

2021-09-13

DESIGNING ADULT MORTALITY PREDICTION MODEL FROM ADMITTED PATIENT RECORDS USING DATA MINING TECHNIQUES

SHUMET, MOLLA WASSIE

<http://ir.bdu.edu.et/handle/123456789/13212>

Downloaded from DSpace Repository, DSpace Institution's institutional repository



BAHIR DAR UNIVERSITY
BAHIRDAR INSTITUTE OF TECHNOLOGY
SCHOOL OF RESEARCH STUDIES
FACULTY OF COMPUTING
INFORMATION TECHNOLOGY PROGRAM

**DESIGNING ADULT MORTALITY PREDICTION MODEL FROM
ADMITTED PATIENT RECORDS USING DATA MINING
TECHNIQUES**

BY SHUMET MOLLA WASSIE

ADVISOR: - GEBEYEHU B. (PHD)

BAHIRDAR, ETHIOPIA

SEPTEMBER 13, 2021

DESIGNING ADULT MORTALITY PREDICTION MODEL FROM ADMITTED
PATIENT RECORDS USING DATA MINING TECHNIQUES

SHUMET MOLLA WASSIE

A [Thesis] submitted to the school of Research Studies of Bahirdar Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in the Information Technology in the faculty of computing.

Advisor: GEBEYEHU B. (PHD)

Bahirdar, Ethiopia

September 13,2021


©2021

SHUMET MOLLA WASSIE

ALL RIGHTS RESERVED

DECLARATION

I, undersigned, declare that the thesis comprises my own work. In compliance with internationally accepted practices, I have acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged.

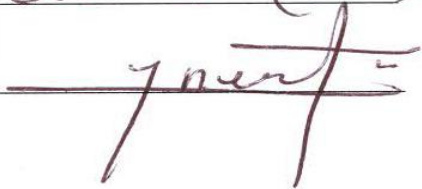
Name of the student Skunet Molla Signature 

Date of submission: 10/02/2014

Place: Bahirdar

This thesis has been submitted for examination with my approval as a university advisor.

Advisor Name: Gebreyehu B (Dr)

Advisor's Signature: 

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF RESEARCH STUDIES

FACULTY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student Shumet Molla Signature [Signature] Date 10/02/2014

As members of the board of examiners, we examined this thesis entitled "Designing Adult Mortality Prediction Model from ^{Admitted Patient 2018}" by shumet molla. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in "Information Technology".

Board of Examiners

Name of Advisor Gebeyehu B (B.S) Signature [Signature] Date _____

Name of External examiner Zewdie Mossie (PhD) Signature [Signature] Date 10/10/2021

Name of Internal Examiner Mekonnen Wasaw (PhD) Signature [Signature] Date 22/10/2021

Name of Chairperson Tamir A. Signature [Signature] Date 22/10/21

Name of Chair Holder _____ Signature _____ Date _____

Name of Faculty Dean Derejau Z. Signature [Signature] Date 26/10/2021

Name of Faculty Dean Asegahun E. Signature [Signature] Date Oct 26, 2021
Faculty Stamp



ACKNOLOGYMENT

First hardly thanks God and His Mother Saint- Merry, for providing me what I do everything to perform successfully.

I would like to thanks to Dr. Gebeyehu Belay, my respected Advisor, thanks for giving me the freedom as a friend, for your encouragement, thoughtful guidance, critical comments, and correction of the thesis.

I also wish to extend my sincere gratitude to all my instructors who have been the source of all my achievements.

Finally, I specially would like to my son Mikiyas Shumet and my sisters Loza Molla and Soliyana Molla for their understanding and support during the time of study.

ABSTRACT

Clinical data refers to health-related information that is associated with regular patient care. It provides information to health care professionals to improve the quality and safety of the care they provide to their patients. Based on huge data mining research is used to improve health care, to plan and make decision policy for satisfy medication process. Hence, health service planning and utilization become limited. Thus, adult mortality levels and trends in the developing countries become hampered. Therefore, in this study, we proposed a data mining prediction model to identify determinant attribute and consideration factors for adult mortality in case of patient dataset in Felege Hiwot Referral Hospital. The study contains 7095 instances from Felege Hiwot referral hospital recorded datasets that age between 15-60 years. To develop the model, we used classification techniques and data mining algorithm such as J48 decision tree algorithm, Support vector Machine, Random Tree, K-Nearest Neighbor and Naïve Bayes algorithm. In this research Attribute selection for better accuracy is performed by using GainRatioAttributeEval with rank. We used data KDD model approach to processed data. K-Nearest Neighbor algorithm is selected algorithm to build the model that predict adult mortality in better accuracy and possible for correct classification with value 88.27% and K-Nearest Neighbor was processed in 0 second speed, 88.3% recall, 88.9% precision and 88.5% F-Measure scored in this research.

Finally, National Classification of Disease, Duration of illness, Length of stay was significantly selected attribute to predict adult mortality. From this research Urinary Tract Infection, tuberculosis, congestive heart failure; renal disease, Road traffic accident and Poisoning were main factors of adult mortality which needed attention to minimize adult mortality by found the cause of establishment of such diseases and provided community awareness.

From this study K-nearest neighbor is recommended Algorithm for build model to predict adult mortality with 88.27% accuracy was possible. Duration of illness was very dominant attribute to adult mortality in Felege Hiwot Referral Hospital in this study.

Keyword: National classification of disease, mortality, length of stay, KNN, clinical data

TABLE OF CONTENTS

CHAPTER ONE	1
INTRODUCTION	1
1.1. Background	1
1.2. Statement of problem	3
1.3. Objectives	6
1.3.1. General objective	6
1.3.2. Specific objective	6
1.4. Scope and Limitation of the Study.....	6
1.5. Significance of the study.....	7
CHAPTER TWO	9
LITERATURE REVIEW	9
2.1. Overview of Data mining.....	9
2.2. The Data Mining and it's Process.....	10
2.3. Data Mining	11
2.4. Classification techniques	11
2.5. Data Mining in Healthcare.....	12
2.6. Application in Healthcare Management	13
2.7. Adult Mortality Conditions.....	14
2.8. Data Mining Techniques for Adult Mortality	15

2.8.1.	Support vector machine (SVM).....	15
2.8.2.	Random Trees classifier	15
2.8.3.	K - Nearest Neighbor (KNN)	16
2.8.4.	Decision Tree algorithm	17
2.8.5.	Naïve Bayes Algorithm	17
2.9.	Data Mining Research Methodology	18
2.9.1.	KDD process Model.....	18
2.9.2.	CRISP Data Mining Model	20
2.9.3.	SEMMA Process Model	23
2.9.4.	Hybrid Methodology	25
2.9.5.	Comparative Analysis of Models for study	28
2.10.	Data mining implementation in adult mortality prediction related work.....	30
CHAPTER THREE		33
3.	Research Methodology	33
3.1.	Understanding of the Problem Domain	34
3.1.1.	Define the Problem Domain FHRH datasets.....	34
3.1.2.	Data Mining Goals.....	35
3.1.3.	Data Mining Tool Selection	35
3.1.4.	Identifying the key Domain Expert	39
3.1.5.	Learning about Current Solutions to the Problem.....	40

3.2.	Data Understanding	41
3.2.1.	Data Source and Data Collection	41
3.2.2.	Description of Raw Data Quality	41
3.2.3.	Attribute Selection	43
3.2.4.	Attributes Rank with Information Gain.....	44
3.3.	Preparation of the Data	46
3.3.1.	Data Cleaning	46
3.3.2.	Missing Values Handling	47
3.3.3.	Handling Outlier Value	48
3.3.4.	Data Transformation.....	49
3.4.	Data Mining	51
3.5.	Methods of Analysis and Evaluation of System Performance	52
3.5.1.	Methods of Training and Testing	52
3.5.2.	Methods of Evaluation.....	53
	CHAPTER FOUR.....	58
	EXPERIMENTATION, RESULT AND DISCUSSION	58
4.1.	OVERVIEW OF EXPERIMENTATION.....	58
4.2.	Model Building	60
4.2.1.	Measurement of model on cross validation fold.....	60
4.2.2.	Percentage Split model measurement with different classifier	64

4.2.3.	Confusion matrix.....	68
4.3.	Selected model performance and evaluations	71
4.4.	Discussion of the study	73
4.5.	Rule extraction by using Random tree algorithm	74
4.6.	Using Discovered Knowledge	77
CHAPTER FIVE		78
CONCLUSION AND RECOMMENDATION		78
5.1.	Conclusion	78
5.2.	Recommendation	79
References.....		81
Appendix 1: Outputs of the Classifiers in Experimentation		87
Appendix 2: Partial Random Tree Rule Generation for Clinical Dataset.....		90
Appendix 3: Outlier in weka.....		90
Appendix 4: Name of Consulted Domain Experts		91

LIST OF ABBREVIATIONS

ARFF	Attribute Relation File Format
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Value
SVM	Support vector machine
FN	False Negative
KNN	K-nearest Neighbors
RT	Random Tree
LOS	Length of stay
FHRH	Felege Hiwot referral hospital
FP	False Positive
KDD	Knowledge Discovery Databases
ROC	Receiver Operating Characteristics
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Learning

List of Figure

Figure 1: KDD process model	19
Figure 2: CRISP-DM process models	22
Figure 3: SEMMA process model by SAS institute	25
Figure 4: Hybrid DM process model	26
Figure 5 : Box Plot to Detect Outliers	48
Figure 6: General adult mortality Prediction in clinical data work flow diagram	52
Figure 7 : Confusion matrix	54
Figure 8: model comparison of cross validation with different classifier.....	64
Figure 9: percentage split comparison with different classifier.....	65
Figure 10: Accuracy comparison of cross validation and percentage split	67
Figure 11: Accuracy of All Models in the Experiments	71
Figure 12: KNN model output	72

List of Table

Table 1: summary of KDD, CRISP-DM, SEMMA and HYBRID processes model	29
Table 2: Summary of related work	32
Table 3: Performance Criteria.....	37
Table 4: Functionality criteria.....	38
Table 5: Auxiliary task support.....	39
Table 6: List of attributes in the initial dataset of FHRH	42
Table 7: Socio demographic Attribute	43
Table 8: Health related attribute	44
Table 9: Attribute Rank on all input Data using information gain ranker	45
Table 10: Selected attribute from the safety care database.....	45
Table 11: Selected attribute after handling missing value	48
Table 12: Selected attribute after data transformation	51
Table 13: Performance Measures of ROC Area	57
Table 14: J49 Classifier Parameter Options.....	59
Table 15: Performance of the Classifier for Different K-Values by classifier	60
Table 16: Detail performance measures in cross validation	61
Table 17: Detail performance measures in percentage split	64
Table 18: Accuracy Comparison Cross validation VS percentage split	66
Table 19: Confusion matrix of cross validation Measurements in different algorithm	68

CHAPTER ONE

INTRODUCTION

1.1. Background

The question of cause-of-death remains of interest among demographers, epidemiologists and public health researchers. Adults in the age group 15-59 years play a significant role in the socio-economic development of a country. The critical need for information on causes of death for health policy, planning, targeting, allocation of resources, monitoring and evaluating population health programs and interventions is well documented in demographic, public health and epidemiological literature (VESPER H. CHISUMPA, 2019). Adult mortality rate represents the probability that a 15 years old person will die before reaching his/her 60th birthday, if subject to age-specific mortality rates between those ages for the specified year. Knowing causes of death facilitates designing and targeting of appropriate health interventions to save lives. Few studies, however, have examined the causes of death in the adult mortality age group in greater depth in Zambia (Kelly, 1998).

Adult mortality rate in the WHO African Region remains very high in 2016. Disease burden from non-communicable diseases among adults - the most economically productive age span – is rapidly increasing in developing countries due to ageing and health transitions. Therefore, the level of adult mortality is becoming an important indicator for the comprehensive assessment of the mortality pattern in a population (WHO, 2002). The death of adults is a neglected health issue about which little is known.

It has been estimated that about eight million avoidable deaths occur in those between ages 15 and 60 each year. Comprehensive and extensive experimentation is needed to substantially describe the loss experience of adult mortality in Ethiopia. The aim of this study is suggestion of healthcare organization to explore factor of adult mortality and plan to treatment patients to reduce mortality rate. In population, adults comprise the great majority of the labor force, and it is to be expected that adult ill health and death become series effects on the productivity and well-being of the population groups since adult women and men are considered as care providers of both family and community (Mitike Molla, 2019).

In Ethiopia, despite a major progress that have been made to improve the health status of the population for the last one and half decades, people still facing a high rate of morbidity and mortality and the health status of population is remained poor (WHO, 2002). According to sub-Saharan countries statistical report the population has remained predominately young i.e., over half (52%) of the population is in the age group of 15- and 60-years Burden of disease in Ethiopia has declined dramatically which has contributed to the improvement in life expectancy, with the highest reduction already recorded in major communicable diseases. Though it is encouraging that mortality from children has reduced in the country, the slow change in mortality and burden of disease in the general adult population needs future public attention (Addisalem Tebeje Zewudie, 2020). Applying the data mining techniques is intending to address different problem associated with adult health and to extract useful knowledge from the patient admission dataset. Discovering pattern data mining technology to determine the risk factors and predict adult mortality based on patient datasets recorded on referral hospital. Clinical data are

the details of the patients that collect from referral hospitals which are used for predicting the death of adult who age between 15-60 years. Living long is a much-desired aspiration by everyone. The aim of this study was designed adult mortality prediction model from admitted patient records using data mining techniques in Felege Hiwot referral hospital.

1.2. Statement of Problem

The reason that initiated this research work is the high rate of adult mortality in sub-Saharan country including Ethiopia according to the (WHO, 2002). And it affects the family, community and also country because adult age is more responsible for family care, hardworking and very productive age of community and countries. Analysis of the impact of adult deaths on households and their members requires a community-based study is essential as data collected by the clinical data at medication of hospital database. Due to this to mad medication policy, health care management plan that used to patient satisfaction, improve adult health and to operate modern health activity, computerized adult mortality prediction is preferable solution.

((US), 2010)Clinical data is a data that collected from medication process of patient in the health care organization. It used view statistical data of patient and distribution of disease that occurred based on medication data. Based on Enormous deposit of clinical data for number of years, human could not alter individual factors on row and parameters under huge data. Therefore, computerized prediction and analysis method is current solution, this is called data mining.

In health care organization include hospital scarcity of full medical process of patient data registration that used for analysis. Due to the lack of vital statistics, adult mortality

estimates for most of Africa have been obtained mainly by indirect estimation (Awoke Misganaw, 2012). Estimates from WHO, accounting for mortality show that in most countries in Sub-Saharan Africa, the probability of an individual aged 15 dying before reaching the age of 60 was about 50% which is in sharp contrast with the developed world. The theory did not highlight or anticipate many issues that are unfolding and influencing mortality and population growth in sub-Saharan Africa.

A variety of model life tables have been developed and used in estimating and understanding adult mortality where no mortality data exist. Model life tables were based on the fact that there were predictable similarities in age-patterns of mortality between populations.

(Kristopher J. Krohn, 2015) Nowadays diseases among adult were rapidly increasing in developing countries due to ageing, health transitions and some other predictors of adult death. In Ethiopian adult mortality was increase day to day in governmental hospital. As world health report Ethiopia is one of high rate of adult mortality in sub-Saharan country. Burden of disease in Ethiopia has declined dramatically which has contributed to the improvement in life expectancy, with the highest reduction already recorded in major communicable diseases. Though it is encouraging that mortality from children has reduced in the country, the slow change in mortality and burden of disease in the general adult population needs more public attention. Understanding this phenomenon has an important implication in terms of health promotion, disease prevention and treatment of illness by intervening at different levels of factors that helps to avoid unnecessary and the unfinished death among the population. However, such conceptualization of primitive and preventive aspects of health care depends on population-based analysis which usually

needs timely, accurate, complete and adequate information on clinical characteristics and influencing factors of health problems in the population. In most developing countries including Ethiopia, because deaths are unregistered and nearly all take place outside health facilities, it is difficult to identify the causes of death among different population groups to found by using data mining process. Consequently, provisions of appropriate interventions and evaluations become a difficult task on deep data researching.

Felege Hiwot Referral Hospital do not have recommended model that suggest more affected area on patients for making plan according to their huge data records dataset. Hospitals still has not model to identify such factors and other basic reports that essential to decision making. Various studies have been conducted on adult mortality conditions and they suggest researcher do on this area. (HAILEMARIAM, 2012)Suggests Clinical data that have been gathered from different health care institutions should pay attention in adult mortality reduction. As we shown on this research clinical data-based research was not work in the area of health care institution. However, this study concerned on data analysis to suggest adult mortality reduction, healthy satisfaction program could process by decision makers at health care institution. Therefore, this study initiates reduce the rate of mortality in population by analyzed health care organization dataset and shown the basic factors of adult mortality problem.

1.2.1. Research Questions

- ▶ What is the major attribute that consider adult mortality prediction from clinical dataset?
- ▶ How to characterize adult death disease to enhance health care organization diagnosis performance?

- ▶ Which data mining Algorithm is best to predict adult mortality from clinical dataset?

1.3. Objectives

1.3.1. General objective

The objective of this study was designing adult mortality prediction model from admitted patient records using data mining techniques

1.3.2. Specific objective

- ▶ To develop predictive model that enhance health care safety optimization.
- ▶ To identify death factors that support to adults' death problem
- ▶ To propose novel ideas that enhances or optimize healthcare towards mortality reduction, and death factors prioritization for medication purpose.

1.4. Scope and Limitation of the Study

The scope of the research is bounded on prediction of adult mortality in case of patient admitted dataset. The research discovers significant knowledge using hospital patient admission data and predicts adult mortality through identifying dominant factors to cause of adult death in the hospital. In this study we develop model that used to predict adult mortality. This study is concern clinical dataset of adults whose age group is 15-60 years but below and above from describe age group is not include in the research.

The major limitation of this research is, due to time limitation unable to apply cluster and association rule discovery techniques to investigate the internal association exists among the different variables considered in this research.

1.5. Significance of the study

Primarily, the research work has an explicit significance in development of knowledge for the researcher and uses as a benchmark for interesting researchers to explore the issues in the area. The research done based on Information on adult mortality rates and causes of death were clearly important to inform regional and national health policy and other stakeholders who were collaboratively done on adult health to monitor and improve adult health of the community. The outcome of the study provides hidden knowledge by extracting large volumes of data and a model used to predict adult mortality by realizing the hidden features from governmental referral hospital dataset. The research output used to make proper health policy, make decision on improves adult health. To enhances business goal by indicating where the emphasis of adult health services may be focus to reduce adult mortality by taking proactive knowledge-driven decisions. These further scales up adult health improvement issues through guiding to prevention policies and programs. The study has a paramount importance for governmental referral hospital to plan and implement health services focus on adult health so as to mitigate the death attribute by avoidable factors through implementing the extract rules form the experimentation.

1.6. Organization of the thesis

The research work was organized in to five chapters. The first chapter deals with introducing the burden of adult mortality conditions, statement of the problem, objective of the study, methodology, scope of the study, limitation of the study and significance of the research work.

The second chapter discusses about data mining and its techniques, methods and algorithms. Application of data mining in health care and some related works with adult mortality condition and the algorithm selected were addressed in this section.

The third chapter mainly focuses on how the research conducted including understand the problem, collect, and analyze the data. Generally, preprocessing tasks including data cleaning, transformation and attribute selection are discussed, the model and method used in this research work discussed in this chapter. Chapter four presents the experimentation done, performance evaluation and the analysis of the result using classification techniques in data mining algorithms.

At the end, chapter five provided conclusion and recommendation to show further research directions.

CHAPTER TWO

LITERATURE REVIEW

2.1. Overview of Data mining

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. The ability to use these data to extract useful and new information for quality healthcare is crucial. In the recent years researchers of machine learning techniques have shown interest in the medical field. Many healthcare applications are developed for enhancement of the healthcare management as well as patients' care.

Data mining attract a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge (M., 2001).

Data mining is a science to discover knowledge from databases or huge sources of data. The database contains a collection of instances (records or case). Each instance use by machine learning and data mining algorithms is formatting using same set of fields (features, attributes, inputs, or variables). Data mining involves discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. Many other terms are used for data mining, for example, knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, and information harvesting [(Lan, 2011)].

Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. With a combination of these techniques, it is possible to find different kinds of structures and relations in the data, as well as to derive rules and models that enable prediction and decision making in new situations. It is possible to perform classification, estimation, forecasts, affinity grouping, clustering and description and visualization.

Data mining approaches solving problem through analysis of massive data. But data mining has been around much longer than analytics; data mining find itself a rather large space within that arc, ranging from descriptive exploration of identifying relationships and affinities among variables to developing models to estimate future values of interesting variable (Delen, 2014).

Obtain results from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better position to meet their long-term needs. Data mining applications also can benefit healthcare providers such as hospitals, clinics, physicians, and patients by predicting healthcare services and plan accordingly for expansion programs.

2.2. The Data Mining and it's Process

Basically, data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. Describing some of the major stages/steps involved in data mining, stated that data mining is more than just applying software, it is a process that involves a series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling results.

According to these authors the process of building and implementing a data mining solution is known as Knowledge Discovery in Databases (KDD) (Mannila, 2002). Argued that discovering knowledge from data should therefore be seen as a process containing several steps like understanding the domain, preparing the data set, discovering patterns (data mining), post processing of discovered patterns, and putting the results into use.

2.3. Data Mining

Data mining techniques are used in healthcare organization for, diagnosis and treatment, healthcare management, customer relationship management and fraud and anomaly sources (Jr, 2009). Data mining uses two strategies; supervise and unsupervised learning. In supervise learning; a training set is using to learn model parameters whereas in unsupervised learning no training set is using. Unsupervised learning refers to modeling with an unknown target variable. The models are solely descriptive. The goal of the process is to build a model that describes interesting regularities in the data.

Tasks of Data mining can be separated into descriptive and predictive. Descriptive tasks have a goal on finding human interpreted forms and associations, after reviewing the data and the entire construction of the model, prediction tasks tend to predict an outcome of interest. The prediction is one of the data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables (J B, 2003).

2.4. Classification techniques

Classification is one of the predictive data mining tasks. It is a technique used to predict group membership for data instances by assigning previously unseen records a class as

accurately as possible. Classification is important for management of decision making. This technique employs a set of pre-classified examples to develop a model that can classify the population of records at large (Benbelkacem, 2014). The goal of classification is to accurately predict the target class for each case in the data. The accuracy of the classification rules is estimated using test data. The derived model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class level of the input data.

Rx: (Condition1 \wedge Condition2 \wedge ... \wedge Condition) \implies Class

Where Rx is the rule-id and the left-hand side (LHS) of the rule represents *conditions* on some or all of the attributes in the dataset except the class label. The right-hand side (RHS) is the class name, which is either “alive” or “died”. Classification rule based on the dataset of condition of 1 is: (NCoD = Tuberculosis) \wedge (Sex = M) \wedge (Du_of_ill > 45) \wedge (payment_status = Free) \wedge (LOS \geq 7) \rightarrow (class = died)

2.5. Data Mining in Healthcare

Data mining techniques is using intensively and extensively by many organizations. In healthcare, data mining is gradually increasing popularity, if not by any case, becoming increasingly essential. Data mining applications can greatly benefit all parties are involving in the healthcare industry (Esfandiari N. B.-M., 2014). Widespread use of medical information systems and explosive growth of medical databases require

traditional manual data analysis to be coupled with methods for efficient computer-assisted analysis.

Data mining can help healthcare insurers detect fraud and abuse; healthcare organizations can make customer relationship management decisions; physicians can identify effective treatments and best practices; and patients receive better and more affordable healthcare services. Data mining can be defined as the process of finding previously unknown patterns and trends in databases or any big sources and using that information to build predictive models (Ngai, 2011).

2.6. Application in Healthcare Management

The key directions in applying data mining for adult mortality prediction used Healthcare datasets management broadly classify into the following categories. **Diagnosis and Treatment:** Research studies have suggested that unaided human analysis of data for decision making is unintentionally flawed. Applying data mining to even small data sets can provide protection against error-prone unassisted human inference and can consequently support improve treatment decisions. Data mining can particularly useful in medicine when there is no dispositive evidence favoring a particular treatment option. Other key areas where data mining is proving as an effective tool are disease diagnosis, detection and prediction (Koh, 2011).

Healthcare Resource Management: The goal here is to effectively manage resource allocation by identifying high risk areas and predicting the need for and usage of various resources. For example, a key problem in healthcare is measuring the flow of patients through hospitals and other healthcare facilities. If the inpatient length of stay could be a

factor to adult mortality, so the planning and management of hospital resources can be greatly enhanced (Azari, 2012).

Applications examples of classification in health sectors are the following.

- A hospital may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness.
- Classifying the type of drug, a patient should be prescribed based on certain patient characteristics in hospital.
- A medical researcher wants to analyze breast cancer data in order to predict which one of the specific treatments a patient should receive.

2.7. Adult Mortality Conditions

Mortality is considered as the most basic health outcome indicator. In a given population, adults comprise the great majority of the labor force, and it is to be expected that adult ill health and death have a deleterious effect on the productivity and well-being of the population groups since adult women and men are considered as care providers of both family and community.

Thus, applying the data mining techniques is intended to address multifarious problem associated with adult health and to extract useful knowledge from clinical data of hospital patient admission. In the case of medical system in hospital end result is completely survive, reduce or elongate patient health from death. If the adult inpatient mortality can be predicted efficiently, the cause diagnosis, medical treatment and planning the management of hospital can be greatly enhancing (Duthe G, 2009).

2.8. Data Mining Techniques for Adult Mortality

2.8.1. Support vector machine (SVM)

A support vector machine is a machine learning model that is able to generalize between two different classes if the set of labeled data is provided in the training set to the algorithm. The main function of the SVM is to check for that hyper plane that is able to distinguish between the two classes. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The algorithm creates a line or a hyper plane which separates the data into classes. SVM often required data to be normalized prior to the training/classification. However, SVM is reported to perform better when the training set is small or unbalanced. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. SVM require much fewer input data than ANN, so as we have a data requirement SVM is more preferable for this research experiment.

2.8.2. Random Trees classifier

Random Trees (RT) belong to a class of machine learning algorithms which does ensemble classification. The term ensemble implies a method which makes predictions by averaging over the predictions of several independent base models

- It is an excellent classifier--comparable in accuracy to support vector machines.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.

- It has an effective method for estimating missing data and maintains accuracy when up to 80% of the data are missing.
- It has a method for balancing error in unbalanced class population data sets.
- Generated forests can be saved for future use on other data.
- It gives estimates of what variables are important in the classification.
- Output is generated that gives information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or by scaling, give interesting views of the data.

In general, the random trees classifier can handle a mix of categorical and numerical variable. The Random Trees is also less sensitive to data scaling and works better and faster with large training sets (Breiman, 2003).

2.8.3. **K - Nearest Neighbor (KNN)**

KNN is a k-nearest-neighbor classifier that uses the same distance metric. The number of nearest neighbors can be specified explicitly in the object editor or determined automatically using leave one-out cross-validation focus to an upper limit given by the specified value. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbors. A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees” (Sundaram S. a., 2012).

2.8.4. **Decision Tree algorithm**

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances (Kaur&Chhabra, 2014). Decision trees constitute a method able to forecast or classify future observations according to decision rules. If the information is divided in classes, it is possible to utilize the data to generate rules able to classify previous cases and new cases with absolute precision. The decisional process behind the model appears clear when observing the tree, with a clear advantage with respect to other techniques whose internal logic is difficult to interpret.

Furthermore, the process includes automatically in the rule only the attributes relevant to the decision, as the irrelevant attributes are ignored and the data dimensions are reduced. Decision trees are converted into “if-then” rules to enhance the comprehension of the model when the relationship among the elements of a group is relevant.

2.8.5. **Naïve Bayes Algorithm**

Naive Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks (E., 1997).

- It is a highly extensible algorithm that is very fast.
- It can be used for both binaries as well as multiclass classification.
- It has mainly three different types of algorithms that are Gaussian, Multinomial, and Bernoulli Naïve Bayes.
- It is a famous algorithm for spam email classification.

- It can be easily trained on small datasets and can be used for large volumes of data as well but it considered all the variables independent that contributes to the probability.

2.9. Data Mining Research Methodology

2.9.1. KDD process Model

Knowledge Discovery and Data Mining (KDD) is a multi-stage process (Qaise, 2014). The overall methodology will follow the main stages of KDD process, as depicted in Figure 1.

1. Developing and understanding of the application domain. Developing and understanding of the application domain is the first stage of KDD process in which goals are defined from customer's view point and used to expand and understanding about application domain and its prior knowledge.
2. Creating a target data set. Creating target data set is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables.
3. Data cleaning and pre-processing. This step is composed of several sub-steps and requires understanding of both the domain to which the KDD is applied and of statistical methods for improving data quality. Hence, the cooperation of experts from various disciplines is most crucial. Data cleaning and preprocessing is the third stage of KDD process which focuses on target data cleaning and pre-processing to complete and consistent data without any noise and inconsistencies. In this stage strategies are develop for handling such type of noisy and inconsistent data.

4. Data Transformation. This is the fourth stage of KDD process which focuses on transformation of data from one to another so that data mining algorithms can be implemented easily. For this purpose, different data reduction and transformation methods are implemented on target data. Data transformation aims to manipulate the data so that its content and its format are most suitable for the data mining process. The transformation process effects the distribution of the various features and the structure in which their values are stored. Various data mining techniques present different requirements regarding these characteristics of the data. The requirements of each technique should be taken into account prior to its application.

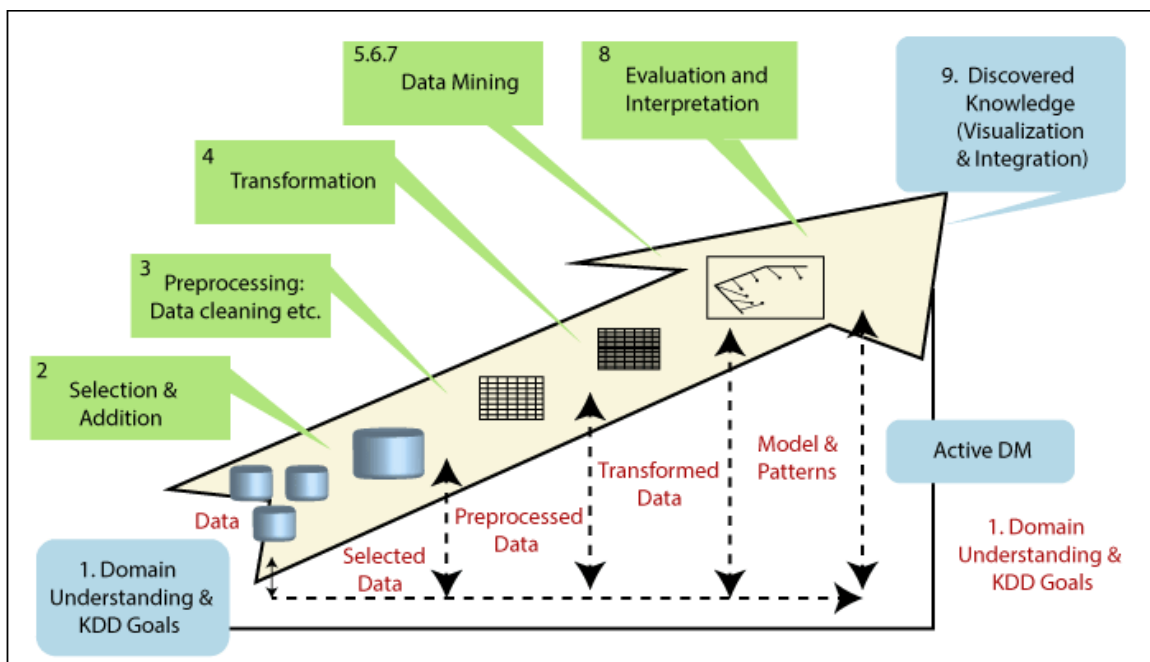


Figure 1: KDD process model

Each one of the above KDD steps has to be performed efficiently and thoroughly, in order to ensure the quality of the next step and the final outcome of the entire process.

5. Choosing the suitable data mining task. Choosing the suitable data mining task is the fifth stage of KDD process in which appropriate data mining task is chosen based on particular goals that are defined in first stage. The examples of data mining method or tasks are classification, clustering, regression and summarization.
6. Choosing the suitable data mining algorithm. Choosing the suitable data mining algorithm is the sixth step of KDD process in which one or more appropriate data mining algorithms are selected for searching different patterns from data. There are a number of algorithms present today for data mining but appropriate algorithms should be selected to produce quality knowledge
7. Employing data mining algorithm. This is the seventh step of KDD process in which selected algorithms are implemented.
8. Interpreting mined patterns. Interpreting mined patterns are the eighth step of KDD process that focuses on interpretation and evaluation of mining patterns. This step involves extracted pattern visualization.
9. Using discovered knowledge. This is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used interested parties or can be integrate with another system for further action

2.9.2. CRISP Data Mining Model

CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem (Weiss Sholom M. Z. T., Performance Analysis and

Evaluation, 2003). The CRISP-DM KDP model consists of six steps, which are summarized below:

1. Business understanding

This step focuses on the understanding of objectives and requirements from a business perspective. It is further broken into several sub steps, namely, determination of business objectives, assessment of the situation, determination of DM goals, and generation of a project plan.

2. Data understanding

This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data and detection of interesting data subsets. Data understanding is further broken down into collection of initial data, description of data, exploration of data and verification of data quality. In addition, once business objectives and the project plan are established, data understanding considers data requirements. In addition, this step can include initial data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics (which includes the visual display of categorical variables) can occur at the end of this phase. Models such as cluster analysis can also be applied during this phase, with the intent of identifying patterns in the data.

3. Data preparation

This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tools in the next step. It includes table, record, attribute

selection, data cleaning, data transformation and construction of new attributes. Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding

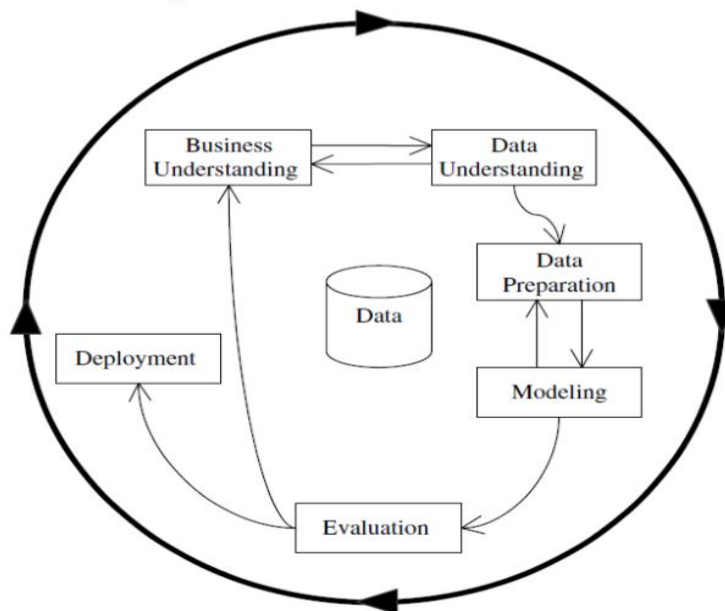


Figure 2: CRISP-DM process models

Adopted from the CRISP-DM KD process model (source: <http://www.crisp-dm.org>)

4. Modeling

This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem. At this point, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same DM problem type and the calibration of their parameters to optimal values. Since some methods may require a

specific format for input data, often reiteration into the previous step is necessary. This step is subdivided into selection of modeling techniques, generation of test design, creation and assessment of generated models.

5. Evaluation

This is the fifth stage of CRISP-DM, after one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the DM results should be reached. The key sub steps in this step include evaluation of the results, process review and determination of the next step.

6. Deployment

Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating report or as complex as implementing a repeatable KDP. This step is further divided into plan deployment, plan monitoring and maintenance, generation of final report, and review of the process sub steps.

2.9.3. SEMMA Process Model

SEMMA stand for (Sample, Explore, Modify, Model, and Access) is data mining method developed by SAS institute (Qaise, 2014). It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing

the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them. It has a cycle of five stages or steps.

Sample: - This is the first and optional stage of SEMMA process which focuses on sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.

Explore: - This is the second stage of SEMMA process which focuses on exploration of data. This can help in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

Modify: - This is the third stage of SEMMA process which focuses on modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.

Model: -This is the fourth stage of SEMMA process which focuses on modeling of data. The software for this automatically searches for combination of data. There are different modeling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.

Access: - This is the fifth and final stage for SEMMA process focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

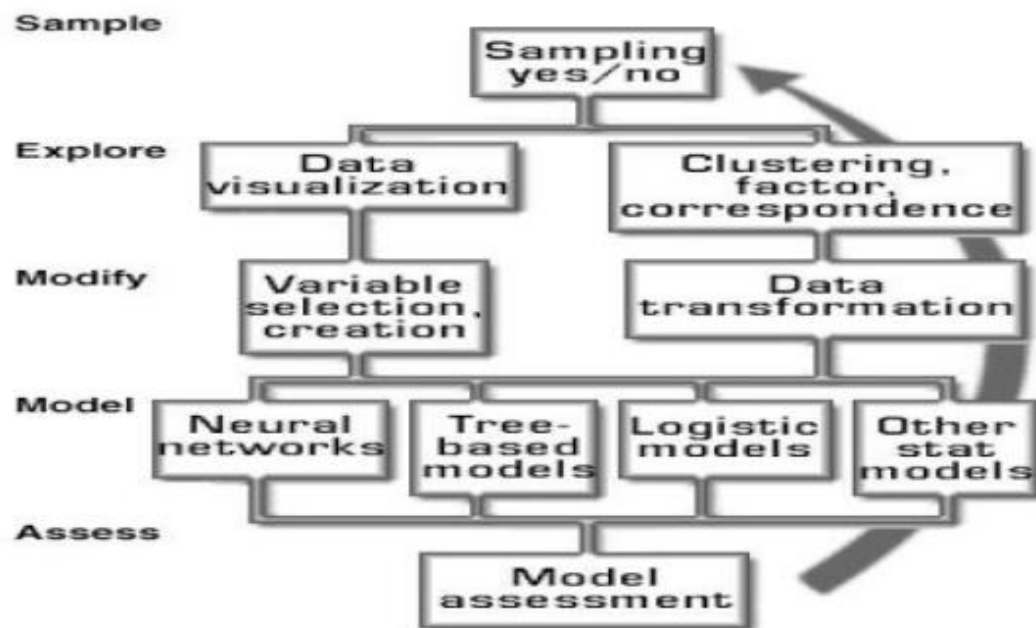


Figure 3: SEMMA process model by SAS institute

2.9.4. Hybrid Methodology

Hybrid process is characterized by providing more general, research-oriented description of the steps (et.a, 2015). The hybrid model also encourages the application of knowledge discovery for a particular domain in other domains and it has a six-step process as depicted in Figure 4.

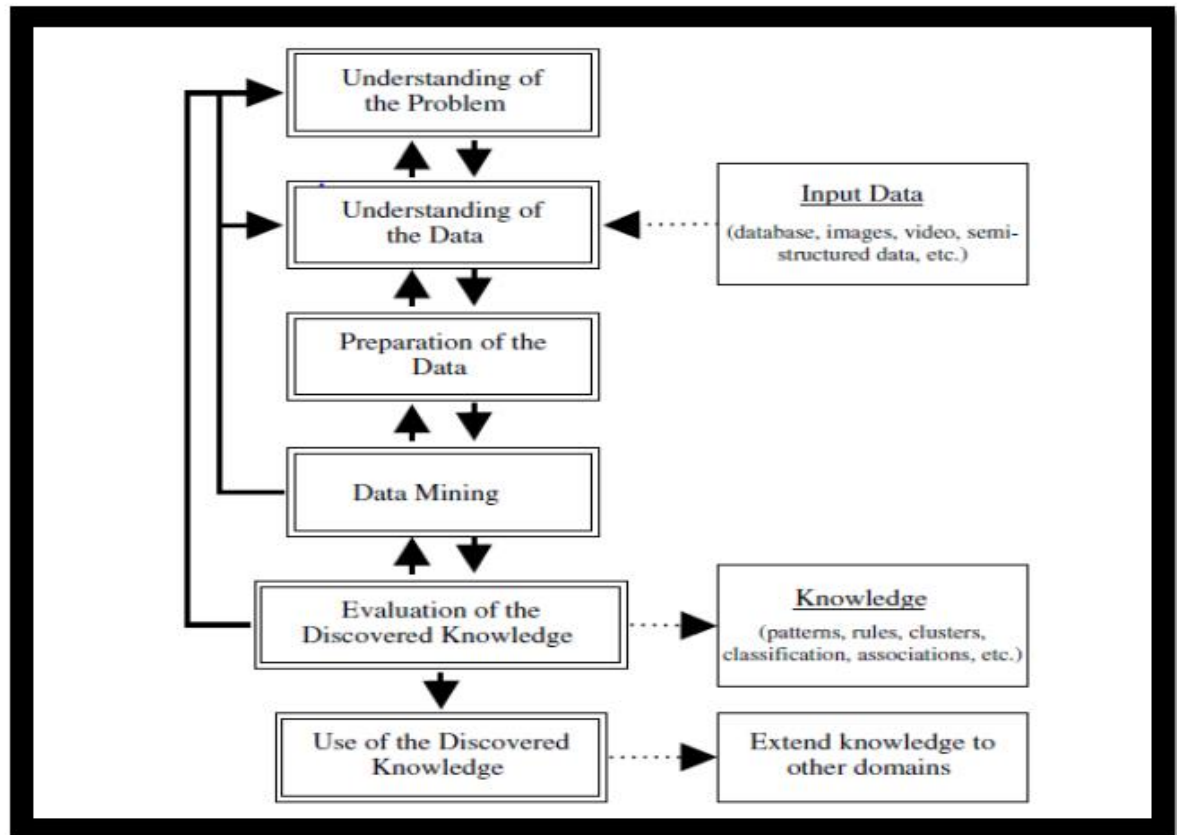


Figure 4: Hybrid DM process model

Description of the six step Knowledge Discovery Process

1. Understanding of the problem domain: This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. A description of the problem, including its restrictions, is prepared. Finally, research goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.
2. Understanding of the data: This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing

- values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.
3. Preparation of the data: This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms to reduce dimensionality by derive new attributes by discretization, and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.
 4. Data mining: The data miner uses various data mining methods to derive knowledge from preprocessed data.
 5. Evaluation of the discovered knowledge: Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.
 6. Use of the discovered knowledge. This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

2.9.5. Comparative Analysis of Models for study

There are nine, six, five and six stages for KDD, CRISP-DM, SEMMA and hybrid process model respectively. By examining all the three data mining process models they clearly shows that they are in somehow equivalent to each other but SEMMA is directly linked to the SAS enterprise miner software (Qaiser, 2014). The comparison between them shows that:

- The KDD process step “Developing and Understanding of the Application Domain” can be identified with “Business Understanding” phase of CRISP-DM process.
- The KDD process steps “Creating a Target Data Set” and “Data Cleaning and Preprocessing” can be identified with “Sample” and “Explore” stages of SEMMA respectively, and these can be identified with “Data Understanding” phase of CRISP-DM.
- The KDD process stage “Data Transformation” can be identified with “Data Preparation” stage of CRISP-DM and “Modify” stage of SEMMA process respectively.
- The three stages of KDD “Choosing the suitable Data mining task”, “Choosing the suitable Data Mining Algorithm” and “Employing Data Mining Algorithm” can be identified with “Modeling” phase of CRISP-DM and “Model” stage of SEMMA process respectively.
- The KDD process step “Interpreting Mined Patterns” can be identified with “Evaluation” phase of CRISP-DM process and “Assessment” stage of SEMMA process respectively.

- The KDD step “Using Discovered Knowledge” can be identified with “Deployment” phase of CRISP-DM process.
- The KDD step “developing and understanding of the application” can be identified with “understanding of the problem domain” phase of hybrid methodology.
- CRISP step “modeling” can be identified with “data mining” phase of hybrid methodology.
- CRISP step “deployment” can be identified with “using discovered knowledge” phase of hybrid methodology.

Table 1: summary of KDD, CRISP-DM, SEMMA and HYBRID processes model

Data mining process model	KDD	CRISP DM	SEMMA	Hybrid
No. of steps	9	6	5	6
1	Developing and Understanding of the Application	Business Understanding	-----	Understanding of the problem domain
2	Creating a Target Data Set	Data Understanding	Sample	Understanding of the data
3	Data Cleaning and Pre-processing		Explore	
4	Data Transformation	Data Preparation	Modify	Preparation of the data
5	Choosing the suitable Data Mining Task	Modeling	Model	Data mining
6	Choosing the suitable Data Mining Algorithm			
7	Interpreting Mined Patterns	Evaluation	Assessment	Evaluation
8	Using Discovered Knowledge	Deployment	-----	Use of the discovered knowledge

2.10. Data mining implementation in adult mortality prediction related work

According to (Hsiang-Yang Chen, 2011) exploring the risk factors of Preterm birth using data mining. The aim of this study is to explore the risk factors of preterm. Hence the Preterm birth is the leading cause of perinatal morbidity and mortality. Data mining with neural network and decision tree C5.0 were used. Through the construction of a decision tree, 17 rules were explored to predict preterm birth. Ten of these rules, with an accuracy of 80% or more the multiple birth was the highest risk factor, with hemorrhage during pregnancy the second most important the specific risk factors related to pregnant women were previous preterm birth, diseases, body height, and weight before pregnancy, while the, paternal risk factors were smoking, drinking, age, and occupation. In this study variety of algorithm to build a model is limited. It might be better used more than such researchers used.

Accordingly, to (Vuda, 2016) Application of Data Mining Techniques on Pre ART-Data: The case of FELEGE HIWOT referral hospital. Antiretroviral therapy (ART) is one of the treatments given to HIV patients Felege Hiwot Referral Hospital (FHRH) to restore patients with severe disease to healthy. The dataset for the study contains pre-ART records of the year 2005 and 2006 E.C produced by the ART office of patients Felege Hiwot Referral Hospital. The dataset has been utilized for the purpose of predicting clients' eligibility for ART. The final goal of this paper is to build ART eligibility predictive model that helps to deciding whether HIV positive individual should start Anti-retroviral treatment or not. For building ART eligibility predictive model, Naive Bayesian Classifier and J48 Decision Tree Classifier are used. J48 classifier using 10-fold cross validation that performs well and can be used as a best predicting model algorithm than Naive Bayesian classifier in predicting clients' eligibility for ART is

created. In this research ART eligibility prediction model used only decision tree and Naïve Bayes algorithm rather researcher might be used KNN, SVM and Part rule etc. for further performance.

According to (HAILEMARIAM, 2012) every year, more than 7.7 million children die before their fifth birthday. However, over three times those of nearly 24 million adults die every year. Less attention has been given to adults which are the most productive phase of life for both economic and social ramification of families and countries. Objective of this research is to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database. Decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869 records of both alive and died adults through three experiments and six scenarios. In this study the algorithm selected was limited. Only decision tree and Naïve Bayes algorithm is performed and the data which is used only demographic data collected in BRHP. So, it was not included clinical data. The performance of J48 pruned decision tree reveals that 97.2% of accurate results are possible for developing classification rules that can be used for prediction. The study suggests that education plays a considerable role as a root cause of adult death, followed by outmigration.

Table 2: Summary of related work

Authors name	Year	Title	Technique	Result
(Benbelkacem,K adri, Chaabane, & Atmani	2014	A Data Mining-Based Approach to Predict Strain Situations In Hospital Emergency Department Systems	decision tree	80%
Shegaw Anagaw	2002	application of data mining technology to predict child mortality patterns: the case of butajira rural health project (brhp)	Decision tree	92%
Tesfahun Hailemariam	2012	application of data mining for predicting adult mortality	J48 pruned decision tree	97.2%
Getaneh Berie Tarekegn	2016	Application of Data Mining Techniques on Pre ART-Data: The Case of Felege Hiwot Referral Hospital	<i>J48 decision tree</i>	95.8%

CHAPTER THREE

3. Research Methodology

In this Chapter, Clinical data sets used in this study were introduced, preprocessed, and analyzed to gain a superficial insight into their attributes. Also, for each data set, the various predictive data-mining techniques will be used to build models in order to compare their predictive abilities with each other.

The techniques used when performing mortality prediction are often a combination of machine learning algorithms and statistical knowledge of medical phenomena. In this study prediction method has been constructed based on Meta classifier SVM, KNN, J48, Naïve Bayes and Random Forest. The models have been evaluated by using mean absolute error and root mean-squared error, accuracy, ROC curve analysis and confusion matrix as performance metrics. Also, the researcher evaluates time taken to build the model on percentage split and cross validation to compare and differentiate among the model result.

The aim of this research is to develop new models to adult mortality, based on hospital safety care database available medical variables data in Felege Hiwot Referral Hospital (FHRH). In this study, we investigate how the technique proposed in prediction method, could be adapted to solve adult mortality problems.

3.1. Understanding of the Problem Domain

Understanding problem domain is working closely with domain experts to define the problem and determine the research goals, identifying key people, and learning about current solutions to the problem (et.a, 2015).

3.1.1. Define the Problem Domain FHRH datasets

Statistical modeling methods have already been developed for producing empirical predictions and also for the standardization of deterministic predictions produced by physically derived dynamical models (Coelho, 2005).

Although, mortality in hospital is measure based on statistical estimation of patient medication but the measurement is not show basic cause of patient medication the case based on necessary factors of hospital medication process. In current statistical estimation the data might be not fully decide for health care management. For simple long term summery of medication prediction is not still a true picture of health care organization. To obtain this requires the analysis of patient medication process using data mining technology have tremendous advantage. So far, Felege Hiwot Referral Hospital was not using analysis tools for adult mortality prediction system.

However, with the availability of clinical Data based adult predictions from different models is clearly the need for development of proposed model for the enhancement of improved adult mortality predictions. Thus, the main aim of this research is to investigate the potential applicability of data mining technology in adult mortality using clinical data-based prediction.

3.1.2. Data Mining Goals

The first data mining goal for this study investigated the potential application of data mining technology for rainfall forecasting by using ensemble and single algorithm method. To effectively accomplish this task, identifying important variables from the total data Collected was necessary. The identified variables were in turn used to build a model by using classification techniques.

The data preparation and data analysis phases help the identification of important attributes that could serve as an input for the model building. Classification model done by using the most appropriate data mining algorithm for adult mortality, i.e., is decision tree, IBK, MLP and PART algorithm. The classification technique the one which has performed better accuracy will be selected.

3.1.3. Data Mining Tool Selection

Now days software's are flooded in the market, there is a need for selection criteria of software package that should be available for individuals and organizations (Bhargava, 2013). As the number of software continues to grow and new features added to the new software the selection of the most suitable software package is becoming has tremendous advantage. Wrong or unformed decision would result in great loss of time and money.

In this thesis WEKA, R-programing, ORANGE and KEEL are compared to select the best appropriate data mining tool. WEKA is java based open-source data mining tool which has collection of data mining algorithms such as lazy, rules, decision trees and so on. On the other hand, R-project is very similar to Matlab and it uses same logic and user can easily load CSV format file and work with functions.

KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. Also, Orange is a component-based data mining and machine learning software suite, featuring a visual programming front end for explorative data analysis and visualization, and Python bindings and libraries for scripting.

However, based on performance KEEL is not efficient and ORANGE is not good for report capabilities but R-programing and WEKA are good. On the other hand, based on functionality WEKA is better than KEEL, R-programing & ORANGE. Finally, based on auxiliary support WEKA, ORANGE and KEEL are better than R-programing. Generally, WEKA is better based on performance, functionality and auxiliary support when we compare to ORANGE, KEEL and R-programing. So, WEKA 3.8 is selected to achieve the objective of this research. The description of selection criteria is described below in the following tables.

Table 3: Performance Criteria

No	Criteria	Criteria group	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Sturdiness	reliability	Capability: run consistently Without crushing.	WEKA can run consistently.	Reliable: can run consistently without crush.	Capability: run consistently Without crushing.	Capability: run consistently Without crushing.
2	Time behavior	Efficiency	Ability: produce results in reasonable amount relative to data size.	Efficient: can produce reasonable output based on data size.	Efficient: can process result reasonable amount of time	can produce reasonable output based on data size.	Efficiency is restricted
3	Report	output	Capability: Standard and customized report from the software package	Standard and customized from software package.	To use R-programing the package must be downloaded	Reporting capabilities are limited to exporting visual representations of data models	Standard and customized from software package

As, we have seen from the above table WEKA and R-programing is more reliable, efficient and generating the output with reasonable amount of time. However, ORANGE is also reliable and efficient but reporting capabilities are limited or not robust like WEKA. Also, KEEL is reliable and good for reporting but KEEL is not efficient like WEKA and ORANGE.

Table 4: Functionality criteria

№	Criteria	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Algorithmic variety	Availability of adequate variety of mining and algorithms available in the software and customization	WEKA has variety of machine learning and data mining algorithm.	Used for only Naïve Bayes, confusion matrix and data summary	Limited list of machine learning algorithms	Limited algorithms
2	Adaptability	Possible customization in general and for the specific company	Possible to customize	Possible to customize	Possible to customize	Possible to customize
3	Data type Flexibility	Variety of data types that are supported	WEKA can support CSV and ARFF.	R-programing support CSV	CSV	CSV

In the above table, WEKA is better in algorithmic variety, familiarity to the researcher and the ability to support different data types. However, R-programming, ORANGE and KEEL are not good for algorithmic variety. They have limited list of machine learning and data mining algorithms.

Table 5: Auxiliary task support

No	Criteria	Criteria meaning	WEKA	R-Programing	ORANGE	KEEL
1	Data Cleansing	Capability to modify spurious values in the data	WEKA have different data cleansing mechanisms	R-Programing cannot used for data cleansing or preprocessing.	ORANGE have different preprocessing mechanisms	KEEL can support preprocessing
2	Data filtering	Capability for selection of subsets based on user defined selection criteria	WEKA used for feature selection	-	Capability to data filtering	Capability to data filtering
4	Record deletion	Capability to delete record which may be biased from entire population of records	WEKA used for deletion of unnecessary records	-	Capability to support deletion of unnecessary records	Capability to support deletion of unnecessary records
5	Handling blanks	Capability to handle blanks and replace the entries	WEKA can handle blanks and replace missing value	-	Capability to missing value replacement	Capability to missing value replacement

As we compare WEKA, R-programing, ORANGE and KEEL based on auxiliary task support WEKA, ORANGE and KEEL are better for data preprocessing. However, R-programing is not supporting data preprocessing. In general, based on the above three criteria WEKA is better and selected for this research domain.

3.1.4. Identifying the key Domain Expert

Domain experts are consulted to have insight into the problem domain. As a result, prediction experts are the key domain expert and the researcher constitute from referral

hospital information and medical experts on the basis of the insight gained from discussion and review of relevant documents; a clear understanding of the data was achieved. Domain experts consult verification of data, result of the research and discuss on missing data and the way of handling the missing data.

3.1.5. Learning about Current Solutions to the Problem

In today's information age, adult mortality prediction has drawn much attention for research community because it helps in safeguarding human life. Beyond that, it is useful in effective prediction of condition of patient life, medication plan growth, better management decision, generally for medication purposes. Also, literatures shows that machine learning Algorithms proved to be good than the existing techniques traditional statistical methods (R.Kanth, 2014).

Generally, two methods are used to predict adult mortality: the empirical approach and the dynamical approach. The first approach is based on the occurrence of analogues and it is often referred to as analogue prediction. This approach is useful in predicting mortality if recorded cases are bountiful. The second case is based upon equations and forward simulations of the medication process and is often referred to as computer modeling. In this study, Naïve Bayes (NB), Support vector machine (SVM), Random Tree, KNN and Decision Trees (DT) were used to analyze clinical data gathered in-order to develop classification rules for the application of data mining techniques in adult mortality prediction.

3.2. Data Understanding

According (et.a, 2015) data understanding focus on collecting sample data and deciding which data type and format size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, missing values, plausibility of attribute values. Finally, in this step includes verification of the usefulness of the data with respect to the data mining goals.

3.2.1. Data Source and Data Collection

To achieve this study, we use patient medication data records of two years period (2012-2013) recorded from Felege Hiwot Referral Hospital. The obtained record includes the drug prescription, laboratory result, disease occurred, discharge condition of patient, sex and primary and secondary diagnosis of patient.

3.2.2. Description of Raw Data Quality

This section addresses the description of the structure of the data, by introducing exploration and verification of the data quality and illustrating selection of data sets. One of the initial problems in checking the data quality was the records from the database has not the full record data of those attributes. With this data heterogeneity problem in mind and with the contemporary objective of analyzing initially thousands of patient records from safety care database records from (15/05/2009 – 30/10/2011) with 7095 instances and 13 attribute were selected for the initial runs of the data mining methods. In this period the coding system was consistent and allowed avoiding potential problems related to the heterogeneity of the data.

Furthermore, the source data employed for this research purpose is safety care clinical data of hospital. The aim of this study is to create a model based on secondary data that was selected on the database medication data in the Felege Hiwot Referral Hospital in Bahirdar Amhara Region. The entire attributes in the original dataset were not concerned for this experimentation. Thus, only relevant attributes were considered to achieve the objectives of the study. Sample data was collected from the whole dataset on which feature selection and preprocessing is conducted, the next task is comprehensive assessment of distal and proximal determinants based on their significance level with respect to adult mortality prediction.

Table 6: List of attributes in the initial dataset of FHRH

No.	Attribute	Description	Data Type
1	MRN	Medical registration number	Numeric
2	Patient_Name	Name of patient in health care	Nominal
3	Age	Age of patient	Numeric
4	Sex	Sex of patient	Nominal
5	Address	Location of patient	Nominal
6	Payment_status	Status of payment of patients fee and free (assured , non-assured)	Nominal
7	(NCoD)	National classification of disease in hospital level	Nominal
8	Dis_group	Categorized of disease to general group	Nominal
9	Date_of_admission	Patients entering date to admission	Date
10	Date_of_discharge	Patient outing date from admission	Date
11	Admission_status	It is the status of patient either admitted or not admitted	Nominal
12	Diagnostic_note	It is the description of the case of the cofounder of disease	Nominal
13	Discharge_condition	Condition of Patient after admission	Nominal

3.2.3. Attribute Selection

In data mining technology not, all attributes are relevant. Therefore, selection of relevant attribute is necessary for data mining, among all original attributes. Many irrelevant attributes may be present in data to be mined. So, they need to be removed. And also, many mining algorithms don't perform well with large amounts of features or attributes. Therefore, feature selection techniques need to be applied before any kind of mining algorithm is applied.

➤ **Socio Demographic attributes to predict adult mortality**

Table 7: Socio demographic Attribute

No.	Attribute	Description	Data type
1	MRN	Medical registration number	Numeric
2	Patient Name	The name of patient in health care	Nominal
3	Age	Age of patient	Numeric
4	Sex	Sex of patient	nominal
5	Address	Location of patient	Nominal
6	Payment_status	Status of payment of patients fee and free (assured, non-assured)	Nominal

➤ **Health related data attribute**

Table 8: Health related attribute

No.	Attribute	Description	Data type
1	NCoD	National classification of disease in hospital level	Nominal
2	Dis_group	Categorized of disease to general group	Nominal
3	Date_of_admission	Patients entering date to admission	Date
4	Date_of_discharge	Patient outing date from admission	Date
5	Admission_status	It is the status of patient either admitted or not admitted	Nominal
6	Diagnostic_note	It is the description of the case of the cofounder of disease	Nominal
7	Discharge_condition	Condition of Patient after admission	Nominal

The main objectives of feature selection are to avoid over fitting and improve model performance and to provide faster and more cost-effective models (Arora, 2012). The literatures consulted and communications with domain experts has given the researcher the knowledge of attributes and significant factors that affect predict adult mortality. As a result, all attributes are not necessary for the experiment; Due to this reason in this thesis information gain ranker is used for attribute selection.

3.2.4. Attributes Rank with Information Gain

The effect of the attributes on the model performance was investigated. The full training set containing a total of 7095 instances and 9 attributes. The attributes are selected by using information gain ranker.

Table 9: Attribute Rank on all input Data using information gain ranker

Relative Importance	Attribute
0.058	NCoD
0.036	Group
0.005	Address
0.00121	Age
0.001030	Sex
0.000101	Du_of_ill
0.000100	LOS
0.000010	payment_status

Based on information gain ranker result the selected attributes for this thesis are listed in the following table.

Table 10: Selected attribute from the safety care database

Number	Field Name	Data Type	Description
01	NCoD	Nominal	Classification name of disease
02	Group	Nominal	Category of disease
03	Address	Nominal	Address of patient
04	Age	Numeric	Age of patient
05	Sex	Nominal	Gender patient
06	Du_of_ill	numeric	Number of day of patient occur upto discharge from hospital
07	LOS	Numeric	Number of Day living in hospital
08	payment_status	Numeric	Payment status of patient

3.3. Preparation of the Data

Data about medication provide in Felege Hiwot referral hospital in Amhara Region Bahir Dar Town. This research utilized Felege Hiwot Referral Hospital safety care databases for clinical data analysis purpose, i.e. the data source contained raw data about clinical data variables. Accumulating and cleaning the data were the initial task in the clinical data analysis with data mining techniques.

The researcher concerns which data is used as input for data mining methods which involves sampling, running and significance tests and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values are performed. To improve our classification accuracy, first analyzes the raw data. Before running any classification algorithms on the data, the data initially cleaned and transformed in what is called a pre-processing stage. In this preparation stage, several activities were performed including evaluating missing values, eliminating noisy data such as outliers, discretization, aggregation, and balancing unbalanced data was performed.

3.3.1. Data Cleaning

Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In order to provide access to accurate and consistent data, missing value replacement, outlier detection, data reduction and data transformation are performed.

3.3.2. Missing Values Handling

According to (J.Kaiser, 2014) the handling of missing values is an important task in KDD process. Mainly, while the dataset contains a large amount of missing data, the handling of missing data can improve the quality of KDD intensely. Selection of missing values using series mean method is highly depends on given data set, structure of attributes and missing data mechanism. There are several strategies that could be used to handle missing values. Instances with missing values could be removed, missing values can be replaced with a certain value not present data can be replaced with a value that is representative for the data set (M., 2001).

A common method for continuous attributes is to replace the missing value with the mean value of instances with no missing values. In the same way nominal missing values can be removed from attribute list and also in this study, removed and replaced methods were applied. From this research data set diagnosis note, admission status are removed because of no data is filled under such parameters. MRN, patient name also removed because of private issue.

Table 11: Selected attribute after handling missing value

No.	Attribute	Description	Data Type
1	Age	Age of patient	Numeric
2	Sex	Sex of patient	Nominal
3	Address	Location of patient	Nominal
4	Payment_status	Status of payment of patients fee and free (0,1)	Nominal
5	NCoD	National classification of disease in hospital level	Nominal
6	Dis_group	Categorized of disease to general group	Nominal
7	Date_of_admission	Patients entering date to admission	Date
8	Date_of_discharge	Patient outing date from admission	Date
9	Discharge_condition	Condition of Patient after admission	Nominal

3.3.3. Handling Outlier Value

Outliers are undesirable entries which affect the data in one or the other hand and misrepresent the distribution of the data and may mislead the algorithm (Shivan, 2008) so, upper and lower extreme values are discarded from independent variables. The following figure 6 shown outlier detection plot (M., 2001).

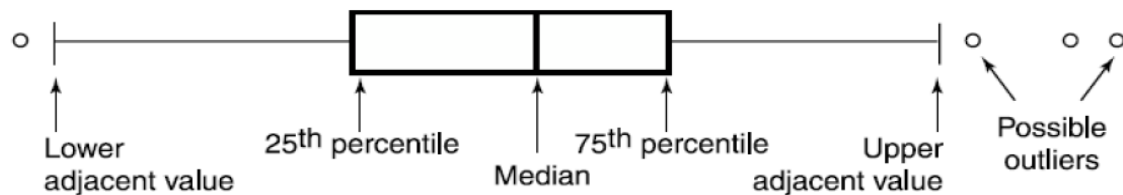


Figure 5 : Box Plot to Detect Outliers

Multiplying the interquartile range (IQR) by 1.5 will give us a way to determine whether a certain value is an outlier. If we subtract $1.5 \times \text{IQR}$ from the first quartile, any data

values that are less than this number are considered outliers. Similarly, if we add $1.5 \times$ IQR to the third quartile, any data values that are greater than this number is considered outliers. The interquartile range is what we can use to determine if an extreme value is indeed an outlier. The interquartile range is based upon part of the five-number summary of a data set, namely the first quartile and the third quartile. The calculation of the interquartile range involves a single arithmetic operation. The researcher used interquartile range method in Weka to calculate outlier values (Appendix 3). Finally after the outliers are detected only 7095 instances with 9 attributes are used for the experiment.

3.3.4. Data Transformation

At this step the researcher changes the data into a format which was suitable for the data mining tool or algorithms. The preprocessing of the data performed in WEKA 3.8 and SPSS and MS Excel. The final dataset was also in MS-Excel format; however, the selected tool WEKA 3.8 doesn't accept the data in Excel format. So, the researcher first convert the data in comma delimited (CSV) text file in ARFF (Attribute Relation File Format). Comma delimited applied for a list of records where the items are separated by commas, whereas ARFF is an extension of a file format that the WEKA software can read.

Now, we have reliable data we can make it more efficient. The uses of feature selection methods to reduce dimensionality combine features into new ones are implemented at this point. Also, as discussed before the data file saved in Comma Separated Value (CSV) file format which is appropriate for WEKA and the datasets are discretized and aggregated to reduce the effect of scaling on the data i.e. discretizing continuous value of clinical data in two classes either alive or dead.

LOS: is an attribute that occur different between two independent variable those are Date_of_admission and Date_of_discharge.

Payment_status:- This is essential variable to decide patient's medical processing, treatment and then last condition of discharging from healthcare organization. This variable represents assured and non-assured value of patient for medication.

NCoD: This is national classification of disease which is used decide the consideration role of disease for this research.

Dis_group: this variable is used to categorize disease into seventeen groups according to expert and Ethiopian Standard Treatment Guidelines for General Hospital. Disease in hospital level is more than 1780 type; to analysis such disease type the categorization into 17 type of group is very important as exert expression.

Du_of_ill: this variable is used to consider the time of disease appear up to the time decision of patient in the health care organization. This very important to decide the characteristic of disease based on the duration. Therefore the value of this variable numeric value which is the different between date of occurred disease and discharge date of patient from health care

Discharge_condition: this variable is very important variable for this research which is target variable that contained alive and dead class values.

MRN and Patient Name are ignored from research parameter in the case of private issue. Finally, WEKA class balancer used to balance the alive and dead class of target variable values.

Table 12: Selected attribute after data transformation

No.	Attribute	Description	Data Type	Value
1	Age	Age of patient	Numeric	15-60
2	Sex	Sex of patient	Nominal	M or F
3	Address	Location of patient	Nominal	List woreda
4	Payment_status	Status of payment of patients to medication	Nominal	Assured or non-assured
5	NCoD	National classification of disease in hospital level	Nominal	List of disease
6	Dis_group	Categorized of disease to general group	Nominal	Category of disease
7	LOS	Date difference between initial and discharge date of patient	Date	No. of day
9	Du_of_ill	Patient Duration the time of illness happen up to time of discharge	Date	No. of duration of date
9	Discharge_condition	Condition of Patient after admission	Nominal	Alive or Died

3.4. Data Mining

For the process of data mining uses various data mining methods to derive knowledge from preprocessed data. There are different tasks performed during the data mining phase, the majors are selection of modeling techniques and building a model.

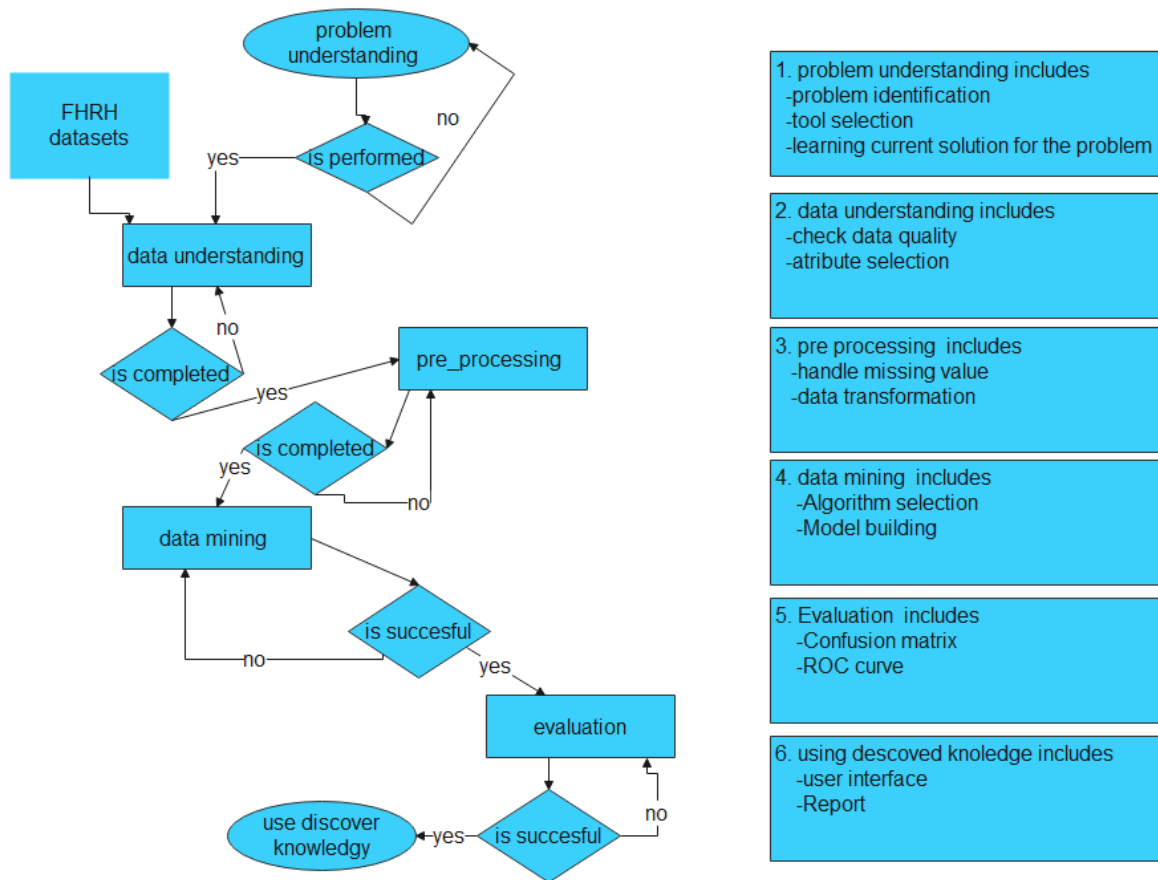


Figure 6: General adult mortality Prediction in clinical data work flow diagram

3.5. Methods of Analysis and Evaluation of System Performance

3.5.1. Methods of Training and Testing

The classifiers are evaluated by cross-validation using the number of folds. K-fold is a natural number used to check the performance of the model through k-times. K-Fold is appropriate whether the size of the data is very large or not. This is because of its extensive tests on numerous datasets with different learning schemes. It is also suggested that in k-fold algorithm, 10-fold is indicated about the right number of folds to get the best estimate of error (Weiss Sholom M. Z. T., 2003).

In 10-folds cross validation, the learning scheme or dataset is randomly reordered and then, split into „n“ folds of equal size. In each iteration fold one is used for testing and the

other n-1 folds are used for training the classifier. The test results are collected and averaged over all folds giving the estimate of accuracy for ten times. In percentage split evaluation the researcher attention, 70% of data is used for training purpose and the remaining 30% for validation of classifier accuracy. Therefore, k-folds minimize the bias effects by random sampling of the training and holdout data samples through repeating the experiments

3.5.2. Methods of Evaluation

After accomplishing model creation, comparing predictive accuracy of the classifiers for unknown tuples is often helpful to evaluate the performance of predictive modeling. It tells us how frequently instances of particular classes are correctly classified as actual class or misclassified as some other classes. To obtain reliable results, the extract knowledge is evaluating by a comparison of results obtain with various supervise classification methods and using several measurement. Performance evaluation of classifiers can be measure by hold-out, random sub-sampling, percentage split, cross-validation and bootstrap (Esfandiari N. B.-M., 2014). Furthermore, performance measures can use to analyze predictive models. They based on four values of the confusion matrix true positives (TP), false positives (FP), true negatives (TN), false negatives (FN).

3.5.2.1. Confusion matrix

A confusion matrix is a very useful tool for understanding results; Confusion matrix shows the counts of the actual versus predict class values. It shows not only how well the model predicts, but also presents the details needs to see exactly where things may do wrong. Give two-class, a confusion matrix may use to summarize the predictive

performance of a classifier on test data. It is commonly encountered in a two-class format, but can generate for any number of classes. Suppose we have a two-class problem with classes refer to as positive and negative.

Confusion matrix is useful tool for analyzing how well classifier recognized the classes. It is body of table with m by m (row and column) matrix the row corresponds to correct classification and the column corresponds to the predicted classifications. For a classifier to have good accuracy, ideally, most of the tuples would be represented along the diagonal of the confusion matrix with the rest of the entries being closed to zero.

In confusion matrix, there is classifier evaluation shows two class classification result simple confusion matrix which contains both predicted and actual classes.

		Observed	
		True	False
Predicted	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

Figure 7 : Confusion matrix

True positives (TP): These refer to the positive tuples that are correctly label by the classifier. Let TP be the number of true positives.

True negatives (TN): These are the negative tuples that are correctly label by the classifier. Let TN be the number of true negatives.

False positives (FP): These are the negative tuples that are incorrectly label as positive (e.g., tuples of class died = no for which the classifier predicted died= yes). Let FP be the number of false positives.

False negatives (FN): These are the positive tuples that are mislabel as negative (e.g., tuples of class alive = yes for which the classifier predicted alive= no). Let FN be the number of false negatives. Four performance measures are using:

Accuracy: is a rate of correct classification defining by:

$$(TP + TN) / (TP+TN+FP+FN)$$

The performance of the model enables it to classify the positive cases correctly is sensitivity. It is defined as the probability of having a positive test result among those with a positive diagnosis for the disease or true Positive recognition rate (Shelly, 2011).

$$\text{True Positive Rate (sensitivity)} = TP / (TP+ FN)$$

The performance of the model to classify the negative cases is specificity. It is defined as the probability of having a negative test result among those with a negative diagnosis for the disease or true negative recognition rate:

$$\text{True Negative Rate (specificity) or Recall for False class} = TN / (TN + FP)$$

Recall is what percent of positive tuples the classifier labeled as positive for both True and False classes (alive and died). Another detailed performance measure for the classifier is precision which measures what percent of tuples that the classifier labeled as positive are actually positive:

$$\text{Precision} = TP / TP+FP \text{-----For True Class}$$

$$\text{Precision} = TN / TN+FN \text{-----For False Class}$$

Recall: is the evaluation and ranking of each sample based on positive class defining by:

$$TP / (TP +FN)$$

Finally, the F measure is the inverse relationship between precision & recall (F1 or F-score): harmonic mean of precision and recall. It is the point to conclude that the precision and recall of the model are significantly balanced (T, 2003).

$$F\text{-Measure} = \frac{2 \times \text{Precision} \times \text{recall}}{\text{Precision} + \text{Recall}}$$

Kappa statistic: measures the agreement of prediction with the true class labels. A score value of 1.0 signifies complete agreement, and a value greater than 0 means that the classifier is doing better than pure random behavior [69].

3.5.2.2. Receiver Operating Characteristic (ROC) Curve

A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) show great progress and they are being developed, practically to aid clinician and to improve patient care in areas such as diagnosis, prognosis, decision support and screening. To test which classifier is highly significant for a given subject is determined by ROC analysis and it becoming widely used tool in medical tests evaluation (Ifeakor C E, 2004).

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified. For example, it can be used to classify adults those who are alive and died correctly based on their previous history.

ROC curve is a useful visual tool for comparing classification models. It shows the trade-off between the true positive (Ifeakor C E, 2004) rate (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model.

Table 13: Performance Measures of ROC Area

ROC Area	Performance
0.9-1.0	Excellent(A)
0.8-0.9	Good (B)
0.7-0.8	Fair (C)
0.6-0.7	Poor (D)
0.5-0.6	Fail(F)

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00). By using ROC analysis one can identify predictors in order to find the one with optimal characteristics and their associated cut-points. Therefore, sensitivity, specificity, precision, F-measure, and ROC area are taken into account when the classifier performance is evaluated.

3.5.3. Selection of Modeling Techniques

Selecting an appropriate model depends on the main goal of the problem to be solved and the structure of the available data (et.al, 2008). Consequently, to attain the objectives of these research four classification techniques have been selected for model building. The analysis was performed using WEKA experimental environment and Explorer. Among the different available classification algorithms in WEKA Naïve Bayes, J48, KNN, SVM and Random Tree are used for experimentation of this study. The researcher selected the above algorithms, easy of understanding and interpretation of the result of the model and appropriateness for adult mortality prediction.

CHAPTER FOUR

EXPERIMENTATION, RESULT AND DISCUSSION

4.1. OVERVIEW OF EXPERIMENTATION

In this study different experiments were conducted using various data mining methods to derive knowledge from preprocessed data to predict unseen data of adults. According to the previous chapter of this study after preparation of the data, the next task is the mining process. As it has been stated in the previous sections, from the both classes alive and died dataset were applied for experimentation. For experimentation, different algorithms were employed considering different parameters for model building such as pruning or unpruning and testing model performance in both training and testing phases as depicted in Table 14.

Table 14: Experiments and Scenarios

Experiments (1-4)	Scenarios (1-5)
Decision tree	J48 decision tree model
Naïve Bayes	Naïve Bayes algorithm
Artificial neural network	Artificial neural network
K - Nearest Neighbor	K - Nearest Neighbor
Random Tree	Random Tree

To build predictive model, 7095 instances and 9 attributes are used through for all algorithm. The models generated with 9 attributes are compared with each scenario. In this study an attempt is made to design a model that enables to predict the adult mortality in hospitals. Hospital admission data is consulting to extract the dataset require for

training and evaluating the models create by the classifiers. For creating prediction model a total size of 7095 datasets are using for training and testing. The researcher used 10-fold cross validation and percentage split to measure the various performances of the classifiers. In 10-fold cross validation, the data is divided in 10 segments, where 9 segments are using in the training phase, and the remaining segment is used in the test phase to measure the performance of the model. This process is repeat 10 times, each time with a different set of 9 segments as the training set, and the remaining segment as the test set. The overall performance measures of the model are then average out over the 10 different runs.

In percentage split different values are taken to measure performance of model. If the values of percentage split are 70%, then 70% for training and the remaining 30% is for testing. Selected predictive model build was based on the hospital patient admission registered data that has been preprocessed and introduce to the Weka tools.

Table 14: J49 Classifier Parameter Options

Parameters	Descriptions	Parameter type
binarySplit	When to use binary split on nominal attributes when building the trees.	Boolean
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning).	Numeric
NumFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	Numeric
saveInstanceData	Whether to save the training data for visualization.	Boolean
Seed	The seed used for randomizing the data when reduced-error pruning is used.	Numeric
Unpruned	Whether pruning is performed.	Boolean
ReplaceMissingValue	Used to replace missing value in training data	numeric
Resample	Used to resample data for training and testing	

4.2. Model Building

To build the model, four different experiments are conducted using J48 decision tree and Bayes classifier (Naïve Bayes), Support vector machine (SVM) and k-nearest Neighbor (KNN) and Random Tree (RT). The intention here is investigated the effect of attribute on classification accuracy as well as model complexity and decision tree size classifiers. The second algorithm Naïve Bayes, Random tree, support vector machine; K-nearest neighbors are evaluated on model performance with selected attributes. To get the best performance of the model, the researcher conducted using different cross validation values on both training and testing schemes operated next experiment.

4.2.1. Measurement of model on cross validation fold

In this method every 10 folds check its measurement except fold 1. Fold 1 used for test only. In this experiment J48 decision tree, Naive Bayes, SVM, and k-nearest network (KNN) and Random Tree (RT) performance measurement is performed.

Table 15: Performance of the Classifier for Different K-Values by classifier

Performance measure	Number of cross validation folds (K values) in %								
	2	3	4	5	6	7	9	9	10
J48 decision tree	81.96	83.86	85.14	86.00	85.84	85.54	86.13	86.21	86.77
Naïve Bayes	78.44	79.53	79.98	80.15	80.31	80.34	80.38	80.68	81.49
KNN	79.93	83.24	84.92	86.20	86.47	86.71	87.23	87.39	88.27
SVM	81.61	81.6	82.01	80.15	81.27	82.45	83.21	83.5	83.81
RT	81.01	83.62	85.72	85.33	84.79	87.15	86.17	86.03	86.89

As indicated in table 15, though the performance variations among different k values are described for decision tree algorithms, Naïve Bayes, KNN,SVM and RT. Higher performance is observed at 10th fold cross validation to J48 decision tree algorithm and RT are 86.77% and 86.89% respectively. Naïve Bayes and KNN are score 81.49% and 88.27% in 10th fold cross validation respectively and also SVM score 83.81% accuracy at 10th fold.

In the experiment KNN is score 88.27% accuracy which means 88.27 % of data correctly classified into right class and 11.73% of data is classified into wrong class. In other ways KNN is performed minimal error to class instance in wrong class rather than J48 decision tree, Naïve Bay, SVM and RT. The results obtained from above table classifier were summarized in Table 16 with their respective performance measures detailed.

Table 16: Detail performance measures in cross validation

Experiment	Model	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Accuracy (%)
I.	J48 decision tree	0.871	0.263	0.868	0.871	0.869	0.631	0.919	0.920	86.77
II.	Naïve Bayes	0.815	0.430	0.801	0.815	0.803	0.439	0.840	0.867	81.49
III.	KNN	0.883	0.165	0.889	0.883	0.885	0.690	0.858	0.859	88.27
IV.	SVM	0.838	0.476	0.39	0.838	0.814	0.497	0.681	0.749	83.81
V.	RT	0.869	0.243	0.868	0.869	0.868	0.634	0.835	0.844	86.89

As we have seen from table 16, five experiments were executed. The first experiment used J48 decision tree classifier; various results were acquired during the experiments. The result of J48 decision tree shows that the model correctly classified 87.12% instances, while 12.98% of the instances were classified incorrectly. Naïve Bayes, KNN 81.49% and 88.27 % respectively classifies instance correctly and naïve Bayes 18.51% and KNN 11.73% classify instance incorrectly. A SVM and RT score 83.81% and 86.89 % instance correctly classified 16.19% and 13.11% incorrectly classified respectively.

With regards to TP rate, J48 decision tree score of 0.871%, Naïve Bayes and KNN were score 0.815 and 0.883 TP rate respectively. Naïve Bayes score 0.165 FP rate and KNN scores 0.165 FP rate. SVM and RT perform 0.838 % TP rate, 0.476 % FP rate and 0.869 % TP, 0.243FP rate respectively.

From the experiment KNN score higher TP and low FP rate when we compare from others. KNN score 0.631 MCC for both class Alive and Died which means the measure of MCC for both class Alive and Died were appropriate and it approach to 1 the target class has better correlation coefficient. Therefore, KNN is better performance measure rather than others classifiers.

Mean absolute error and Root mean squared error and time processed for J48 (0.1287, 0.305, 0.06 second) respectively, KNN (0.1177, 0.3426, 0 second) respectively, SVM (0.1619, 0.4024, 397.08 second) respectively, Naïve Bayes (0.2562, 0.3598, 0 second) respectively and RT (0.1434, 0.3531, 0.02 second) respectively. This shown KNN, J48, RT, SVM and Naïve Bayes were better scored in Mean absolute error. Thus KNN is better than others with mean absolute error 0.1177 value. J48 perform 0.305 Root mean

squared error and KNN 0.3426 Root mean squared error. So, J48 is better than KNN in Root mean squared error values. And KNN took the time to processed 0 second but J48 taken time 0.02 second to process. In most processed KNN is performed well rather than others.

The next task in experiment the model to decide which one of the models constitutes a better model/classifier of the data is evaluated using ROC analysis. This analysis in which the curve the more to the upper left would indicate a better (INDRAYAN, 2014). Here in our case, the ROC area performance of the algorithms show that J48 decision tree model score the area 91.9 %. In experiment III and IV score 84.00% and 85.8% respectively. Experiment IV and V perform 0.681% and 0.835% respectively. As we show this result J48 decision tree model is high score area than other model in the experiment (appendix 1).

As a whole KNN is performed better in number measurement but J48 decision tree is the second better performed model in the experiment. J48 decision tree model and KNN classifier are appeared with competent predictive performance for adult mortality. From five the experimented, both models expose the better performance in predicting True positive cases or sensitivity (alive); than predictive performance of false positive case or specificity (died). This is the fact due to the model committing a bias to majority class over the minority class (alive and died) respectively. The KNN model shows an experiment is capable in adult mortality predicting with performance of 88.27% accuracy. As shown in figure 8, x-axis represent algorithm in different measurement and y-axis represent value of algorithm in different measurements.

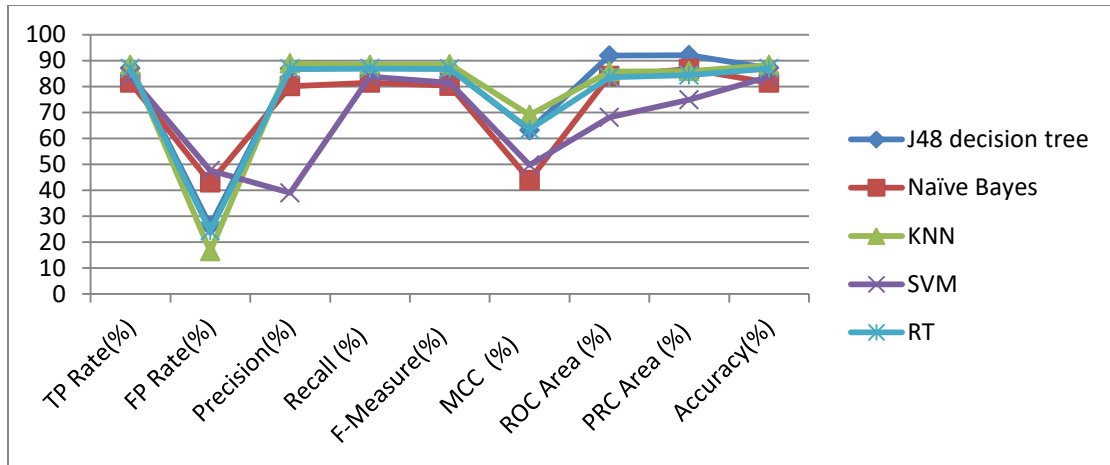


Figure 8: model comparison of cross validation with different classifier

4.2.2. Percentage Split model measurement with different classifier

In this measurement J48 decision tree, Naïve Bayes, KNN, SVM and RT are performed as classifier. In percentage split experiment 70 percent of instance from the whole dataset for training and the remaining 30 percent of instance from dataset used for test is performed experiment.

Table 17: Detail performance measures in percentage split

Model	TP Rate	FP Rate	Precision	Recall	F- Measure	MCC	ROC Area	Accuracy y (%)
J48 decision tree	0.861	0.334	0.855	0.861	0.854	0.586	0.890	86.09
Naïve Bayes	0.805	0.469	0.788	0.805	0.789	0.395	0.831	80.45
KNN	0.860	0.233	0.863	0.860	0.862	0.618	0.813	86.04
SVM	0.825	0.511	0.825	0.825	0.797	0.445	0.657	82.51
RT	0.861	0.311	0.855	0.861	0.856	0.594	0.830	86.14

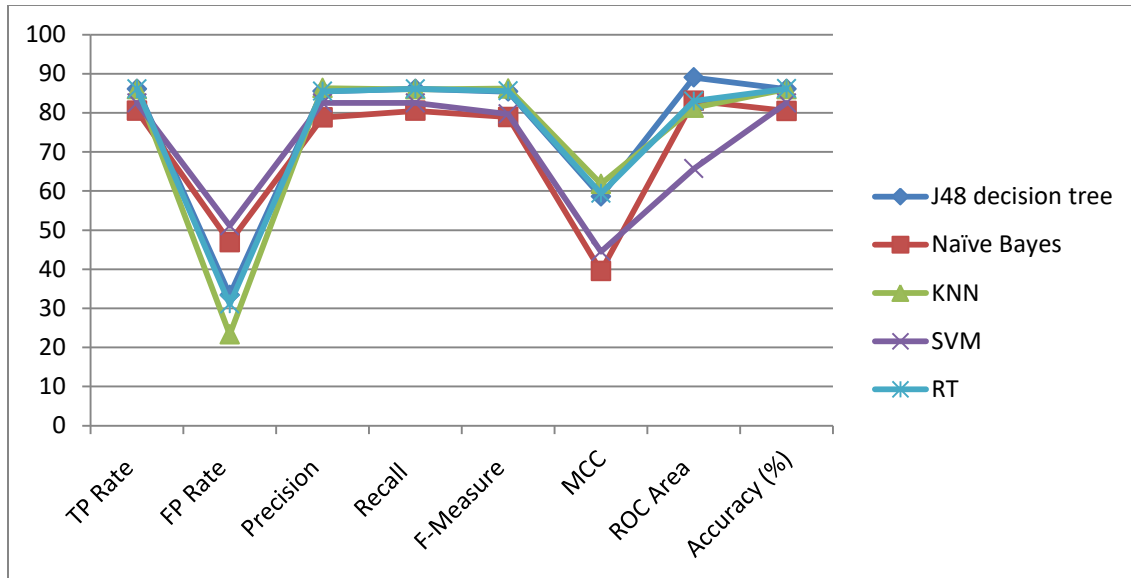


Figure 9: percentage split comparison with different classifier

As shown in figure 9, y-axis indicates values of algorithm in different measurement and x-axis indicates the algorithms with different measurement. And table 17 states, J48 decision tree algorithm performed 86.09% accuracy, Naïve Bayes is performed 80.45% accuracy, and KNN also performed 86.04% accuracy. SVM and RT performed 82.51% and 86.14% respectively.

In this experiment J48 decision tree score higher next to RT, accuracy is J48 scored 86.09% performed and RT scored 86.14% accuracy performed which means 86.09 and 86.14 percent of instance classify correctly and 13.91% and 13.86% of instance classified incorrectly respectively for J48 decision tree and RT. KNN algorithm score better accuracy next to J48 decision tree and RT with score 86.04% accuracy which means 86.04 percent of instance classified correctly and 13.96 percent of instance classified incorrectly. To comparison each classifier based on TP, FP and ROC were as follows, RT score 86.1% TP rate, 31.1 % FP rate and J48 decision tree score 86.1% TP rate, 33.4 % FP rate. Naïve Bayes score 80.5% TP rate and 46.9% FP rate and SVM score 82.5%

TP rate, 51.1% FP rate and KNN score 86.0% TP rate, 0.233 FP rate. When we compare each other J48 decision tree and RT are better to perform true classifier and minimal classification error than other but RT is better than J48 decision tree according to minimize error classification in relative way. Next to J48 decision tree and RT, KNN is appropriate classifier for prediction of adult mortality. ROC area of J48 decision tree 0.89 and ROC area for Naïve Bayes is 0.831 which is less than 0.89 and also KNN ROC area score 0.813 which is less than from the others ROC area values. SVM 0.657 ROC area score and RT 0.83 ROC area, so both RT and SVM performed ROC area less than KNN performed. Summarily KNN is scored better accuracy, TP rate, FP rate and MCC next to J48 decision tree.

As we show in this experiment J48 decision tree score high accuracy and better to minimize error classification of instance in wrong classes. J48 decision tree is score better accuracy with value of 86.09% in percentage split measurement.

Table 18: Accuracy Comparison Cross validation VS percentage split

No.	Performance measure	Cross validation	Percentage split
1	J48 decision tree	86.77%	86.09%
2	Naïve Bayes	81.49%	80.45%
3	KNN	88.27%	86.04%
4	SVM	83.81%	82.51%
5	RT	86.89%	86.14%

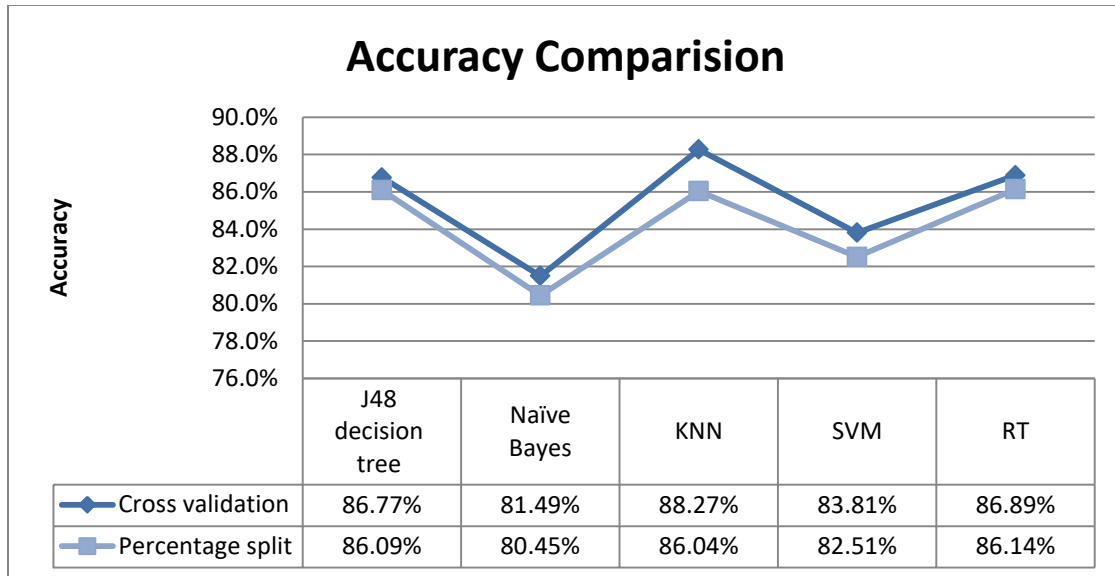


Figure 10: Accuracy comparison of cross validation and percentage split

Figure 10 shown as x-axis stated algorithm and y-axis stated also accuracy of algorithm. As indicated in table 18, KNN shows better performance 88.27% accuracy in cross validation measurement and 86.04% in percentage split. J48 decision tree with cross validation measurement is performed 86.77% and also 86.09% perform in percentage split. Random tree performed 86.89% in cross validation and 86.14% in Percentage split. From this summery table KNN, RT and decision tree score high performance accuracy than other experimented algorithm in cross validation with value 88.27%, 86.89% and 86.77% accuracy respectively. Both algorithm KNN, RT and J48 decision tree score appropriate accuracy for both measurement cross validation and percentage split. In summery cross validation measurement methods is better measurement than percentage split measurement methods to predict adult mortality with KNN algorithm based on this experiment.

4.2.3. Confusion matrix

Table 19: Confusion matrix of cross validation Measurements in different algorithm

Model s		Actual class	Prediction class	
			Alive	Died
Model 1	J48 decision tree	Alive	5047(93.1%)	374(6.9%)
		Died	540(32.3%)	1134(67.7%)
Model 2	Naïve Bayes	Actual class	Predicted class	
			Alive	Died
		Alive	5012(92.46%)	409(7.54%)
		Died	904(54%)	770(46%)
Model 3	KNN	Actual class	Predicted class	
			Alive	Died
		Alive	4902(90.4%)	519(9.6%)
		Died	313(18.7%)	1361(81.3%)
Model 4	SVM	Actual class	Predicted class	
			Alive	Died
		Alive	5304(97.8%)	117(2.2%)
		Died	1032(61.6%)	642(38.4%)
Model 5	RT	Actual	Prediction	
			Alive	Died 1682
		Alive	4976(91.9%)	437(8.1%)
		Died	493(29.3%)	1189(70.7%)

As we have seen in table 19, the first model of the number of True-positives correctly classified instances are 5047(93.1%) adult out of 5421 adult as Alive and the remaining 374(6.9%) adult were incorrectly classified as Died while in fact they belong to Alive. This model also correctly classified/True-negative/ 1134 (67.7%) adult Died and the remaining 540(32.3%) adult were incorrectly classified as Alive which should be categorized as Died.

The second model the number of True-positives correctly classified instances are 5012(92.46%) adult out of 5421 adult as Alive and the remaining 409(7.54%) adult were incorrectly classified as Died while in fact they belong to Alive. This model also correctly classified/True-negative/ 770 (46%) adult Died and the remaining 904(54%) adult were incorrectly classified as Alive which should be categorized as Died.

The third model the number of True-positives correctly classified instances are 4902(90.4%) adult out of 5421 adult as Alive and the remaining 519(9.6%) adult were incorrectly classified as Died while in fact they belong to Alive. This model also correctly classified/True-negative/ 1361 (81.3%) adult Died and the remaining 313(18.7%) adult were incorrectly classified as Alive which should be belong to as Died.

The fourth model the number of True-positives correctly classified instances are 5304 (97.8%) adult out of 5421 adult as Alive and the remaining 117(2.2%) adult were incorrectly classified as Died while in fact they belong to Alive. This model also correctly classified/True-negative/ 642(38.4%) adult Died and the remaining 1032(61.6%) adult were incorrectly classified as Alive which should be belong to as Died.

The fifth model the number of True-positives correctly classified instances are 4976 (91.9%) adult out of 5421 adult as Alive and the remaining 437(8.1%) adult were incorrectly classified as Died while in fact they belong to Alive. This model also correctly classified/True-negative/ 1189(70.7%) adult Died and the remaining 493(29.3%) adult were incorrectly classified as Alive which should be belong to as Died. In general, J48 decision tree, Naïve Bayes, KNN, support vector machine and RT are appeared with competent predictive performance for adult mortality prediction according to confusion measurement. From all the experimented, all models expose the related performance in predicting True positive cases or sensitivity (alive); than predictive performance of false positive case or specificity (died). This is the fact due to the model committing a bias to majority class over the minority class (alive and died) respectively. From these output, based on research objective we identify the cause of adult mortality, due to this the fourth model is performed better classification for class alive and Died class rather than other model. Therefore KNN model is better model in terms of minimized incorrectly classified instances.

The predictive performances of the models were also compared using different performance measures Cross validation fold, Percentage split and confusion matrix in different criteria with different classifier algorithm. As this experiment output K-nearest neighbor (KNN) is better model prediction algorithm for adult mortality prediction with 88.27% accuracy.

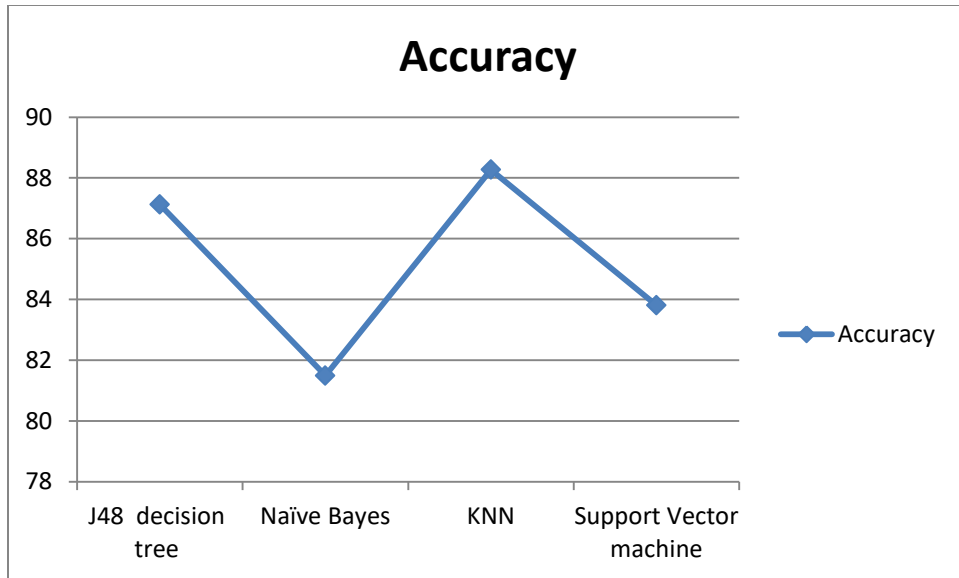


Figure 11: Accuracy of All Models in the Experiments

4.3. Selected model performance and evaluations

The full training set containing a total of 7095 instances are used in 9 attributes. In addition to above performance metrics used, KNN model generation with 9 attribute is more understandable and less complex to human than others model generated. Therefore, the performance of KNN model valuable information in predicting adult mortality as compared to other models. The following Figure 12 shows the output of KNN model.

```

Time taken to build model: 0 seconds

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

Correctly Classified Instances      6263           88.2734 %
Incorrectly Classified Instances    832            11.7266 %
Kappa statistic                    0.688
Mean absolute error                 0.1177
Root mean squared error            0.3426
Relative absolute error            32.645 %
Root relative squared error        80.6942 %
Total Number of Instances         7095

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.904   0.187   0.940     0.904   0.922     0.690   0.858    0.926    Alive
                0.813   0.096   0.724     0.813   0.766     0.690   0.858    0.642    Died
Weighted Avg.   0.883   0.165   0.889     0.883   0.885     0.690   0.858    0.859

=== Confusion Matrix ===

  a  b  <-- classified as
4902 519 |  a = Alive
 313 1361 |  b = Died

```

Figure 12: KNN model output

As we shown in figure 12, KNN predictive model, the predictive performance of the model is 88.27%.As we shown in the matrix 4902 instances are classified correctly while 519 instances are wrongly misclassified to other class. The model classified 4902 instances as alive out of 5421 instances that in fact they are alive as tested on the test data or which are classified correctly in the class of alive. The remaining 519 instances are misclassified to another class as died actually they are alive.

The model classified 1361 instances as died out of 1674 instances that in fact died and wrongly classified 313 instances to other class as alive while actually they had died. The model has a good performance in classifying the instances in True class (TP) than false positive class (FP) (alive and died) with predictive performance of 90.4% and 18.7%

respectively. Thus, it is possible to conclude that the model is in a good performance to classify True positive than false positive.

From the precision score of the model, the precision of this model for „alive“ class is a bit higher than precision of „died“ class (94% and 72.4%) respectively. With an average precision of 88.9%, instances labeled as belonging for each class alive and class died. From harmonic mean of precision and recall which is F-score, with value 88.5%. it can be concluded that the precision and recall of the model are relatively balanced.

4.4. Discussion of the study

In this study, comparison of classification techniques is performed. Five experiments have been conducted using four data mining classification algorithms i.e. J48 decision tree algorithm, Naïve Bayes, KNN, SVM and RT in order to build a model that predicts adult mortality status. Furthermore, different experiments are done using 70 % split and cross validation. The experimental result shows that with 70% percentage split decision tree and IBK achieved better result. J48 Decision tree produce 86.09% accuracy followed by IBK with 86.04% accuracy. IBK produce 88.27% accuracy followed by J48 decision tree with 87.12% accuracy in cross validation. The goal in this study is to explore patient dataset in order to build the model that can predict adult mortality and to discover knowledge which is not interpreted in human being under clinical dataset, finally to notice the attributes that existed as determinant factors of adult mortality.

KNN model is selected with its performance that could predict status of adult (alive, died) i.e. 88.27% accurate prediction with the respective (True positive and True Negative) with the lower mean absolute error (0.1177) which measures the error between actual and predicted value and with high kappa statistic measures (0.688); it is usually 1.0

which implies complete agreement. In this study, the model created using KNN model registers good performance and hence selected for further analysis.

Hence, IBK algorithm selected for further adult mortality prediction. In general, based on information gain ranker experiment NCOD, age, sex, du_of_ill, LOS and payment status are the major attributes to predict Adult mortality. However, based on domain expert suggestion, Address and disease group is not necessarily relevant for Adult mortality prediction model building rather its relevant for trend analysis. Beyond this, the data and method used in this study can utilize as baseline in future related researches.

4.5. Rule extraction by using Random tree algorithm

To make a decision rule to human-readable each path from root to leaf can be transformed into an IF-THEN rule. If the condition is satisfied, the conclusion follows. The algorithm Random Tree is known method for deriving rules from classification trees. In Random tree simply traversing any given path from the root node to any leaf. The numbers at the end of each leaf indicates the number of value in the leaves. The number of misclassified value would also be given, after a slash and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong.

From the entire models that are generated, Random tree model is selected as the best model for rule generation. This is due to the rules provided by Random tree models can be easily assimilated by human without any difficulty. Random tree model produced different rules. However, the researcher selected best rules that cover most of the data points in the study. The partial Random tree generated is presented (appendix 2).

After the rule extraction, the researcher turns back to domain experts to discuss up on the generated rules. Some of the rules generated by the selected models are:

Rule #1: NCoD = Urinary Tract Infection and Du_of_ill < 2011 and LOS >= 4 and Age >= 53.5 and Age < 56.5: Died (1/0)

Rule #2: NCoD = Urinary Tract Infection and Du_of_ill < 511 and LOS < 3.5 and Sex = M and Age >= 26.5 and Age < 30.5: Died (1/0)

Rule 1 and 2 state that NCoD is more respected attributes to predict and consider adult mortality with other attribute such as Du_of_ill, LOS and Sex and Age. In this rule shows Urinary Tract Infection disease is main factor to decide adult mortality based on duration illness and age.

Rule #3: NCoD = Tuberculosis and LOS < 7.5 and Du_of_ill >= 59: Died (3.0/2.0)

Rule 3 states that NCoD, LOS and duration of illness is a major considerable attribute adult mortality in the case of Tuberculosis disease. Tuberculosis is major factor of adult mortality when duration of illness >59 days and assessed in hospital under 8 days, then the patient may be died

Rule #4: NCoD = Road traffic accident (Person injured in unspecified vehicle accident) and payment_status = assured and Sex = M and Age >= 26.5 and Du_of_ill >= 5 and LOS > 2.5: Died (2/0)

Rule 4 states that NCoD, Sex, Payment status, Age, LOS and Du_of_ill are considerable attribute for adult mortality in the cause of Road traffic accident case. If the case is traffic accident duration of illness/accident >=5 and Age >26 and duration of medical process >2 but the patient is not payable, then patient may be died

Rule #5: NCoD = Renal Disease (Chronic kidney disease) and LOS < 12.5 and Sex = M and Du_of_ill >= 186 and Du_of_ill <= 2511.5: Died (1/0)

Rule #6: NCoD = Poisoning and Age >= 44 and Sex = F and LOS >= 2.5 and Du_of_ill < 4.5: Died (4/0)

Rule 6 states that if the case is Poisoning Age, Sex, LOS and duration illness is more considerable attribute. For Poisoning disease duration of illness >4.5 and Age > 44 and LOS >2.5, then the patient may be died

Rule #7: NCoD = Congestive heart failure and Age >= 36.5 and Du_of_ill >= 158: Died (1/0)

Rule #8: NCoD = Congestive heart failure and Age >= 36.5 and Du_of_ill >= 158: Died (1/0)

As rule 7 and 8 shows that NCoD, Age, Duration of illness are major considerable attribute of adult mortality in the case of Congestive heart failure disease. Patient stay more than 158 day without medication treatment with Age >36, then the patient will be died.

As shown rule extraction NCoD is major attribute for considerable adult mortality. Finally, NCoD, Duration of illness, LOS, sex and Age are more determinant Attribute of adult mortality. As rule extraction shown Urinary Tract Infection, tuberculosis, congestive heart failure; renal disease, Road traffic accident and Poisoning are needed more attention to minimize adult mortality.

4.6. Using Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. This step is to put discovered knowledge in practical use either by documenting it and reporting it or by embedding it in a computer system. The conclusions drawn from the KDD process often reveal the complex nature of the problem and its solutions. Hence, the implementation of the new knowledge should often be done gradually, while continuously monitoring the result achieved and the degree to which they fulfill the expectations.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

Data mining is extracting meaningful patterns and rules from large quantities of data. It is clearly useful in any field where there are large quantities of data and something worth learning. In this respect, widespread use of medical information systems and explosive growth of medical data records require manual data analysis to be coupling with methods for efficient computer-assist analysis.

This research investigates the potential applicability of data mining technology in improving healthcare resource management and developing a model to predict the occurrence of mortality case in hospitals. This investigation was conducted according to the KDD process model. The data is collected from hospital patient admission dataset from 15/05/2009 to 30/10/2011 for the research purpose. Analyzing the large volume of patient data and extracting useful information and knowledge for decision making about death of adult is done. First the data is preprocessed for data cleaning, attribute and feature selection, and data transformation. This experimental research, which engages a KDD methodological approach and use predictive modeling techniques, in the result of experiment KNN model to address the problem. The experiment result shows that KNN model suitable for predicting adult mortality. Hence KNN model with 88.27% accuracy prediction model building is selected to extract interesting pattern to mention.

Knowledge gain with the use of techniques of data mining can be using to make successful decisions that will improve success of healthcare care organization and health

of the patients. In addition, results from this study show that the problem of adult mortality can be supporting by the use of data mining, in particular KNN model technique. Moreover, further extensive experiments at district level and using various data sources available with those organizations working in relate public health will enhance the result obtain in this study.

Finally, NCOD, Du_of_ill, sex, LOS are more determinant attribute of adult mortality prediction. As rule extraction shown Urinary Tract Infection, tuberculosis, congestive heart failure; renal disease, Road traffic accident and Poisoning are needed more attention to minimize adult mortality.

5.2. Recommendation

This research work is caring out for academic purpose and it reveal the potential applicability of data mining technology to improve healthcare and reduce adult mortality through developing a model to predict adult mortality using patient dataset in hospital. Accordingly, based on the findings of this research work, the researcher forwards the following recommendations for future work particularly in relation to the possible application of data mining technology in supporting the effort to efficient health care management

- In this study, an attempt has been made to assess the applicability of data mining technology to predict the probability of mortality for adult by using some set of important variables that are considered by experts. For a number of other variables, however, it remains to investigate further the effect of those variables to build models will be better accuracy and performance than the models built in this study.

- This research work has used a small percentage of the overall dataset to build Naïve Bayes, decision tree models and support vector machine, J48 decision tree and KNN is appropriate to build more all-inclusive models by using large training and testing datasets taken from the main dataset, so research with more huge data will be done further result that different from this study.
- All Naïve Bayes, decision tree, KNN and Support vector machine approaches resulted in an encouraging output, the application of other data mining techniques such as Cluster, association rule and other also prove to be important techniques in the healthcare sectors. it is recommended that other data mining techniques should also be tested to see if they could be more applicable to the problem domain
- This research work, data collected from hospital dataset but the database is not fully structured to generate data for research. Since hospital database used only for report format that accept only numbering data and not include all variable of this research has been done. We recommended researcher could work framework for possible output of performance with further variable. It became better performance of healthcare area and sufficient clinical data for researchers.
- Health care organization who concern health issue should be attention for Urinary Tract Infection, tuberculosis, congestive heart failure; renal disease and asthma are need more treatment ,found the source of them and provide community awareness to minimize adult mortality.

References

- (US), N. A. (2010). *Clinical Data as the Basic Staple of Health Learning*. washington(DC): Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care.
- Addisalem Tebeje Zewudie, A. A. (2020). Determinants of Under-Five Child Mortality in Ethiopia: Analysis Using Ethiopian Demographic Health Survey, 2016. *international journal of pediatrics*.
- Arora, B. a. (2012). Classification and Feature Selection Techniques in Data Mining. *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, 1-6.
- Awoke Misganaw, D. H. (2012). The Double Mortality Burden Among Adults in Addis Ababa, Ethiopia, 2006-2009. *public health research, practice and policy*.
- Azari, A. J. (2012). Predicting hospital length of stay (PHLOS). *2012 IEEE 12th International Conference on*.
- Benbelkacem, S. K. (2014). (DATA MINING-BASED APPROACH TO PREDICT STRAIN SITUATIONS IN HOSPITAL EMERGENCY DEPARTMENT SYSTEMS. *10ème Conférence Francophone de Modélisation*.
- Bhargava, N. (2013). Selection Criteria for Data Mining Software: A Study. *International Journal of Computer Science Issues*, vol. 10, no. 2, 308-312.
- Breiman, L. (2003). Using And Understanding Random Forests. *University of California, Berkeley. Department of Statistics*, .

- Chhabra, K. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications, Vol. 98, No.22, 13-17.*
- Coelho. (2005). Forecast Calibration and Combination: Bayesian Assimilation of Seasonal Climate Predictions. *PhD. Dissertation, dept. Meteorology.*
- D., A. (2011). Application of Data Mining Techniques to Predict Household Health Seeking Patterns: The Case of BRHP. *Addis Ababa University.*
- Delen, D. (2014). Real-World Data Mining: Applied Business Analytics and Decision Making. *FT Press.*
- Duthe G, P. G. (2009). Adult mortality in a rural area of Senegal: Non-communicable diseases have a large impact in Mlomp. *Max Planck Institute for Demographic Research, GERMANY.*
- E., C. (1997). Bayesian Networks without Tears. *Publication of the American Association for Artificial Intelligence.*
- Esfandiari, N. B.-M. (2014). Knowledge discovery in medicine. *Current issue and future trend. Expert Systems with Applications, 41(9), 4434 - 4463.*
- et.a, C. (2015). Knowledge Discovery Process: The Next Step for Knowledge Search. *International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp., 4277-4283.*
- et.al, K. G. (2008). On the role of pre and post-processing in environmental data mining. *International Congress on Environmental Modeling and Software, Modeling for Environment's Sake, Fifth Biennial Meeting.*

- F., M. (2009). Mortality and Survival from Childhood to Old Age in Rural Ethiopia. *Umeå University Medical Dissertations, Sweden, Umeå University, SE-901 97 Umeå.*
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (2009). "From Data Mining to Knowledge Discovery in Databases" . . 6. .
- HAILEMARIAM, T. (2012). APPLICATION OF DATA MINING FOR PREDICTING ADULT MORTALITY . *ADDIS ABABA UNIVERSITY, health informatics.*
- Hsiang-Yang Chen, C.-H. C.-J.-P. (2011). Exploring the risk factors of preterm birth using data mining. *Journal of Expert Systems with Applications, Exploring the risk factors of preterm birth using data mining .*
- Ifeachor C E, H. B. (2004). Receiver Operating Curve Analysis in The Evaluation of Intelligent Medical Systems. *UK: University of Plymouth. Drake Circus Plymouth PL4 9AA, Devon.*
- INDRAYAN, R. K. (2014). Receiver Operating Characteristic (ROC) Curve for Medical Researchers. *Neuropsychiatric Disease and Treatment, .*
- J B, M. K. (2003). Data Mining-Concepts, Models, Methods and Algorithms. *John Wiley & Sons Publication Inc.*
- J.Kaiser. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration, vol. 1, 42-51.*
- Jr., u. D. (2009). Data Mining in Healthcare: Current Application and Issues. *Thesis, Australia: Carnegie Mellon University;.*
- Kaur&Chhabra. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications, Vol. 98, No.2, 13-17 .*

- Kelly, P. R. (1998). High adult mortality in Lusaka. *Lancet* 351(9106), 883.
- Koh, H. C. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 65.
- Kristopher J. Krohn, T. A. (2015). National mortality burden due to communicable, non-communicable, and other diseases in Ethiopia, 1990–2015: findings from the Global Burden of Disease Study 2015. *Population Health Metrics*.
- Lan, H. F. (2011). Data mining: Practical machine learning tools and techniques. *Morgan Kaufman, Boston*.
- M. Hall, e. a. (n.d.). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, vol. 11, no. 1, pp. , 10-19. .
- M., H. J. (2001). Data Mining: Concepts and Techniques. 11.
- M.Divya, V. a. (2011). An Efficient Algorithm for Classification Rule Hiding. *International Journal of Computer Applications* vol. 33, no.3, 39-45.
- Mannila, H. (2002). Methods and Problems in data Mining.
- Mitike Molla, P. B. (2019). Mortality Decreases among Young Adults in Southern. *Centre for International Health, University of Bergen*.
- Ngai, E. H. (2011). The application of data mining techniques in financial fraud detection. *A classification framework and an academic review of literature*, 559-569.
- Qaise, U. S. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, vol. 12, no. 1 , 217-222.

- Qaiser, U. S. (2014). "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)" . *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp., 217-222.
- R.Kanth, N. a. (2014). Weather analysis of guntur district of andhra region using hybrid SVM data mining techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 3, 133-136.
- Shelly, G. D. (2011). Performance Analysis Of Various Data Mining Classification Techniques On Healthcare Data. *International Journal of Computer Science and Information Technology*, 3(4), 155-169.
- Shivan, V. a. (2008). Robust Outlier Detection Technique in Data Mining: A Univariate Approach. *Faculty of Engineering and Technology, Mody Institute of Technology and Science*, i, class lecture, Topic: "", Lakshmanagarh, Sikar, Rajasthan, India.
- Sundaram, S. a. (2012). Knowledge Discovery from Real Time Database using Data Mining Technique. *International Journal of Scientific and Research Publications*, vol. 2, pp., 1-3.
- VESPER H. CHISUMPA, C. O. (2019). MORTALITY IN SUB-SAHARAN AFRICA: WHAT IS KILLING ADULTS AGED 15-59 YEARS. *Vienna Institute of Demography Austrian Academy of Sciences*.
- Vuda, G. B. (2016). Application of Data Mining Techniques on Pre ART Data: The Case of Felege Hiwot Referral Hospita. *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)*, 1-9 .
- Weiss Sholom M., Z. T. (2003). Performance Analysis and Evaluation. *Lawerence Erlbaum Associates Inc*.

WHO. (2002). Revised Global Burden of Disease (GBD) Estimates. *World Health Organization*, Geneva.

Yemane B, S. W. (1999). Establishing an Epidemiological Field Laboratory in Rural Areas Potentials for Public Health Research and Intervention. *The Butajira Rural Health Program 1997-99. Ethiop J Health Dev*, 13:1-47.

Appendix 1: Outputs of the Classifiers in Experimentation

J48 Decision Tree classifier in cross validation

```
Time taken to build model: 0.06 seconds

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

Correctly Classified Instances      6181           87.1177 %
Incorrectly Classified Instances    914            12.8823 %
Kappa statistic                    0.63
Mean absolute error                 0.1287
Root mean squared error            0.305
Relative absolute error             35.6886 %
Root relative squared error        71.8374 %
Total Number of Instances          7095

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.931   0.323   0.903     0.931   0.917     0.631   0.919    0.962    Alive
                0.677   0.069   0.752     0.677   0.713     0.631   0.919    0.782    Died
Weighted Avg.   0.871   0.263   0.868     0.871   0.869     0.631   0.919    0.920

=== Confusion Matrix ===

  a    b  <-- classified as
5047  374 |  a = Alive
 540 1134 |  b = Died
```

J48 decision tree in percentage split

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances      1818           85.4323 %
Incorrectly Classified Instances    310            14.5677 %
Kappa statistic                    0.5711
Mean absolute error                 0.1623
Root mean squared error            0.3297
Relative absolute error             43.8261 %
Root relative squared error        76.5673 %
Total Number of Instances          2128

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.944   0.421   0.873     0.944   0.907     0.581   0.885    0.942    Alive
                0.579   0.056   0.771     0.579   0.662     0.581   0.885    0.726    Died
Weighted Avg.   0.854   0.331   0.848     0.854   0.847     0.581   0.885    0.889

=== Confusion Matrix ===

  a    b  <-- classified as
1515   90 |  a = Alive
 220  303 |  b = Died
```

Naïve Bayes in cross validation

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances      5782           81.494 %
Incorrectly Classified Instances    1313           18.506 %
Kappa statistic                    0.4283
Mean absolute error                 0.2562
Root mean squared error             0.3598
Relative absolute error              71.0385 %
Root relative squared error         84.732 %
Total Number of Instances          7095

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.925   0.540   0.847     0.925   0.884     0.439   0.840    0.944    Alive
                0.460   0.075   0.653     0.460   0.540     0.439   0.840    0.616    Died
Weighted Avg.   0.815   0.430   0.801     0.815   0.803     0.439   0.840    0.867

=== Confusion Matrix ===

  a    b  <-- classified as
5012  409 |  a = Alive
  904  770 |  b = Died
```

K-nearest Neighbors Cross validation

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances      6263           88.2734 %
Incorrectly Classified Instances     832           11.7266 %
Kappa statistic                    0.688
Mean absolute error                 0.1177
Root mean squared error             0.3426
Relative absolute error              32.645 %
Root relative squared error         80.6942 %
Total Number of Instances          7095

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.904   0.187   0.940     0.904   0.922     0.690   0.858    0.926    Alive
                0.813   0.096   0.724     0.813   0.766     0.690   0.858    0.642    Died
Weighted Avg.   0.883   0.165   0.889     0.883   0.885     0.690   0.858    0.859

=== Confusion Matrix ===

  a    b  <-- classified as
4902  519 |  a = Alive
  313 1361 |  b = Died
```

Support vector SVM in cross validation

Time taken to build model: 397.08 seconds

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

Correctly Classified Instances	5946	83.8055 %
Incorrectly Classified Instances	1149	16.1945 %
Kappa statistic	0.4462	
Mean absolute error	0.1619	
Root mean squared error	0.4024	
Relative absolute error	44.9112 %	
Root relative squared error	94.7804 %	
Total Number of Instances	7095	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.978	0.616	0.837	0.978	0.902	0.497	0.681	0.836	Alive
	0.384	0.022	0.846	0.384	0.528	0.497	0.681	0.470	Died
Weighted Avg.	0.838	0.476	0.839	0.838	0.814	0.497	0.681	0.749	

=== Confusion Matrix ===

```
  a   b  <-- classified as
5304 117 |   a = Alive
1032 642 |   b = Died
```

Random tree in cross validation

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances	6165	86.8922 %
Incorrectly Classified Instances	930	13.1078 %
Kappa statistic	0.6334	
Mean absolute error	0.1434	
Root mean squared error	0.3531	
Relative absolute error	39.6361 %	
Root relative squared error	83.0181 %	
Total Number of Instances	7095	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.919	0.293	0.910	0.919	0.915	0.634	0.835	0.911	Alive
	0.707	0.081	0.731	0.707	0.719	0.634	0.835	0.631	Died
Weighted Avg.	0.869	0.243	0.868	0.869	0.868	0.634	0.835	0.844	

=== Confusion Matrix ===

```
  a   b  <-- classified as
4976 437 |   a = Alive
493 1189 |   b = Died
```


Appendix 2: Partial Random Tree Rule Generation for Clinical Dataset

```
NCoD = Urinary Tract Infection (Urinary tract infection and site not specified)
| Address = dangla
| | LOS < 3
| | | Du_of_ill < 398.5 : Died (4/0)
| | | Du_of_ill >= 398.5 : Alive (1/0)
| Address = bdr
| | Du_of_ill < 1.5 : Died (3/0)
| | Du_of_ill >= 1.5
| | | LOS < 0.5 : Alive (4/0)
| | | LOS >= 0.5
| | | | Age < 59
| | | | | LOS < 1.5 : Died (2/0)
| | | | | LOS >= 1.5
| | | | | | LOS < 4
| | | | | | Du_of_ill < 2011
| | | | | | | Age < 53.5 : Alive (7/0)
| | | | | | | Age >= 53.5
| | | | | | | | Age < 56.5 : Died (1/0)
| | | | | | | | Age >= 56.5 : Alive (1/0)
| | | | | | | Du_of_ill >= 2011 : Died (1/0)
| | | | | | | | LOS >= 4
| | | | | | | | | LOS < 8 : Died (3/0)
| | | | | | | | | LOS >= 8 : Alive (2/0)
| | | | | | | Age >= 59 : Died (2/0)
| Address = enemey : Alive (2/0)
| Address = huletijunese : Died (2/0)
| Address = goncha
| | LOS < 3
```

Appendix 3: Outlier in weka

weka.filters.unsupervised.attribute.InterquartileRange

SYNOPSIS

A filter for detecting outliers and extreme values based on interquartile ranges. The filter skips the class attribute.

Outliers:

$$Q3 + OF * IQR < x \leq Q3 + EVF * IQR$$

Or

$$Q1 - EVF * IQR \leq x < Q1 - OF * IQR$$

Extreme values:

$$x > Q3 + EVF * IQR$$

Or

$$x < Q1 - EVF * IQR$$

Key:

$$Q1 = 25\% \text{ quartile}$$

Q3 = 75% quartile

IQR = Interquartile Range, difference between Q1 and Q3

OF = Outlier Factor

EVF = Extreme Value Factor

OPTIONS

ExtremeValuesFactor -- The factor for determining the thresholds for extreme values.

Outlier Factor -- The factor for determining the thresholds for outliers.

ExtremeValuesAsOutliers -- Whether to tag extreme values also as outliers.

Debug -- If set to true, filter may output additional info to the console.

DetectionPerAttribute -- Generates Outlier/Extreme Value attribute pair for each numeric attribute, not just a single pair for all numeric attributes together.

OutputOffsetMultiplier -- Generates an additional attribute 'Offset' that contains the multiplier the value is off the median: $value = median + 'multiplier' * IQR$

DoNotCheckCapabilities -- If set, the filter's capabilities are not checked before it is built. (Use with caution to reduce runtime.)

AttributeIndices -- Specify range of attributes to act on; this is a comma separated list of attribute indices, with "first" and "last" valid values; specify an inclusive range with "-", eg: "10 , 11 and last".

Appendix 4: Name of Consulted Domain Experts

No	Name	Sex	Field	Institute	Role in research
1	Dr.Melash Belachew	M	Medical Lecture and health researcher	BDU	Approve result, Data quality assurance
2	Mr.Minichil	M	HIT and Data analyst	Felege Hiwot hospital	Data quality assurance
3	Tizazu Gigar	M	Health Expert	ANRS health Office	Data quality assurance

