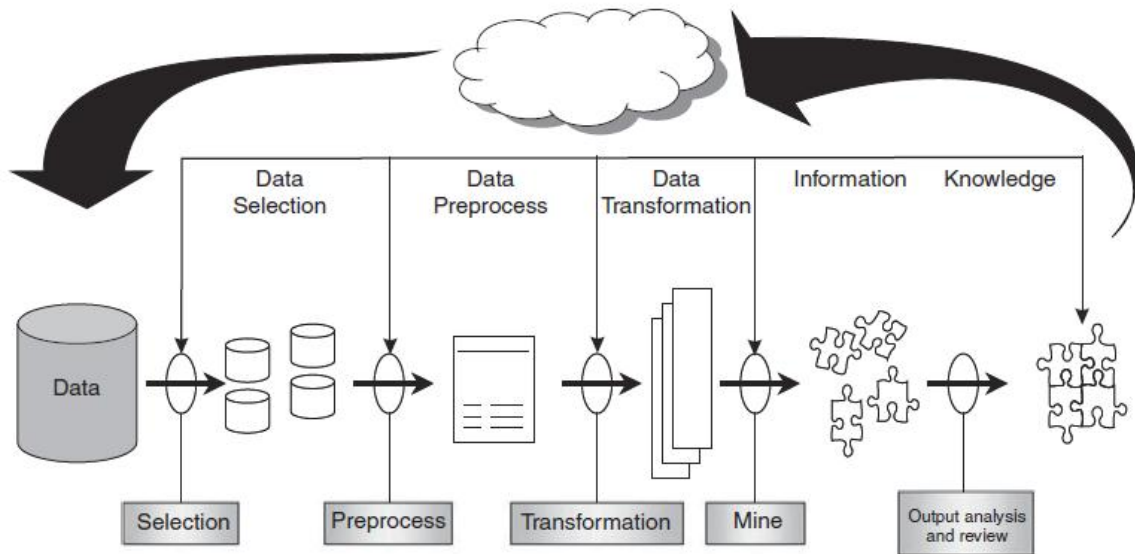


Figure 1: Data mining process

Prostate cancer is among the top killing cancers. It has been ranked sixth globally and third in Kenya as a cause of death after infectious and cardiovascular diseases. Family history, age, diet, weight (obesity), high-risk habits (smoking, heavy drinking), and exposure to environmental pollutants help to predict an individual's risk of developing cancer. Prostate cancer is a non communicable disease that only affects men. It is common among men who are above 50 years

but a few cases reported in men below 50 years. Most cases are diagnosed in men over 65 years of age (Magoha, 2000).

The epidemiology of prostate cancer is complex and has few established risk factors. However, most established risk is associated with family history, age, race, country and testosterone deficiency (Magoha, 2000). American Cancer Society shows more modifiable risk factors, which can be changed. These modifiable risks include smoking, obesity, workplace exposure, sexually transmitted diseases, and diet.

According to (KNCCS, 2011) the disease cannot be eradicated but its effects can be significantly reduced if effective measures are put in place to control risk factors, detect cases early and offer good care to those with the disease.

Algorithms have been used successfully recently to learn the cancer pattern. In this study, two tools were used i.e. Waikato Environment for Knowledge Analysis (WEKA) software and R - which contains several state-of-the-art learning algorithms- to establish prostate cancer prevalence patterns. Policy makers, medical practitioners (physicians), and patients can use these patterns to provide essential input into the rational planning of cancer control programmes.

The use of data mining techniques and knowledge management technologies in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine. This movement towards predictive medicine is important, not only for patients (in terms of lifestyle and quality-of-life decisions) but also for physicians (in making treatment decisions) and health economists and policy planners (in implementing large scale cancer prevention or cancer treatment policies) (Cruz, J. A. and Wishart, D. S, 2006).

1.2 Statement of the Problem

An ideal situation would be a cancer free continent but the reality is that this disease is becoming common and leading to several deaths each year. The vast medical data on prostate cancer patients should be carefully studied so that good decisions can be applied to reduce this epidemic.

This academic research study carried out a comparative analysis of data mining techniques in prostate cancer, which were able to detect patterns. These patterns were used to check prostate cancer prevalence by incorporating major variables such as age, location and stage of cancer.

We finally identified the patterns that are to assist the medical practitioners and the government to determine where to spread out available treatment options (watchful waiting, radical prostatectomy and radiation therapy) in most constituencies hospitals within Nairobi County. We feel that these patterns can be used to get donor funds to cater for the equipment required in diagnostic centers and establishment of palliative cares and hospices in the areas that showed high cases of advanced prostate cancer.

1.3 Research Objectives

The main goal of this research was to use datamining to derive patterns from the four different zones (from Nairobi County) which will be used in building a prognostic tool that helps oncologist in identification of the Gleason score once screened and deciding the treatment technique. This prognostic tool will assist in early detection of the prostate cancer thus cut back on high rates of reported cases of advanced prostate cancer and deaths.

The objectives of this study:

- i. Identify data mining techniques appropriate for prostate cancer analysis.
- ii. Propose a model for retrieving patterns of prostate cancer incidences.
- iii. Develop a prototype for community awareness (prognostic tool).

1.4 Justification and Significance of the Study

The Government of Kenya has set ambitious targets for providing essential health services to all its citizens. Improving accessibility and equity of health services requires a well trained, managed, and motivated workforce and programmes that will push for community wellness (Intrahealth International, 2012)

This research purpose was to check the incidence, trends against different factors such as age and demographics. This way the policy makers will be able to arrive at good decision or judgment in establishing centers and also developing programmes that will create awareness in a community.

1.5 Scope of the Study

This study focused on cases of prostate cancer that have been reported and treated in Nairobi County in both private and public hospital.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction To prostate cancer

The prostate is a gland in the male reproductive system, which is located under the bladder. It produces seminal fluid that carries sperm. Prostate function is regulated by testosterone, a male sex hormone produced by the testicles. There are several types of cells in the prostate but research evidence demonstrates that prostate cancer develops from gland cells.

Churchill Livingstone's Dictionary of Nursing (2006) defines prostate cancer as a common disease that affects elderly men and it's characterized by various factors which include dysuria (painful urination), cystitis (urine infection in the bladder), prostatitis (inflation of the prostate and its surroundings), and urine retention.

Prostate cancer can spread to any part of the body but most mainly affects the bones and lymph nodes. It can also spread to the urethra, a tube that carries urine through the penis, the bladder, ureters, which carry urine from the kidneys to the bladder (ureters) and part of the bowel (prostate cancer UK, 2013).

The major decision point in care delivery is the Gleason score and age because it helps to evaluate prognosis. Prostate cancer is first diagnosed, graded, and staged. Treatment is then delivered depending on the situation. Treatment options for men with prostate cancer include: watchful waiting or active surveillance, surgery, radiation therapy, cryosurgery (cry therapy), chemotherapy and also vaccine

2.2 Prostate cancer Risk Factors

The main cause of prostate cancer is not clearly known but we have some modifiable and non-modifiable risk factors that are linked to the disease. Several studies show that some factors like family history, age, diet, weight (obesity), high-risk habits (smoking, heavy drinking), and exposure to environmental carcinogens, have a positive association with prostate cancer (Cruz, J. A. and Wishart, D. S, 2006; Quinn M and Babb P, 2002; Jemal et. al, 2012). The leading cause of death among prostate cancer patients is the spread of the disease to other parts of the body especially the bones (metastatic cancer).

Platz et. al. (2000) states that the established risk factors for prostate cancer are primarily non-modifiable such as age and race. However, modifiable risk factors beginning to emerge (Quinn M and Babb P, 2002) found a positive association between risk of prostate cancer and moderate consumption of spirits. Middle aged men who have several female sexual partners are also prone to this form of cancer.

2.3 Cancer In Africa

Cancer is becoming an epidemic in Africa with the number of death increasing every year. In a study by Jemal et., al (2012), about 75,000 new cases and 542,000 cancer deaths occurred in 2008 in the African continent alone. There is fear that these numbers could double in the next 20 years because of an aging population and increased prevalence of risk factors. The growing cancer burden seems to receive a low public priority in Africa largely because of very limited resources and other public issues such as communicable diseases such as HIV and AIDS, malaria and tuberculosis (Jemal et., al, 2012)

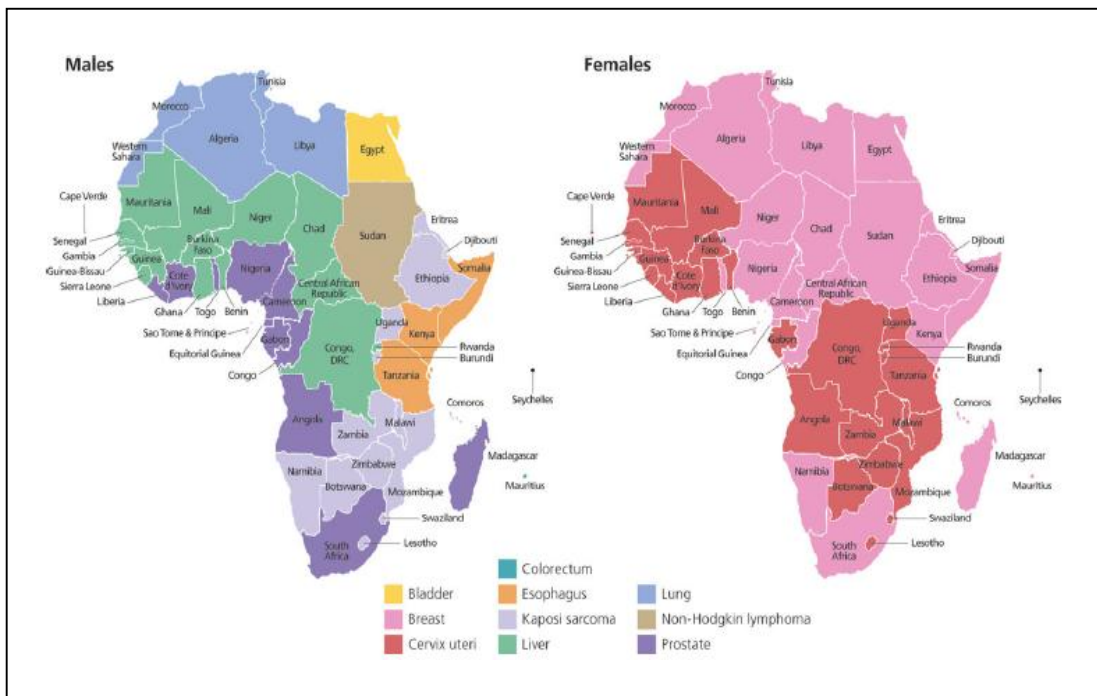


Figure 2: Most common cancer sites in Africa by Sex and Country are shown for 2008.

Source: Jemal et al, 2012

Breast cancer was the most diagnosed cancer and the second leading cause of death in African women in the year 2008. Southern African women reported a higher rate of breast cancer in the African region. This is because of the late childbearing and early menarche (early puberty). Cervical cancer is the second most diagnosed cancer and leading cause of death in African women. The rates though vary across Africa, with incidence and death rates being highest in East Africa and West Africa as compared to North Africa. The reason of such high rates is because of lack of screening services for early detection that would prevent these cancers. (Jemal et al, 2012)

In African men, prostate cancer is the commonly diagnosed cancer. Southern Africa (South Africa) and Western Africa (Nigeria and Cameroon) record the highest rates of prostate cancer (Wasike and Magoha, 2007). Kaposi Sarcoma is the most common cancer in African males where more than 70 % cases are recorded in East Africa compared to only 1% in Northern Africa. The epidemic of HIV/AIDS is the reasons of the high rates in East African Countries. (Jemal et al, 2012)

In a study conducted in a period of two years (Ohwole et al., 2012) characterized and determined the pattern and distribution of prostate cancer. The results indicate that prostate cancer emerged as a major cause of morbidity and mortality amongst men in Lokojo-Nigeria. This cancer has been shown (Jemal et. al, 2012) to be high in Western African countries and Southern African countries. The findings from (Ohwole et al., 2012) show that rare cases are reported at the ages below 50, and predominantly high in men above 65 years which seems to be the mean age of cases globally. In this city in Nigeria and other regions in Africa men learn about the cancer when it is in its advanced stage because the services are new in Hospitals. Lack of health education, absence of routine screening programs, inadequate diagnostic facilities, and poverty also contribute to the late diagnosis.

In a period of one year (Wasike and Magoha, 2007) studied Kenyan patients at Kenyatta National Hospital and confirmed findings by (Ohwole et al., 2012) in Nigeria. This study observed that most patients were presented with late (advanced) prostate cancer Majority who made 58.5% was presented with stage IV (advanced stage) of the disease. The earlier it gets detected the better and a treatment can be advised according to the patient's age. The deaths caused by late diagnosis in Africans explain the failures to reduce mortality rate in African men.

In Africa, there is a need of greater awareness of prostate cancer through health education, affordable routine screening programs and political will by the government to increase budgetary allocation to the healthcare system. The government also needs to partner with renowned centers abroad to engage in meaningful research to highlight prostate cancer among African men.

2.4 Predictive models used in prostate cancer

Predictive modeling uses the data sets that one has collected so as to derive a mathematical model which will be used to predict outcomes of potential patients. (Tewari et. al., 2001). The main goal of a predictive model is that it should be very accurate in its results as they are used in decision making for potential users.

Tewari et. al. (2001) shows that in earlier years, majority of predictive models used in management of prostate cancer utilized traditional statistical techniques. They continue to shine more light to the investigators on the emerging techniques such as Artificial Neural Networks which complements the traditional methods. ANN and other emerging techniques can be used independently or can combine with the traditional methods when building an accurate predictive model.

Predictive models can be built using different approaches which (Tewari et. al., 2001) defines some of them as stated below.

Mechanistic Modeling

These are mathematical tools that describe the cause and effect relationships of a proposed action in terms of a mathematical function. These models though, have a downfall when they are to describe a system that has strong stochastic (probabilities) elements which are contributing to the model output. Some statistical models may not though be useful to medical predictions as biological systems most often are complex thus bringing in approximations and simplified assumptions which may introduce further uncertainty and error.

Stochastic Modeling

In statistical models, the dependent variable, (in our case it's the type of treatment to administer to patient) is predicted by one or more independent variables, e.g. age, type/stage of tumor.

Statistical models incorporate methods to determine how much of the observed variability can be explained by the independent variables, and how much is due to "chance"; i.e., is unexplained.

Statistical methods have various advantages like use of actual observations which could be either random or forced values and availability of rich set of analytical methods. The forced or unusual observations may cause problems with the model and also its not obvious which independent variable should be included in the relation.

Simple Linear Regression

Simple linear regression assumes a linear relation, within a certain range, between a single independent predictor and the dependent variable. A single predictor such as age isn't enough to explain most of the observed response in a biological system. Calibration is the process of fitting a regression equation to observe data which yields a goodness-of-fit measure. If the sample is truly representative of the desired sample space, we would expect to obtain the same parameters, within experimental and observational error, in similar repeated studies. The SLR are the most commonly used predictive models

Artificial Neural Networks

Artificial Neural Networks (ANNs) model data to provide an outcome prediction based on a set of given independent variables. They are nonlinear mathematical models that are characterized by an intricate structure of interconnected computational elements. These computational elements put together a series of inputs using a summation operation and produce an output, such as the probability of 10-year survival. Inputs to each neuron are multiplied by a weight factor that reflects the excitatory or inhibitory strength of the connection from the input source to the neuron. The sum of the weighted inputs plus a bias term then goes through an activation function that behaves like a "switch" to determine whether or not the neuron will "fire" and thus send an output signal. The bias term may be viewed as the threshold that the weighted sum of inputs must exceed before the neuron sends an output signal. The information-processing capacity of an ANN is a function of the type and quantity of nodes in a given network and by the arrangement of interconnections between nodes. The "knowledge" or the "processing capability" of an ANN is made possible by the actual values of the interconnection.

The knowledge is acquired through a learning phase, during which examples of data are repeatedly fed through the ANN, and the connection weights in the ANN are adjusted adaptively for the ANN as a whole to satisfy some predetermined performance goals.

Existing ANNs are used primarily to evaluate outcome and have been introduced as an alternative to classical statistical models.

As a tool for predicting outcomes they can be judged against classical statistical methods by comparing their performances on a receiver operator characteristic (ROC) curve. ANN is seen to perform way much better when predicting outcomes for prostate cancer patients.

2.5 Application of datamining in cancer

Healthcare data is massive therefore there is need to use data mining to answer several important and critical questions related to health care. The patterns or trends that are discovered guide decision making such as forecasting treatment costs and demand for resources and anticipating patient's future behavior. Data mining is an essential step in the process of knowledge discovery in database in which intelligent methods are applied in order to extract patterns (Kharya, 2012). Data mining and knowledge based management are now applied so as to create knowledge rich in healthcare environment. Traditionally, decisions were basically made through ground information and lessons learnt lessons from previous experiences (Kaun H and Wasan, S.K, 2006).

Medical data analysis has often been performed using standard statistical methods because clinicians understand them better as they are familiar with some of the statistical packages that are widely available. However, despite of their popularity, many statistical techniques are based on very simple models, which often fail to catch data complexity. In this light, data mining can provide quite useful tools because its models are usually much more powerful and flexible and can actually deal with problems that contain complex data. (Chang et. al, 2013),

Interestingly, most of the algorithms used in cancer prediction and prognosis employ supervised learning i.e. supervisor gives the learning algorithm a labeled set of training data or examples. Among the available options in the data mining field, the most popular models in medicine are logistic regression (LR), artificial neural network (ANN), and decision tree The algorithms have their own challenges and advantages. In the case between Artificial Neural Networks and Decision trees, the latter has numerous advantages such as they are simple to understand and interpret, they require little data preparation, they can handle many types of data including numeric, nominal (named) and categorical data, they are quick to learn and they can be validated using statistical tests. (Cruz, J.A. and Wishart, D.S, 2006). The decision trees though generally do not perform well in larger dataset as compared to Artificial Neural Networks.

2.5.1 Data Mining In Breast Cancer

In a breast cancer study (Changa,W.P, and Liou, D.M.) used the artificial neural network, decision tree (J4.8), logistic regression, and genetic algorithm. Accuracy and positive predictive value of each algorithm were used as the evaluation indicators. The dataset consisted of 699 patients records of the breast cancer patients at the University of Wisconsin. Among them, about 241 had breast cancer while the rest did not have breast cancer. 466 patients were randomly picked to be the model building set while the remaining were retained as testing set. A 10 fold crossover validation is used so as to reduce errors resulting from random sampling. The study indicated that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification.

In comparing Logistic Regression and Artificial Neural Network Models, (Ayer et. al..2009) estimates breast cancer risk on the basis of mammographic descriptors and demographic risk factors. Mammography is a process that uses low energy X- rays to examine human breast, it's used as both a diagnostic and screening tool. SPSS statistical software was used to construct a mammographic logistic regression model and a mammography ANN model (Multiple layers Perceptron) was also built for this study to review the two models in predicting the risk of breast cancer. They found out that the two models demonstrated high discrimination accuracy and are similar in performance. They claim that once these models are integrated into clinical practices, they both have a good potential to be used as decision support tools.

2.5.2 Data Mining In Cervical Cancer

In a cervical cancer study, two advanced data mining techniques are applied to check the recurrence of cervical cancer (Chang et. al, 2013). Multivariate Adaptive Regression Splines (MARS) and C5.0 are the two techniques applied. MARS has in recent years been applied in modeling variety of data that includes speech modeling, mobile radio channels prediction and intrusion detection in information systems security. Results from these studies are promising and these techniques can be employed in forecasting. On the other hand, C5.0 classifier is a process for the classification and analysis of information hidden in large datasets/databases, which retrieves useful information in the form of a decision trees. The cervical cancer datasets provided are used in order to verify the feasibility and effectiveness of the algorithms. In C5.0 only two independent variables are used while in MARS a number of variables are used as inputs. The

classification results show that C5.0 is more effective with a rate of 96% as compared to MARS, which had effectiveness of 86%. The independent variables that were picked (pT and pStage) when applying C5.0 show that they are very important in prognosis thus reduction in recurrence of cervical cancer.

2.5.3 Data mining in prostate cancer

In developing a predictive model for prostate cancer, (Delen D, Walker G, & Kadam A, 2005), applied three popular algorithms including Logistic Regression, Decision trees (CART algorithm) and ANN architecture known as multi-layer perceptron (MLP). The MLP is known to be a robust function approximator for prediction/classification problems. They argue MLP is the most commonly used and well-studied ANN architecture which was actually proved by their experimental runs showing that it performs better than any other ANN architectures such as radial basis function (RBF), recurrent neural network (RNN), and self-organizing map (SOM). In their results, the models were evaluated based on the accuracy, sensitivity and specificity). The results were achieved using tenfold cross validation for each model, and are based on the average results obtained from the test data set.

The results showed that the decision tree model achieved a classification accuracy of 0.9000 with a sensitivity of 0.9188 and a specificity of 0.7375. The logistic regression model achieved a classification accuracy of 0.8961 with a sensitivity of 0.9130 and a specificity of 0.7361.

The Artificial Neural Network was the best performer of the three models as it achieved an accuracy of 0.9107 with sensitivity of 0.9310 and a specificity of 0.7383. The development of an accurate prediction model for prostate cancer is achieved by Delen and Patil (2006). It shows that the use of Artificial Neural Network gives us an accurate prediction of 91.07%

In analyzing treatment of prostate cancer, multiple techniques of Data mining, (Rahman and Sarma, 2013) chose an existing algorithm C4.5 where they attempted to increase the efficiency so that it can be more practically feasible and applicable. They note that C4.5 has probably become the most widely used and studied decision tree construction algorithm. C4.5 is well known and used in for classifying dataset but it is inefficient when introduced to a larger dataset. To make the algorithm more practical to the dataset, another unsupervised data mining technique known as K-means clustering is introduced. It integrates with C4.5 algorithm in a modified way to increase its efficiency. Although K-means clustering has its own drawback, it overcame the earlier prerequisite of number of clusters by using Calinski Harabasz Index to make the right decision of

taking the number of clusters efficiently. They used a known tool, WEKA data mining tool where experimental dataset was tested using the C4.5 algorithm and K-means algorithm and the findings were that age is really crucial in the treatment of prostate cancer thus should be a determining factor when advising on treatment options for the patient. On the basis of the different ages of the patients, different treatments can be given to them though their stages of the disease are same, thus getting more consistent and accurate rules.

(Saritas, Ozkan & Sert, 2010,) applied Artificial Neural Networks that yields a prognostic results indicating whether a patient has prostate cancer or not. They acknowledge that although this model does not diagnose cancer conclusively, it helps the doctor to decide whether a biopsy is necessary by determining if a patient has prostate cancer or not. ANN achieved a success rate of 94.11% and 94.44% for prognosis of disease and validity respectively. They conclude that ANN will help the doctors make a quicker and reliable diagnosis which allows them to offer the best treatment.

2.6 Preferred Algorithm

Most of the research papers highlighted above show good results with the use of ANN and Decision trees. ANN are considered to have more complex structures and this complexity gives this algorithm the power to classify any data with complex relationships. In evaluating computer models used in risk estimation, logistic regression and ANNs enjoy the most widespread use. This is because they are relatively easy to build and often have excellent predictive ability in medical domain (Ayer et. al..2009). Kenya has minimally explored this option in prostate cancer and thus this project has brought out certainty by making sound decisions using the patterns we got after analysis.

CHAPTER 3

METHODOLOGY

3.1 Research Design

In this study, CRISP-DM methodology was used. CRISP-DM is a comprehensive data mining methodology used by both data mining experts and novice users in accomplishing data mining projects. This model has been recommended as the best model by various data miners as it encourages best practices and offers organizations the structure needed to realize better, faster results from data mining (Shearer, 2000). CRISP-DM methodology takes us through six stages, which include understand the problem in the domain, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment as shown in Figure 3.

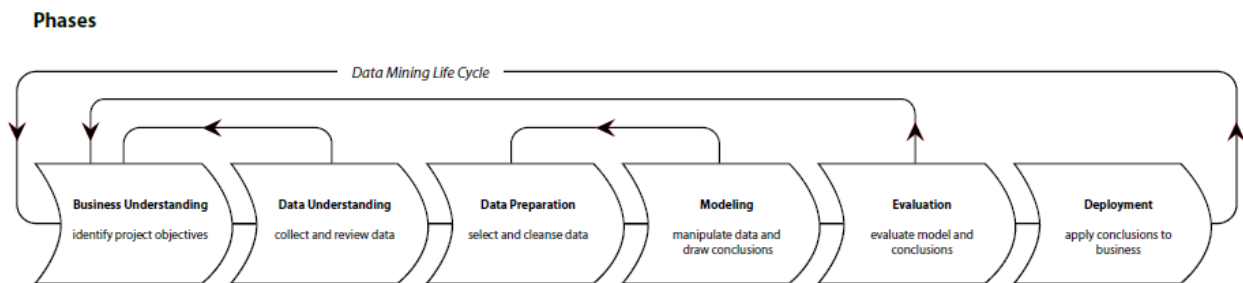


Figure 3: A Visual Guide to CRISP-DM Methodology:

Source: CRISP-DM 1.0. Available from: <http://www.crisp-dm.org/download.htm>

The dataset of prostate cancer for Nairobi County was used to acquire knowledge of the variables used. We used R Algorithm and WEKA which enabled us to compare the results of both techniques.

3.2 Overview of CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) was first proposed in the year 2000 (Chapman et al., 2000). CRISP-DM is the most widely used methodology for developing data mining projects (KdNuggets.Com, 2002, 2004, 2007b) and is considered the de facto standard (Chapman et al., 2000). This model describes the activities that must be done to develop a data mining project. Every activity is composed of tasks. For every task, generated outputs and needed inputs are detailed.

The main objectives of CRISP (Presutti, 1999) include: Ensure quality of data mining projects results, reduce skills required for data mining, capture 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 tool supportable

Although CRISP is widely used, its use is not becoming any more widespread because of rivalry with other in-house methodologies developed by work teams and sample, explore, modify, model, assess (SEMMA) methodology. This decrease in the use of CRISP-DM is due to the fact that it just defines what to do and not how to do. This has forced work teams to use its own methodologies. Another inconvenience is that CRISP-DM does not include project management activities such as quality management or change management (Marbá'n et al., 2008). On the other hand, the use of SEMMA methodology has lightly increased because of the growth in the use of its data mining support tool, Enterprise Miner, developed by SAS and based in SEMMA methodology. SAS is a leader company in Business Intelligence (BI) and it has the most comprehensive BI platform in the industry with the most advanced analysis capabilities.

3.3 Overview of WEKA

WEKA is an acronym for Waikato Environment for Knowledge Analysis, which is a free and open source software used to mine data. WEKA implements different algorithms which include Decision Trees, Artificial Neural Networks, and Logic Regression. WEKA allows the GUI user to select the four different ways to work with. These four ways include Explorer, Experimenter, KnowledgeFlow or a simple CLI. This study will apply explorer as it is an easy to use graphical user interface that harnesses the power of the WEKA as a software.

WEKA only accepts data in ARFF (Attribute-Relation File Format) formats which is an ASCII text file that describes a list of instances sharing a set of attributes. These files have two sections: Header information sector and Data information sector. The header (Figure 6) contains the name of the relation, a list of attributes and their data types.


```

% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}

```

Figure 4: Header information in ARFF Source:

<http://weka.wikispaces.com/ARFF+%28book+version%29>

For data the ARFF is as shown below:

```

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa

```

Figure 5: Data in ARFF

Source: <http://weka.wikispaces.com/ARFF+%28book+version%29>

This data is processed by the different algorithms exhibited in the WEKA GUI chooser and from these different outcomes one is able to know which algorithm is best for the predictive model.

3.4 Sources of Data and Target Population

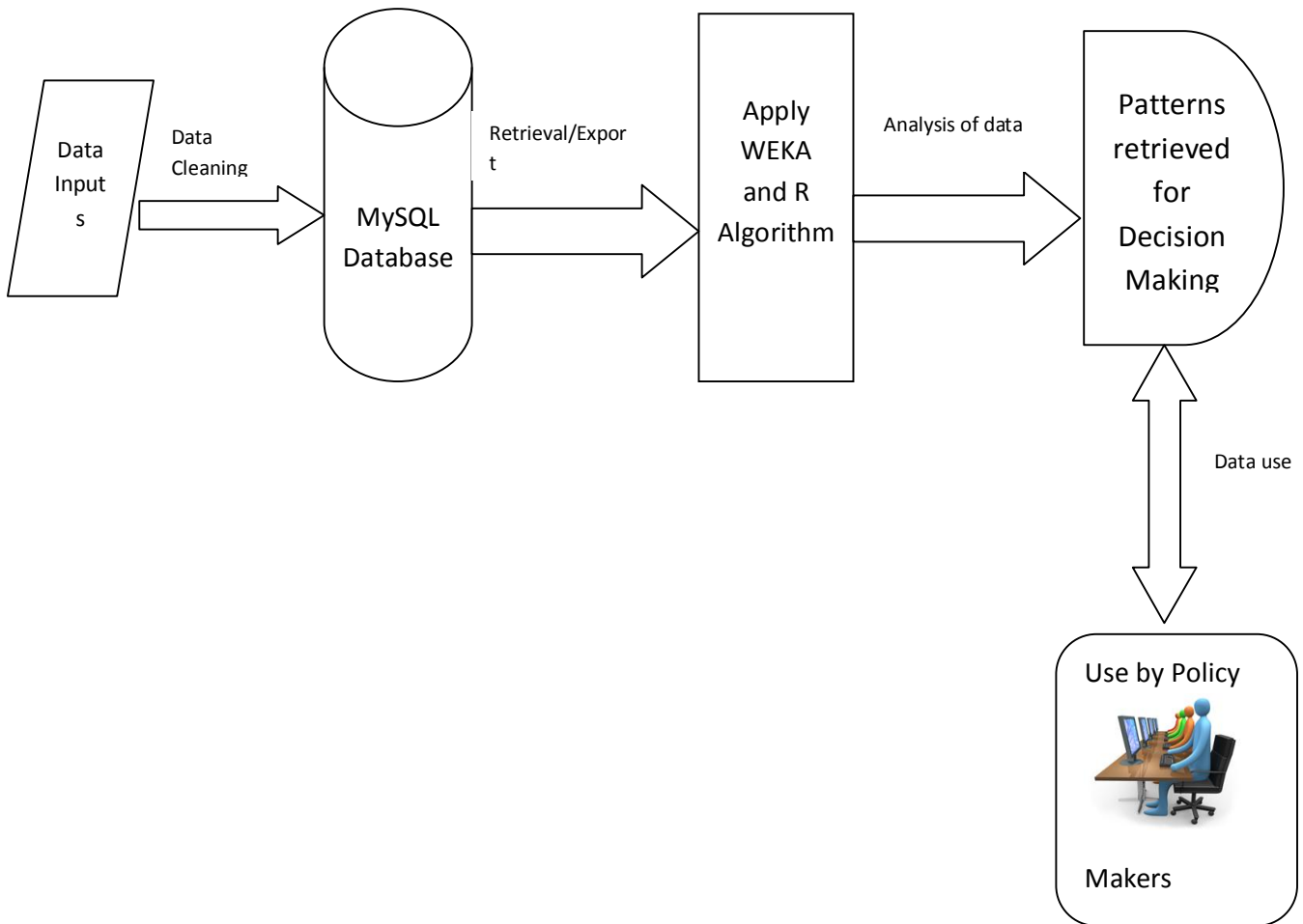
This study used survey-based secondary data provided by Nairobi Cancer Registry (NCR). NCR is reliable source as medical experts as well as World Health Organization (WHO) use data collected by NCR so as to analyze and deduce information from the data which later enables

them to make sound decisions. NCR carries out surveys in Nairobi County hospitals through interaction with patients and reviewing their prognosis records. The prostate cancer data was a great measure in achieving the objectives of this study. The outcome of the patterns are to help policy makers, medical practitioners and the affected, to be able to make timely and rational decisions.

3.5 Data Collection and Analysis

The data obtained from NCR was passed through two different data mining tools namely R Algorithm and WEKA where we were able to discover patterns that were helpful in decision making.

3.6 Proposed Model



3.7 Requirements

The resources required for the proposed system are: Operating systems - Windows 7, Windows XP; R Algorithm; WEKA; MySQL and PHP for System Development; and Laptop with 4 GB RAM.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data Pre-processing

The dataset was issued by NCR and had five attributes namely: year, age, detailed location, stage of cancer, and cases found in each location having the same age in a given year. We had 720 instances which had numerous cases with need of being cleaned as there were duplicates, blanks as well as mistakes which needed rectification. Some unknown instances were also reported due to many of these cases being referred to private health facilities thus their records not being traced.

Data Cleaning

The data availed to us was raw data thus having a lot of impurities that needed to be removed. Using Microsoft Excel, we did data cleaning where we initially had a dataset of 720 but were reduced to 484. Most of the deleted data had no detailed location as it had a general location as either Nairobi or Nairobi unknown. Any data that was blank in any of the attributes was also removed.

Table 1: Part of the Raw Data

	A	B	C	D	E	F
1		YEAR	AGE	DEATAILAD	STAGE	CASES
60	59	2004	70	UNK	3	1
90	89	2005	31	NAIROBI	4	1
100	99	2005	45	ARIOBANGI SOUT	3	1
107	106	2005	51	AREHE BOYS CENT	2	1
122	121	2005	60	SOUTH B	2	1
125	124	2005	65	NAIROBI	4	1
172	171	2006	25		3	1
180	179	2006	38	UMOJA I	3	1
197	196	2006	48		2	1
220	219	2006	60	NGUMMO	3	1
286	285	2006	78	SOUTH C	3	1
294	293	2006	80	KARAI	3	1
309	308	2006	87	NAIROBI UNK	3	1
355	354	2007	63	KANGEMI	4	1
361	360	2007	64	KAYOLE	4	1
363	362	2007	65	MATHARE	4	1
373	372	2007	67	NGARA	4	1
376	375	2007	68	EMBAKASI	4	1
378	377	2007	68	KILIMANI	4	1
403	402	2007	72	RIRUTA	4	1
447	446	2007	87	RUNDA	4	1
511	510	2008	73	DONHOLM	4	1
550	549	2009	46	SOUTH B	3	1
552	551	2009	49		2	1
570	569	2009	60	SOUTH B	3	1
579	578	2009	64		4	1
610	609	2009	79	ZIMMERMAN	3	1
615	614	2009	84	WOOD AVENUE	3	1
628	627	2010	55	WOOFLEY	2	1
632	631	2010	57	BURUBURU	4	1
635	634	2010	57	PANGANI	4	1

Data Generalization

Duplicate data was removed and a new attribute, named Zones, was added. These zones are directives given in the guide of Nairobi City Development Ordinances and Zones from the Department of City Planning by the City Council of Nairobi. They have categorized the different locations in zones (Table 2):

Table 2: Zones of Nairobi County according to City Council of Nairobi

Zones	Area Covered
1E	Kenyatta National Hospital, Community
2	Eastleigh, Starehe, Pumwani
3	Westlands, Parklands
4	Woodley, Kileleshwa
6	Muthaiga,
7	Mathare, Kariobangi, Huruma
8	Jericho, Jerusalem, Donholm, Buruburu, Umoja (1-2), Komarock , Kayole
10	Nairobi West,South B and C, Ngummo, Langata
10E	Embakasi
11	Ayany, Kibera Slums
15	Dagoretti
17	Kahawa West
18	Kasarani, Mwiki

The zone attribute assists in knowing the central points for building the hospice or palliative cares as well as know what type of equipment, which will be leased by the Government of Kenya, should be installed. The zones are further classified (Figure 6) into four clusters namely; Cluster A Nairobi Central, Cluster B (Nairobi East), Cluster C (Nairobi West) and Cluster D (Nairobi North). This is data generalization according to major hospitals.

Table 3: Data Generalization according to major Hospitals

	YEAR	AGE	LOCATION	ZONES (according to CCN)	STAGE	CASES
1	2004	52	STAREHE BOYS CENTRE	2	3	1
2	2004	71	EASTLEIGH	2	3	1
3	2004	75	EASTLEIGH	2	2	1
4	2004	99	PANGANI	2	1	1
5	2004	43	PARKLANDS	3	1	1
6	2004	52	PARKLANDS	3	2	1
7	2004	76	NGARA FLATS	3	3	1
8	2004	86	WESTLANDS	3	4	1
9	2004	67	KILELESHA	4	1	1
10	2004	40	LAVINGTON	5	1	1
11	2004	68	MUTHAIGA	6	3	1
12	2004	76	MUTHAIGA	6	3	1
13	2005	51	STAREHE BOYS CENTRE	2	2	1
14	2005	53	KARIOKOR	2	3	1
15	2005	53	SHAURI MOYO	2	3	1
16	2005	71	PANGANI	2	2	1
17	2005	72	EASTLEIGH	2	4	1
18	2005	72	PANGANI	2	4	1
19	2005	79	EASTLEIGH	2	4	1
20	2005	84	EASTLEIGH	2	3	1
21	2005	85	EASTLEIGH	2	4	1
22	2005	55	RACE-COURSE	20G	1	1
23	2005	24	PARKLANDS	3	1	1
24	2005	65	PARKLANDS	3	1	1
25	2005	72	WESTLANDS	3	3	1
26	2005	58	LAVINGTON	5	2	1

4.2 Modeling Techniques, Tools and Modeling Infrastructure Used

The software tools used i.e. software-R statistical tool and WEKA are open-source and downloadable from the World Wide Web and used under the GNU license. The presentation of results and the development of the prototype were done using PHP while the data will be stored in MySQL Server database.

4.3 Data Analysis and Results

4.3.1 Basic Classification Results using WEKA

In all these clusters we used the artificial neural network and decision tree (J4.8). Accuracy and positive predictive value of each algorithm were used as the evaluation indicators. A 10 fold crossover validation is used so as to reduce errors resulting from random sampling. In all the cases below we used the attribute stage to be our class identifier for model accuracy. Stage was the only attribute that would assist in knowing where to locate the equipment We analyzed the data using both J48 and ANN and below are the results.

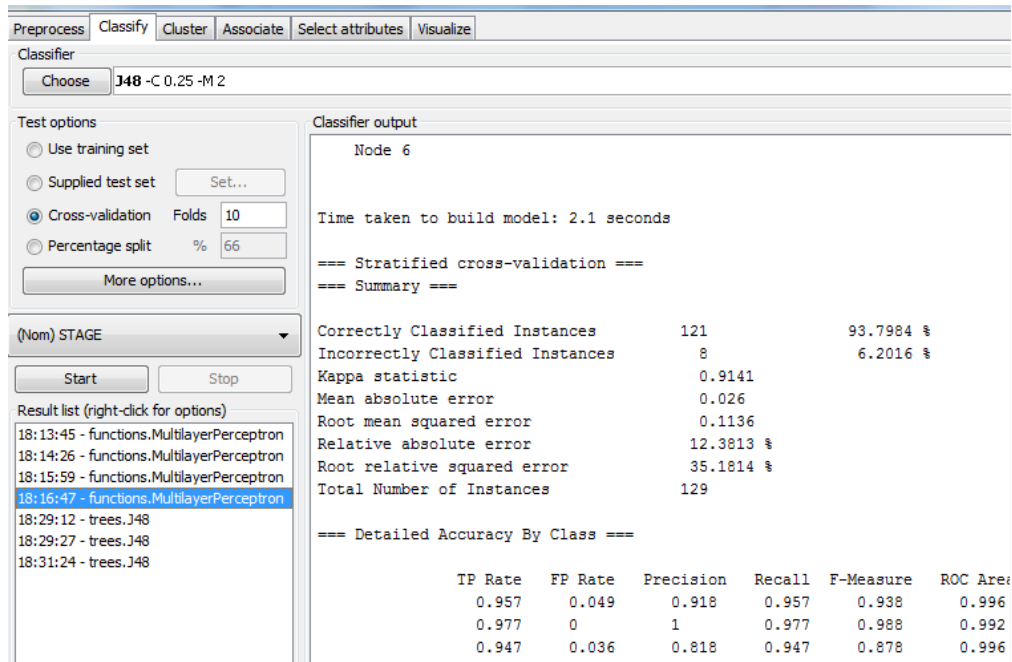


Figure 7: MLP for Nairobi Central

When building a model for Cluster A using Cross Validation of 10 folds for Artificial Neural Networks, we see the model performing well where it correctly classifies 121 instances with an accuracy of 98%.

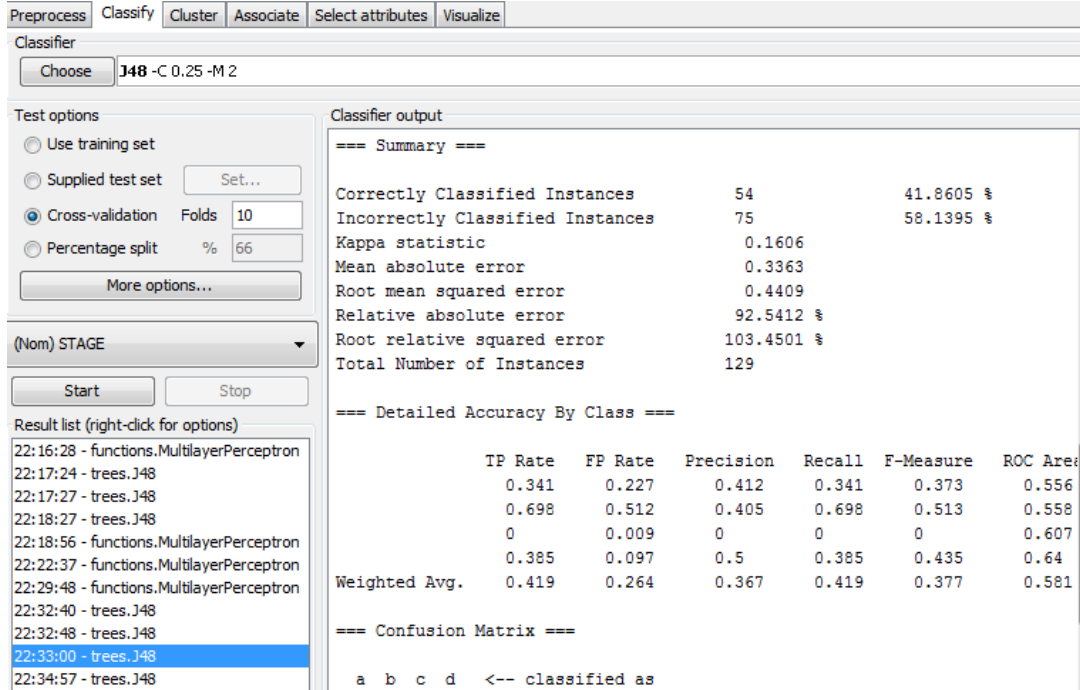


Figure 8: Decision tree for Nairobi Central

For Decision trees using J48, the accuracy goes down to 41% as it only correctly classified only 54 instances.

Using the Clustering Algorithm in our Dataset

Clustering allows users to make groups of data to determine patterns. We loaded our four sets of data separately and used the clustering algorithm SimpleKMeans which gave similar results to what we got when using R Algorithm. The Euclidean distance is used for this clustering method where any other mentioned areas are engulfed by the cluster nearest to them.

Dataset A

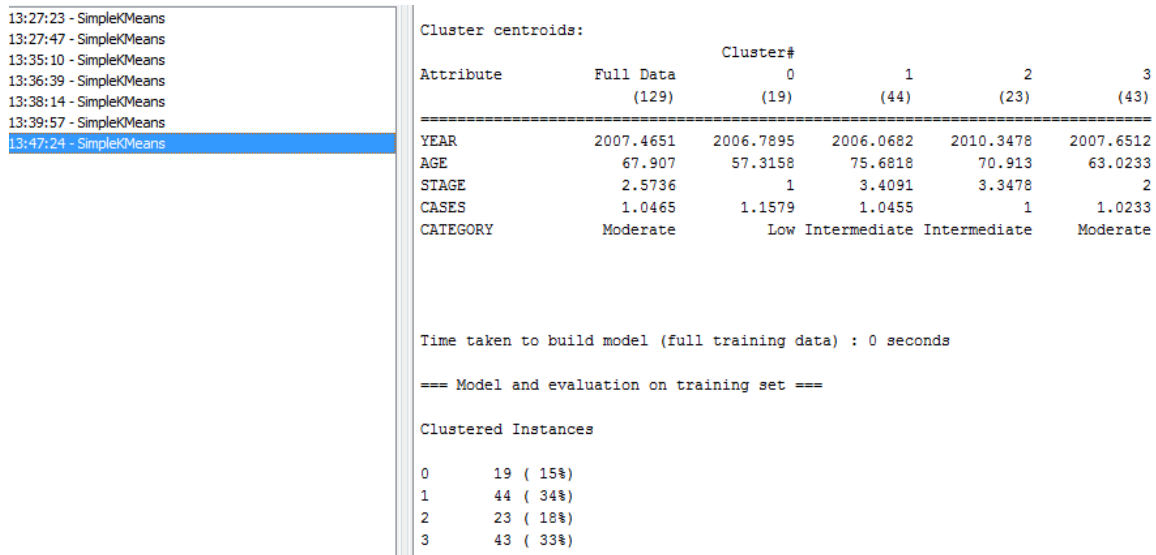


Figure 9: Clustering Dataset A

Interpretation of results for Dataset A

In this workout for clustering, we ignored attributes of Location where full data reveals that the most prevalent case is Stage 2 (moderate) of cancer. The mean age for these four clusters is 67 and highly affected cluster is 1. This will assist in knowing where to place the equipment to be leased especially in highly hit clusters hospitals.

Figure xx below shows the visualization of the clusters. Cluster one has varying cases in both Intermediate and High stage while cluster 3 which had the second highest cases were in the median stage. This shows that many cases are slipping into stage 3 and 4 when not taken well care of in the 2nd stage.

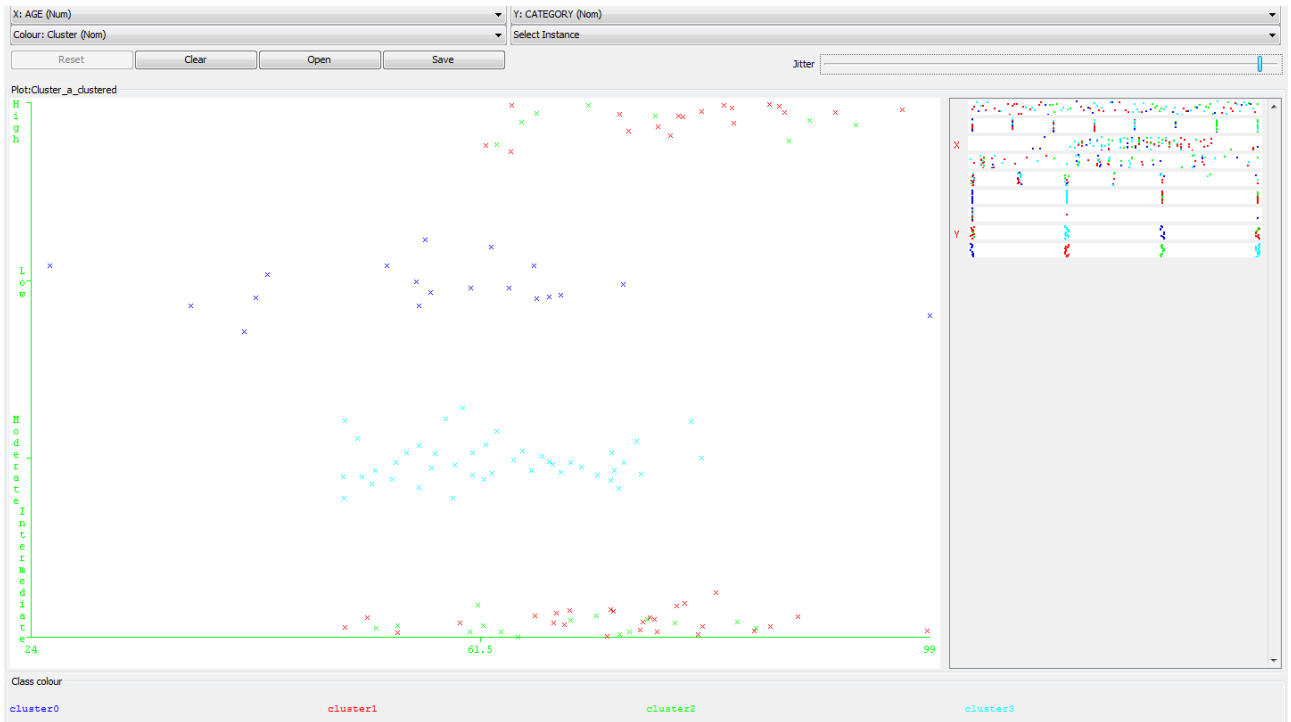


Figure 10: Data Visualization for DatasetA

Dataset B

```

13:35:10 - SimpleKMeans
13:36:39 - SimpleKMeans
13:38:14 - SimpleKMeans
13:39:57 - SimpleKMeans
13:47:24 - SimpleKMeans
15:09:51 - SimpleKMeans
  
```

Attribute	Full Data (171)	Cluster#			
		0 (20)	1 (33)	2 (49)	3 (69)
YEAR	2007.6491	2004.75	2010.2727	2008.7347	2006.4638
AGE	67.1754	68.7	67.8485	70.1224	64.3188
STAGE	Intermediate	High	Intermediate	High	Intermediate
CASES	1.0292	1	1.0303	1	1.058

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	20 (12%)
1	33 (19%)
2	49 (29%)
3	69 (40%)

Status
OK

Figure 11: Clustering Dataset B

Interpretation of results for Dataset B

This run shows that the intermediate stage (stage 3) of prostate cancer is most prevalent in cluster3 which has taken 40%. This will clearly bring out the need of Stage 3 treatment in the hospitals that are within cluster 3 or any other easy access point.

Data Visualization

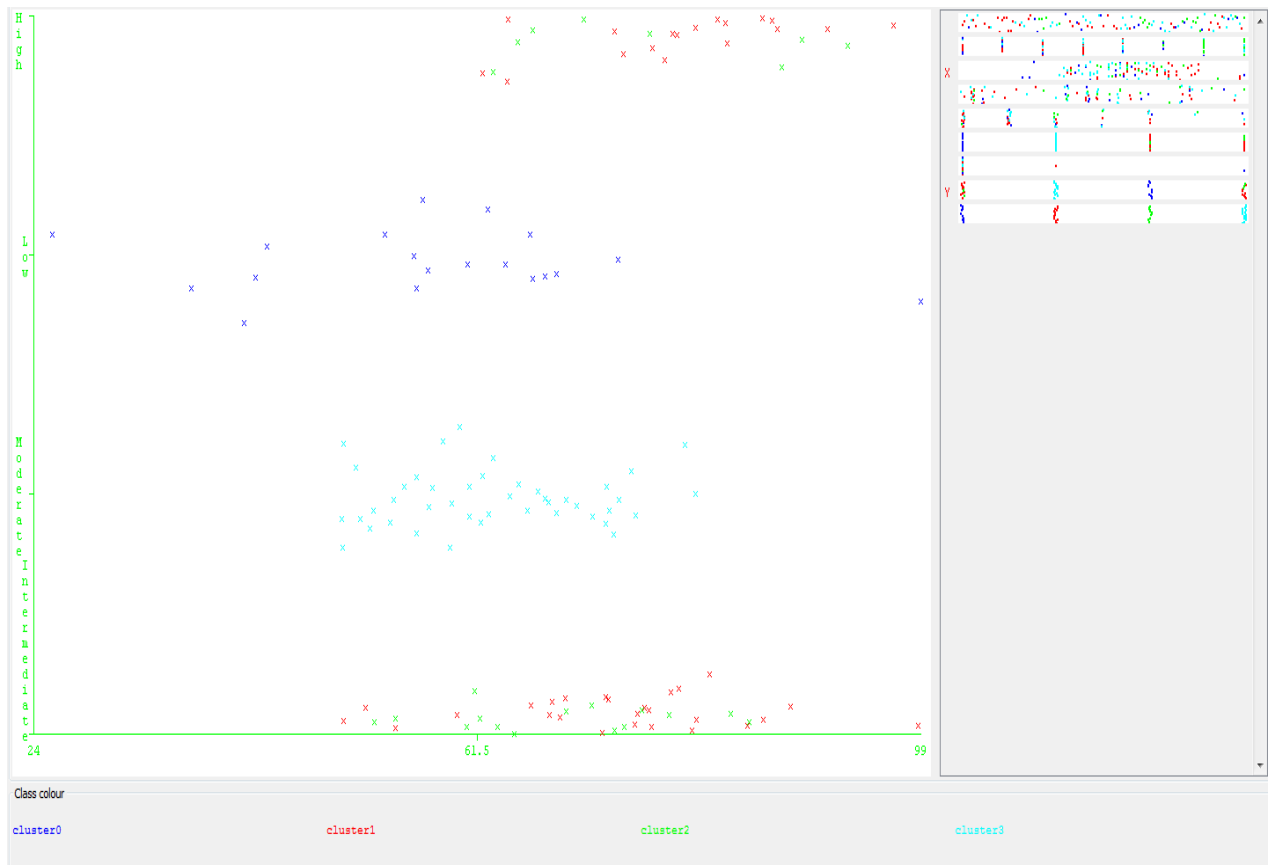


Figure 12: Data Visualization for DatasetB

Dataset C

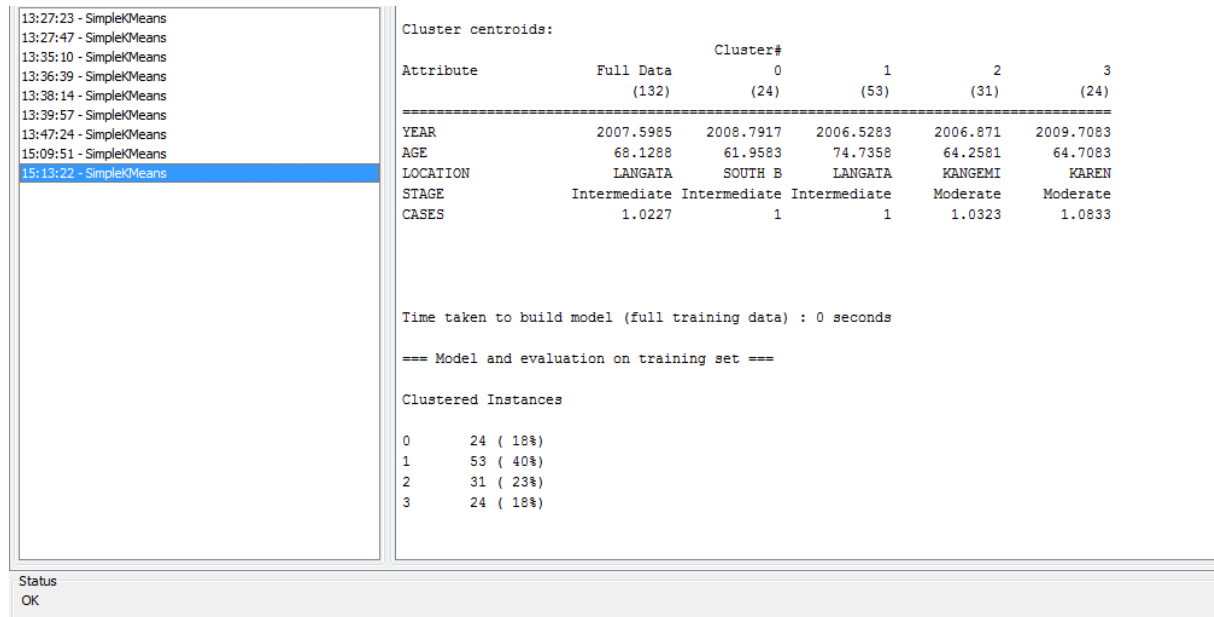


Figure 13: Clustering Dataset C

Interpretation of results for Dataset C

This run brought in the Location attribute which shows the different clustering as well as the location that is highly hit. This run shows that the intermediate stage (stage 3) of prostate cancer is most prevalent in cluster1 which has taken 40%. Langata is the highly hit thus this will clearly bring out the need of Stage 3 treatment in the hospitals that are within Langata or any other easy access point.

Data Visualization

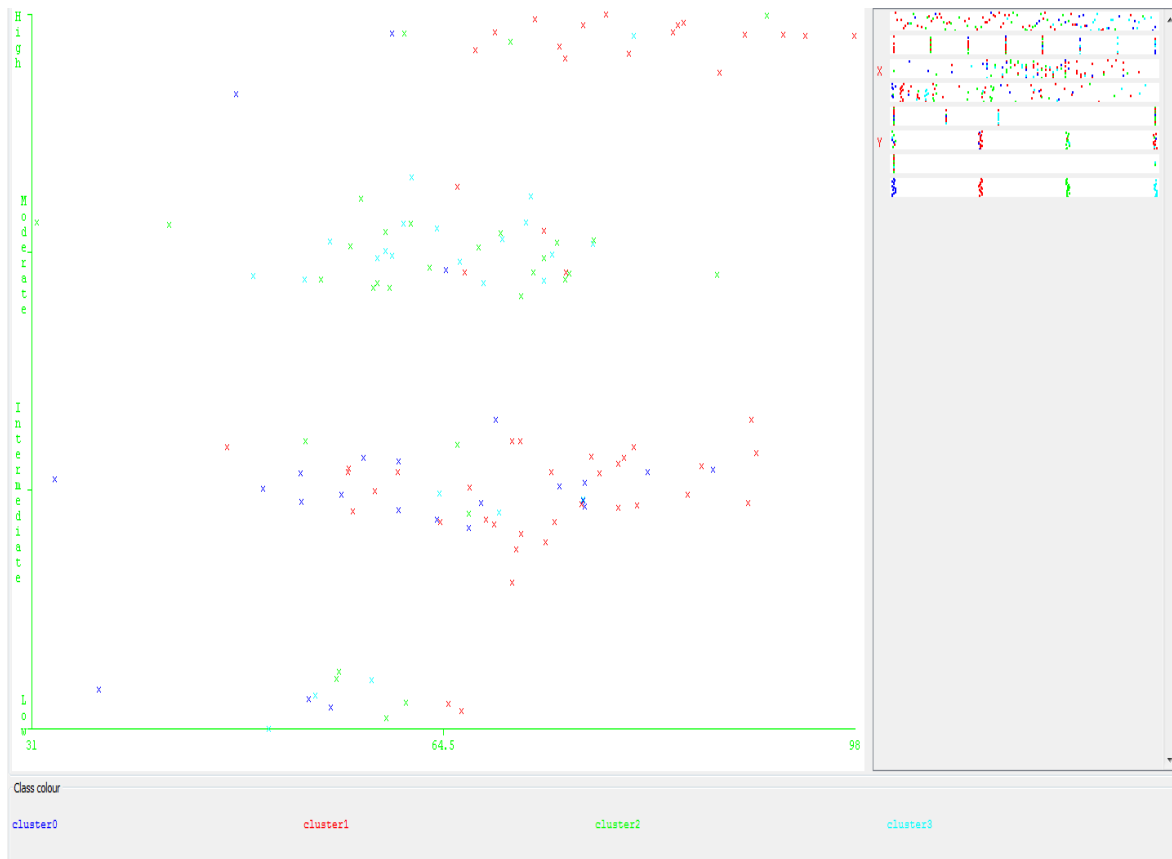


Figure 14: Data Visualization for DatasetC

4.3.2 R Algorithm Analysis

This analysis is performed so as to get patterns from each and every cluster to get the prevalence (widespread presence) of prostate cancer in terms of numbers and the stage these numbers belong to. Once patterns are derived- as it will be seen later as we progress- we will be able to make sound conclusions which will show from our patterns. The Analysis of Variance (ANOVA) is to be used in calculation of our F-Value and P-value. The primary research question for this study was: Are there significant differences of number of cases and stages of cancer through all the four clusters?

H_0 - There is no significant difference of number of cases being same in all the stages of cancer.

H_0 : false = stage1 = stage2 = stage3.

H_A - There is a significant difference of number of cases not being same in all the stages of cancer.

$H_A: \mu_{\text{true}} = \mu_{\text{stage1}} = \mu_{\text{stage2}} = \mu_{\text{stage3}}$.

The testing was done and below is the summary.

Analyzing Each Cluster

When the first cluster was subjected to R, the output was as shown in Figure 15. The plot shows the maximum and minimum numbers in each case. The mean is depicted by the thick line in the boxes. F statistics is a ratio of different measure of variance of the data and if the null hypothesis is true then they are both estimates of the same thing and the ratio will be around 1. The P value is computed from the F ratio which is computed from the ANOVA table. If the overall P value is large, the data do not give you any reason to conclude that the means differ and If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean only that at least one differs from the rest.

CLUSTER A (Nairobi Central)

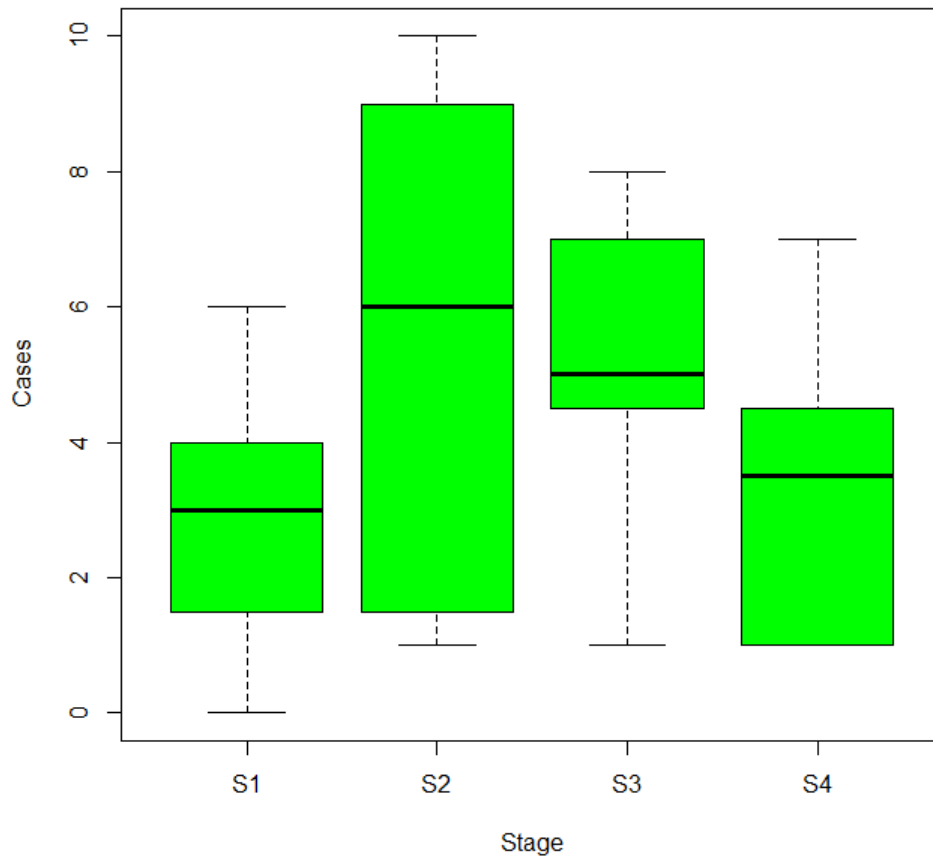


Figure 16: BoxPlot of Cluster A

Table 4 : F- Value Table for Cluster A

```
>  
> anova.cluster1=aov(patients1~stage,data=data)  
> summary(anova.cluster1)  
          Df Sum Sq Mean Sq F value Pr(>F)  
stage      3  43.59  14.531    2.056  0.129  
Residuals 28 197.87   7.067  
>  
> qf(0.95,3,28)  
[1] 2.946685  
> |
```

Summary of cluster A

In the study of cluster A, the Null Hypothesis (Ho) was that there is no significance in the number of cases thus the null hypothesis is then accepted as the F δ Value (2.056) is less than the preset level of significance (2.95) that was assumed to be at 0.05 (95% Confidence Interval). In cluster A, stage of prostate cancer is on the X axis while cases are recorded in the Y axis as shown in figure 4 above.

Table 5: The mean, variance and standard deviation of cluster A

```
>
> sapply(split(data$patients1,data$stage),mean)
  S1  S2  S3  S4
2.875 5.500 5.250 3.250
> sapply(split(data$patients1,data$stage),var)
  S1      S2      S3      S4
3.553571 15.142857 4.785714 4.785714
> sqrt(sapply(split(data$patients1,data$stage),var))
  S1      S2      S3      S4
1.885092 3.891382 2.187628 2.187628
```

The mean in Table 2 shows that stage 2 has the highest cases thus one may seem to comfortably conclude that this is the stage that needs a lot of keenness in treatment. The main issue is that its variance (i.e. The average of the squared differences from the Mean.) is larger compared to the other variances thus conclusion is that the numbers in stage 2 are sparsely distributed compared to cases in stage 3 which are concentrated above the mean line.

The patterns that come out strongly in this analysis is that most cases in stage 2 are exponentially growing to be stage 3 cases thus the concentration exhibited in stage 3 where we see a larger number is exhibited above the mean line. The policy makers and the government of Kenya (GOK) should be able to decide on putting up treatments of stages 2 and 3 in this zone. Most cases in stage 3 are not advancing to stage 4 thus no need of building a palliative care or hospice in this zone.

Table 6: F-Value of Cluster B

```
      Df Sum Sq Mean Sq F value    Pr(>F)
stage   3  261.8   87.28    11.1 5.64e-05 ***
Residuals 28  220.1    7.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Summary of cluster B

In the study of cluster B, the Null hypothesis (Ho) was that there is no significance in the number of cases thus we refuse to accept the null hypothesis as the F value (11.1) is more than the preset level of significance (2.95) that was assumed to be at 0.05 (95% Confidence Interval). This shows that the cases in each stage do vary and thus we go ahead to check which has a higher concentration than the others so as we can administer the right kind of treatment in this cluster.

Table 7 : the mean, variance and standard deviation of cluster B

```
> apply(split(data1$patients2,data1$stage),mean)
      S1      S2      S3      S4
0.875 7.250 8.375 5.375
> apply(split(data1$patients2,data1$stage),var)
      S1      S2      S3      S4
1.267857 7.928571 19.696429 2.553571
> sqrt(apply(split(data1$patients2,data1$stage),var))
      S1      S2      S3      S4
1.125992 2.815772 4.438066 1.597990
> |
```

From Table 4 above , stage 3 has a higher mean but the distance from mean which is the variation is large thus we take up much concentration being in case 2 as shown in Figure 17 below. The policy makers and the government of Kenya (GOK) should be able to decide on putting up treatments of stages 2 which will curb the few cases which advance to stage 3. The few cases in stage 3 should also be considered as there is a possibility of an alarming rise of stage 4 cases if they pass undetected or are presented late (Figure 18).

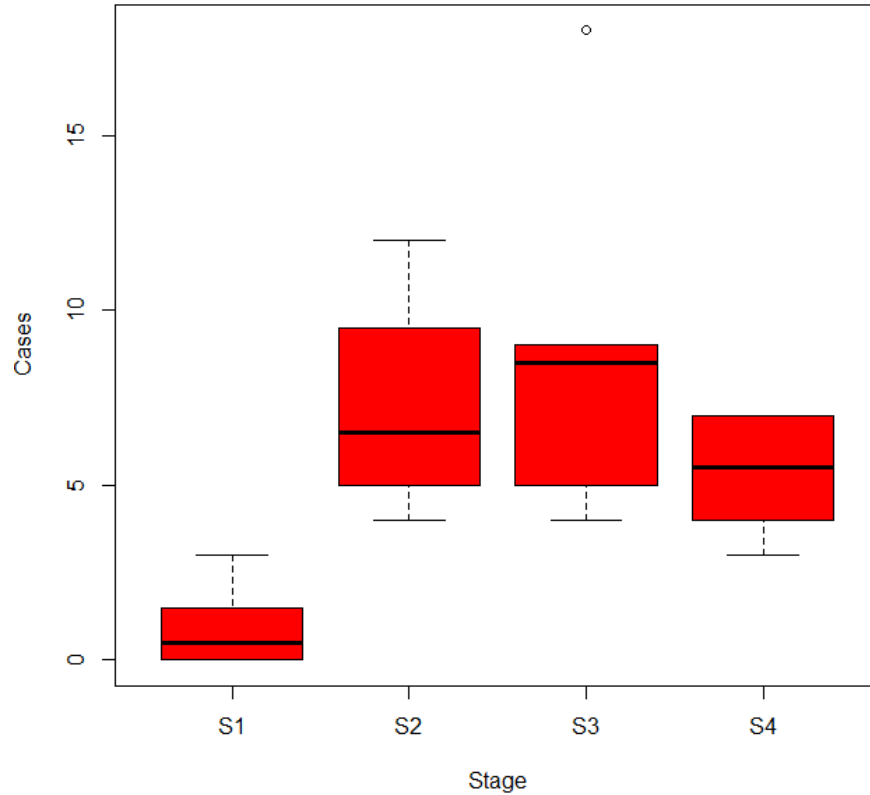


Figure 19 : Boxplot of cluster B (Nairobi East)

CLUSTER C (Nairobi West)

In cluster C, there is a significance difference in number of cases for each stage thus we fail to accept the null hypothesis. The boxplot shows that though not many cases reported in stage 2, there is a possibility of these cases advancing to stage 3. In such a case, G.O.K and policy makers can make a decision of placing both stage 2 and stage 3 treatments in this cluster. Below are the tables for the F- Value which is higher than our expected value thus though registering same number of cases in each stage, the boxplot gives us a significant difference.

Table 8: The F-Value table for cluster C

```
      Df Sum Sq Mean Sq F value    Pr(>F)
stage   3  261.8   87.28   11.1 5.64e-05 ***
Residuals 28  220.1    7.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 9: The mean, variance and standard deviation for Cluster C

```
>
> sapply(split(data2$patients3, data2$stage), mean)
  C1    C2    C3    C4
1.625 6.375 8.250 3.125
> sapply(split(data2$patients3, data2$stage), var)
  C1          C2          C3          C4
1.410714 13.696429 51.928571  8.982143
> sqrt(sapply(split(data2$patients3, data2$stage), var))
  C1          C2          C3          C4
1.187735 3.700869 7.206148 2.997022
>
```

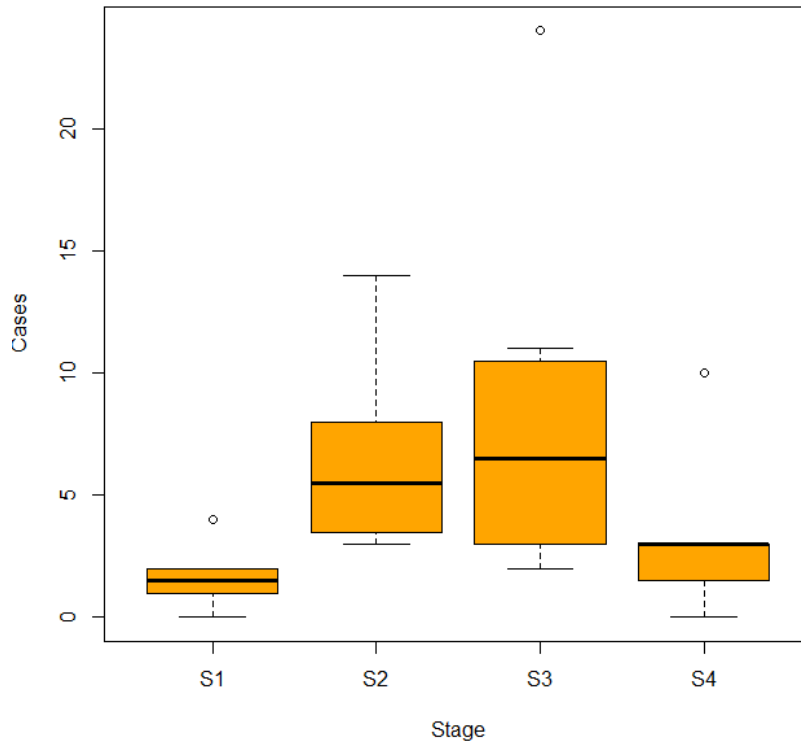


Figure 20 : Boxplot for Cluster C

CLUSTER D (Nairobi North)

In the study of cluster D, the Null hypothesis (H_0) was that there is no significance in the number of cases thus we refuse to accept the null hypothesis as the F value (11.1) is more than the present level of significance (2.95) that was assumed to be at 0.05 (95% Confidence Interval). This shows that the cases in each stage do vary and thus we go ahead to check which has a higher concentration than the others so as we can administer the right kind of treatment in this cluster.

Table 10 :The F-Value table for cluster D

```

Df Sum Sq Mean Sq F value Pr(>F)
stage      3  261.8   87.28   11.1 5.64e-05 ***
Residuals 28  220.1    7.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

Table 11: The mean, variance and standard deviation for Cluster C

```
<
> sapply(split(data3$patients4,data3$stage),mean)
  C1  C2  C3  C4
0.750 2.250 3.125 0.500
> sapply(split(data3$patients4,data3$stage),var)
  C1      C2      C3      C4
0.7857143 3.0714286 0.9821429 0.2857143
> sqrt(sapply(split(data3$patients4,data3$stage),var))
  C1      C2      C3      C4
0.8864053 1.7525492 0.9910312 0.5345225
>
> |
```

From the mean, variance and standard deviation table, Stage 3 is prevalent thus meaning cases are developing from stage 2 to 3. The Policy makers and G.O.K should establish equipment for both stage 2 and 3 so as to curb cases in earlier stages.

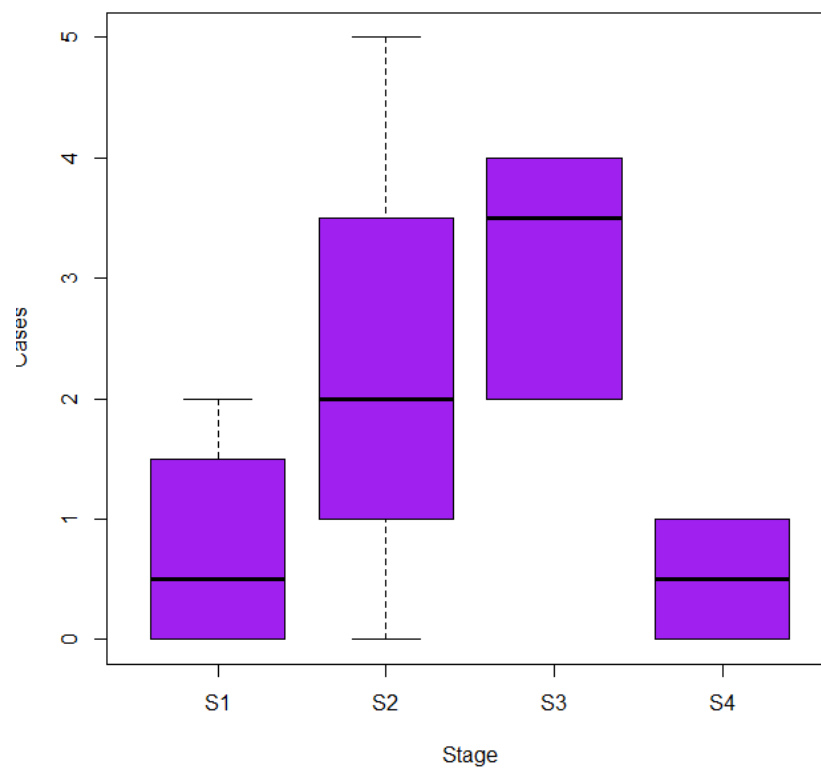


Figure 21 : Boxplot for Cluster D

Discussion on R analysis

Results in R in all these clusters are recommending stage 2 and stage 3 treatment thus this clearly shows that most cases pass through stage 1 undetected. The prototype that we are proposing for use by NCR as well as the hospitals, is to assist in awareness where such cases can be detected earlier thus early treatment which will curb the alarming rise of deaths of prostate cancer.

Overall Discussion of the two tools and their results

The results showcased by these two tools are set out to know how datamining will assist in bringing solutions fit for distribution of the equipment in Hospitals of Nairobi County.

The two tools i.e R Algorithm and WEKA performed quite well where the Multiperception Layer (Artificial Neural Network) and J48 (Decision Tree) were used in WEKA while in R Algorithm we ran a script and gave as output in form of graphs which were much easier to interpret especially to a naive person in DataMining.

A clustering Algorithm known as SimpleKMeans was used to form the different clusters using four different datasets. The Manhattan and Euclidean distance is used to determine where each instance will fall in the most significant clusters. This clustering really helps to know which location is most prime to equip as well as for cases where alot of stage IV cancer ttype is reported they can bring up a Hospice or a Palliative care.

R Algorithm can be tedious especially when working with a huge database as one has to key in each and every instance for it to be run through to produce any meaningful information. WEKA on the other hand is pulling data from a CSV or ARFF file format and then run the data through the different algorithms available.

The results as discussed in each of the interpretation above shows that cases in Nairobi are prime in stage II and Stage III. These two stages need Chemotherapy (Chemo), Cryosurgery and Radiation Therapy. With this in mind, the policy makers and the proffessionals in this field can assist in helping the G.O.K to do a rational distribution of the cancer equipment (especially for this case of prostate cancer)

The resultant of this project is a system which based on results, we can apply it as an awareness tool which will be used by NCR workers on site of screening as well as other willing partners like Churches, self help groups in the community etc

Screening is not diagnosis thus any screening done by either of the partners should have an advisory comment in relation to the PSA level. The protein tested if its in small quantities, one can be recommended for atleast a screen test every two years majorly in contribution to the area one comes from. Unfortunately this study did not go further to check why the prevalence of the stage of cancer in those areas as some may be due to a number of reasons. This way one would be adviced on what to avoid to not put himself as risk of prostate cancer. The causes of prevalence of prostate cancer will be a future work.

4.4 Proposed Prototype

The output of the analysis performed was patterns that will enable the GOK as well as policy makers on making sound decisions. These decisions may include equitable placement of prostate cancer equipment in prevalent areas and creating awareness.

We propose a prototype that will assist the Nairobi Cancer Registry (NCR), Eldoret Cancer Registry and level 4, 5, and 6 hospitals to create awareness in the society. This system will be used to collect data on prostate cancer as well as referring prostate cancer patients to the health facilities where they will get adequate assistance and intervention.

The application which will is stand alone, based on both web and mobile platforms, and will run on any Android devices i.e. Mobile phones and tablets. The users of this application will be the hospitals and awareness groups expected to go round the different zones on an awareness drive.

The application will start by entering the patients age then the detailed location e.g. Kayole. The application is supposed to locate where Kayole lies in the zones and display it for further cause. The patient will then go ahead for screening. If the patient results are negative i.e. no prostate cancer, the patient is be adviced further according to prevalence of prostate cancer in their zones. For the patients who have prostate cancer, the stage is entered and the application refers one to hospital in the zone or elsewhere for treatment. Figure 22 outline how the System works

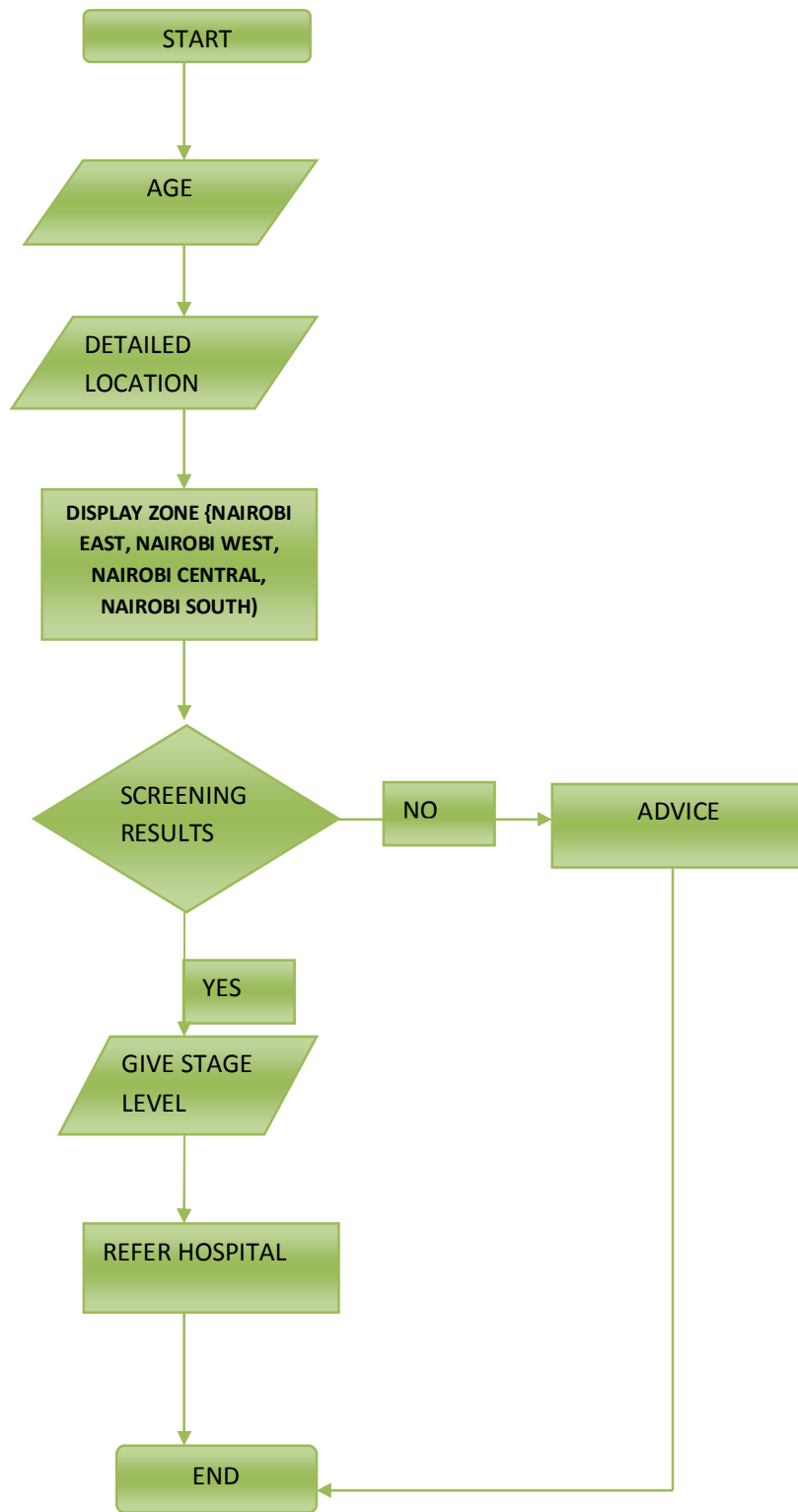


Figure 22: How Prototype Works

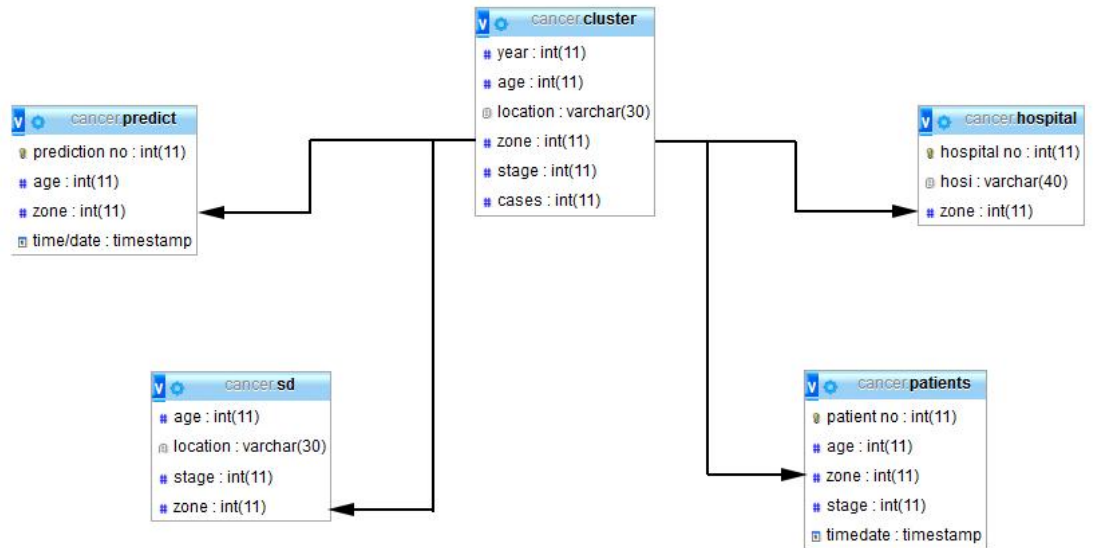


Figure 23: Entity Relationship Diagram for the Database

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Achievements

Data mining is increasingly used for analysis to give out new knowledge which can assist in decision making. This research sought to assist Nairobi Cancer Registry, the Government of Kenya (GOK), and policy makers establish the prevalence of prostate cancer. Data from Nairobi Cancer Registry was used in the study.

This research concentrated on the possibility of building a prognostic model. This research found out that many cases are reported in stage 2 thus an indicator that stage 1 passes without any knowledge of one being sick. According to American Cancer Society, the month of June is Men's Health /Cancer Awareness month but we don't see this being as active as the breast cancer month slated for October. The NCR will be able to use this system which can be used as a website or on mobile phone as they conduct a community drive just as blood drive which takes place frequently.

The WEKA and R used gave almost the same results but the R is an easier tool to learn and its representation of data is much efficient and easy to interpret. The patterns achieved will assist the GOK to know what type of equipment is to be placed and accessed in the different hospitals in Nairobi County.

The prototype system that has been developed will be used as an awareness tool by hospitals and the Nairobi Cancer Registry. This will facilitate collection of more data after screening. The data will be stored in a database and used to do analysis for the different zones.

5.2 Contribution of the Study

This study will contribute to society in terms of awareness which will assist in stigma eradication and also avoidance of late detection of prostate cancer. The Nairobi Cancer Registry in particular will be able to work with a database that is more accurate. This registry is well entrusted with data collected by World Health Organization (WHO) which uses the data for world cancer statistics.

In academic, this study has successfully utilized the two datamining tools which have proved to be of high contribution in deriving of patterns that were used in decision making.

In cancer research, this study proposes a prognostic tool which will be used during the screening drives. The active screening will reduce on number of cases being detected in late stages. This tool also collects data from the different patients and thus a benefit of having a vast database which can assist in other research works is beneficial.

5.3 Limitations of the Study

The data collected from the Nairobi Cancer Registry had a number of blanks thus using means we were able to fill some blanks while some we deleted as they would have made analysis a challenge. A good case is where we had Nairobi as the detailed location and yet we are analyzing the different locations within Nairobi County. Another limitation for now is that this system cannot be used by any other person other than the screening team, as one has to go through screening so as to know which stage patients are in.

5.4 Recommendation for Future work

Future work will research in details on picking up the patterns and knowing the cause of escalating cases of prostate cancer in the different zones. Environmental and genetic factors are a possible link with prostate cancer; the dietary as well as lifestyle are seen as some factors that cause prostate cancer.

References

1. Abernethy M. (2010). Data mining with WEKA, Part 1: Introduction and regression. Available from: <http://www.ibm.com/developerworks/library/os-weka1/>. Accessed on 12th February, 2014
2. American Cancer Society <http://www.cancer.org> : Available from: <http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-treating-general-info>
3. ARFF. Available from: <http://weka.wikispaces.com/ARFF+%28book+version%29>. Accessed on 12th February, 2014
4. Ashutosh Tewari, Christopher Porter, James Peabody, E. David Crawford, Raymond Demers, Christine C. Johnson, John T. Wei, George W. Divine, Colin O'donnell, Eduard J. Gamito, and Mani Menon. (2001). Predictive Modeling Techniques in prostate cancer. *Molecular Urology*. Volume 5, Number 4.
5. Ayer Turgay, Jagpreet Chhatwal, Oguzhan Alagoz, Charles E. Kahn, Jr, Ryan W. Woods, and Elizabeth S. Burnside. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics* 2010; 30:136-22. Published online 10.1148/rg.301095057
6. Benoît, G. (2002). Data mining. *Annual Review of Info. Sci. Tech.*, Vol 36, Issue 1, :265-310
7. Changa, W.P, and Liou, D.M. (n/d). Comparison of Three Data mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data.
8. Chi-Chang Chang, Sun-Long Cheng, Chi-Jie Lu, and Kuo-Hsiung Liao. (2013). Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification. *International Journal of Machine Learning and Computing*, Vol. 3, No. 1,
9. Churchill Livingstone Dictionary of Nursing (2006) retrieved from http://search.credoreference.com/content/entry/ehscl字典nursing/prostate_cancer/0
10. Cruz, J. A and Wishart, D. S.(2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics* 2006:2 596-77
11. Delen D, Walker G, & Kadam A, (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*. 34(2):113-127

12. Delen, D. (2009), Analysis of cancer data: a data mining approach. *Expert Systems*, 26: 1006112. doi: 10.1111/j.1468-0394.2008.00480.x
13. Dursun H, and Nainish P. (2006). Knowledge extraction from prostate cancer data. *Proceedings of the 39th Hawaii International Conference on System Sciences*.
14. Intrahealth International. (2012). Country brief, Available from: http://www.intrahealth.org/files/media/kenya/Country_Brief_Kenya_web%20ready.pdf. [Accessed 10 January 2014]
15. Jalloh, M, Niang, L, Ndoye, M, Labou, I and Serigne M. Gueye. (2013). Prostate cancer in Sub Saharan Africa. *Journal of Nephrology and Urology Research*, 1, 15-20
16. Jemal, A, Bray, F, Center, M. M, Ferlay, J, Ward, E. and Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61: 69690. doi: 10.3322/caac.20107
17. Jones, S.J. Prostate cancer Diagnosis. (2013). [Springer eBook version]. Retrieved from <http://link.springer.com/book/10.1007/978-1-62703-188-2> [Accessed 15 January 2014].
18. Kaur, H. and Wasan, S.K. (2006). Empirical Study on Applications of Data mining Techniques in Healthcare. *J. Comput. Sci.*, 2: 194-200
19. Kenya-National-Cancer-Control-strategy. (2011-2016). Available from: <http://www.ipcrc.net/pdfs/Kenya-National-Cancer-Control-strategy.pdf>. [Accessed 15 January 2014].
20. Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, Vol.2, No.2, DOI : 10.5121/ijcseit.2012.2206 55
21. Magoha G. A. (2000) Management and Survival In Advanced prostate cancer In Nairobi, *East African Medical Journal* Vol. 77, No. 5.
22. Mariscal, G., Marba, O., and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, Vol. 25:2, 1376166
23. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.
24. Murray S. A, Grant . E, Grant . A, and Kendall .M. 2002. Dying from cancer in developed and developing countries: lessons from two qualitative interview studies of

- patients and their careers. Available from: <http://www3.interscience.wiley.com/cgi-bin/fulltext/112098736/PDFSTART> [Accessed 10 January 2014].
25. Ngugi, P.M and Magoha, G.A.O. (2007). The Management Of Early prostate cancer: A Review. *East African Medical Journal* Vol. 84 No. 9 (Supplement)
 26. Njihia, N, Saidi, H, and Mwaura, A. (2010). Towards establishing a National Clinical Cancer Registry ó lessons learnt from the Kenya Oncological Research Database (KORDA). *Distributed Computing Systems Research and Application with focus on Resource sharing*, 49-54.
 27. Oluwole O.P, Taiwo J.O, Awani K.U and Adugba E.O. (2012). Studies on the prevalence of prostate cancer in Lokoja Metropolis. *Adv Trop Med Pub Health Int.*; 2(1): 15 - 20.
 28. Platz, E. A, Kantoff, P.W, and McGeady, E. G. (2000). *Management of prostate cancer*. Oxford: Humana Press.
 29. Prostate cancer UK, 2013: Advanced prostate cancer: Managing symptoms and getting support <http://prostatecanceruk.org/media/1823882/advanced-prostate-cancer-booklet.pdf> [Accessed 13 January 2014].
 30. Quinn M and Babb P. 2002. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. *BJU Int.* 2002 Jul;90(2):162-73
 31. Rahman N and Sarma P (2013). "Analysis of Treatment of prostate cancer by Using Multiple Techniques of Data Mining" *International Journal of Advanced Research in Computer Science and Software Engineering*, 3 (4) [Internet]. Available from: http://www.ijarcsse.com/4_April2013.php
 32. Saritas, I., Ozkan, I. A., and Sert, I. U., (2010). Prognosis of prostate cancer by artificial neural networks, *Expert Systems with Applications*, Volume 37, Issue 9, Pages 6646-6650, ISSN 0957-4174, <http://dx.doi.org/10.1016/j.eswa.2010.03.056>.
 33. Shearer Colin. (2000). *The CRISP-DM Model: The New Blueprint for Data Mining*. *Journal of Data Warehousing* Vol 5. No 4.
 34. Wasike, R.W and Magoha, G.A.O. (2007). Descriptive case series of patients presenting with cancer of the prostate and their management at Kenyatta National Hospital, Nairobi. *East African Medical Journal* Vol. 84 No. 9 (Supplement)

Appendices



Fig 1: Home Page

We built a responsive site for our prognostic tool which will be used in the screening drives within the different zones as assigned herein the report. This tool will key in the age and location of every man to be tested then proceeds for screening. PSA test is carried out where they check the level of protein which is given a gleason score between 1-10 (low risk to high risk relatively)

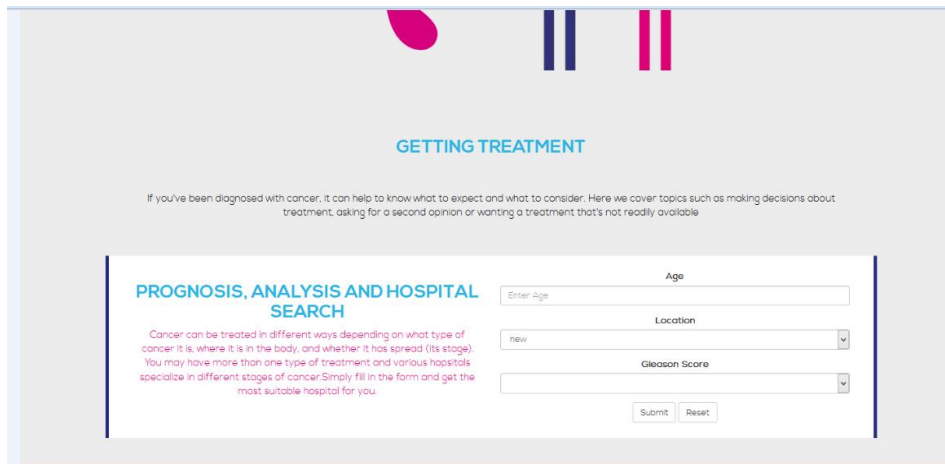


Fig 2: Patient's details keyed in

Once details are submitted, one gets a prognosis which will briefly be interpreted to you by the one who screened you and advised further on what to do. We incorporated GIS where one can be

able to visit the recommended hospitals (according to proximity) to seek further advice or biopsy. The fig 1 . shows a sample of mapping of the hospitals on the maps.

```

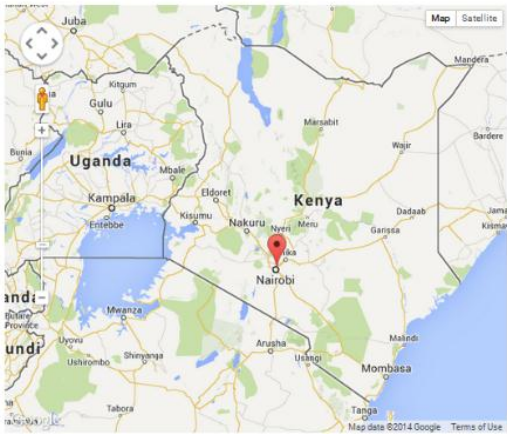
- <markers>
  <marker hosi=" JAMAA HOSPITAL" lat="-1.281757" lng="36.873169"/>
  <marker hosi="NAIROBI WEST HOSPITAL" lat="-1.306558" lng="36.825668"/>
  <marker hosi=" JAMAA HOSPITAL" lat="-1.281757" lng="36.873169"/>
  <marker hosi="NAIROBI WEST HOSPITAL" lat="-1.306558" lng="36.825668"/>
  <marker hosi=" JAMAA HOSPITAL" lat="-1.281757" lng="36.873169"/>
  <marker hosi="NAIROBI WEST HOSPITAL" lat="-1.306558" lng="36.825668"/>
  <marker hosi=" JAMAA HOSPITAL" lat="-1.281757" lng="36.873169"/>
  <marker hosi="NAIROBI WEST HOSPITAL" lat="-1.306558" lng="36.825668"/>
</markers>

```

Fig 3: Mappings of Hospitals

SEARCH RESULTS

Here are the hospitals you can visit



Data Mined

Based on your age there 2 cases that have similar

Hospitals to go to

- JAMAA HOSPITAL
- NAIROBI WEST HOSPITAL

Prognosis: Very Low Risk

Patients with a life expectancy < 20y, should be treated with active surveillance; they should also be referred for observation

Surveillance includes periodic PSA, prostate exam, and prostate biopsy

PSA as often as every 3mo or at least every 6mo; digital rectal examination (DRE) as often as every 6mo but at least every 12mo; repeat biopsy within 18mo but as often as every 12mo

For treatment recommendations for patients with a life

Fig 4: Search Results



Data Mined

Based on your age there 2 cases that have similar

Based on your Location 52 cases that have similar

Total Number of cases 624

Based on location 8.3333% of the cases are similar to your diagnosis

Based on age 0.3205% of the cases are similar to your diagnosis

Patients with a life expectancy $< 20y$, should be treated with active surveillance; they should also be referred for observation

Surveillance includes periodic PSA, prostate exam, and prostate biopsy

PSA as often as every 3mo or at least every 6mo; digital rectal examination (DRE) as often as every 6mo but at least every 12mo; repeat biopsy within 18mo but as often as every 12mo

For treatment recommendations for patients with a life expectancy $\geq 20y$, PSA at least every 6mo; digital rectal examination (DRE) at least every 12mo

Fig 5: Data Mined

The prediction bit comes in where data mined shows some patterns and based on what found in that zone, one's case is predicted based on location as well as based on age to check the likelihood of being a prostate cancer victim.

Sample Code

```
<!doctype html>
<html>
<head>
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="viewport" content="width=device-width, initial-scale=1">
<title>Cancer Support Centre</title>
<link rel="stylesheet" href="bootstrapvalidator-dist-0.4.5/dist/css/bootstrapValidator.min.css">
<link rel="stylesheet" href="css/bootstrap.css">
<link href="css/typography.css" rel="stylesheet">
<link href="css/new.css" rel="stylesheet">

<script src="http://code.jquery.com/jquery-latest.min.js" type="text/javascript"></script>
<script src="js/bootstrap.min.js"></script>
<script src="js/jquery-1.9.1.min.js"></script>
<script src="js/jquery.scrollTo.js"></script>
<script type="text/javascript" src="bootstrapvalidator-dist-
0.4.5/dist/js/bootstrapValidator.min.js"></script>

<!--[if lt IE 9]>
<script src="dist/html5shiv.js"></script>
<![endif]-->

</head>

<body id="home">
<div class="container-fluid">
<header class="row">
<div class="col-md-3" style="text-align:left"></div>
<nav class="navbar navbar-default col-md-9" role="navigation" style="background-color:#fff;
border:#fff;">
  <div class="navbar-header">
    <button type="button" class="navbar-toggle" data-toggle="collapse"
      data-target="#example-navbar-collapse">
      <span class="sr-only">Toggle navigation</span>
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
      <span class="icon-bar"></span>
    </button>
    <a class="navbar-brand" href="#"></a>
  </div>
  <div class="collapse navbar-collapse" id="example-navbar-collapse">
    <ul class="nav navbar-nav" style="float:right">
      <li class="active"><a href="index.php">Home</a></li>
      <li><a href="#">About Us</a></li>
      <li><a href="#">Contact Us</a></li>
      <li><a href="#">Help</a></li>
    </ul>
  </div>
</div>
```

```

</nav>
</header>
<div id="landing" class="row">
  <div class="col-md-8"><h1 style="color:#df037f;">Welcome to <br>
  Cancer <br>
  <span style="color:#162a83;">Support Centre</span></h1>
  no one should face cancer alone. We want to reach and improve the lives of every one of those
  people.</div>
  <div class="col-md-4"></div>
</div>
<div id="call" class="row">
  <div class="col-md-2"></div>
  <div id="serv" class="col-md-4"><h4>Simply Search and...</h4><br><a href="#section1"></a></div>
  <div id="express" class="col-md-4"><h4>Get the nearest Cancer Support Centre</h4><br><br></div>
<div class="col-md-2"></div>
</div>
<div id="section1" class="row">
  <div id="ourservices" class="col-md-12">
    <h2>GETTING TREATMENT</h2><br><br>
    <p>If you've been diagnosed with cancer, it can help to know what to expect and what to
    consider. Here we cover topics such as making decisions about treatment, asking for a second opinion or
    wanting a treatment that's not readily available</p>
  </div>
  <div id="numbers" class="row"></div>
</div>
<div id="services">
  <div id="a4c" class="row">
    <div id="a4ca" class="col-md-6">
      <h2>PROGNOSIS, ANALYSIS AND HOSPITAL SEARCH</h2>
      Cancer can be treated in different ways depending on what type of cancer it is, where it is in the
      body, and whether it has spread (its stage). You may have more than one type of treatment and various
      hospitals specialize in different stages of cancer. Simply fill in the form and get the most suitable hospital
      for you.<br>
      <br>
    </div>
    <div id="a4cb" class="col-md-6">
      <form role="form" class="search" method="post" action="datastuff.php">
      <div class="form-group">
        <label for="age">Age</label>
        <input type="text" class="form-control" name="age" id="age" placeholder="Enter Age">
      </div>
      <div class="form-group">
        <label for="exampleInputEmail1">Location</label>
        <select class="form-control" name="loc" id="loc" >
        <?php
        $db = mysql_connect('localhost','root','') or die ('Unable to connect. Check your connection parameters. ');
        mysql_select_db('cancer', $db) or die(mysql_error($db));
        $sql="SELECT DISTINCT location,zone FROM cluster";
        $result =mysql_query($sql);
        while ($data=mysql_fetch_assoc($result)){
        ?>

```

```

<option value = "<?php echo $data['zone'] ?>" selected="selected" ><?php echo $data['location']
?></option>
    <?php } ?>

</select>

</div>
<div class="form-group">
  <label for="exampleInputEmail1">Gleason Score</label>
  <select class="form-control" id="score" name="score">
    <option selected></option>
    <option value="1">1</option>
    <option value="2">2</option>
    <option value="3">3</option>
    <option value="4">4</option>
    <option value="5">5</option>
    <option value="6">6</option>
    <option value="7">7</option>
    <option value="8">8</option>
    <option value="9">9</option>
    <option value="10">10</option>
  </select>
  </div>
  <button type="submit" class="btn btn-default">Submit</button>
  <button type="reset" class="btn btn-default">Reset</button>

</form>

<script>
$(document).ready(function() {
  $(' .search').bootstrapValidator({
    message: 'This value is not valid',
    feedbackIcons: {
      valid: 'glyphicon glyphicon-ok',
      invalid: 'glyphicon glyphicon-remove',
      validating: 'glyphicon glyphicon-refresh'
    },
    fields: {
      age: {
        message: 'age',
        validators: {
          notEmpty: {
            message: 'The age is required and cannot be empty'
          },
          stringLength: {
            min: 1,
            max: 2,
            message: 'Invalid Age'
          },
          regexp: {
            regexp: /^[0-9]/,
            message: 'Can only be a number'
          }
        }
      }
    }
  });
});

```

```

    }
  },
  loc: {
    validators: {
      notEmpty: {
        message: 'The location is required and cannot be empty'
      }
    }
  }
},

    score: {
  validators: {
    notEmpty: {
      message: 'The gleason score is required and cannot be empty'
    }
  }
}
}
});
</script>

```

```
<!--<script>
```

```

function sendData(){
    var age = $("input#age").val();
    var loc = $("select#loc").val();
    var score = $("select#score").val();

// Returns successful data submission message when the entered information is stored in database.
    var dataString = 'age=' + age + '&loc=' + loc + '&score=' + score;

$.ajax({
  url: "datastuff.php",
  method: "POST",
  data: dataString,
  success: function() {
    $('#myForm').append("<div id='message1'></div>");
    $('#message1').html("")
    .append("<p>dataString is now "+dataString+"</p>")
    .hide()
    .fadeIn(1500);
  }
});
return false;
}

```

```

</script>-->
<script language="javascript">
<!--

<?php
//Include database connection details
require_once('config.php');

//Array to store validation errors
$arrmsg = array();

//Validation error flag
$errflag = false;

//Connect to mysql server
$link = mysql_connect(DB_HOST, DB_USER, DB_PASSWORD);
if(!$link) {
die('Failed to connect to server: ' . mysql_error());
}

//Select database
$db = mysql_select_db(DB_DATABASE);
if(!$db) {
die("Unable to select database");
}

//Function to sanitize values received from the form. Prevents SQL injection
function clean($str) {
$str = @trim($str);
if(get_magic_quotes_gpc()) {
$str = stripslashes($str);
}
return mysql_real_escape_string($str);
}

//Sanitize the POST values
$page = clean($_POST['age']);
$loc = clean($_POST['loc']);
$score = clean($_POST['score']);

function parseToXML($htmlStr)
{
$xmlStr=str_replace('<','&lt;',$htmlStr);
$xmlStr=str_replace('>','&gt;',$xmlStr);
$xmlStr=str_replace('"','&quot;',$xmlStr);
$xmlStr=str_replace("'",'&#39;',$xmlStr);
$xmlStr=str_replace("&","&amp;",$xmlStr);
return $xmlStr;
}

$query9 = "SELECT * FROM hospital WHERE zone = '$loc' ";
$result9 = mysql_query($query9);

```

```
if (!$result9) {
    die('Invalid query: ' . mysql_error());
}

header("Content-type: text/xml");

// Start XML file, echo parent node
echo '<markers>';

// Iterate through the rows, printing XML nodes for each
while ($row = @mysql_fetch_assoc($result9)){
    // ADD TO XML DOCUMENT NODE
    echo '<marker ' ;
    echo 'hosi="' . parseToXML($row['hosi']) . "' ";
    echo 'lat="' . $row['lat'] . "' ";
    echo 'lng="' . $row['lng'] . "' ";
    echo '>';
}

// End XML file
echo '</markers>';

?>
```