UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A COMPARATIVE STUDY OF DECISION TREE AND NAÏVE BAYESIAN CLASSIFIERS ON VERBAL AUTOPSY DATASETS

BY

GORDON OUMA ONDEGO

This report is submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

`

OCTOBER 2015

DECLARATION

I hereby declare that this report is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.

……. ……

Ondego Gordon Ouma **Date**

P58/76292/2012

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

……..……...…… …….

Dr. Lawrence Muchemi **Date**

School of Computing and Informatics,

University of Nairobi

DEDICATION

I dedicate this project my beloved mother, late dad and family for their love, continued support, encouragement and understanding and always being there for me when I needed them. Above all I dedicate it to God for His blessings all through.

ACKNOWLEDGEMENTS

Above all, I would like to glorify the almighty God for giving me the ability to be where I am. You have done so much for me, O Lord. No wonder I am glad! I sing for joy, Amen!

Secondly, my sincere thanks and appreciation to my supervisor Dr. Lawrence Muchemi and the panel chairman Dr. Elisha Opiyo for their constructive comments and overall guidance.

I would also like to thank Dr. Samuel Danso, a Machine Learning Researcher and Senior lecturer at the Leeds University-UK for his invaluable support and professional guidance during this research

I am very much grateful to those who provided me with verbal autopsy samples during this research, Dr. Abraham Flaxman and Sean Green of the Institute of Health Metrics and Evaluation, Washington University, USA.

At last, but by no means the least, I would like to thank my friends and workmates for the constant assistance and encouragement they rendered to me since the time of my admission to the postgraduate program.

ABSTRACT

With an increased effort to reduce mortality rate in most developing countries, accurate information on the causes of such mortalities is a very crucial component for the development and formulation of health policy, strategies and other key critical decisions in the health sector. However there is lack of complete, accurate and reliable vital registration system that is expected to generate and report accurate causes of death information for health intervention policies and other programs. This research sets out to make a comparative evaluation of two most common supervised machine learning approaches Naive Bayes (NB) and J48 decision tree which builds a decision tree in the context and with the aid of Institute for Health Metrics and Evaluation (IHME) Verbal Autopsy (VA) dataset.

This research also focuses on experimental comparison of these two state of art supervised learning techniquues with respect to their accuracy of correctly classified instances, incorrectly classified instances and very important Receiver Operating Characteric (ROC) Area which helps in understanding the classification model and their results, which can also help other researchers in making decision for the selection in classification model based on their data and number of attributes.

With reference from several conference papers published recently, journals and other resources, the research was accomplished by training and testing the selected algorithms with the same datasets using a 10 fold cross validation method in Waikato Environment for Knowledge Analysis (WEKA) platform. The experiments carried out in this research are about classification accuracy, sensitivity and specificity using true positive (TP) and false positive (FP) in confusion matrix generated by the respective algorithms. The results obtained shows that J 48 decision tree algorithms out performs Naïve Bayes in terms of accuracy, recall, precision and F score. The perfection of these algorithms in the classification task is further explained with the analysis of ROC curve. The results obtained from the study indicate that J48 decision tree algorithm performs better than the Naïve Bayes classifier. A prototype has been developed based on the J48 decision tree algorithm because it exhibits good performance in the prediction of cause of death from the verbal autopsy data set.This prototype can be used by medical experts both in the private and public hospitals to make more timely and consistent diagnosis of the causes of death from the verbal autopsy for those deaths occuring outside health institutions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN-Artificial Neural Networks

ARFF-Attribute Relation File Format

AUC-Area Under Curve

CoD-Cause of Death

CRISP-DM-Cross Industry Standard Process for Data Mining

CSV-Comma Separated Values

DSS-Demographic Surveillance System

DM-Data Mining

DT-Decision Tree

FN-False Negative

FP-False Positive

GBD-Global Burden of Disease

ICD 10-International Classification of Diseases version 10

IHME-Institute for Health Metrics and Evaluation

KDD-Knowledge Discovery in Databases

KDP- Knowledge Discovery Process

ML-Machine Learning

MAP–Maximum a posteriori

NB-Naïve Bayes

NBC-Naïve Bayes Classifier

PHMRC-Population Health Metrics Research Consortium

ROC-Receiver Operating Characteristic

SEMMA- Sample, Explore, Modify, Model, and Assess

SQL-Structured Query Language

SSP-Simplified Symptom Pattern

SVM-Support Vector Machines

TN-True Negative

TP-True Positive

VA-Verbal Autopsy

WEKA-Waikato Environment for Knowledge Analysis

WHO-World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Background

Machine learning is the study of computer algorithms that improve automatically with experience. That is, the ability of the computer program to acquire or develop new knowledge or skills from examples for optimizing the performance of a computer or a mobile device. (Beyene 2011).The application of machine learning is growing in various applications widely like analysis of organic compounds, medicals diagnosis, product design, targeted marketing, credit card fraud detection, financial forecasting, automatic abstraction, education, computational linguistics, bio-informatics, stock market prediction, predicting shares of television audience etc. (Patil and Sherekar 2013). The field of Machine learning deals with developing programs that learn from past data and is also a branch of data processing. Machine learning includes the stream in which machines learn for knowledge gain or understanding of some concept or skill by studying the instruction or from experience (Archana, Raj and Savita 2013).Machine learning therefore can be used to develop systems resulting in increased efficiency and effectiveness of the system. Machine learning task can be categorized as either supervised on unsupervised. In supervised learning, the learning algorithm is given a labeled training set to build the model on. It is called “supervised “as it could be thought of as the teacher providing the patterns and their true classes on the basis of which the model learn show to return the best solution to the given problem. The term Verbal Autopsy is used to denote the process that involves interviewing people such as caregivers who were very close to the deceased prior to death and may have witnessed the events prior to the death and are able to clearly narrate them. The interview is always in the form of a standard questionnaire designed by the WHO with information containing signs and symptoms of the possible ailment that led to the death. The VA data is then studied analyzed and interpreted by physicians to ascertain the true cause of death. Verbal autopsies rely on the assumption that most causes of death have distinct symptoms and signs that can be recognized, recalled, and reported by household members or associates of the deceased to a trained, usually nonmedical field worker.

The physician’s approach is characterized by several limitations: high cost; intra-physician reliability; repeatability; and the time consumed. Consequently, research into computational

techniques to explore and analyze verbal autopsy data is being studied to address these limitations (Danso et al, 2013), (Murray et al, 2014)

This paper carries out at a comparative study between two popular text classification algorithms which are useful in solving classification problems to identify which approach is most suitable for the verbal autopsy data in terms of the predictive accuracy on the selected datasets. The paper also investigates various feature value representation schemes, machine learning algorithms and the effect of feature reduction on the overall performance accuracy of the machine learning algorithms

To the best of my knowledge this is the first paper that reports on a comparative study between two state of art supervised machine learning approaches based on this particular data set which is less than a year, other data sets which have been used for this study before were not of gold standard.

1.2 Classification

Classification is a supervised technique in machine learning which is a task of predicting the value of a categorical variableby building a model based on one or more numerical and/or categorical variables (predictors or attributes) (Deepajothi.S & Selvarajan.S., 2012) In classification, training examples are used to learn a model that can classify the data samples into known classes. The classification process involves creating training data set, identifying the class attribute and classes, identifying useful attributes for classification (relevance analysis),learn a model using training examples in training set and finally use the model to classify the unknown data samples

There are various machine learning classification techniques and they have been employed to tacke various classification problems.The only major differences that exists between these techniques is the philosophy behind the learning process. Classification methods refer to classes and attributes; in the context of VA, classes are the validated Cause of Death (CoD) and attributes are signs, symptoms and other data about the deceased which are collected using the VA questionnaire.

The application of machine learning to classify cause of death from verbal autopsy data has been proved to be useful (Danso, et al., 2010).VA is a technique recommended by the World Health Organization (WHO) as an alternative to accurately determine the true cause of death in resource poor countries where death may have occurred outside a health facility and with poor death registration systems (Danso et *al*., 2013).

1.3 Prediction Algorithms

1.3.1 J48 Decision Tree Algorithm

This is a simple graphic structure where non-terminal nodes represent tests on one or more attributes and terminal nodes give decision outcomes. This tree consists of one root, branches, internal nodes and leaves. This tree is drawn from left to right or beginning from the top root to downward nodes, so that it is easy to draw it. (Archana, et al., 2013).The classifier is an important model to realize the classification with a flowchart like structure in which the internal nodes i.e. non-leaf node denotes a test on an attribute and each leaf node denotes a class label (Jeyarani, et al., 2013).J48 decision trees based algorithm learns from training examples by classifying instances and sorting them based on feature values.

The algorithm has been employed successfully in many traditional applications in different domains (Jeyarani, et al., 2013) eg it hasrecently been employed as a machine learningtechnique to develop classification models that automatically classify pancreatic cancer data (Danso, et al., 2013).However, decision tree techniques are known to have scalabilityand efficiency problems, such as substantial decrease in performance and poor use of availablesystem resources

The figure below is an illustration on how the Decision Tree works in classification task within the feature space.

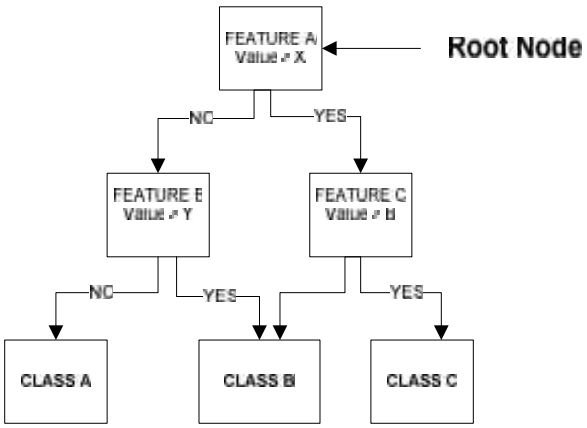


Figure 1: Graphical representation of a Decision Tree learning algorithm

The algorithm starts the whole process of classification at a root node the tree. The root is the feature that best divides the feature space. The classes are assigned based on the weights that are computed on the features during the process of classification (learning) and these weights are used to classify future unseen data (Parmar and Shah 2013)

1.3.2 Naïve Bayes Algorithm

A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier based on Bayes "rule with strong (naive) independence assumptions i.e. given a class label the value of each attribute is independent to each other. Considering D to be the data that has been seen so far and h being a possible hypothesis, then Bayes‟ theorem definition is given by:

$$P(h|D) = \frac{P(D|h)P(h))}{P(D)}$$

.

Where:

P (h): Prior probability of hypothesis h - Prior

P (D): Prior probability of training data D - Evidence

P (D|h): Conditional Probability of D given h – Likelihood

P (h|D): Conditional Probability of h given Posterior probability

The conceptual framework for the naïve bayes is based on joint probabilities of features and classes to estimate the probabilities of a given document belonging to a given Class.

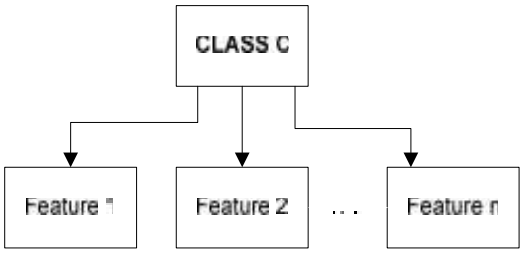


Figure 2: Naïve Bayes Conceptual representation

During training, the probability of each class is computed by counting how many times it occurs in the training dataset known as the "prior probability". In addition to the prior probability, the algorithm also computes the probability for the instance 'x' given a class 'c' with the assumption that the features are independent.

Naïve Bayes (NB) has been used in this study because it is considered to be a relatively simple machine learning technique based on probability models (Danso et al. 2013). This classification technique analyses the relationship between each feature and the class for each instance to derive a conditional probability for the relationships between the feature values and the class. The attribute conditional independence assumption of naive bayes essentially

ignores attribute dependencies and is often violated. On the other hand, although a Bayesian network can represent arbitrary attribute dependencies, learning an optimal Bayesian network classifier from data is intractable. Thus, learning improved naive bayes has attracted much attention from researchers and presented many effective and efficient improved algorithms (Deepajothi.S and Selvarajan.S. 2012)

In this study, a model using Naïve Bayes classifier has been developed since the technique is popular in machine learning applications, due to its simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes naïve bayes techniques attractive and suitable for many domains including verbal autopsy data classification

1.4 Problem statement

The information about the exact cause of death and its usefulness to the WHO, local and international community has been cited and acknowledged as a pertinent issue and globally over 60% deaths occur outside the health facilities hence their true causes go unrecorded and uncertified (Baiden et al. 2007). This is in itself a tragedy and it is therefore not easy to realize the full potential of health systems if what people die from cannot be properly ascertained.
However the accuracy and efficiency of this information is what counts for the formulation of sound and solid health care strategies and policies. Moreover, understanding the current determinants of child mortality is essential to inform policies and strategies to accelerate the reduction of child mortality.

There is lack of experiments that have been done to identify the most suitable learning algorithm for classifying verbal autopsy data. This is a weakness of the existing systems and this study addresses this weakness through a comparative study and evaluation of the prototype results of the two classification techniques.
Since no single machine learning classifier is adequate, perfect and accurate for all possible learning problem in general and VA data classification in particular. This study therefore performs an evaluation of these two common algorithms and chooses the best that best reflect the predicted class.

The other problem revolves around the learning time, the accuracy, the data requirements and the imperfect data presence in the verbal autopsy data since the data is characterized by these

features. The existing methods have not addressed these features as well and this calls for a study. Also there had been many researches that compared different machine learning techniques including Naïve bayes, J48 decision trees and Support Vector Machines. However they used small data sets which are not gold standard.

Effective tools are required to help in correctly classify verbal autopsy data.However the accuracy and efficiency of this information is what counts for the formulation of sound and solid health care strategies and policies.Utilizing the capability of Naïve Bayes and J48 decision tree classifiers can help handle the complexity of these processes. In view of this, presented here is a model based on the use of J48 decision tree and Naïve Bayes classifiers to help speedup and improve accuracy and efficiency in verbal autopsy classification

1.5 Objective of the Research

1.5.1 General Objective

The general objective of this research was to classify verbal autopsy data sets using machine learning algorithms and techniques and predict the accurate cause of death in a population so that the information can aid decision- or policy-making processes in the health sector.

1.5.2 Specific Objectives

To achieve the general objective, the specific objectives for this research are:

- To assess and compare the performance of a Naïve Bayes classifier against J48 decision tree classifiers based on the verbal autopsy dataset.
- To identify a suitable machine learning algorithm that implements the techniques identified
- To build a prototype based on the best classifier, test and evaluate the prototype performance using a set of experiments

1.6 Significance of the study

This paper aims to highlight the important role of computer science in general and in specific machine learning in classifying verbal autopsy data to predict the cause of death from such data and propose a basic model based on some machine learning classifiers. The information about the cause of death in a population is of a greater benefit to the health policy makers,

strategist, planners and the decision makers at all levels in the health sector to know what kills its people so that prior interventions can be made to reduce such deaths and also the mortality statistics are a widely-used resource for setting spending priorities.

With the increased demand for accurate information about the cause of death amongst all age groups in the world and acquiring such information is always not easy due to poor vital registration systems in most developed countries, this study develops a model that demonstrates the capabilities of Naive Bayes and J48 decision tree classifiers as a tool to classify cause of death from the Verbal autopsydata sets so as to help improve the efficiency of the process,this model helps in verbal autopsy data classification problem so that the exact mortality cause can be predicted from a set of verbal autopsy data .

1.7 Scope of the Project

The study examines the application of Naive Bayes classifier and J48decision tree and their relevance to Verbal Autopsy data classification

The applicability of machine learning in this research is limited to development and testing of the model instead of deploying the model at health care centres since the study is being carriedout for academic achievement. That is, the scope of the current experimental researchundertaking is strictly limited to appraising the potential applicability of machine learning technology to support primary health care activities at the area of study.

1.8 Definition of Terms

Classification: The systematic grouping of like things or objects into classes or categories according to some shared quality or characteristic

J 48 Decision Tree Classifier: A classifier that builds decision trees from a set of labeled training data using the concept of information entropy

Naïve Bayes Classifier (NBC): A probabilistic classifier based on applying Bayes Theorem with strong (naive) independence assumptions.

Conditional Independence: A simplifying assumption that attribute values are independent given a target value.

Maximum posterior (MAP): This is the maximally probable hypothesis from amongst a set of generated hypotheses

ConfusionMatrix: It is ann-dimensional square matrix,where n is the number of distinct target values

Training set: A set of examples used for learning. It is used to obtain the pattern in data

Validation set: A set of examples used to tune the parameters of a classifier

Testing Set: A set of examples used only to assess the performance (generalization) of a fully-specified classifier.

Sensitivity: Measures the proportion of actual positives which are correctly identified as such (i.e. accuracy on the class Positive)

Specificity: Measures the proportion of negatives which are correctly identified (i.e. accuracy of classifier)

Gold Standard: Is a diagnostic test or benchmark that is regarded as definitive

1.9 Assumption and Limitations

There are limitations involved in this study as indicated below:-

i) Despite the vigorous attempts led by the WHO to standardise almost all the verbal autopsy tools and coding procedures, there is no unified format of the questionnaires used. They vary in both content and length, with some using open questions, some only closed questions and some a mixture of the two.This becomes a limitation when doing studies using different questionnaire format

ii) The data about the the deceased may not be a true respresentation of the general population.This could affect the answers given at the VA interview and also the cultural issues affects the quality and accuracy of the verbal autopsy data.The willingness of the relative of the deceased to agree to an interview,narrate the way the symptoms and disease is an important major contributing factors to the attainment of specific cause of death. Also the attitude of a particular community towards a particular cause of death eg HIV/AIDS limits the quallity of the data obtained

CHAPTER 2

BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Introduction

The term “Verbal Autopsy” is the collection of post-mortem information about a deceased individual through questionnaire or interview of household members, friends and others (including health care workers) who cared for the person at home or is familiar with the circumstances of the death. Verbal autopsy methods are most often used in locales where formal medical care is difficult to access. Verbal autopsy procedures are widely used for estimating cause-specific mortality in areaswithout medical death certification. In such locales, deaths often occur at home and official records are inconsistently available.(Danso, et al., 2011) Verbal autopsies may provide important public health information about factors related to deaths and actions taken to address the medical problems and prevent the death.

2.2 Verbal Autopsy Background

Interest in causes of death for public health purposes goes back to the 17th century in London, when “death searchers”were recording deaths in the population by weekly household visits, with the main target being to estimate mortalityfrom the plague(Gary & Ying, 2008).The first simplified lists of causes of death for use in developing countries were published by the WHO in 1978 (Mathers, et al., 2005) and since then the needs to have an accurate assessment of causes of premature deaths have only increased. Such needs are well covered in developed countries by a combination of routine compulsory deathregistration and medical diagnosis of each death. In many developing countries, however, death registration is still in complete and causes of death remain largely undocumented because many deaths occur outside health facilities. the leading causes of death can help formulate policies to combat these and evaluate current strategies and health programs.(James, et al., 2011) Verbal autopsies were developed to bridge this gap. At first, they were conducted in research settings by an in-depth interview with the family of the deceased person.

A good example is the Narangwal research project inIndia, where the term “verbal autopsy” was coined in theearly 1970s (Garenne, 2014) This approach was limited by its cost and by the potential bias of a single observer. The next step was to use systematic questionnaires on a detailed history of the disease, signs, symptoms, treatments and any contextual information, including risk factors. This approach was less costly, more objective and allowed for some kind of proof for the final diagnosis. Several questionnaires were developed in the late 1970s

and early 1980s for maternal deaths in Egypt (Ruzicka & Lopez, 1990)for neonatal and children deaths in Bangladesh (Peter, et al., 2003)and for all causes in Senegal (Quigley, et al., 1999)which were further developed and adapted to a great variety of situations. They were used in research projects, in Demographic Surveillance System (DSS) sites such as Agincourt in South Africa (Boulle, et al., 2001) and soon were tried on a few Demographic and Health Surveys (DHS) (Ghana 2007; Afghanistan 2010), and now on a very large scale in countries such as Mozambique, India and China (James, et al., 2011)

However it is the work of Garenne & Fontaine who are considered the founders of the VA technique through the development of a VA questionnaire used in studies in Senegal. This technique has been adopted worldwide (Murray, et al., 2011)

Figure 3: The use of verbal autopsy across the world

World map of countries (grey shading) where verbal autopsy methods are applied.**Source**:Fottrell/Byass.2010. *Verbal Autopsy: The Tools*

A standard verbal autopsy tool as shown in the figure below consists of a questionnaire, cause of death classification system and diagnostic criteria (physician review, expert or data driven algorithm) (Mathers, et al., 2005).The actual questionnaire itself contains 10-100 questions (see Appendix D for an example). There are two different interview methods; one uses an in-depth, open-ended history of the final illness asking the care giver to outline the events in their own words. This is a descriptive account which will then be read and coded. The other technique is interviewer asking closed questions often pre-coded for use with an algorithm. Most VA's are

conducted using a mixture of the both the closed and open-ended approach (Murray, et al., 2011)

The interview is conducted by a well-trained lay person, medically trained interviewer or health professional. Much debate has taken place on the pros and cons of using lay and medical trained personnel. Although to date, the effects and outcomes of different interviewers are not known to have been formally studied (Murray, et al., 2014). Those conducting the interviews do receive training, although it is argued that the process would benefit from standardized guidelines. The understanding of local customs/culture, terminology and concepts of illness and their symptoms are seen as key in the process of acquiring a quality questionnaire (Leitao, et al., 2014). The most common interpretation method of the questionnaire is local physician review without algorithms (Peter, et al., 2003).When the VA questionnaire is complete it is sent to a local health facility. On arrival the VA is annotated using the International Classification of Disease version 10 (ICD-10) coding standards by a "coder" and then entered onto a computerized system either by the coder or a data entry clerk. If consensus can be gained a cause of death is decreed. If not, the death is recorded as "indeterminate". The second approach is expert algorithm. "The algorithm can be developed from text book description, existing clinical algorithms, local experience of a combination of both "The third approach is data driven algorithm (Soleman, et al., 2006).This requires an additional sample of deaths from a medical facility where each cause is known and symptoms are collected from relatives. Then a parametric statistical classification method (logistic regression, neural networks and support vector machines)is trained on the hospital data and used to predict each cause of death in the community (D.Flaxman & T.Green, 2010)

Source: Soleman etal 2006

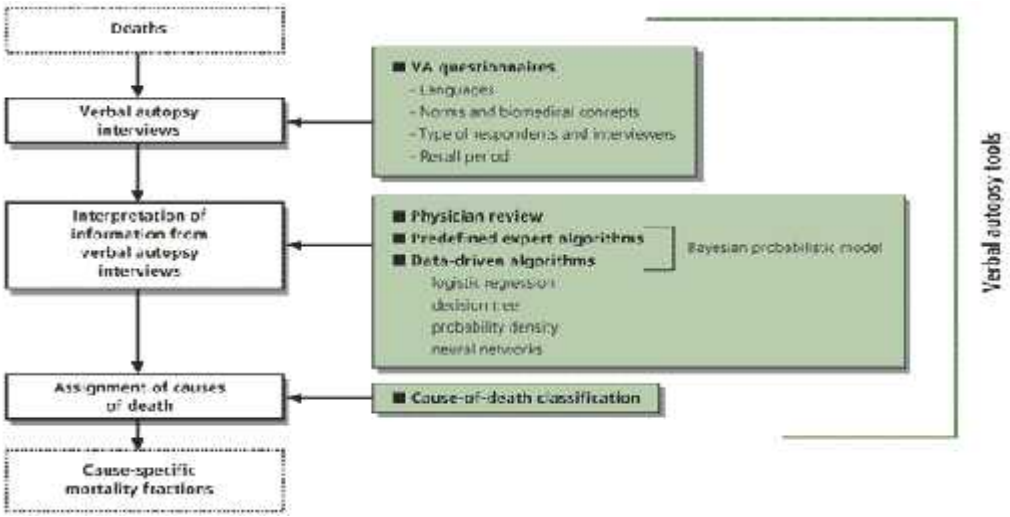


Figure 4: Verbal Autopsy Tools and Process.

2.4 Methodology of Administering Verbal Autopsy

The main purpose of VA data collection is to analyze health in the community with the goal of determining individual cause of death and community/specific mortality fraction in a population without vital registration system (Murray, et al., 2011). The information is gathered through interviews with family, or friends or caregivers of the deceased. Thereafter the interpretation is done by a coder. The interpretation of VA data provides an opportunity for health planners, policy makers, and epidemiologists to understand better the patterns and implications of mortality in the community. A questionnaire is administered to obtain health data which later on are used to ascertain the cause of death when a death event is reported (Leitao, et al., 2014). A baseline census is usually the source of data. A baseline census is conducted initially to provide a denominator of the population. Within the enumeration area, individuals are registered in their respective households. Any member who intends to stay in the house for more than six months is registered. A community integrated system is defined such that whenever there is a vital event within the community, that event is reported by a key informant (KI) using a mobile phone (Vitalis, et al., 2014)

Data for cause of death is a critical input in formulating good public health policy. However, data need to be collected reliably and interpreted consistently to serve as a global indicator (James, et al., 2011). For those countries with no vital registration, VA is a reliable method that is commonly used to study the pattern of cause of death. Regardless of the methodology and tool used, the process of collecting, interpreting and processing VA data is very involving and uncertain (Gary & Ying, 2008)(Mathers, et al., 2005), (Murray, et al., 2011). It is pinpointed in (Murray, et al., 2011) that rigorous validation of VA procedure is needed to establish confidence in the data collection. Additionally, in order to understand the operational characteristics of VA in the population under study and to identify misclassification patterns, a controlled method of information collection is indispensable. Furthermore, the significance of collecting VA is to improve country and regional global health information. The VA information is vital for public health, decision making, health sectors reviews, planning and resource allocation as well as program monitoring and evaluation (Ruzicka & Lopez, 1990).Also, the cause of death statistics is useful to understand which disease kill and how many people die (Soleman, et al., 2006).Collection of cause of death requires strong collaboration between the ministry of health, department of civil registration, and national bureau of statistics as well as the health research institutions.

2.5 Machine Learning Application in Verbal Autopsy and Related Works

There have been many papers written and research work done in the field of classification and most work is based on Naïve Bayes,J48 decision trees, Artificial Neural Networks (ANN),Support Vector Machines (SVM).As mentioned earlier verbal autopsy is an indirect method of ascertaining cause of death from information about symptoms and signs obtained from bereaved relatives. This method has been used in several settings to assess cause-specific mortality. However, cause-specific mortality estimates obtained by VA are susceptible to bias due to misclassification of causes of death. One way of overcoming this limitation of VA is to employ other computational approaches in classifying cause of death from the dat.

Many reseachers have proposed the use of various data and expert-driven algorithms to analyse Verbal Autopsy data and they have successfully made tremendous impact in the cause of death prediction.A comparative study and analysis of various machine learning methods for classifying verbal autopsy data sets have been studied by (Danso et al. 2013),(Murray et al. 2014).The authors (Danso et al. 2013) explored various machine learning classification techniques & algorithms and presented a comparative study that explores various aspects of machine learning approaches suitable for classifying verbal autopsy data:feature value reduction;machine learning algorithms;and the effect of future reduction.Their study discussed and investigated some of the methods that have been used in text classification and the performance evaluated:NB,SVM and decision Trees.The experiment found out that SVM was best performing algorithm and most suitable for verbal autopsy data.However Naive Bayes perfomed better than SVM when explored with binary feature representation which is appropriate for data with limited vocabulary size.

This study however as reported (Danso et al. 2013) did not make efforts to compare the results of the experiment with others reseachers who have explored the closed part of the verbal autopsy data sets in their research.This was because the main aim of their research was to build the perfect obtainable baseline results from the methods explored using a Bag-of-Words (A bag of words representation of a document assigns a weight value for each term occurring in the document. It is a simplified representation of a document, because it assumes that the document's terms are independent of each other) approach for bulding a classifier with the highest accuracy using machine learning algorithms.The authors recommended that

future work should explore the possibility of employing feature reduction approaches and compare the results with the approaches used in their experiment.

Some researchers (James et al. 2011) in their study proposed a techique called tarriff method as a way of validating a simple additive algorithm for analysis of verbal autopsy data.The method works on the principle of identifying signs and symptoms collected in the Verbal Autopsy and these are the main pointers to the cause of death.It assigns a tarriff for each sign and symptom for each cause of death to show and reflect how informative that sign and symptom is for a particular cause.For a given death,the tarriffs are summed resulting into an item-specific tarriff score for each death for each cause.The cause that results into the highest tarriff score for a particular death is assigned as the predicted cause of death for that individual.The method uses data sets where the cause of death is known and the tarriff is computed as a function of the fraction of deaths for each variable having a positive response.The authors argued that the tarriff method the physician certified verbal autopsy,however it does not take into account the interdepedencies of signs and symptoms conditional on particular cause.

Gary and Ying (2008) experimented with a probabilistic model using symptom profiles to determine the mortality fractions for all causes of death in the community at once without individual case of death attribution. In this model, multiple causes for an individual are handled by joining two or more causes together into a single category.The major drawback as reported in (Rebecca 2010) is that it requires a high quality health facility mortality data. According to the study by (Peter et al. 2003) experimented on a model based on Bayes theorem that identified various disease indicators and defined the probability of a particular cause based on the presence of specific indicators. This study reported consensus for 75% of cases between the model-assigned and physician review-assigned causes of death.

Finally, combining the methods of King and Lu and Byass and InterVA method a new method called Simplified Symptom Pattern was proposed (Murray et al. 2011) and validated with the standard physician coded verbal autopsy and the results showed that the simplified Symptom Pattern correctly estimated cause specific mortality fractions with less error than physician coded verbal autopsy at both the population and individual level. These methods have advantages in that they do not rely on algorithms, require less time to analyze and do not require the time, effort and cost of physician reviewers. It is still unclear how these methods will vary across cultural and language barriers, however, validation studies are currently being conducted in multisite global field settings (Murray et al, 2011)

Abraham et al. (2011) in their study proposed a random forest method to analyse verbal autopys to examine the accuracy of the method compared to a data set with known causes and with physician certified verbal autopsy.The authors argued and reported that the method performs better than the usual physician certified verbal autopsy method in accuratetly the cause specific mortality at both the individual and population levels.This method was based on Decision Tree whereby the decisions between two possibilities was made starting from the top level and systematically progresses to the next level,following the branch to the right if the symptom is endorsed and vice versa.

There was a study done using Artificial Neural Networks for classifying mortality cause from a verbal autopsy data (Boulle et al. 2001) .The authors argued that this method outperforms other data derived techniques such as the random forest and tarriff methods.However the method had limitations too: the number of hidden nodes, inputs and training time all affect the training time;it is time-consuming to build and train multiple networks for each ANN model.

A research by (D.Flaxman and T.Green 2010) described how a study and an experiment was done with SVM classification algorithm in R programming environment.It was realized that the algorithm was not able to classify the cause of death for all causes with an average generalization error below 60%. These researchers proposed a model that combines the outputs of multiple classifiers since some classifiers appear to predict some cause of death than others. They also recomended making adjustments to the lis of casues so that a generalization arror could be reduced by clustering together causes with similar signs and symptoms.
Peter et al. (2003) experimented with with a probabilistic approach to interpret verbal autopsy data.They described and developed a Bayesian model for verbal autopsy interpretation as an attempt to find a better approach.The results of their experiment proved to be much better than physician certified verbal autopsy

In this study, a model using Naïve Bayes Classifier and J48 Decision Tree is developed to help overcome these overheads. Naive Bayes (NB) models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes NB techniques attractive and suitable for many

domains including verbal autopsy data classification. The conditional independence assumption, even when violated, does not degrade the model's predictive accuracy significantly and this makes NB-based systems offer quick training, fast data analysis and decision making, as well as straight forward interpretation of test results. All these algorithms differ greatly in the characteristics and the approach they use for learning and are popular algorithms for solving supervised learning problems (Jeyarani, et al., 2013) as exposed from the literature search (section 2.4

2.6 Naïve Bayes Classifier

A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier based on Bayes "rule with strong (naive) independence assumptions. Naive Bayes Classifier is used mainly for performing classification tasks. Considering D to be the data we've seen so far and *h* being a possible hypothesis, then Bayes "theorem definition is given by:

$$P(h|D) = \frac{P(D|h)P(h))}{P(D)}$$

Where:

P (h): Prior probability of hypothesis h- Prior

P (D): Prior probability of training data D - Evidence

P (D|h): Conditional Probability of D given h - Likelihood

P (h|D): Conditional Probability of h given D- Posterior probability

In the general case, we have *K* mutually exclusive and exhaustive classes $_i$,i=1...K;$P(D|h_I)$ is the probability of seeing *D* as the input when it is known to belong to class h_I.The posterior probability of class *hi* can be calculated as-

$$P(h_i|D) - \frac{P(D|h_i)P(h_i)}{P(D)} - \frac{P(D|h_i)P(h_i)}{\sum_{i-1}^{K} P(D|h_i)P(h_i)}$$

Source: (Alpaydin, 2004)

The above formula can be summarized as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Since the denominator of the fraction does not depend on the class variable, the numerator is considered and thus the latter is equivalent to the joint probability model. This is represented as:

$$P(D|h_i)P(h_i)$$

The Naïve Bayes Classifier is based on the simplifying assumption that the attribute values are conditionally independent given target value; Mitchell (1997).This assumption is called class conditional independence. It is made to simplify the computation involved and this is why it is considered "naive". In other words, the assumption is that given the target value of the instance, the probability of observing the conjunction a_1, *a2*.....a_n *is* just the product of the probabilities for the individual attributes:

Incorporating the assumption, the Naïve Bayes Classifier is given by:

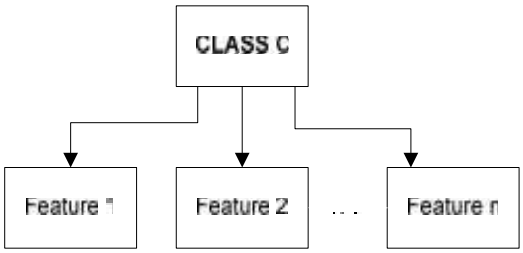


Figure 5: Structure of a Naïve Bayes Classifier

2.6.1 Advantages of Naïve Bayes Classifier

Suitability and extensive use of NBC as an enabling tool for health care decisions and planning such as cause of death prediction from verbal autopsy data sets have been attributed to certain contributing factors some of which include the following:

- Easy to implement - It requires a small amount of training data to estimate the parameters necessary for classification
- The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. Naïve assumption of class conditional independence helps reduce computational cost.
- The algorithm is good for large data sets
- Highly practical Bayesian learning method and is particularly suited when dimensionality of the input is so high
- Its operation is simple and intuitive, relying only on basic laws of probability
- It accommodates limited information as encountered in the problem domain, thus allows a broader set of model parameters to be used, since the model does not require observations for all independent variables.
- Being explicitly probabilistic, it reports results in a form that can easily be interpreted.
- It is robust to outliers

2.7 J48 Decision Tree Classifier

A J48 decision tree is a classifier model that works with recursive partition of the instance space. It is used to represent a supervised learning approach (Dewan Md, et al., 2010). It is a simple graphic structure where non-terminal nodes represent tests on one or more attributes and terminal nodes give decision outcomes.J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.J48 examines the normalized information gain (difference in entropy) those results from choosing an attribute for splitting the data. The entropy is low, and the attribute value is very useful for making a decision (S.Deepajothi & Dr.S.Selvarajan, October 2012). Entropy measures the amount of randomness or surprise or uncertainty.i.e. when entropy = 0 implies there is no disorderliness in the item or dataset.

This classifier has been employed successfully in many traditional applications indifferent domains(Jeyarani, et al., 2013) Despite the fact that it can be regarded as relatively old technique, it has stood the test of time. For example, decision tree has recently been employed as a machine learning technique to develop classification models that automatically classify pancreatic cancer data(Danso, et al., 2013).Decision based algorithm ‘learns’ from training examples by classifying instances and sorting thembased on feature values. Each node in a decision tree represents a feature of an instance to be classified,and each branch represents a

value that the node can include in making a decision. The algorithm starts the process at a root node ofthe tree. This root node is established by finding the feature that best divides the feature space, and there are numerous approaches to identifying the best feature(Jeyarani, et al., 2013).The classes are assigned based on weights that are computed on the features during the processes of learning and these weights areused to classify unseen data. Due to the approach J48 decision tree uses to search for a solution within the problem space, efficiency tends to be an issue, especially when dealing with large datasets.

Decision tree is one of the easier data structure to understand in machine learning. Rules from the training data set are first extracted to form the decision tree which is then used for classification of the testing dataset. A decision tree is necessarily a tree with an arbitrary degree that classifies instances (Patil & Sherekar, 2013)

2.7.1 Advantages of J48 Decision Tree Classifier

The major benefits of using a decision tree are:

- It is a simple model that helps in decision making.
- It is relatively easy to interpret and understand.
- It can be easily converted into a set of production rules.
- It can classify both categorical and numerical data but the resultant attribute is categorical.
- It requires no prior assumptions about the nature of the data

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Design

The study used the Cross Industry Standard Process for Data Mining (CRISP-DM) model to achieve the goal of building predictive model using machine learning techniques. This methodology was selected among different methodologies like KDD, SEMMA, and KDP etc. due to the benefits and the needs of the academic research community, providing a more general, research-oriented description of the steps (Beyene, 2011)

3.1.1 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 and machine projects and is considered the de facto.This methodology is the leading interms of its usage by the data miners based on the polls conducted in 2002,2004,and 2007(Pete et al. 2011)

This methodology is also an excellent fit to this project because the subject area has been identified and also the requirements and aims of the project as identified are flexible enough. To obtain the best outcome it was of key importance to build and refine as knowledge grows.As a result, other methodologies like the spiral and water fall has been rendered less useful because the type of problem that is handled in this research involves understanding the problem space and through this building a model with a number of iterations to understand the issue and draw conclusions,hence CRISP-DM was considered to be the perfect methodology suitable for this research. Overall this approach is the most perfect for this project. Why? Strong emphasis needed to be placed on a thorough understanding of the dataset and its preparation. (Rebecca 2010), (Samuel 2006).The model comprises of six stages as shown in the figure below:-

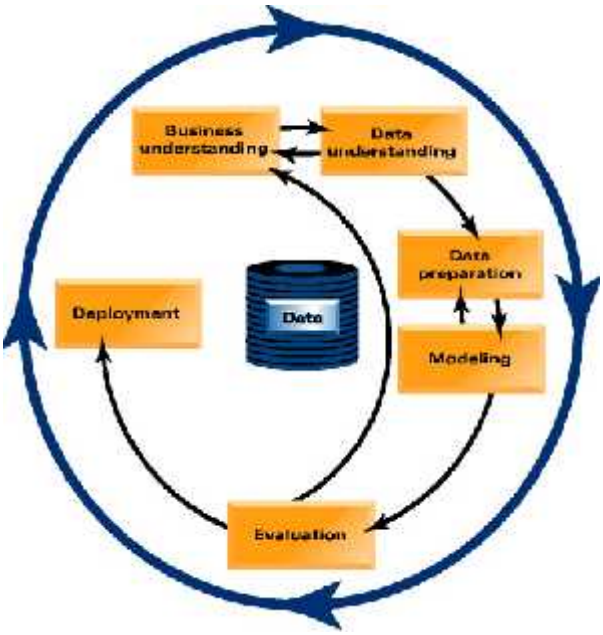


Figure 6: A Visual Guide to CRISP-DM

Source : CRISP-DM 1.0 available from http://www.sv-europe.com/crisp-dm methodology.html)

Business understanding: This phase focuses on understanding the projectobjectives and requirements from a business perspective and convert it to a machine learning problem.

Several literatures were reviewed to assess machine learning technology, both concepts and techniques, and researches in this field and also to gain an insight of what was required. Various books, journals, magazines, and papers from the internet pertaining to the subject matter of machine learning were reviewed to understand the potential applicability of machine learning to classify verbal autopsy data set with a view of predicting cause of death.

Data understanding: The stage is about data collection and proceedswith activities in order to get familiar with the data, to identify data qualityproblems.The major goal here is to understand the data sources, data parameters and quality of data. The potential source of data that was used to undertake this research was the from the IHME database and as a result, one main source of verbal autopsy data was identified which was in a CSV format.

Data preparation and Transformation: This covers all the activities required to construct the final dataset from the initial IHME verbal autopsy raw data. The data derived from the

questionnaire was transformed into the proper format in order to be analyzed based on selected algorithms. The data was represented by numbers and stored in the form of a CSV file.WEKA toolkit was used to preprocess the data software for specific machine learning These files are prepared and converted to (arff) format compatible with the WEKA data mining toolkit (Abraham, et al., 2011; Bharat & Manan, 2012) which is used in building the model.The activities that were carried out include attribute selection whereby non-relevant attributes such site of data collection, details of the interviewee during the verbal autopsy process and other socio-demographic information about the deceased were removed, re-sampling, replacing missing values using the arithmetic mean was also applied to the data and formatting data in order to apply specific machine learning tasks. This was applied using replace missing values feature under unsupervised filter option available in Weka toolkit

As mentioned previously, the data has been divided into two datasets. The first one includes the data for the training and the the second includes data for the testing of the model.Each dataset has two arff files containing its data, with the class attribute (performance). Each of these datasets was used in a separate training and test experiments respectively.

This is the stage in which the selected data were transformed into forms acceptable to Weka data mining software. The data file was saved in Comma Separated Value (CSV) file format in Microsoft excel and later converted to Attribute Relation File Format (ARFF) file inside Weka for easy use.

Modeling: To build a predictive model from the cleaned verbal autopsy data, WEKA tool was used and two classification algorithms J48 Decision Tree and Naïve Bayes were applied to classify cause of death from the verbal autopsy datasets.

Creation and test of the data classification model were conducted by WEKA program with the algorithms J48 and Naïve Bayes. The model was tested by means of 10 -fold cross-validation to find out the values of Correctly Classified, Precision, Recall and F-Measure.Then, the results of the tests were compared in terms of efficiency of each data classification technique

Evaluation: Although machine learning task reveal patterns and relationships, this by itself is not sufficient. Domain knowledge and machine learning expertise is required to interpret,

validate and identify interesting and significant patterns. The machine learning team in corporate domain expertise and data mining expertise in evaluating and visualizing models in order to identify interesting patterns and trends.

In this phase, the researcher evaluated the performance of J48 decision tree and Naïve Bayes by means of confusion matrix as well as ROC analysis and also discussion on the generated rules or models with domain experts from the health sector. The results of this particular machine tasks were visualized and interpreted.

Deployment: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

3.2 Algorithms Considered and Justification

The Decision Tree learning algorithm and naïve bayes were considered in this project due to their popularity and usefulness in solving data mining classification problems as exposed from the literature search (section 2.6 and 2.7). Moreover, because they have all been applied to various datasets and disparity, results were obtained (Gopala and Bharath 2013), (Qasem et al. 2014).And finally, this is due to the fact that they differ in their characteristics (Murray et al. 2014). These reasons make it possible to have a true representation of various techniques and to ensure whatever results that will be obtained from the experiments will be accurate and authentic reflection of what is established in the literature. A good comparison can then be made between the outcome of this project and what is said in the literature.

3.3 Overview of WEKA Machine learning tool and justification

WEKA is an acronym for Waikato Environment for Knowledge Analysis and the workbench is a collection of state-of-the-art machine learning algorithms for solving real-world problems(E. & V. R. , 2013).The tool contains general purpose environment tools for data pre-processing, regression, classification, association rules, clustering, feature selection and visualization. It contains 41 different algorithms for classification and numeric prediction (Srivastava 2014)

It provides a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset and the data format for WEKA is ARFF

The reason why this tool is specially selected is that it is the only toolkit that has gained widespread adoption and survived for an extended period of time and it is open source software as well it offers many powerful features (sometimes not found in commercial data mining software), Weka also became one of the favorite vehicles for data mining and machine learning research and helped to advance it by making many powerful features available to all (Sushilkumar, 2015)

Figure 7: Weka GUI Application Main Window

3.3.1 Justification

This workbench was the chosen tool to build the classifiers primarily because it was a known entity and is well established and well regarded both in academia and the commercial arena across the world (Srivastava, 2014). Finally, it supports process models of data mining including CRISP-DM which is the chosen methodology for this project (Bharat & Manan, 2012)(Pete, et al., 2010)

For the applicability issue, the WEKA toolkit has achieved the highest applicability followed by Orange, Tanagra, and KNIME respectively. The toolkit has achieved the highest improvement in classification performance; when moving from the percentage split test mode to the cross validation test mode, followed by Orange, KNIME and finally Tanagra respectively (Appendix 1)

Importantly, WEKA can handle the problem of the multiclass data set, which is not the case in other data mining and machine learnig tools. Moreover, applicability (run specific algorithm on a selected tool) is highest in WEKA. Furthermore, WEKA is able to run 6 selected classifiers using all data sets(Qasem, et al., 2014).

One way of using WEKA is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances and the third way is to apply several different learners and compare their performance in order to choose one for prediction. The learning methods are called classifiers, and in the interactive WEKA interface you select the one you want from a menu lists. Many classifiers have tunable parameters, which you access through a property sheet or object editor. A common evaluation module is used to measure the performance of all classifiers (Beyene 2011)

3.4 Description and Exploration of the Data Sets

The verbal autopsy dataset used in this study was obtained from the Institute of Health Metrics and Evaluation (IHME) which was collected as part of the Population Health Metrics Research Consortium (PHMRC) project. The files contain verbal autopsies (VAs) that were collected at six sites in four countries (India, Mexico, Tanzania, and the Philippines) using a standardized VA questionnaire developed by the WHO. The original data set contains a total of 7841 instances and 946(945 continuous input attribute and 1 nominal class label target attribute which is the known cause of death from the data set).The data was in a CSV format and a gold standard cause of death diagnosis is included within the CSV fileand there are some special values like “1” and “0” meaning yes and no respectively. The VA questionnaire was used to collect information about the symptoms of the deceased, demographic characteristics and other potentially contributing characteristics. Other components of the data e.g. the signs and symptoms that led to the death, history of any ailments and care seeking and treatments of the deceased were included in the dataset.

3.5 Feature Value Representation

The data consist of the closed part which uses a binary approach to represents feature occurrence of a disease symptom in a verbal autopsy as‘1’and non- occurrence of such a disease symptom as‘0’.The open narrative uses the frequency counts of certain words or phrases in the narrative which suggest weights based on frequency counts of either the

feature or the documents containing the feature. The basic assumptions here are that the importance of a feature is based on its frequency of occurrence in a given document, and a count of documents of which that feature occurs (Abraham, et al., 2011)

A cause list was constructed based on the WHO Global Burden of Disease (GBD) estimates of the leading causes of death, potential to identify unique signs and symptoms, and the likely existence of sufficient medical technology to ascertain gold standard cases (Danso, et al., 2011)

The individual verbal autopsys' are matched with "gold standard" diagnoses of underlying known causes of death, which were established from medical records using stringent diagnostic criteria, including laboratory, pathology and medical imaging findings. All "open narrative" portions of the verbal autopsy were parsed for individual words or stems, which are included as variables in the final dataset, to remove any potentially identifying information in that portion of the interview. Variables that were analyzed as "health care experience" in past research are identified in the codebook (Murray et al. 2011)

3.6 Preprocessing and Feature selection

The original data set being a real world data included noisy, missing and inconsistent data. Many instances had missing attribute values.

Data preprocessing improved the quality of the data and facilitated efficient machine learning.Before the experiment,data suitable to next operation was prepared as follows:-

- Delete or replace missing values;
- Delete redundant properties (columns);
- Data Transformation;
- Export data to a required format from .csv format to .arff format

The most common method of filling the attributes quickly and without too much computation is to replace all the missing values with the arithmetic mean or the mode with respect to that attribute.In this project, this was handled using WEKA tool filter named replace missing values.This filter replaces the missing attribute values by means and modes for numeric and nominal attributes respectively. This filter was used for J48 decision tree classification algorithm which needs fully filled dataset. Missing values for numeric attributes were replaced with the global mean of each numeric attribute and missing values for nominal

attributes were replaced with the global mode of each nominal attribute. This filter replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data. It handles both numeric and nominal attributes.The less sensitive or irrelevant attributes like the age,date of birth,level of education,sex,date of death of the deceased were removed since they have no value in classifying cause of death. So the number of attributes were reduced to 34.

The original and modifed formats of data set are shown in Table 1 and Table 2 below:

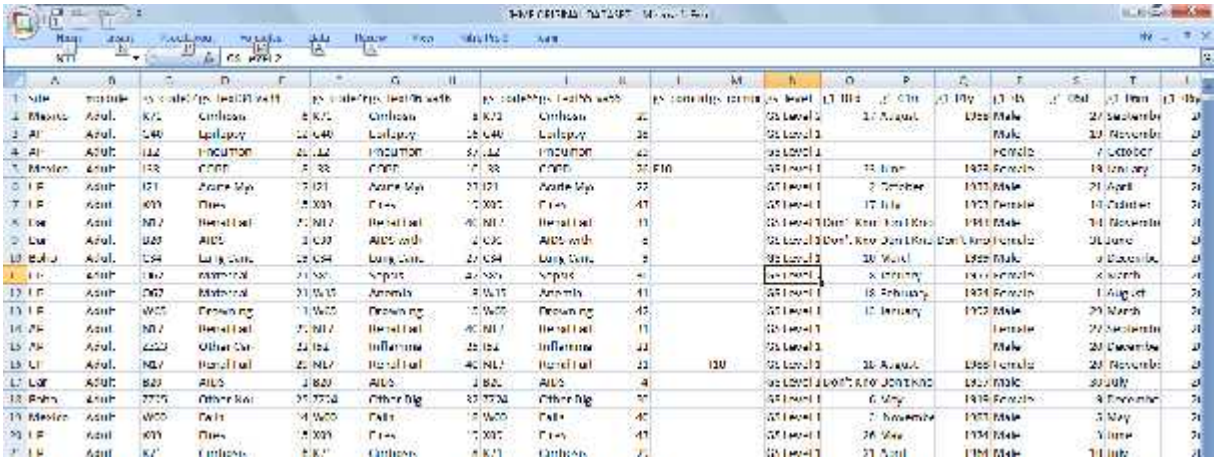


Table 1: Sample original .CSV data set

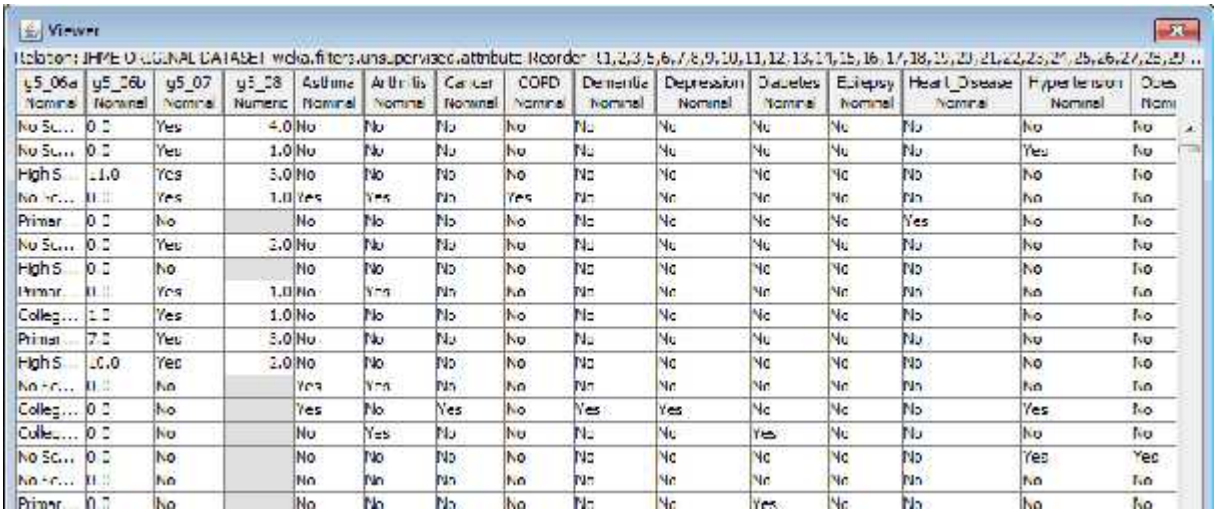
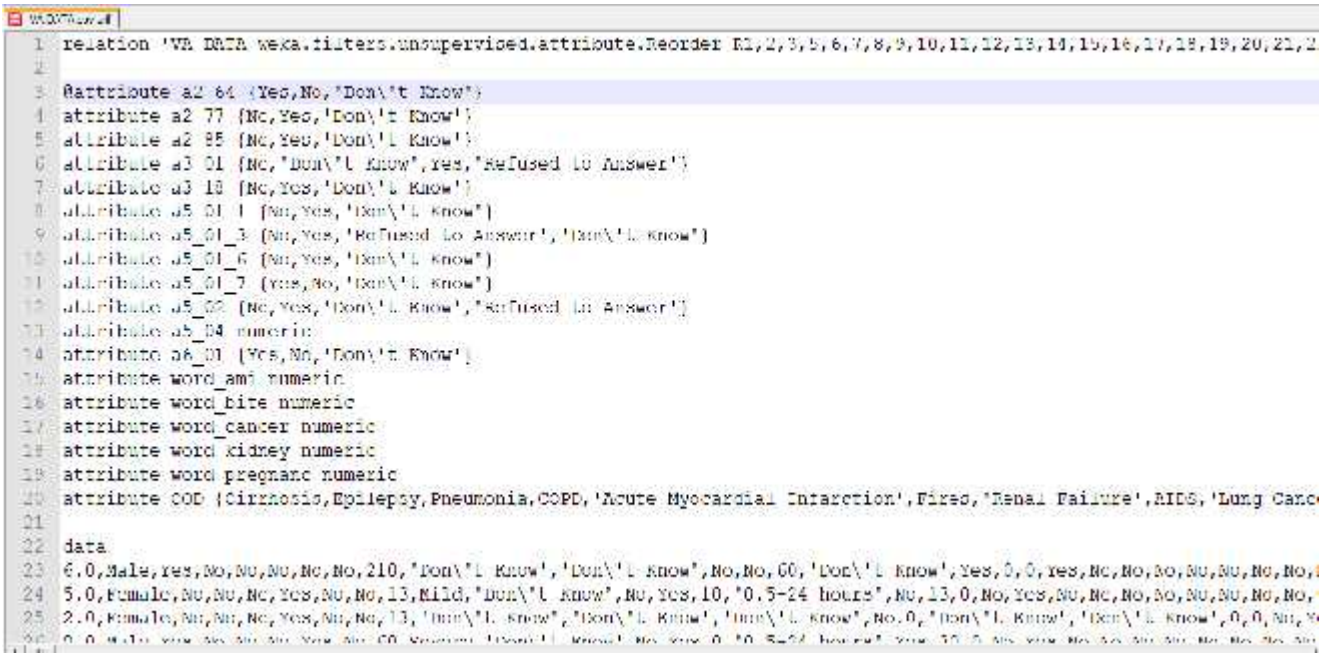
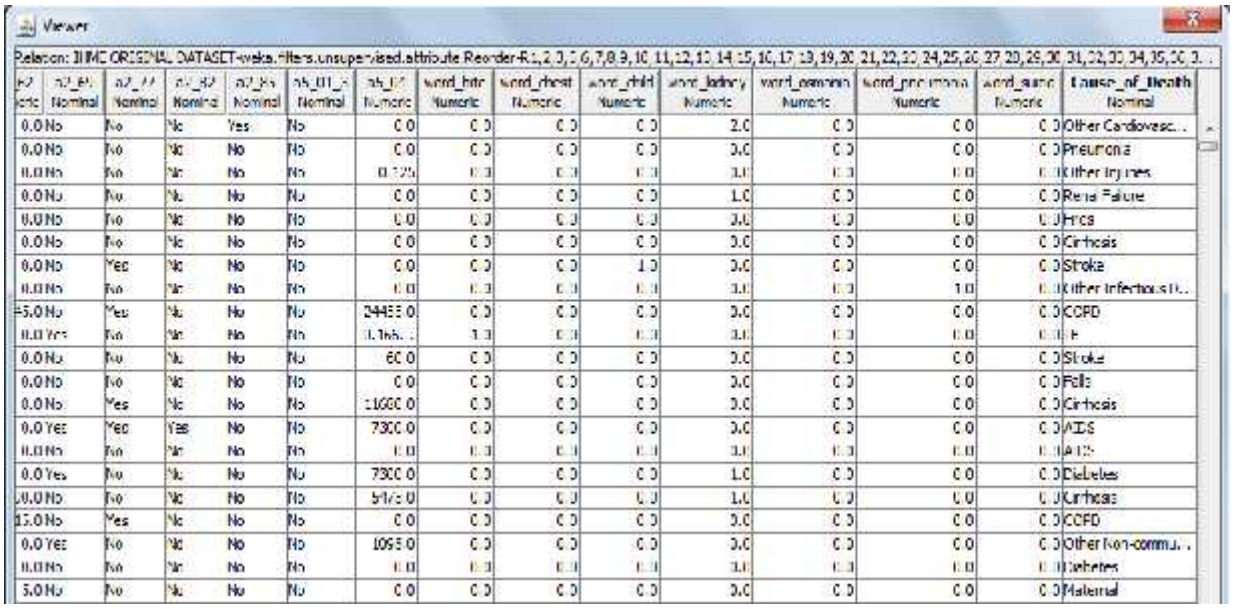


Table 2: Sample data converted to .ARRF file using WEKA Arff viewer tool



```
1  relation 'VA DATA weka.filters.unsupervised.attribute.Reorder R1,2,3,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,2
2
3  @attribute a2_64 {Yes,No,'Don\'t Know'}
4  attribute a2_77 {No,Yes,'Don\'t Know'}
5  attribute a2_85 {No,Yes,'Don\'t Know'}
6  attribute a3_01 {No,'Don\'t Know',Yes,'Refused to Answer'}
7  attribute a3_18 {No,Yes,'Don\'t Know'}
8  attribute a5_01_1 {No,Yes,'Don\'t Know'}
9  attribute a5_01_3 {No,Yes,'Refused to Answer','Don\'t Know'}
10 attribute a5_01_6 {No,Yes,'Don\'t Know'}
11 attribute a5_01_7 {Yes,No,'Don\'t Know'}
12 attribute a5_02 {No,Yes,'Don\'t Know','Refused to Answer'}
13 attribute a5_04 numeric
14 attribute a6_01 {Yes,No,'Don\'t Know'}
15 attribute word_ami numeric
16 attribute word_bite numeric
17 attribute word_cancer numeric
18 attribute word_kidney numeric
19 attribute word_pregnanc numeric
20 attribute COD {Cirrhosis,Epilepsy,Pneumonia,COPD,'Acute Myocardial Infarction',Fires,'Renal Failure',AIDS,'Lung Canc
21
22 data
23 6.0,Male,Yes,No,No,No,No,No,210,'Don\'t Know','Don\'t Know',No,No,60,'Don\'t Know',Yes,0,0,Yes,No,No,No,No,No,No,No,
24 5.0,Female,No,No,No,Yes,No,No,13,Mild,'Don\'t Know',No,Yes,10,'0.5-24 hours',No,13,0,No,Yes,No,No,No,No,No,No,No,
25 2.0,Female,No,No,No,Yes,No,No,13,'Don\'t Know','Don\'t Know','Don\'t Know',No,0,'Don\'t Know','Don\'t Know',0,0,No,Y
```

Figure 8: Sample data in ARFF



Viewer

Relation: BMC ORIGINAL DATASET-weka.filters.unsupervised.attribute.Reorder-R1,2,3,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,3...

a2_64 Nominal	a2_77 Nominal	a2_85 Nominal	a3_01 Nominal	a5_01_1 Nominal	a5_02 Numeric	word_bite Numeric	word_chest Numeric	word_child Numeric	word_kidney Numeric	word_pneumon Numeric	word_pneumonia Numeric	word_pregnanc Numeric	Cause_of_Death Nominal
0.0 No	No	No	Yes	No	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	Other Cardiovasc...
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Pneumonia
0.0 No	No	No	No	No	0.125	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Other Injuries
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	Renal Failure
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Fires
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Cirrhosis
0.0 No	Yes	No	No	No	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	Stroke
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	1.0	0.0	Other Infectious D...
65.0 No	Yes	No	No	No	24455.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	COPD
0.0 Yes	No	No	No	No	0.166...	1.0	0.0	0.0	3.0	0.0	0.0	0.0	F
0.0 No	No	No	No	No	60.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Stroke
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Falls
0.0 No	Yes	No	No	No	11680.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Cirrhosis
0.0 Yes	Yes	Yes	No	No	7300.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	AIDS
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	AIDS
0.0 Yes	No	No	No	No	7300.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	Diabetes
0.0 No	No	No	No	No	5475.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	Cirrhosis
15.0 No	Yes	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	COPD
0.0 Yes	No	No	No	No	1095.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Other Non-commu...
0.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Diabetes
5.0 No	No	No	No	No	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	Maternal

Table 3: Sample Preprocessed datasets with missing values replaced

3.7 Training and Test data

For the purpose of this study the dataset has been split into two parts: some has been used for training and some for testing. Two-thirds (75%) of it has been used for training and one-third(25%) of it for testing For any classification task in machine learning, it's really important that the training data is different from the test data The dataset contains good mix of attributes continuous, nominal with small numbers of values, and nominal with larger

numbers of values. The tens cross-validation method is used for testing the accuracy of the classification of the selected classification methods.

A ten-fold cross-validation method was used in this experiment. In ten folds cross-validation, a data set is equally divided into 10 folds (partitions) with the same distribution. In each test 9 folds of data are used for training and one fold is for testing (unseen data set). The test procedure is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

3.8 Data Analysis

As mentioned earlier in the report there exists so many free machines learning tools that are available in the market today e.g.Scikit-learn(Python),Rapid Miner,R, and ELKI.However,WEKA was chosen as a better tool to build the classifiers primarily because it is a landmark system in the history of machine learning. The data obtained from IHME was therefore passed through WEKA toolkit where patterrns were discovered and were helpful in decision making.

3.9 Overall Architecture of the Proposed Model

The overall design of the proposed model is given in figure.9 below and each of these components is addressed in the following sections briefly.

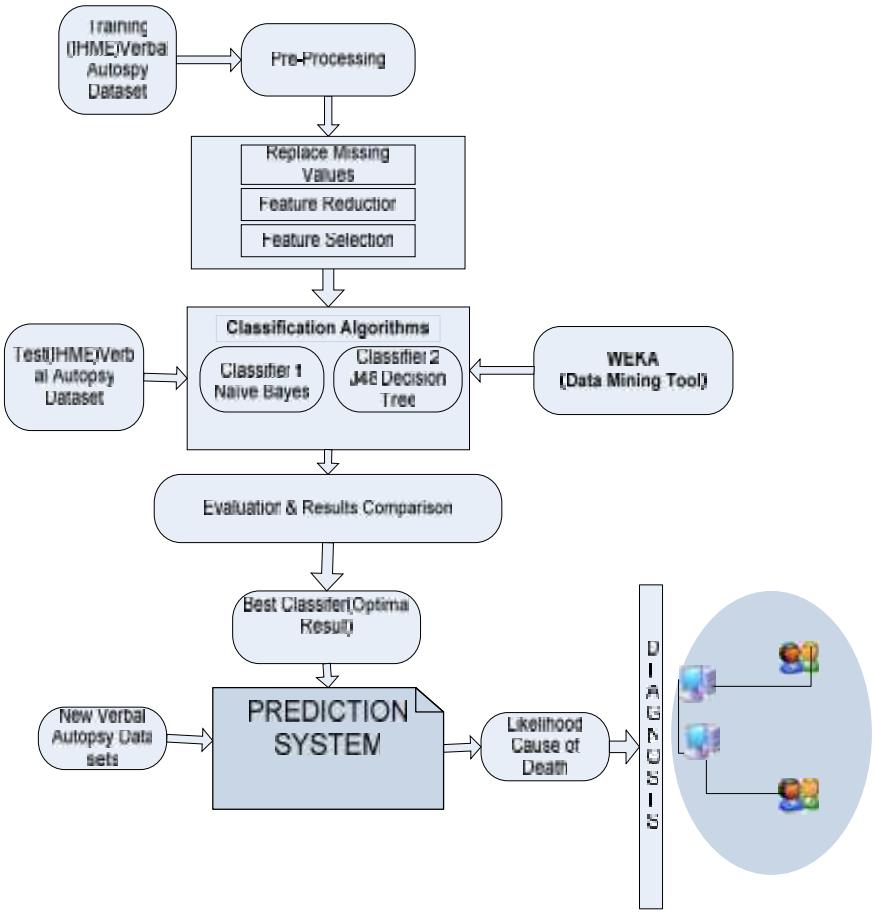


Figure 9: The Overall Architecture of the proposed VA classification system

Data Acquisition component-The component is responsible for storing the verbal autopsy data, gathered from different data sources in a data warehouse.

Data Preprocessing component-The component is responsible for cleaning the verbal autopsy data set.The preprocessing activities involve replacing missing values,feature selection and reduction

Model building and comparison component-The component responsible for obtaining knowledge about the cause of death, through appropriatesclassification algorithms such as decision trees and naive bayes and compare the two algorithms

Prediction System component-The prediction system component responsible for mapping the pattern in the rules generated with the newverbal autopsy data to predict likely cause of death

CHAPTER 4

DESIGN OF EXPERIMENTS,RESULTS AND ANALYSIS

4.1 Overview

In this study, a series of classification experiments were set up focusing on two supervised learning algorithms which are Naïve Bayes and J48 Decision Tree.The task is to classify verbal autopsy data and predict cause of death based on the given symptoms and other information from the Institute for Health Metrics and Evaluation (IHME) data sets.This was done to evaluate the selected classification algorithms using the given datasets based on some evaluation metrics.

4.2 Experimental Setup

The CSV(comma separated values) format dataset was imported into WEKA using an import tool ArffViewer available in WEKA so that it could be converted to ARFF (atribute relation file format) file format to use it with WEKA software.

a5_01 Nominal	a3_18 Nominal	a5_01_1 Nominal	a5_01_3 Nominal	a5_01_6 Nominal	a5_01_7 Nominal	a5_02 Nominal	a5_04 Numeric	a6_01 Nominal	word_pain Numeric	word_bite Numeric	word_cancer Numeric	word_kidney Numeric	word_pregnant Numeric	Cause of Death Nominal
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	1.0	0.0	0.0	Leukemia/Lymph ...
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Stroke
no	Yes	No	No	No	no	no	11.0	Yes	0.0	0.0	1.0	0.0	0.0	Maternal
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	AIDS
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	1.0	0.0	0.0	Other/Unk comm...
no	no	No	No	No	Yes	no	180.0	Yes	0.0	0.0	0.0	0.0	0.0	Kidney failure
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Malaria
no	no	Yes	No	No	no	no	2.0	Yes	0.0	0.0	0.0	0.0	0.0	Road traffic
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	AIDS
no	no	Yes	No	No	no	no	0.0	Yes	0.0	0.0	1.0	0.0	0.0	Stomach Cancer
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	AIDS
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	AIDS
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Leukemia/Lymph ...
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Asthma
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	COPD
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	AIDS
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	1.0	0.0	0.0	Prematurity
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	1.0	0.0	Pneumonia
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Stroke
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Maternal
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Diarrhea/Dysent ...
no	no	No	No	No	Yes	no	0.166...	Yes	0.0	0.0	0.0	1.0	0.0	Cirrhosis
no	no	No	No	No	no	no	0.0	Yes	0.0	0.0	0.0	0.0	0.0	Diabetes

Table 4: Sample verbal autopsy data set used for training

4.2.1 Model Building

The model building supported in this study is a classification in the search for the perfect model. The population for which a model is built is further divided into two sets: training and testing .The ratio of the sample population is set at approximately 75%: 25%: with the motivation to avoid occurrence of over-fitting and thus increase model accuracy and applicability in the performance dataset.

4.2.2 Modeling Techniques and Tools Used

The machine learning predictive model considered in this study was based on supervised learning (classification) technique. The softwaretool used was WEKA an open-source and free software usedfor knowledge analysis and downloadable from the internet and used under the GNU license. WEKA implements different machine learning algorithms. The presentation of results and the development of the prototype were done using JAVA while the data is stored in Mysql database.

4.3 Performance Evaluation for Predictive Model

4.3.1 Prototype Results

The performance of the classifiers was measured in terms of different standard metrics like accuracy, precision, recall, 10-fold validation, and ROC curve and time complexity.

Sensitivity and specificity are statistical measures of the performance which were also employed in the project. Sensitivity is often also known as the recall rate and measures the proportion of actual positives which are correctly identified as such; the percentage of people who are correctly identified as having a disease. Specificity measures the proportion of negatives which are correctly identified the percentage of well people who are correctly identified as not having the disease (D.Flaxman & T.Green, 2010)

Predictive models are evaluated in terms of correctness, often referred to as performance, and applicability. The performance measures are almost always geared towards the evaluation of an instance of a model type, and are almost always realization method independent. Applicability measures also contain measures that apply to the model type itself, pertaining to the need of models to be evaluated in terms of their context (Beyene, 2011)

Once a predictive model was developed using the verbal autopsy dataset, the model was checked as to how it will perform for the future data which, it has not seen during the model building process. The researcher used two differentmachine learning classifiers, techniques and tool to build the predictive model and in order to evaluate the performance of the model, confusion matrix and ROC analysis were used.

4.3.2 Model Validation using Confusion Matrix

To validate the results of the model, a confusion matrix was used. A confusion matrix is an n-dimensional square matrix, where n is the number of distinct target value.It is used to represent the test result of a prediction model. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class as indicated in the figure below. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another). A confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models (Badgerati, 2010)

Source :(Badgerati, 2010)

Figure 10: Confusion Matrix

As shown above, a confusion matrix table of size two by two, the following measures can be calculated to measure the predicted cause of death from the verbal autopsy IHME dataset's accuracy of the model, True Positive Rate, False Positive Rate, Accuracy, Precision, Recall and ROC curve.

Moreover, the confusion matrix is a useful tool for analyzing how well the researcher's classifier can recognize tuples of different classes. The following procedures and rules were implemented to confirm the model performance evaluation for the results of the model to classify cause of death from the verbal autopsy data sets.

In building a classification model, the confusion matrix provides a quick understanding of model accuracy and the types of errors the model makes when scoring records. It is the result of a test task for classification models

The Accuracy of a classifier is projected by dividing the total correctly classified positives and negatives instances by the total number of samples.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

$$\text{sensitivity} \quad \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}$$

To explain in simple terms; the "True Positive Rate "is the cases of disease where the classifier shows that they have the disease and they actually do. The "False Positive Rate "is the cases of disease where the classifier shows that they have the disease when actually they do not. The below table below explains the terms succinctly.

		Actual Disease	
		Disease Present	Disease Absent
Test	Positive	Disease Present + Positive result – True Positive	Disease absent + Positive result – False Positive
Result	Negative	Condition present + Negative result – False (p.value) Negative	Condition absent + Negative result – True (accurate) Negative

Table 5: Explaining Disease Result Outcomes

Source: http freedictionary.com/sensitivity

Other classifier measurements that are examined are "Precision" which is the number of true positives correctly labeled as belonging to the class. The equation below makes this a simple concept to understand.

$$\text{Precision} = \frac{tp}{tp + fp}$$

"Recall" which is the total number of true positives divided by the total number of elements that actually belong to the positive class ie.the sum of true positives and false negatives which were not labeled as belonging to the positive class but should have been. In this context Recall also refers to as the true positive rate. Therefore relating back to the above the true negative rate is also known as the "specificity" and false negative rate is known as the "sensitivity"

$$\text{Recall} = \frac{tp}{tp + fn}$$

Before the results were discussed it was recognized that due to small sample size the validity of the results in terms of offering definite and exacting conclusions are problematic. A larger sample would have significantly increased the statistical validity of the findings. However, the results despite this provide an interesting proof-of-concept and again bring out the computational issues and challenges associated with the verbal autopsy process.

4.4 Basic Classification Results and Predictive model using WEKA

Experiments were conducted under the framework of Weka to study the various kinds of classification algorithms on the verbal autopsy datasets. The experiments compare various results in terms of classification measured by percentage accuracy of no. of correctly classified instances.The environmental variables are same for each algorithm and dataset. The algorithms are compared by using various parameters like tprate, fprate, precision, recall, time taken etc. TP rate is the true positive rate and the FP rate is the false alarming rate. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database.

The algorithms that were used for the experiments are Naive Bayes and J 48 Decision Tree and they were selected due to their popularity and usefulness in solving classification problems as highlihted in the literature and also because they have all been applied to many dataset and disparity according to Putten *et al* (2000) as cited in (Sam1)These algorithms produces a decision tree and Naive Bayes data structures respectively of the correctly and incorrectly classified results.The experiments were based on the IHME data comprising of the adult verbal autopsies on deaths with gold standard diagnoses that were collected 7,836 adults(Murray, et al., 2011)

In order to train the classifier of verbal autopsy data, 75% of the dataset were used for training and the rest 25% for testing. For creating a cause of death predictive model J48 and Naive Bayes algorithm are used. To evaluate the performance of the model; 10 cross validation was used due to its relative low bias and variations. This means the data were randomly partitioned equally into ten parts. The learning scheme is trained ten times using nine-tenths of the total data and the remaining is used for testing. Therefore the learning procedure is executed a total of 10 times on different training and testing sets. The experiment was done

using WEKA data mining tool version 3.6.11. The tool takes the data in .arff format in a single table, before that the prepared data in excel format is changed to CSV

The results aims to understand the issues,evaluation metrics and successes of using data and expert-driven algorithms and to understand how effective a computational hybrid approach from supervised approaches by using an ensemble model would replace the physician's certified verbal autopsy in predicting the specific cause of death for those deaths whose cause are not medically certified

4.5 Evaluation

The above stated supervised learning algorithms were implemented and evaluated using WEKA tool kit on the selected verbal autopy data.As a rule of thumb accuracy of classification is used as the metric for deciding the best suited classifier. According to Patrick and Sampson,cited in(Jeyarani, et al., 2013), accuracy is determined as the ratio of instances correctly classified during testing to the total number of instances tested.The accuracy of the classifers were evaluated through precision,recall and ROC analysis where appropropriate in the performance analysis.

4.6 Cross-validation

To evaluate the robustness of the classifiers in this project, the normal methodology is to perform cross validation on the classifier. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier (Witten and Frank, 2000). In ten-fold cross validation, the training set is equally divided into 10 different subsets. Nine out of ten of the training subsets are used to train the learner and the tenth subset is used as the test set. The procedure is repeated ten times, with a different subset being used as the test set. This can be implemented directly using weka toolkit under the test options

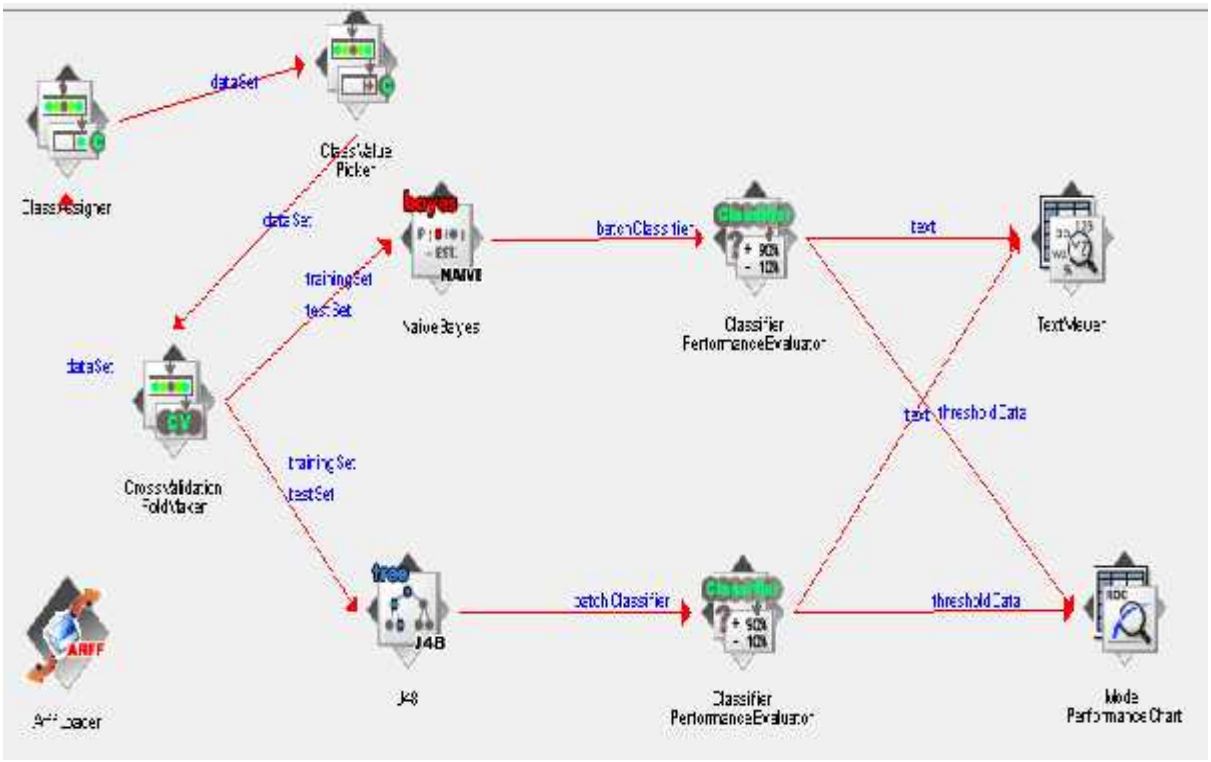


Figure 11: Models knowledge flow environment design of the model

4.7 Training data set

To produce the model a training data was used, a data set with known output values was used to build the model for both the J48 and Naive bayes classifiers. Then, whenever there is a new datapoint, with an unknown output value, the data is put through the model and produce our expected output. The models produced by the training sets are as below

```
Classifier output

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1892               37.1714 %
Incorrectly Classified Instances      3199               62.8286 %
Kappa statistic                          0.2978
Mean absolute error                      0.042
Root mean squared error                  0.1664
Relative absolute error                 74.7956 %
Root relative squared error             99.3449 %
Total Number of Instances             5091

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.302     0.111       0.514     0.302      0.381     0.796  Cirrhosis
                 0.383     0.208       0.319     0.383      0.352     0.857  Epilepsy
                 0.073     0.213       0.26      0.073      0.114     0.745  Pneumonia
                 0.214     0.107       0.307     0.214      0.252     0.801  CHF
                 0.532     0.216       0.349     0.532      0.422     0.891  Acute Myocardial Infarction
                 0.175     0.203       0.438     0.175      0.25      0.838  Fibro
```

Figure 12: Evaluation on the Training Set for the NBC

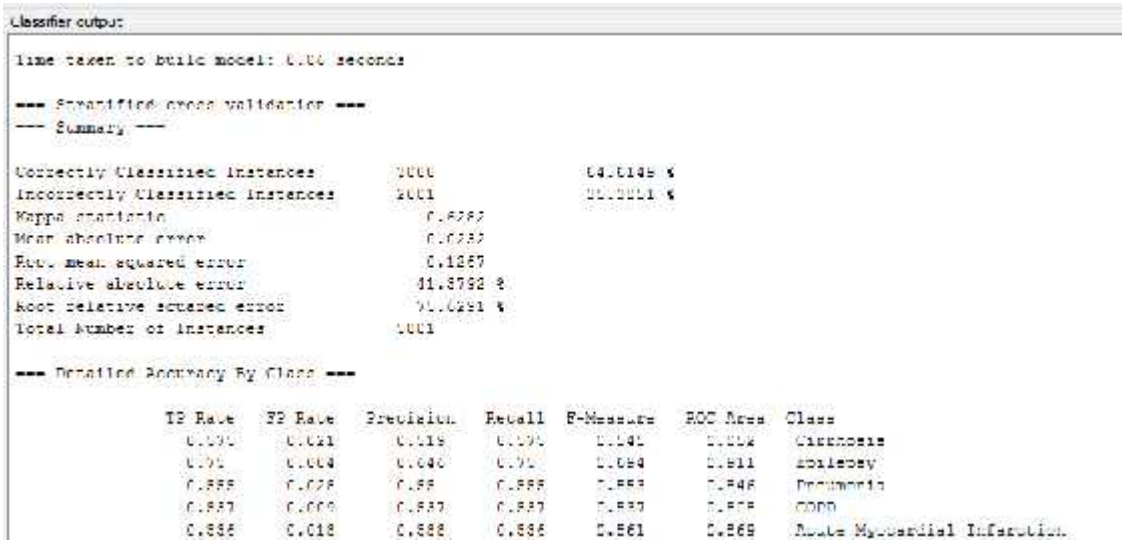


Figure 13: Evaluation on the Training Set for the J48 Classifier

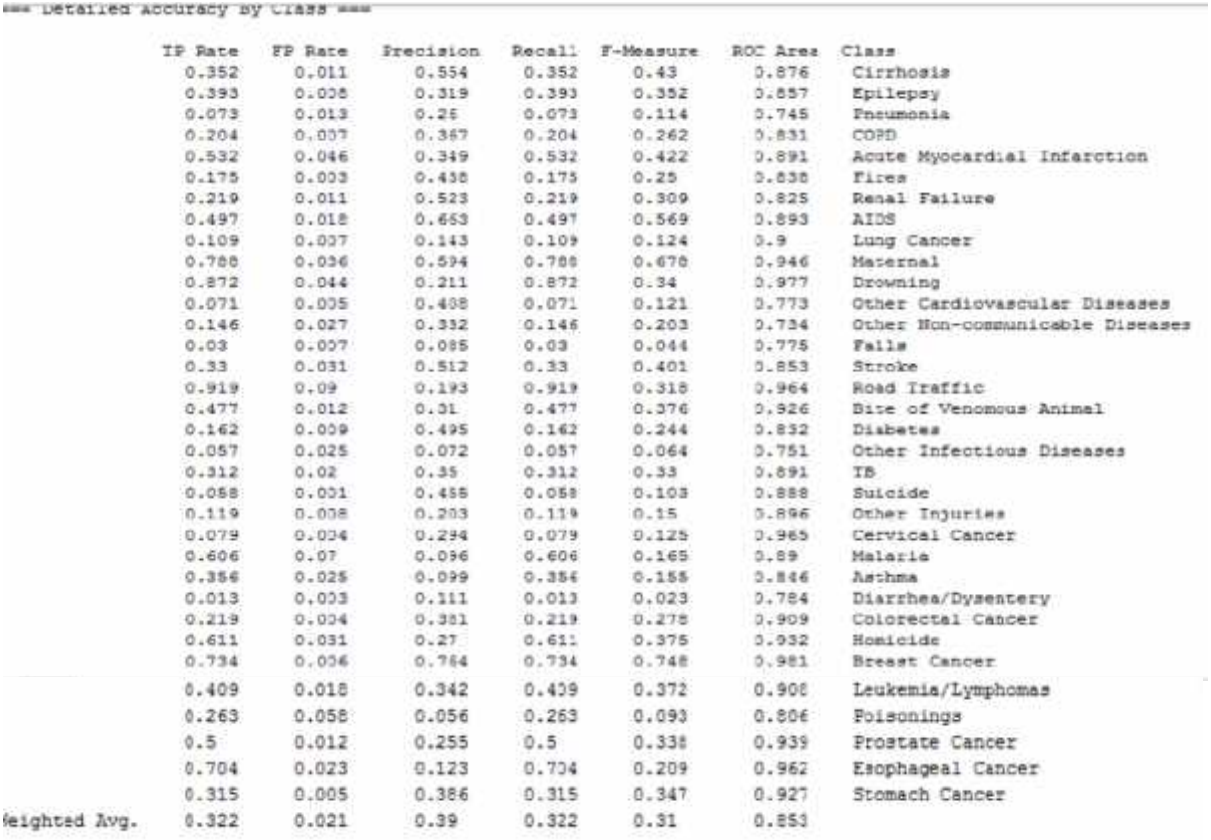
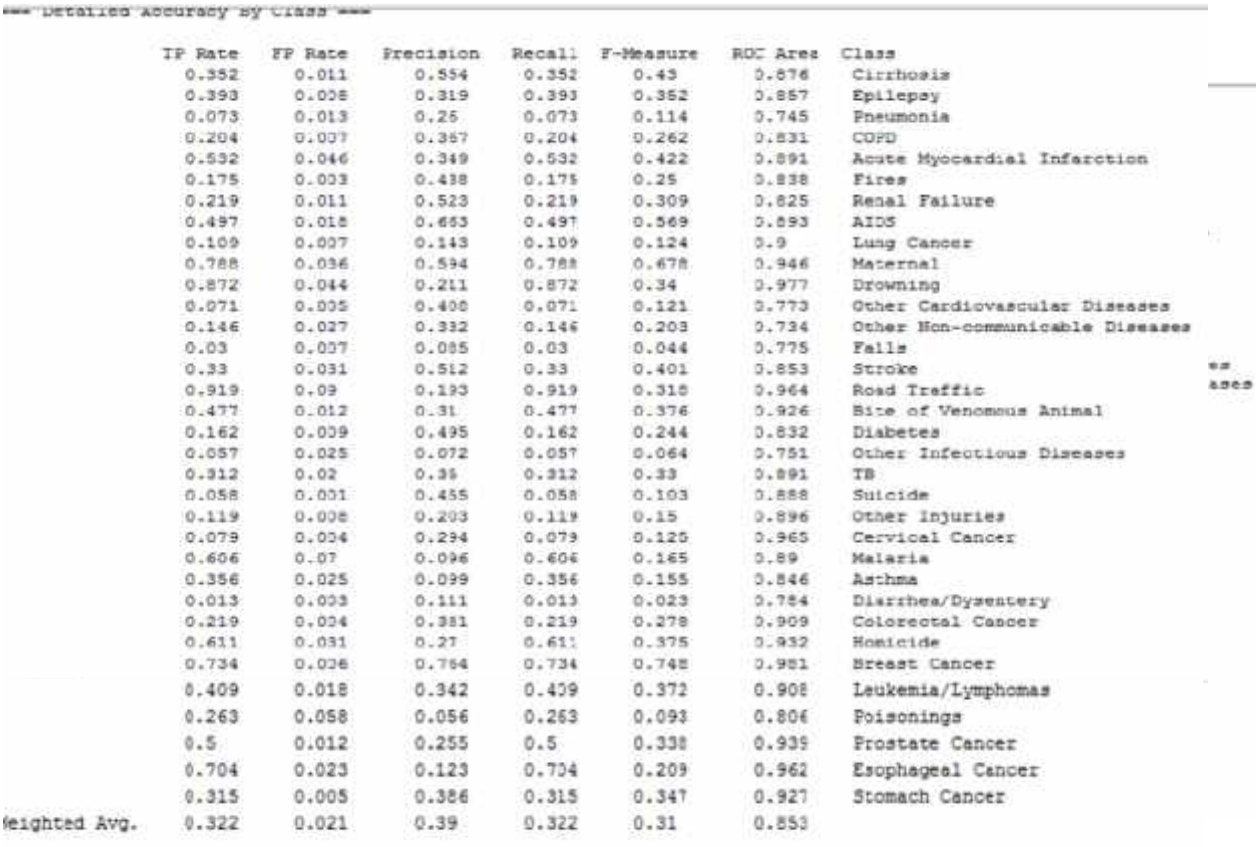


Figure 14: Detailed accuracy by class in a Naive Bayes Algorithm



=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.352	0.011	0.554	0.352	0.43	0.876	Cirrhosis
0.393	0.036	0.319	0.393	0.352	0.857	Epilepsy
0.073	0.013	0.25	0.073	0.114	0.745	Pneumonia
0.204	0.007	0.367	0.204	0.262	0.831	COPD
0.532	0.046	0.349	0.532	0.422	0.891	Acute Myocardial Infarction
0.175	0.003	0.438	0.175	0.25	0.838	Fires
0.219	0.011	0.523	0.219	0.309	0.825	Renal Failure
0.497	0.018	0.663	0.497	0.569	0.893	AIDS
0.109	0.007	0.143	0.109	0.124	0.9	Lung Cancer
0.788	0.036	0.594	0.788	0.678	0.946	Maternal
0.872	0.044	0.211	0.872	0.34	0.977	Drowning
0.071	0.005	0.400	0.071	0.121	0.773	Other Cardiovascular Diseases
0.146	0.027	0.332	0.146	0.203	0.734	Other Non-communicable Diseases
0.03	0.007	0.085	0.03	0.044	0.775	Falls
0.33	0.031	0.512	0.33	0.401	0.853	Stroke
0.919	0.09	0.193	0.919	0.318	0.964	Road Traffic
0.477	0.012	0.31	0.477	0.376	0.926	Bite of Venomous Animal
0.162	0.009	0.495	0.162	0.244	0.832	Diabetes
0.057	0.025	0.072	0.057	0.064	0.751	Other Infectious Diseases
0.312	0.02	0.36	0.312	0.33	0.891	TB
0.058	0.001	0.455	0.058	0.103	0.888	Suicide
0.119	0.008	0.203	0.119	0.15	0.896	Other Injuries
0.079	0.004	0.294	0.079	0.125	0.965	Cervical Cancer
0.606	0.07	0.096	0.606	0.165	0.89	Malaria
0.356	0.025	0.099	0.356	0.155	0.846	Asthma
0.013	0.003	0.111	0.013	0.023	0.784	Diarrhea/Dysentery
0.219	0.004	0.381	0.219	0.278	0.909	Colorectal Cancer
0.611	0.031	0.27	0.611	0.375	0.932	Homicide
0.734	0.026	0.764	0.734	0.748	0.981	Breast Cancer
0.409	0.018	0.342	0.409	0.372	0.908	Leukemia/Lymphomas
0.263	0.058	0.056	0.263	0.093	0.806	Poisonings
0.5	0.012	0.255	0.5	0.338	0.939	Prostate Cancer
0.704	0.023	0.123	0.704	0.209	0.962	Esophageal Cancer
0.315	0.005	0.386	0.315	0.347	0.927	Stomach Cancer
Weighted Avg. 0.322	0.021	0.39	0.322	0.31	0.853	

Figure 15: Detailed accuracy by class in a J 48 decision tree Algorithm

Measurement - Cross Validation	J48 Decision Tree	Naïve Bayes
Number of Attributes	34	34
Total Number of Instances	5881	5881
No: Correctly Classified Instances	3800	1892
No: Incorrectly Classified Instances	2081	3989
% Correctly Classified Instances	64.6%	32.2%
% Incorrectly Classified Instances	35.3%	67.8%
TP Rate Pneumonia	0.073	0.073
TP Rate Acute Myocardial Infarction	0.532	0.532
TP Rate Chronic Obstructive Pulmonary Disease	0.204	0.204
FP Rate Pneumonia	0.013	0.013
FP Rate Acute Myocardial Infarction	0.046	0.046
FP Rate Chronic Obstructive Pulmonary Disease	0.007	0.007
Precision Pneumonia	0.26	0.26
Precision Acute Myocardial Infarction	0.349	0.349
Precision Chronic Obstructive Pulmonary Disease	0.367	0.367

Recall Pneumonia	0.073	0.073
Recall Acute Myocardial Infarction	0.532	0.532
Recall Chronic Obstructive Pulmonary Disease	0.204	0.204

4.8 Interpretation of results of the training data set

The table above contains the results of efficiency analysis of each data classification technique, showing correctly classified instances and incorrectly classified instances. In addition, the table presents the values of Precision, Recall,True Positive rate and False Positive rate.

The J48 model classifies 3800 instances correctly with an accurate rate of 64.6 %, this indicates that the results obtained from training data are optimistic and can be relied on for future or new predictions.However the Naive Bayes classifies 1892 instances correctly translating to 32.2% for the correctly classified instances,this result informed the choice for the selection of the best classification algorithm which is J48 in this case.

4.8.1 Test data set

After the model was created testing was done to ensure that the accuracy of the model built does not decrease with the test set. Thisensures that the model will accurately predict future unknown values

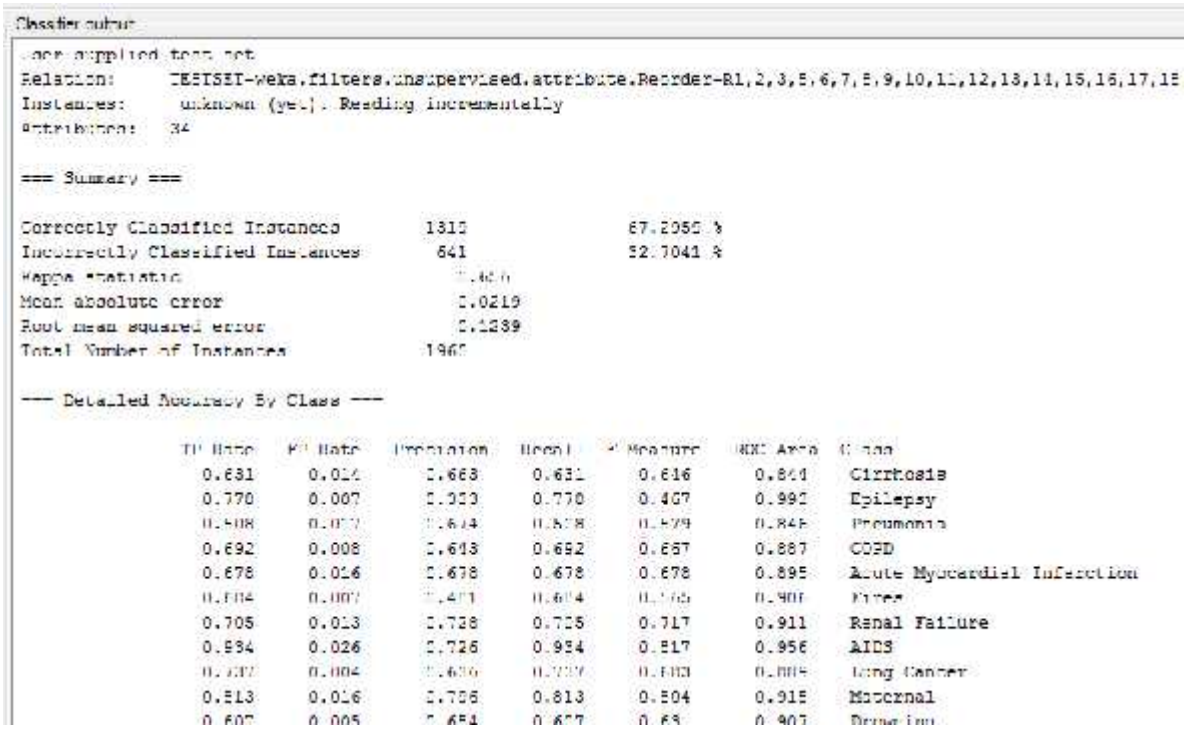
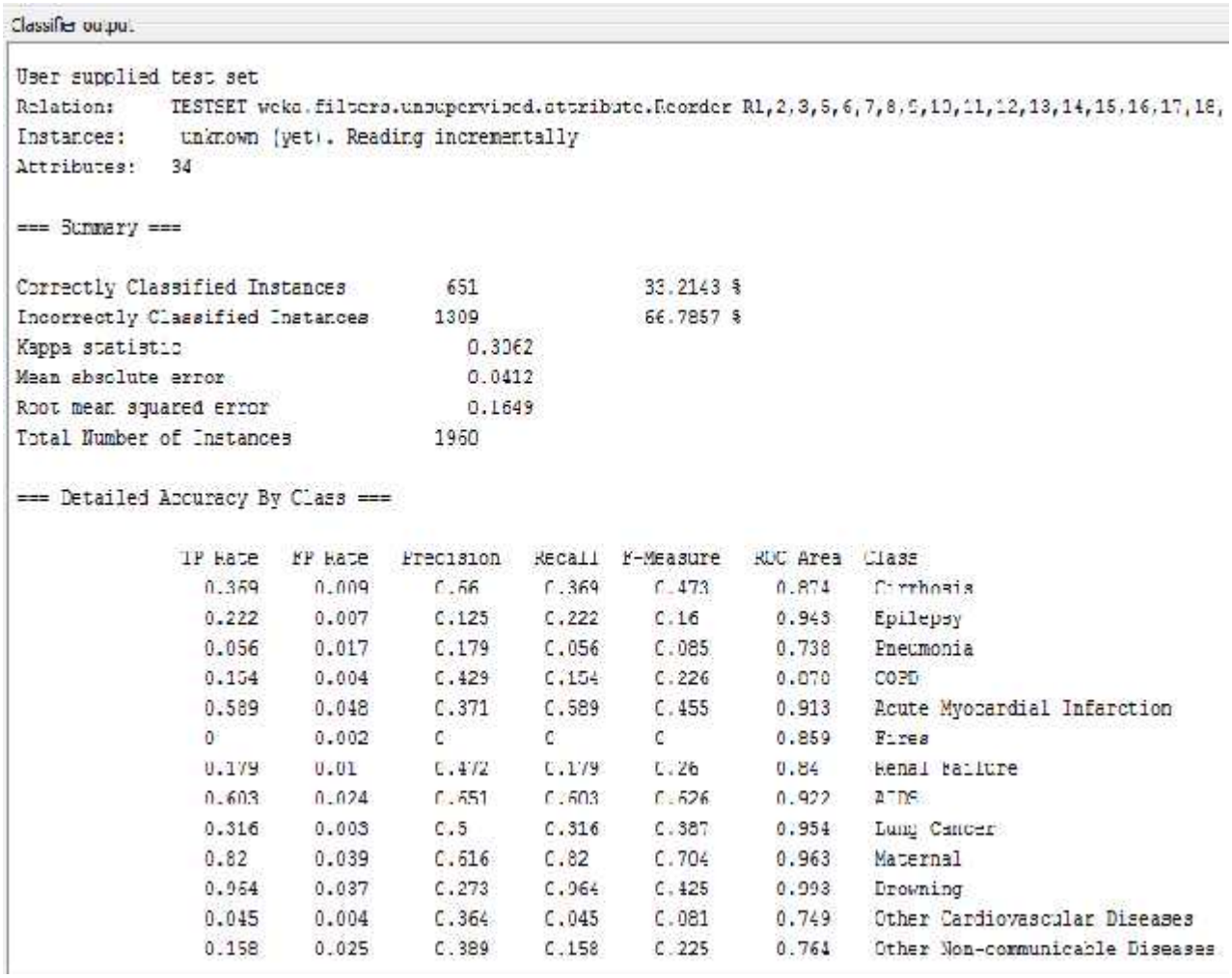


Figure 16: Evaluation on the user supplied test set for J48 classifier

4.8.2 Interpretation of results of the test data set

The model classifies 1319 instances correctly with an accurate rate of 67.3%, this indicates that the model will accurately predict future unknown values..The naive bayes classier however classifies 33.2% correctly,this is a very low accuracy which is below the threshold of any classification algorithm and cannot be relied upon as in the figure below.



Classifier output

```
User supplied test set
Relation:     TESTSET weka.filters.unsupervised.attribute.Reorder R1,2,3,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
Instances:    unknown (yet). Reading incrementally
Attributes:   34

=== Summary ===

Correctly Classified Instances         651               33.2143 %
Incorrectly Classified Instances      1309               66.7857 %
Kappa statistic                          0.3062
Mean absolute error                      0.0412
Root mean squared error                  0.1649
Total Number of Instances             1960

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.369     0.009       0.66     0.369     0.473      0.874    Cirrhosis
                 0.222     0.007       0.125    0.222     0.16       0.943    Epilepsy
                 0.056     0.017       0.179    0.056     0.085      0.738    Pneumonia
                 0.154     0.004       0.429    0.154     0.226      0.878    COPD
                 0.589     0.048       0.371    0.589     0.455      0.913    Acute Myocardial Infarction
                 0         0.002       0        0         0          0.859    Fires
                 0.179     0.01        0.472    0.179     0.26       0.84     Renal Failure
                 0.603     0.024       0.651    0.603     0.626      0.922    AIDS
                 0.316     0.003       0.5      0.316     0.387      0.954    Lung Cancer
                 0.82      0.039       0.616    0.82      0.704      0.963    Maternal
                 0.964     0.037       0.273    0.964     0.425      0.993    Drowning
                 0.045     0.004       0.364    0.045     0.081      0.749    Other Cardiovascular Diseases
                 0.158     0.025       0.389    0.158     0.225      0.764    Other Non-communicable Diseases
```

Figure 17: Evaluation on user supplied test set for NBC

4.9 Models performance.

The performance of models were evaluated using a Receiver Operating Characteristic curve (or ROC curve.) It is a plot of the true positive rate against the falsepositive rate for the different possible cutpoints of a diagnostictest. The closer the curve follows the left-hand border and thenthe top border of the ROC space, the more accurate the model.Based on the threshold curves used to measure the algorithmsemployed in this study, it is discovered that J48 performance is better than the naive bayes algorithm.

4.9.1 Comparison of learning algorithms

No single learning algorithm can uniformly outperform other algorithms over all datasets. Features of learning techniques are compared in Table 6 below from the models built.

SN	Parameters on a 10 fold cross validation	J48	Naïve Bayes
1	TP Rate	0.646	0.322
2	FP Rate	0.019	0.021
3	Precision	0.646	0.39
4	Recall	0.646	0.322
5	F-Measure	0.643	0.31
7	Execution Time	0.86	0.06

Table 6: Comparison of the final statistics of the learning algorithms

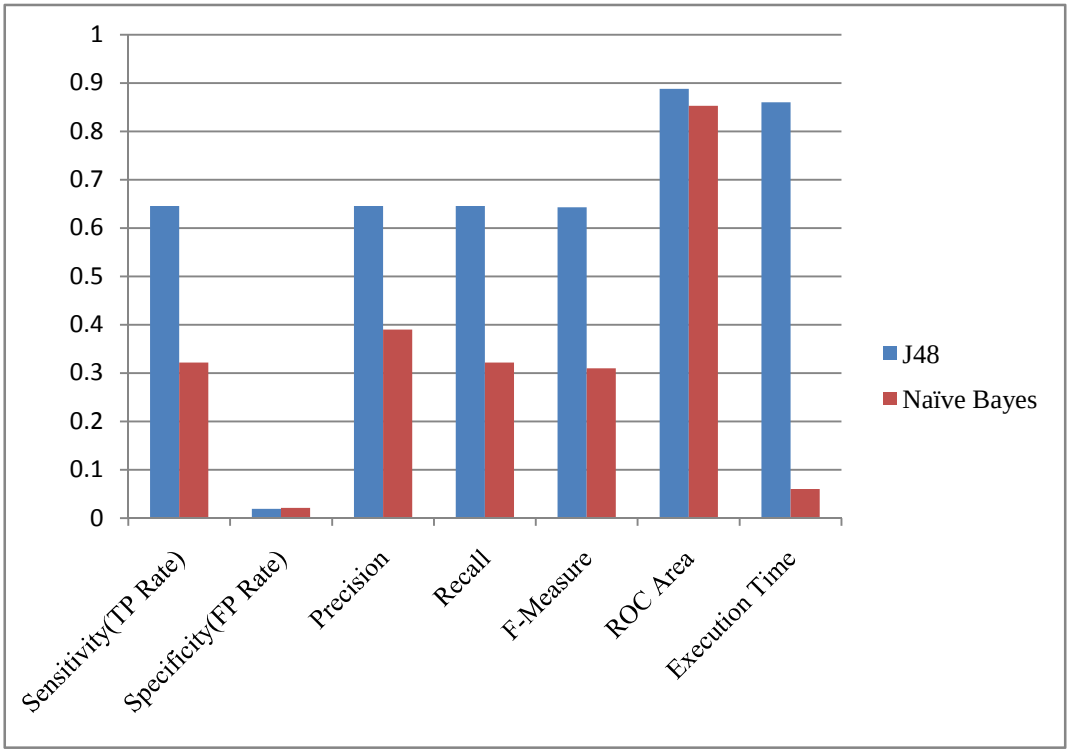


Figure 18: Graphical representation of the performance metrics for the classifiers

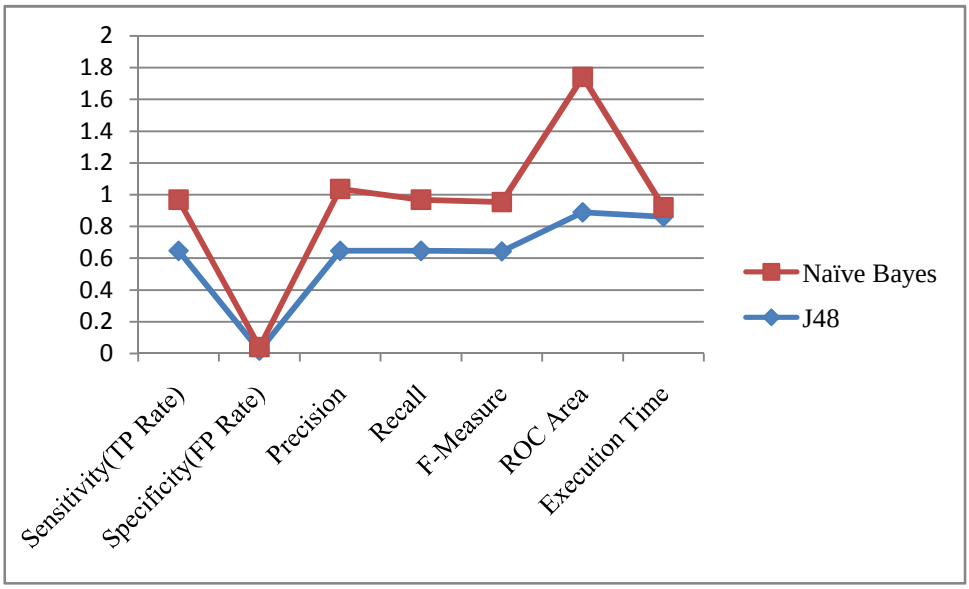


Figure 19 : A chart depicting the performance metrics for the classifiers

Performance Rate / **Classifier**	**Correctly Classified Instances**		**Incorrectly Classified Instances**	
	Training Set %	Test Set %	Training Set %	Test Set %
J48	64.6%	67.3%	35.4%	32.7%
Naïve Bayes	32.2%	33.2%	67.8%	67.8%

Table 7: Classified instances on the Verbal Autopsy Data Set

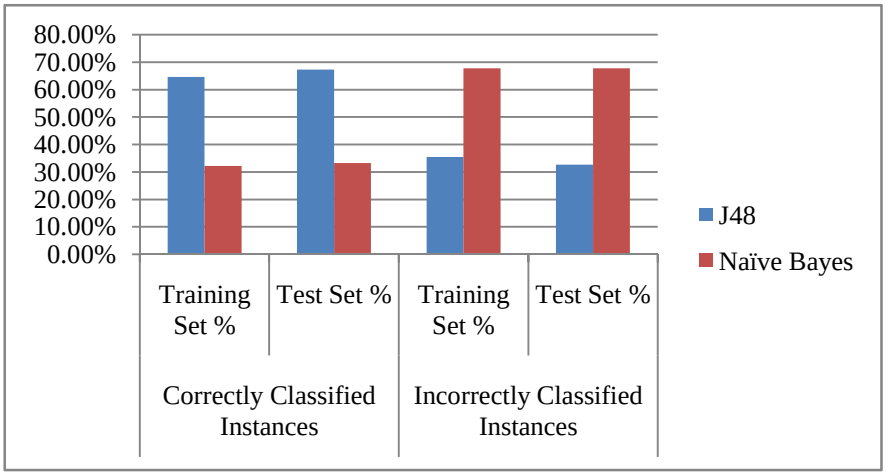


Figure 20: Graphical representation of the classified instances

In Fig. 20 above the researcher visualized the bar graph of the performance evaluation obtained for the different tools. The highest accuracy is found by the J48 decision tree method. Thus, it is considered also the base case. All the J48 decision tree algorithm tools tested have performed much better than the Naïve Bayes classifier method.

The result scores of the Naïve Bayes classifier for time taken to execute the model have better than the J48 decision tree model. However, the overall result scores of the J48 decision tree model higher than that of the Naïve Bayes classifier model.

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, time taken for execution,AUC,sensitivity and specificity).The results were achieved using 75 % split test which is used for the training the model and then supply the unseen remaining part of the record for testing the performance of the model.

4.9.2 Using the classification Algorithm in the data set

Classification is used to find a model that segregates data into predefined classes and this is based on the features present in the data. The result is a description of the present data and a better understanding of each class in the database.

Thus classification provides a model for describing future data. Prediction helps users make a decision. Predictive modeling for knowledge discovery in databases predicts unknown or future values of some attributes of interest based on the values of other attributes in a database as in figure below:-

4.9.3 Prediction using the J48 Classifier

Given the unseen verbal autopsy data set the J48 decision tree classifier predicts the data and produces the following predicted classes:-

a5_11 Nominal	a5_06_1d Nominal	word_ami Numeric	word_cancer Numeric	word_fire Numeric	word_kidney Numeric	word_pregnanc Numeric	predicteds_text34(Cause of Death) Nominal	**gs_text34(Cause of Death)** Nominal
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	AIDS	AIDS
No	Don't Know	0.0	0.0	0.0	0.0	0.0	Drowning	Drowning
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	Fires	Fires
No	Don't Know	0.0	0.0	2.0	0.0	0.0	Fires	Fires
Yes	Don't Know	0.0	2.0	0.0	0.0	0.0	Leukemia/Lymphomas	Stomach Cancer
Yes	15.0	0.0	0.0	0.0	0.0	0.0	Other Cardiovascular Diseases	Other Cardiovascular Diseases
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	Pneumonia	Pneumonia
Yes	Don't Know	0.0	1.0	0.0	0.0	0.0	Cervical Cancer	Cervical Cancer
Yes	Don't Know	0.0	0.0	0.0	1.0	0.0	Asthma	Renal Failure
Yes	Don't Know	0.0	0.0	0.0	1.0	0.0	Renal Failure	Renal Failure
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	Esophageal Cancer	Cirrhosis
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	Stroke	Stroke
Yes	Don't Know	0.0	0.0	0.0	0.0	0.0	COPD	COPD

Table 8: Prediction of unseen data sets using J48 decision tree Classifier

4.9.4 Prediction using the Naïve Bayes Classifier

Given the unseen verbal autopsy data set the naïve bayes classifier predicts the data and produces the following predicted classes:-

word_ami Numeric	word_cancer Numeric	word_fire Numeric	word_kidney Numeric	word_pregnanc Numeric	predictedgs_text31(Cause of Death) Nominal	gs_text34(Cause of Death) Nominal
0.0	0.0	0.0	0.0	0.0	Pneumonia	Other Cardiovascular Diseases
0.0	0.0	0.0	0.0	0.0	Poisonings	Pneumonia
0.0	1.0	0.0	0.0	0.0	Other Non-communicable Diseases	Cervical Cancer
0.0	0.0	0.0	1.0	0.0	Acute Myocardial Infarction	Renal Failure
0.0	0.0	0.0	1.0	0.0	Poisonings	Renal Failure
0.0	0.0	0.0	0.0	0.0	Leukemia/ Lymphomas	Cirrhosis
0.0	0.0	0.0	0.0	0.0	Stroke	Stroke
0.0	1.0	0.0	0.0	0.0	Leukemia/ Lymphomas	Other Infectious Diseases
0.0	0.0	0.0	0.0	0.0	Other Infectious Diseases	COPD
0.0	0.0	0.0	0.0	0.0	AIDS	TB
0.0	0.0	0.0	0.0	0.0	Stroke	Stroke
0.0	0.0	0.0	0.0	0.0	Drowning	Falls
0.0	0.0	0.0	0.0	0.0	Asthma	Cirrhosis
0.0	0.0	0.0	0.0	0.0	AIDS	AIDS
0.0	0.0	0.0	0.0	0.0	AIDS	AIDS

Table 9: Prediction of unseen data sets using NBC.

4.9.5 Overall Discussion of the two algorithms and their results

One of the purposes of this study was to compare the J48 decision tree algorithm and Naïve Bayes classifier machine learning model and to select the one, which performs the best.

Accordingly, each experiment carried out in this research had employed both J48 decision tree and Naïve Bayes classifier. In all experiments the same data sets were used. The output of these experiments indicates that J48 performs better than Naïve bayes classifier.

Based on all the benchmarks used to measure the algorithms employed in this study, it was discovered that J48 performance is the most appropriate interms of accuracy based on this data Focus was therefore laid on designing a predictive system on the most suitable algorithm which is J48 in this particular domain.

4.9.6 Proposed Protoype Development and Implementation

The prototype based of the J48 decision tree alrgorithm has been developed using java and Mysql for the database. Jdeveloper integrated development environment(IDE) was used to design the graphical user interface (GUI). Using the GUI, the user is able to select the

provided data sets and the J48 classifier. Upon the selection of the symptoms of the diseased from the front end, the prototype loads the respective dataset for filtering and classification. The user selection from the front end is taken as input.

Some of the inputs required for this project are defined at the java class level and some user selected inputs are directly been used in the required methods. As mentioned in the third chapter, the data sets were collected from the IHME data sets

To run the project, one should install java on their local machine, integrated development environment (IDE) Jdeveloper and Oracle server weblogic to load the project.

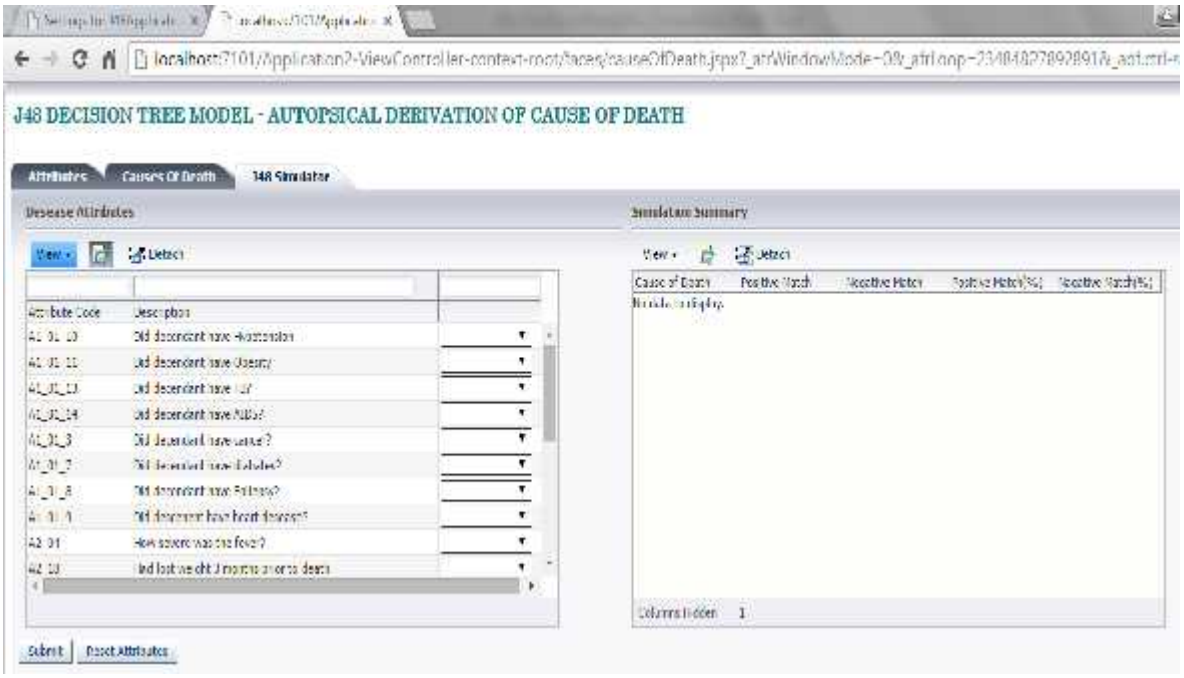


Figure 21: The GUI of the proposed verbal autopsy classification system

The figure above is the main interface for the system where there are defined causes of death and the symptoms, users select a combination of symptoms and then submit so that the system can predict the probable cause of death. This can be obtained by observing the positive and negative matches with the cause having a higher percentage match is picked to be the most probable cause of death based on the combination of symptoms. New (attributes) symptoms can also be added plus the cause list can be updated from the system by the user

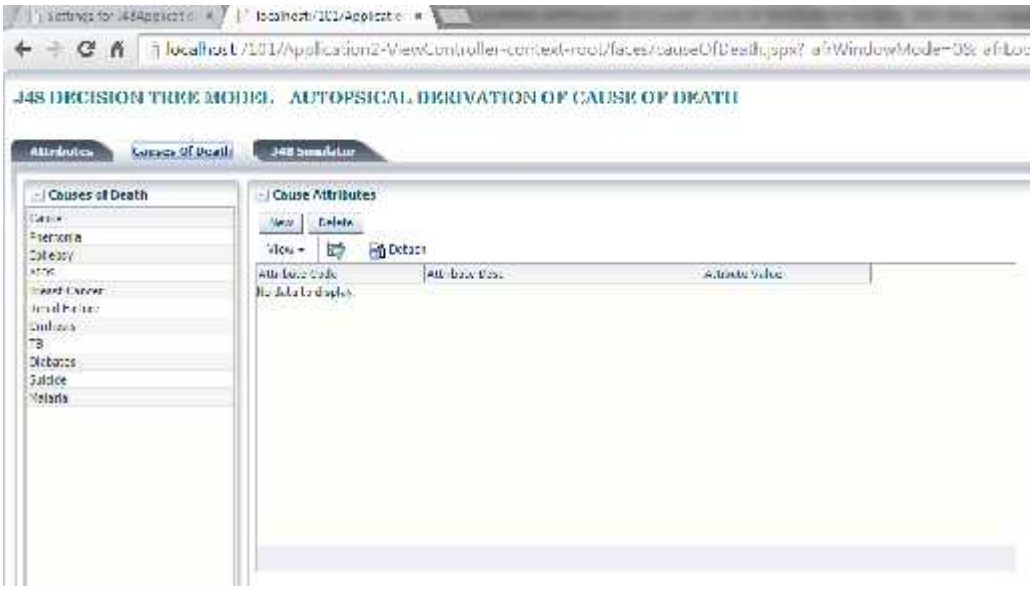


Figure 22: Sample Cause of Death list.

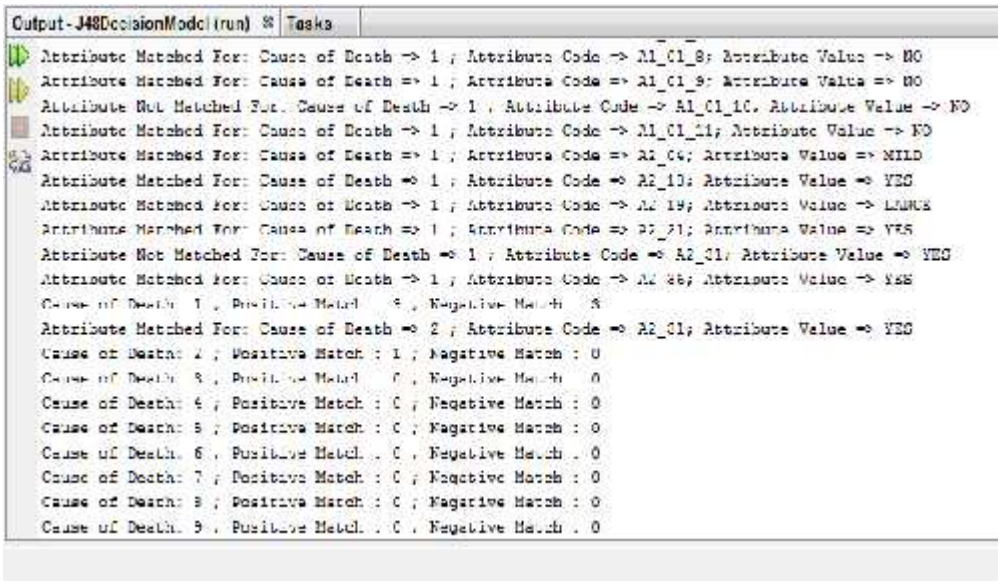


Figure 23: Sample classification results based on J48 Decision Tree classifier

CHAPTER 5

DISCUSSIONS, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the findings of the study, communicate the recommendations and conclusion and suggest areas of further research. The first section provides a summary of the research findings including the achievements accomplished by conducting this study. The second section of this chapter outlines the recommendations and conclusion. The aim is to prove that the suggested recommendations and conclusion are logically derived from the analysis of the findings. Limitations of the study are also identified. The last section is a list of suggestions for further research.

5.2 Summary of research findings

It is important to note that the objectives of this undertaking have been realized. One of the objectives was to examine whether J48 decision tree performs better than Naïve Bayes Classifier when applied to verbal autopsy data and accurately classify the true cause of death. The results of the study have shown that J48 decision tree algorithm is better than Naïve Bayes Classifier and it can effectively be used in verbal autopsy text classification. The classification accuracy obtained indicates that the J48 has the ability to correctly classify more instances in terms of the percentage than the naïve bayes classifier. Feature selection has proven to be vital in improving the performance accuracy of the classifier.

A model has been designed and used to evaluate the verbal autopsy data. This ensures efficiency in cause of death classification is free of bias and ensuring that the results are obtained in a short period of time.

5.3 Conclusion

Machine intelligence algorithms are improving as the number of ML tools, techniques and algorithms increase. A great deal of data in health care is still being gathered and organized using pen and paper. Indeed, the data contains and reflects activities and facts about the organization. But the data's hidden value, the potential to predict health trends, has largely gone unexploited. The increase in data volume causes great difficulties in extracting useful information and knowledge for decision support. It is to bridge this gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as ML or KDD has emerged in recent years.

The application of ML technology has increasingly become very popular and proved to be relevant for many sectors such as healthcare sectors. Particularly, in the public health, ML technology has been applied for predicting the cause of death from verbal autopsy for effective and efficient predictive model

This research has tried to assess the application of ML technology to predict the cause of death from the verbal autopsy and correctly classify the cause of death, for developing a classification model. Such a classification model could enable the public health departments as well as for the governmental and non-governmental organizations to implement predictive model.

5.4 Recommendations

This study and investigation has been conducted mainly for an academic purpose. However, it revealed the potential applicability of ML technology to classify cause of death from the IHME dataset. Moreover, it is the researcher's belief that the contribution of this research work could be a good experience for a competitive study in public health sector as well as computer science field of verbal autopsy in the future.

Apart from this, it is the researcher's faith that the findings of the research would encourage public health sector to work on the application of ML technology in health sector in general and cause of death classification in particular.

Therefore, the researcher strongly recommends the following:

• In this research encouraging results were obtained, further investigation should be done by integrating the numerous verbal autopsy data sources.

• There is a need to develop an operational application prototype verbal autopsy classification system.

•Further extensive experiments should be required by using large amounts of dataset and applying different classification techniques.

5.5 Limitations of the Study

Obtaining comprehensive set of actual data from the health institutions was difficult as such information is considered confidential and thus should be hidden from un-authorized entities

5.6 Future Work

Research has shown that data sample sizes together with an associated gold standard is a major issue overall in this problem space. To be able to take this forward from a computational approach, larger samples need to be gathered and importantly conducted under the same protocols so that comparability can be assessed. Only then can computational processes start to move forward. Standardization is also key so that machine learning becomes a viable option not only to assist in developing more accurate predictors of cause of death but also to assist with cost control.

Alternatives are needed to physician review as it is relatively cost ineffective and not feasible when assessing large numbers of questionnaires. More research needs to be carried out using the data driven methods of Logistic Regression, ANN and Bayesian approaches to provide a real alternative that can handle volume case load and predict with a high degree of accuracy and consistency cause of death.

In final conclusion, data driven research may feedback into improved design of standardized questionnaires. If we have a better understanding of which features and questions are useful in automated diagnosis, this can inform the design of questionnaires, so that the VA can be simplified.

REFERENCES

1. ABRAHAM, Flaxman D, et al. (2011). *Random forests for verbal autopsy analysis:multisite validation study using clinical diagnostic gold standards.Population Health Metrics*, **9:29**,.

2. ARCHANA, Chaudhary, RAJ, Kamal and SAVITA, Kolhe (2013). Machine Learning Classification Techniques:A Comparative Study. *International Journal on Advanced Computer Theory and Engineering (IJACTE)*, **2** (4), 2319 – 2526.

3. ASHOK, Kumar D and GOVINDASAMY, R. (2015). Performance and Evaluation of Classification Data Mining Techniques in Diabetes. *International Journal of Computer Science and Information Technologies(IJCSIT)*, **Vol. 6** (2), 1312-1319.

4. BAIDEN, Frank, et al. (2007). Setting international standards for verbal autopsy. *Bulletin of the World Health Organization*, **85**, 570-571.

5. BEYENE, Tadesse (2011). *"Mining Vital Statistics Data: The Case of Butajira Rural Health Program",Msc Thesis, Addis Ababa University*. Ethiopia, Tadesse Beyene.

6. BHARAT, Chaudhari and MANAN, Parikh (2012). A Comparative Study of clustering algorithms using weka tools. *International Journal of Application or Innovation in Engineering & Management(IJAIEM)*, **Volume 1** (Issue 2),.

7. BOULLE, Andrew, et al. (2001). A Case study of using Artificial Neural Networks for classifying cause of death from Verbal Autopsy. *International Journal of Epidemiology*, **30**, 515-520.

8. CHANG, C. L. (2007). A study of applying DM to early intervention for developmentally-delayed children. Expert Systems with Applications.

9. CHITRAA, V. and THANAMANI, Antony Selvadoss (2013). REVIEW OF ENSEMBLE CLASSIFICATION. *IJCSMC International Journal of Computer Science and Mobile Computing*, **2** (5), 307-312.

10. D.FLAXMAN, Abraham and T.GREEN, Sean (2010). Machine Learning Methods for Verbal Autopsy in Developing Countries. In: *Proceedings of Artificial Intelligence for Development, Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-01*, California,USA, March 22-24,2010. Stanford.

11. DANSO, Samuel, et al. (2013). A comparative analysis for verbal autopy using machine learning methods. *International Journal Of Computer Science Issues*, **Volume 1** (45), 34-67.

12. DANSO, Samuel, et al. (2010). A Verbal Autopsy Corpus Annotated With Time and Cause of Death. In: *Conference Linguistics Conference*, Faculty of Engineering,

School of Computing, Institute for Artificial Intelligence and Biological Systems,Leeds University, UK, Natural Language Processing Group.

13. DANSO, Samwel, et al. (2013). A Comparative Evaluation and Analysis of Machine Learning Methods for Verbal Autopsy Text Classification. *IJCSI International Journal of Computer Science Issues*, **10** (2), 1-10.

14. DANSO, Samwel, et al. (2011). A verbal autopsy corpus for machine learning of cause of death. In: *Corpus Linguistics Conference Proceedings*, 2011. Birmingham.

15. DEEPAJOTHI.S and SELVARAJAN.S. (2012). A Comparative Study of Classification Techniques On Adult Data Set. *International Journal of Engineering Research & Technology (IJERT)*, **1** (8), 1-8.

16. DEWAN MD, Farid, et al. (2010). Combining Naive Bayes Decision Tree for Adaptive Intrusion Detection. *International Journal of Network Security & Its Applications (IJNSA)*, **Volume 2** (2), 12-25.

17. E. , Bhuvaneswari and V. R. , Sarma Dhulipala (2013). The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset . *Information Engineering*, **2** (1), 6-13.

18. GARENNE, Michel (2014). Prospects for automated diagnosis of verbal autopsies. *Medicine for Global Health*, 12-18.

19. GARY, King and YING, Lu (2008). Verbal Autopsy Methods with Multiple Causes of Death. *Statistical Science*, **Volume 23** (No.1), 78-91 doi: 10.1214/07-STS247.

20. GOPALA, Krishna Murthy Nookala and BHARATH, Kumar Pottumuthu (2013). Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification. *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*, **2** (5), 49-55.

21. HAL, Daumé III (2012). A Course in Machine Learning. [online]. Hal, Daumé III, 149-155. last accessed 15 March 2014 at: http://www.ciml.info.

22. IWAN, Syarif, et al. (2007). Machine Learning and Data Mining in Pattern Recognition. [online]. In: *Application of Bagging, Boosting and Stacking to Intrusion Detection*. University of Southampton, UK, Springer, (2012), volume 7376 of Lecture Notes in Computer Science, page 593-602. last accessed 25 March 2014 at: http://dx.doi.org/10.1007/978-3-642-31537-4_46.

23. JAMES, SL, et al. (2011). Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. Population Health Metrics. *Population Health Metrics Research Consortium (PHMRC)*, **9:** (31),.

24. JEYARANI, D Sheela, et al. (2013). A Comparative Study of Decision Tree and Naive Bayesian Classifiers on Medical Datasets. *International Journal of Computer Applications*, (ISSN 0975 – 8887), 5-7.

25. LEITAO, Jordana, et al. (2014). Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low- and middle-income countries: systematic review. *Medicine for Global Health*, 12-22.

26. MATHERS, CD, et al. (2005). Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ 2005*, **Vol 83**, 171-177.

27. MULUGETA, Mahlet and BORENA, Berhanu (2013). Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia. *HiLCoE Journal of Computer Science and Technology, Vol. 2, No. 2 37*, **2** (2), 37-43.

28. MUNDE, M Kusum and MANGRULE, A Rupali (2013). A Review on Various Classification Algorithms for An Incremental Spam Filter. *International Journal of Application or Innovation in Engineering & Management(IJAIEM)*, **Volume 2** (11), 325-331.

29. MURRAY CJL, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, James SL, Vahdatpour A, Atkinson C, Freeman MK, Ohno SL, Black R, Ali SM, Baqui AH, Dandona L, Dantzer E, Darmstadt GL, Das V, Dhingra U, Dutta A, Fawzi W, Gómez S, Hernández B, Joshi.

30. MURRAY, CJL, et al. (2011). Simplified Symptom Pattern Method for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, **9:30**,.

31. MURRAY, CJL, et al. (2011). Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics*, **9:27**, 1726 - 1735.

32. MURRAY, CJL, et al. (2011). Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population Health Metrics*, **9:28**,.

33. MURRAY, JL Christopher, et al. (2014). Using verbal autopsy to measure causes of death:the comparative study of existing methods. *BMC Medicine*, **12(5).** , 1-19.

34. NGEMU, Joseph Mutuku, et al. (March 2015). Student Retention Prediction in Higher Learning Institutions: The Machakos University College Case. *International Journal of Computer and Information Technology*, **04** (02), 489-497.

35. PARMAR, Hitesh H and SHAH, Glory H (2013). Experimental and Comparative Analysis of Machine Learning Classifiers. *International Journal of Advanced Research in Computer Science and Software Engineering*, **Vol 3** (10), pp. 955-963.

36. PATIL, Tina. R. and SHEREKAR, S,S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*, **Vol. 6** (No.2), 256-261 ,ISSN: 0974-1011 (Open Access) Available at:www.researchpublications.org.

37. PETE, Chapman (NCR), et al. (2010). *CRISP-DM 1.0,Step-by-step data mining guide*. SPSS.

38. PETER, Byass, et al. (2003). A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam. *Scandinavian Journal of Public Health*, **Volume 3** ((Suppl. 62):), 32–37.

39. QASEM, A. Al-Radaideh, et al. (2014). A Comparison Study between Data Mining Tools over some Classification Methods. *(IJACSA) International Journal of Advanced Computer Science and Applications*, (Special Issue on Artificial Intelligence), 18-26.

40. QUIGLEY, M, et al. (1999). Diagnostic accuracy of physician review, expert lgorithms and data-derived algorithms in adult verbal autopsies. *International Journal of Epidemiology*, **28**, pp 1081-1087.

41. REBECCA, West (2010). "HealthCare Tagging of Verbal Autopsies using SNOMED-CT",Msc Thesis,Leeds University,UK. [online].. last accessed 10th February 2014 at: http://www.comp.leeds.ac.uk/mscproj/reports/0910/west.pdf.gz

42. ROY, Sankhadeep and MOHAPATRA, Anjali (2013). Performance Analysis of Machine Learning Techniques in Micro Array Data Classification. *International Journal of Software and Web Sciences (IJSWS)*, **Vol 4** (1), 20-25.

43. RUZICKA, LT and LOPEZ, AD (1990). The use of cause-of-death statistics for health situation assessment: national and international experiences. *World Health Stat Q 1990*, **43**, 249-258.PubMed Abstract.

44. S.DEEPAJOTHI and DR.S.SELVARAJAN (October 2012). A Comparative Study of Classification Techniques On Adult Data Set. *International Journal of Engineering Research & Technology*, **1** (8), 1-8.

45. SAMUEL, Danso Odei (2006). "An Exploration of Classification Prediction Techniques in Data Mining:The insurance domain",Msc Thesis,Bournemouth University,UK. [online].. last accessed 15th February 2014 at: http://www.comp.leeds.ac.uk/scsod/MSc%20Dissertation.pdf

46. SEEMA, S, et al. (2012). Ensemble Classifiers with Stepwise feature Selection for Classification of Cancer Data. *International Journal of Pharmaceutical Science and Health Care*, **6** (2), 48-61.

47. SOLEMAN, Nadia, et al. (2006). Verbal Autopsy:Current practices and Challenges. *Bulletin of the World Health Organization*, March, 239-245.

48. SRIVASTAVA, Shweta (2014). Weka: A Tool for Data preprocessing, Classification,Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Applications (0975 – 8887)*, **Vol 88** (10), 26-29.

49. SUGANDHI, C, et al. (2011). Analysis of a Population of Cataract Patients Databases in Weka Tool. *International Journal of Scientific & Engineering Research*, **Volume 2** (Issue 10),.

50. SUSHILKUMAR, Rameshpant Kalmegh (2015). Comparative Analysis of WEKA Data Mining Algorithm Random Forest, Random Tree and LAD Tree for Classification of Indigenous News Data. *International Journal of Emerging Technology and Advanced Engineering*, **5** (1), 507-517.

51. TING, S. L., IP, W.H. and ALBERT, H.C. Tsang (2011). Is Naïve Bayes a Good Classifier for Document Classification? *International Journal of Software Engineering and Its Applications*, **Vol 5** (3), 37-46.

52. VITALIS, Ndume, YAW, Nkansah- Gyekye and JESUK, Ko (2014). The Integrated Model for e-Health Data Collection and Sharing under Distributed Environments in Tanzania. *International Journal of Computer Applications*, **96** (2), 18-25.

53. *WEKA*. (2014). [online]. Last accessed 25th March 2015 at: http://www.cs.waikato.ac.nz/~ml/weka/

54. WITTEN, Ian H, EIBE, Frank and MARK, Hall A (2011). *DATA MINING:Practical Machine Learning Tools and Techniques*. Burlington,USA, Morgan Kaufmann Publishers.

55. ZHI-HUA, Zhou, et al. (2013). Multiple Classifier Systems. In: ZHI-HUA, Zhou, FABIO, Roli and JOSEF, Kittler, eds., (eds.). *In Prooceedings of the 11th International Workshop,MCS 2013*, Nanjing, China, May 15-17, 2013. Springer.

APPENDICES

Appendix A: Check list for machine learning tools evaluation

Features	Common Machine Learningtools				
	Weka3.6	**Tanagra**	**KNIME**	**Orange**	**iDataAnalyser**
Platform independence	Yes	Yes	Only Windows	Only Windows	Only Windows
Abilitytohandle largedataset	Yes	Limited		Limited	
Rangeof data miningalgorithms inthetool	Implements mostall machine learning techniques	Decision treesand Association rules	Cluster algorithm	Implements most statistical functions	
DataSources	CSV, Standard RDBMS,C4.5, Serialised instances,Arff	XML, Oracle, MySql,SAP DB,MS Access	-	CSV,C4.5	Excel
Output	Summary Text,Graphs	_		Summary Text,Graphs	_
Technicalsupport	Yes	-	-	-	-
Multiclass Support	Yes	No	No	No	No
SourceAvailable	Yes,including synopsisof allalgorithms	No	No	Yes	No

Appendix B: Sample IHME dataset for model building

a2_04 Nominal	a2_19 Nominal	a2_21 Nominal	a2_32 Nominal	a2_37 Numeric	a2_44 Nominal	a2_47 Nominal	a2_62 Numeric	a2_64 Nominal	a2_77 Nominal	a2_85 Nominal	a3_01 Nominal	a3_18 Nominal	a5_02 Nominal	a5_03 Nominal	a5_04 Numeric	a6_01 Nominal	gs_text34(Cause of Death) Nominal
Moder..	Don't..	No	No	1.0	Don't...	No	14.0	Yes	No	No	No	No	No	No	0.0	Yes	Other Non-communicable Disea..
Severe	Don't..	No	No	5.0	0.5-24..	Yes	0.0	No	Yes	Yes	No	No	No	No	0.0	Yes	AIDS
Mild	Don't..	Yes	No	0.0	1-24hr	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Acute Myocardial Infarction
Don't ..	Don't..	No	No	3.0	Don't...	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Other Non-communicable Disea..
Don't ..	Don't..	No	No	0.0	Don't...	No	6.0	No	No	No	No	No	No	No	0.0	Yes	Other Non-communicable Disea..
Moder..	Large	No	No	4.0	Don't...	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Other Non-communicable Disea..
Don't ..	Slight	No	Yes	3.0	Don't...	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Maternal
Moder..	Large	No	Yes	0.0	1-24hr	No	0.0	No	No	No	No	No	No	No	0.0	Yes	AIDS
Severe	Don't..	No	Yes	21.0	(30mi..	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Other Cardiovascular Diseases
Moder..	Moder...	No	No	0.0	Don't...	No	0.0	No	No	No	No	No	No	No	0.0	Yes	Diabetes
Don't ..	Large	No	No	1.0	Don't...	No	0.0	No	No	No	No	No	No	No	1825.0	Yes	Poisonings
Don't ..	Don't..	No	No	0.0	Don't...	No	0.0	No	Yes	No	No	No	No	No	0.0	Yes	Stroke
Severe	Don't..	Yes	No	5.0	Don't...	No	0.0	Yes	No	No	No	No	No	No	0.0	Yes	Renal Failure
Severe	Moder...	Yes	No	2.0	Don't...	No	0.0	No	No	No	No	No	No	No	0.0	Yes	AIDS

Appendix C: A partial decision tree generated for IHME training dataset

=== Classifier model ===

J48 pruned tree

```
a6_01 = Yes
|  a1_01_3 = No
|  |  a1_01_14 = No
|  |  |  a5_02 = No
|  |  |  |  a3_18 = No
|  |  |  |  |  a6_06_1d = Don't Know
|  |  |  |  |  |  a5_04 <= 0
|  |  |  |  |  |  |  a2_85 = No
|  |  |  |  |  |  |  |  a1_01_7 = No
|  |  |  |  |  |  |  |  |  a1_01_9 = Don't Know
|  |  |  |  |  |  |  |  |  |  a2_04 = Severe: Other Non-communicable Diseases (6.0)
|  |  |  |  |  |  |  |  |  |  a2_04 = Mild: Maternal (0.01)
|  |  |  |  |  |  |  |  |  |  a2_04 = Don't Know
|  |  |  |  |  |  |  |  |  |  |  word_pregnanc <= 1
|  |  |  |  |  |  |  |  |  |  |  |  g5_02 = Male: Acute Myocardial Infarction (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  g5_02 = Female: Other Cardiovascular Diseases (5.02/1.02)
|  |  |  |  |  |  |  |  |  |  |  word_pregnanc > 1: Maternal (2.0)
|  |  |  |  |  |  |  |  |  |  a2_04 = Moderate
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  g1_07a = 50.0: Other Cardiovascular Diseases (1.0)
|  |  |  |  |  |  |  |  |  |  |  g1_07a = 72.0: Stroke (1.0)
|  |  |  |  |  |  |  |  |  a1_01_9 = Yes
|  |  |  |  |  |  |  |  |  |  g1_07a = 51.0: Other Cardiovascular Diseases (0.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 26.0: Epilepsy (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 60.0
|  |  |  |  |  |  |  |  |  |  |  a2_21 = Yes: Renal Failure (2.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  a2_21 = No
|  |  |  |  |  |  |  |  |  |  |  |  word_ami <= 0: Other Cardiovascular Diseases (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  word_ami > 0: Acute Myocardial Infarction (4.0)
|  |  |  |  |  |  |  |  |  |  |  a2_21 = Don't Know: Acute Myocardial Infarction (0.0)
|  |  |  |  |  |  |  |  |  |  |  a2_21 = Refused to Answer: Acute Myocardial Infarction (0.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 80.0
|  |  |  |  |  |  |  |  |  |  |  g5_02 = Male
|  |  |  |  |  |  |  |  |  |  |  |  a2_01 <= 49: Other Cardiovascular Diseases (5.0)
|  |  |  |  |  |  |  |  |  |  |  |  a2_01 > 49: Renal Failure (2.0)
|  |  |  |  |  |  |  |  |  |  |  g5_02 = Female: Acute Myocardial Infarction (3.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 76.0: COPD (2.0/1.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 16.0: Other Cardiovascular Diseases (1.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 65.0
|  |  |  |  |  |  |  |  |  |  |  a2_37 <= 2: Acute Myocardial Infarction (10.0)
|  |  |  |  |  |  |  |  |  |  |  a2_37 > 2
|  |  |  |  |  |  |  |  |  |  |  |  a2_19 = Large: Renal Failure (1.0)
|  |  |  |  |  |  |  |  |  |  |  |  a2_19 = Don't Know: Prostate Cancer (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  a2_19 = Moderate: COPD (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  a2_19 = Slight: Prostate Cancer (0.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = Don't Know: Other Non-communicable Diseases (8.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 68.0: Other Cardiovascular Diseases (5.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 32.0: Other Cardiovascular Diseases (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  g1_07a = 35.0
|  |  |  |  |  |  |  |  |  |  |  a2_21 = Yes
|  |  |  |  |  |  |  |  |  |  |  |  g5_02 = Male: Acute Myocardial Infarction (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  g5_02 = Female: Pneumonia (4.0)
```