ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

PREDICTING TUBERCULOSIS TREATMENT OUTCOMES USING DATA MINING TECHNOLOGY

SAMSON KIFLOM

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

PREDICTING TUBERCULOSIS TREATMENT OUTCOMES USING DATA MINING TECHNOLOGY

A Thesis Submitted to the School of Graduate Studies of
Addis Ababa University in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Health Informatics

BY

SAMSON KIFLOM

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AND

SCHOOL OF PUBLIC HEALTH

PREDICTING TUBERCULOSIS TREATMENT OUTCOMES USING DATA MINING TECHNOLOGY

BY

SAMSON KIFLOM

	Name	**Title**	**Signature**	**Date**
1.	_____	Chairperson	_____	_____
2.	_____	Advisor(s)	_____	_____
3.	_____	Advisor(s)	_____	_____
4.	_____	Examiner	_____	_____
5.	_____	Examiner	_____	_____

DECLARATION

I declare that this thesis is my original work and has not been presented for a degree in any other university.

_____

Samson Kiflom

May 2013

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

_____

Dr. Alemayhu Mekonnen

May 2013

_____

Mr. Workshet Lamenew

May 2013

ACKNOWLEDGMENT

This thesis would not have been possible without the guidance and the help of few individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, my utmost gratitude goes to my two supervisors, Dr. Alemayhu Mekonnen (School of Public Health) and Mr. Workshet Lamenew (Information Science Department) for their advice, insight and careful comments. Dr. Million was my inspiration in doing research in data mining area. I also would like to express my gratitude to Mr. Tibebe Beshahe (Information Science Department) for his unreserved support whenever I needed it.

In my attempt to understand the domain area and the interpretation of the results, two key individuals were extended their supports. I am deeply grateful to Dr. Andargachew Kumssa (FMoH/ICAP) and Mr. Biruck Kebede (FMoH).

The difficult task of encoding a huge collection of data for the mining task was done by my student Trahas Taddesse. Your patience was amazing. Without your effort this work cannot be possible.

I owe many thanks to my classmates and all of my friends, especially Misganaw Taddese, Wondwossen Shiferaw, Mekonnen Mulugeta and Henok Teferi that always support and give attention for me to solve my problem. You always listen to me and give enjoyable studying environment.

At last, I would like to thank W/ro Meseret Ayano (Health Informatics Coordinator) for her encouragement and understanding in times of confusion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

TB	Tuberculosis
HIV	Human Immunodeficiency Virus
WHO	World Health Organization
FMoH	Federal Ministry of Health
HSDP	Health Sector Development Program
TBL	Tuberculosis Leprosy
DOT	Directly Observed Treatment/Therapy
MDG	Millennium Development Goals
CRISP	CRoss-Industry Standard Process
AAHB	Addis Ababa Health Bureau
SMOTE	Synthetic Minority Oversampling Technique
PTB	Pulmonary Tuberculosis
EPTB	Extra pulmonary tuberculosis
KDD	Knowledge Discovery in Databases
CHAID	Chi-squared Automatic Interaction Detection
CART	Classification and Regression Trees
SQL	Structured Query Language
FN	False Negative
FP	False Positive
WFP	Weighted False Positive
TN	True Negative
TP	True Positive
WTP	Weighted True Positive
ROC	Receiver Operating Characteristics
WROC	Weighted Operating Characteristics
AUC	Area Under Curve
WEKA	Waikato Environment for Knowledge Analysis
ART	Antiretroviral Therapy

CPT	Cotrimoxazol Preventive Therapy
SVM	Support Vector Machine
MLP	Multi layer perceptions
EDA	Exploratory Data Analysis
AACA	Addis Ababa City Administration
SMO	Sequential Minimal Optimization
KDP	Knowledge Discovery Process
CRISP	CRoss-Industry Standard Process

ABSTRACT

Background: Tuberculosis is the second most common causes of death throughout the world next to HIV/AIDS. Ethiopia is also among the high burden countries. Though the disease has been a cause of death for millions of people around the globe, it is curable. Prediction of treatment outcome of TB patients using data mining techniques help the effort to stop TB-health problem.

Objective: The objective of this research was to prepare a predictive model for TB treatment outcomes that assist clinical decisions in connection with TB treatment.

Method: The six steps Ciso et al Hybrid Model were used**.** A total of 6332 instances were collected from five health centers of Addis Ababa City Government that provide tuberculosis treatment. A pre-processed the data was fed in to data mining tools with selected classification algorithms. These algorithms were J48, Naïve Bayes, SMO and PART. Accuracy and Area under ROC were the metrics used to compare models generated by the algorithms.

Result: After successive experiments using the four algorithms, PART algorithm revealed best performance. An accuracy of 81.32% and area under ROC=0.89. The algorithm generated five rules for the three treatment outcomes and the rules were found to be interesting for experts. The rules contain the following predictor variables for treatment outcome: HIV Status, Sex, Age, Initial Weight with second month weight and Patient Category.

Conclusion: The findings from the research indicated that for the tuberculosis dataset with class imbalance PART found to be the best learner algorithm and most importantly clinical decisions such as diagnosis, prognosis and resource allocation can be supported by data mining techniques.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Health care, health management, health policy and health planning all depend on having useful information to make decisions. Health Informatics provides useful information for such decision making. Therefore, informatics leads to better health [1].

Health Informatics is a rapidly growing field that is concerned with applying Computer Science and Information Technology to medical and health data [2].

Zaiane [2] provides an even more specific definition, which divides Health Informatics into four subfields:

> Health Informatics is the computerization of health information to support and optimize (1) administration of health services; (2) clinical care; (3) medical research; and (4) training. It is the application of computing and communication technologies to optimize health information processing by collection, storage, effective retrieval (in due time and place), analysis and decision support for administrators, clinicians, researchers, and educators of medicine.

The ever-increasing of health informatics tools such as computers, the Internet and Tele-health and other potentially powerful technologies results in the explosion of medical data.

In order to generate useful information from this massive amount of medical data, data mining should be applied to the field of health informatics.

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns [2].

One sub-field of Health Informatics, according to the definition given above, is clinical care. In clinical settings, applying data mining techniques on the centralized database would provide doctors with analytical and predictive tools that go beyond what is apparent from the surface of the data.

For instance, a new practitioner can query for all the decisions that previous practitioners have made on a similar case. Similarly, a data mining predictive model can advise doctors whether a certain case would be better treated as an outpatient or an inpatient [2].

Tuberculosis (TB) is a common and often deadly infectious disease caused by mycobacterium; in humans it is mainly *Mycobacterium tuberculosis*. It usually spreads through the air and attacks everyone especially low immune bodies such as patients with Human Immunodeficiency Virus (HIV) [4].

It is a disease which can affect virtually all organs, not sparing even the relatively inaccessible sites. The microorganisms usually enter the body by inhalation through the lungs. They may spread from the initial location in the lungs to other parts of the body via blood stream. They present a diagnostic dilemma even for physicians with a great deal of experience in this disease. Hence, Tuberculosis is a contagious bacterial disease caused by mycobacterium which affects usually lungs and is often could be found as co-infection with HIV/AIDS [4].

Tuberculosis is second only to HIV/AIDS as the greatest killer worldwide due to a single infectious agent. Tuberculosis is a major cause of morbidity and mortality in Ethiopia [5].

Ethiopia is one of the 22 High Burden Countries (HBCs). According to the WHO global TB report 2012, there were an estimated 220,000 (258 per 100,000 populations) incident cases of TB in Ethiopia in 2011. According to the same report the prevalence of TB was estimated to be 200,000 (237 per 100,000 populations). There were an estimated 15,000 deaths (18 per 100,000 populations) due to TB, excluding HIV related deaths, in Ethiopia during the same period [5].

Various data mining tasks can be applied on different treatment dataset. Prediction of treatment outcome has also been done on chronic diseases such as cancer, diabetes and liver

disease. There are a number of literatures are available for these diseases but very few articles related to Tuberculosis treatment outcome [2].

There is vast potential for data mining applications in predicting treatment outcome. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective [4].

In order to meet this need of useful information for clinical decision, data mining prediction task should be applied to the current TB treatment dataset found in health centers. This research shows how prediction is possible from dataset found in health centers for treatment outcome using data mining algorithms.

1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION OF THE STUDY

As it is indicated in the background section, Tuberculosis is the second greatest killer disease in the world and Ethiopia is also one of the countries among the 22 High Burden Countries (HBCs)[5]. According to the recent FMoH guideline:

> Cognizant of the burden of TB, Leprosy and TB/HIV co-infection in the country, the prevention and control of TBL and TB/HIV remains the priority health program in all phases of Health Sector Development Program (HSDP). In line with HSDP IV, a five-year TBL and TB/HIV strategic plan (2010/11 – 2014/15) is developed. Giving much focus on the scale up of community based TB care among others, the strategic plan will guide the successful implementation of effective interventions to reduce the burden of TB, Leprosy and TB/HIV co-infections.

In order to alleviate this serious health challenge, the country has invested a lot on TB treatment. The most important component of TB treatment strategy is still the expansion of high quality Directly Observed Treatment (DOT) [4].

In the DOT clinics around the country, there is a vast amount of treatment data in paper records.

The attempt to maintain medical records as electronic version has succeeded in many countries and the work began in Ethiopia recently [6]. Though Electronic Medical Records (EMR) is implemented in some government hospitals, due to various reasons the data found in the EMR is not that massive. Electronic medical records are not only for converting paper based information in to digital form but the data accumulated in the EMRs can help in finding useful information or knowledge for clinical decision making if data mining techniques are used.

According to the review of related literatures, data mining prediction is employed for treatment outcome prediction of surgical repair, mental health and treatment prescription of hypertension. In the current study, an attempt has been made applying the same techniques for communicable disease namely Tuberculosis. Since TB control is one of the Millennium Development Goals (MDG), treatment outcome prediction research assists the effort of achieving the goal in relation to the TB eradication goal [7-9].

Thus, in this study, two questions, that is, what factors, according to the variables in the dataset, determine treatment outcome of TB patients and what data mining algorithms best fit for medical records such as tuberculosis would be investigated?

1.3 OBJECTIVE OF THE STUDY

1.3.1 General Objective

The general objective of this research is to build predictive data mining model from the existing TB Treatment data at DOT clinics in public health facilities of Addis Ababa City Government.

1.3.2 Specific Objectives

This research aims at meeting the following specific objectives to achieve the stated general objective:

- To explore literature of TB treatment and data mining application to health care
- To encode and prepare dataset for the mining task
- To experiment and build predictive model for treatment outcome using data mining techniques

- To evaluate the knowledge extracted from the dataset
- To develop a prototype to predict treatment outcome of Tuberculosis

1.4 RESEARCH METHODOLOGY

The research was guided by the six steps Cios et al Hybrid Model. The model is based on CRISP Model and adopted to an academic research. The model has research oriented description of the steps, many new feedback mechanisms within the steps unlike the CRISP model which has only three feedback mechanisms and it is widely used in the field of medicine [10].

1.4.1 Understanding Tuberculosis Treatment Outcome

In order to gain a thorough understanding of the problem domain, the researcher attempted to review as many literatures as possible. The primary document that guided the review work was "GUIDELINES FOR CLINICAL AND PROGRAMMATIC MANAGEMENT OF TB, LEPROSY AND TB/HIV IN ETHIOPIA, FIFTH EDITION APRIL, 2012 ADDIS ABABA" [5]. Discussion was made with TB advisors at FMoH, TB focal persons at Addis Ababa Health Bureau (AAHB) and clinicians in health facilities. Based on this, data mining goals were identified.

1.4.2 Understanding and Preparation of TB Treatment Dataset

There were a few hundred records in the SmartCare Software found in Addis Ababa City Government (AACG) hospitals. As a result the researcher went to encode the TB treatment data from DOT clinics in health facilities. The data has 24 attributes and before the encoding work began 12 attributes were selected. Since the data is collected for the purpose of treatment outcome prediction, the dataset had a reasonable quality and further cleaning works were done.

1.4.3 Experimentation, Mining, Evaluation and Use of Discovered Knowledge

To deal with the problem of imbalanced TB dataset, SMOTE was used before the experiments were conducted. Four algorithms are selected for the experiments: the decision tree based J48, the function based Naïve Bayes, Support Vector Machine SMO and inductive decision rule based PART.

After the experiments were conducted, accuracy and ROC area were used to compare the performance of the findings considering the imbalanced outcome class, which is also common in medical dataset. The final step was an evaluation made by TB experts for the novelty and interestingness of the discovered knowledge.

1.5 SCOPE AND LIMITATION OF THE STUDY

The scope of this research is limited to generating predictive model for treatment outcome and developing a prototype as it is indicated in specific objective section. This research is conducted on electronic data collected from five health centers and the mining task is done on only the features found in the TB registry book. But the data found in the facilities is in paper format and encoding took much time of the work. The second limitation is related to the newness of the field of data mining and its application in the health arena, so that much literature is not available for data mining treatment outcome. The third limitation is the interdisciplinary nature of the study that poses a great challenge in understanding both areas: data mining and medical treatment.

1.6 SIGNIFICANCE OF THE STUDY

The significance of the study would be:

- Uncovering hidden patterns or knowledge related with tuberculosis prognosis and in that expanding the knowledge base of TB treatment.
- Helping physicians to predict outcome of the treatment they provided to the patient ahead of time, helps them to select treatment packages and propose the best treatment package for the patient.
- Helping policy makers and NGOs in tuberculosis treatment decisions.
- Supporting the effort to stop death due to Tuberculosis.
- Helping researchers in the area of health informatics, software development, and data mining and machine learning for further research in the area.

1.7 ETHICAL CONSIDERATIONS

The use of data—particularly data about people—for data mining has serious ethical implications, and practitioners of data mining techniques must act responsibly by making themselves aware of the ethical issues that surround their particular application.

Understanding the implications on the current research at hand, the researcher has kept all information that identifies a patient secret. There are a total of 24 attribute values and only 12 attributes values are selected, without including the personal identifiers such as name, address and contact person address, to meet the research objective.

Finally, any attempt to access the data requires the permission of ethical clearance committee and ethical clearance was obtained from Addis Ababa Health Bureau.

1.8 THESIS ORGANIZATION

This thesis report contains seven chapters. The first chapter deals with the general overview of the study including background, statement of the problem, objectives and methodology of the research.

The second chapter has attempted to explore the domain area which is Tuberculosis. Its meaning, types and diagnosis and treatment and the impact of the disease on Ethiopia is explored in brief.

The third chapter of the study has discussed data mining, knowledge discovery in databases, data mining tasks and algorithms, performance measurements of algorithms and finally data mining models.

The fourth chapter is devoted to see related works. The chapter includes discussion of the application of data mining in health care and the process of treatment prediction in data mining in general, chronic disease and data mining prediction and TB treatment outcome prediction.

Under chapter five, based on the Cios et al process model, domain area understanding, data understanding and data preparation work have been discussed step by step. The discussion ends with the presentation of the final dataset for data mining task.

Chapter six comprises experimental setup, post treatment of the dataset, experimentation and its evaluation for each algorithm selected for the purpose of classification, comparison of performance measurements by each algorithms, rules selected by the best algorithms with their discussion and prototype development.

Finally, conclusions and recommendations are forwarded in chapter seven.

CHAPTER TWO

UNDERSTANDING TUBERCULOSIS

This chapter provides basic information such as the meaning and types of tuberculosis, TB treatment and strategy in administering the treatment (DOT), TB treatment outcomes and TB situation in Ethiopia such as the burden of the disease and efforts to deal with the problem.

2.1 DEFINITION AND CLASSIFICATION OF TUBERCULOSIS

Tuberculosis (TB) is caused by a bacterium called *Mycobacterium tuberculosis*. The bacteria usually attack the lungs, but TB bacteria can affect virtually all organs, not sparing even the relatively inaccessible sites such as the kidney, spine, and brain [11], [4].

Based on the anatomical site, TB can be grouped in to two:

Pulmonary Tuberculosis (PTB): It refers to a case of TB involving the lung parenchyma. Miliary tuberculosis is classified as pulmonary TB because there are lesions in the lungs. Tuberculous intrathoracic lymphadenopathy (mediastinal and/or hilar) or tuberculous pleural effusion, without radiographic abnormalities in the lungs, constitutes a case of extra pulmonary TB. A patient with both pulmonary and extra pulmonary TB should be classified as a case of *pulmonary TB* [5].

Extra pulmonary Tuberculosis (EPTB): It refers to a case of TB involving organs other than the lungs such as pleura and larynx. Diagnosis should be based on at least one specimen with confirmed mycobacterium tuberculosis or histological or strong clinical evidence consistent with active EPTB, followed by a decision by a clinician to treat with a full course of tuberculosis chemotherapy.

The case definition of an EPTB case with several sites affected depends on the site representing the *most severe* form of disease. Unless a case of EPTB is confirmed by culture as caused by M. tuberculosis, it cannot meet the “definite case” definition given above [5].

There are also other types of classification such as bacteriological results (including drug resistance); history of previous treatment. Irrespective of site, treatment regiment is all the same [5].

2.2 TB TREATMENT AND TREATMENT STRATAGY

2.2.1 TB Treatment

TB treatment has a number of aims given below and the chemotherapy received by the patient has to be at adequate level and adequacy level is also defined in treatment guidelines.

As per the guideline of WHO [13], TB treatment is given to the patient with the following aims: The aims of TB treatment is that to cure the TB patient and restore quality of life and productivity, to prevent death from active TB or its late effects, to prevent relapse of TB, to prevent the development and transmission of drug resistance, and to decrease TB transmission to others.

To achieve the aims of TB treatment, the patient should receive adequate chemotherapy and the Chemotherapy is considered to be adequate when it: Rapidly and substantially reduces the number of actively multiplying bacteria, Cures patients, Prevents relapse of the disease and Prevents the development of resistance to the drugs.

The requirements for adequate chemotherapy are therefore: An appropriate combination of drugs, prescribed in the correct dosage, taken regularly by the patient, for a sufficient period of time

2.2.2 Directly Observed Treatment (DOT)

DOT stands for directly observed therapy. DOT means that a health care worker meets with a person who has TB to help him or her remember to take the medicines to treat TB. The health care provider supervises the patient take each drug dose. DOT is convenient and easy to arrange, and it can fit into one's daily routine [12].

Effective treatment of tuberculosis requires adherence to a minimum of 6 months treatment with multiple drugs. To improve adherence and cure rates, directly observed treatment (DOT) is a recommended strategy for the treatment of tuberculosis [14]. TB treatment can seem difficult -- it requires taking 2 or more medicines for at least 6 months. Most people have trouble remembering to take their medicines, or they stop taking their medicines when they start to feel better. When this happens, a person with TB could get sick again, and the TB bacilli could become resistant to the medicines. With DOT you don't have to worry about remembering to take your medicines [12].

The success of directly observed treatment requires the patient's cooperation as well as motivation. Health workers and treatment supporters should have the appropriate communication skills when interacting with patients. They should also provide all the necessary information about their treatment so that patients understand the disease and adhere to the treatment [5].

2.3 TB TREATMENT OUTOCOMES

TB treatment outcomes are:

Table 2.1 TB Treatment outcomes

Outcome	**Definition**
Cure	A patient whose sputum smear or culture was positive at the beginning of the treatment but who was smear- or culture-negative in the last month of treatment and on at least one previous occasion.
Treatment Completed	A patient who completed treatment but who does not have a negative sputum smear or culture result in the last month of treatment and on at least one previous occasion
Treatment Failure	A patient whose sputum smear or culture is positive at 5 months or later during treatment. Also included in this definition are patients found to harbour a multidrug-resistant (MDR) strain at any point of time during the treatment, whether they are smear-negative or -positive.
Died	A patient who dies for any reason during the course of treatment.
Defaulter	A patient whose treatment was interrupted for 2 consecutive months or more.
Transfer out	A patient who has been transferred to another recording and reporting unit and whose treatment outcome is unknown.
Treatment Success	A sum of cured and completed treatment

2.4 SITUATION of TUBERCULOSIS

2.4.1 Magnitude of the problem

TB is a major public health problem throughout the world. About a one third of the world's population is estimated to be infected with tubercle bacilli and hence at risk of developing active disease.

According to the WHO Global TB Report 2012, there were an estimated 8.7 million incident cases and 12 million prevalent cases of TB globally, in 2011, of which 1.1 million (13%) were among people living with HIV.

About 26% of the incident TB cases occurred in Africa in 2011. The proportion of TB cases co-infected with HIV is highest in countries in the African region; overall, the African region accounted for 79% of TB cases among people living with HIV [5].

According to health and health related indicators (2009/10) of the FMoH, tuberculosis is the third leading cause of death in Ethiopia. During the year 2010/11, a total of 159,017 TB cases were notified in Ethiopia. Among these 151,866 (95.5%) were new cases of TB, all forms. The proportion of new smear-positive, smear negative and EPTB among all new cases were 32.7%, 34.8%, and 32.5% respectively. Re-treatment cases represent about 2.9% of all notified TB cases [5].

2.4.2 Efforts to deal with the problem

To build on the achievements of DOTS and address the remaining challenges, the STOP TB strategy was launched by WHO in 2006 to help achieve the millennium development goals for TB in 2015. Ethiopia also adopted this strategy to achieve the national TBL and TB/HIV targets [5].

This strategy has six components where DOTS remains the most important component of the strategy. The components are [13]:-

1. **Pursue high quality DOTS expansion and enhancement**

It includes secure political commitment with adequate and sustained financing, ensure early case detection and diagnosis through quality assured bacteriology, provide standardized treatment with supervision and patient support, ensure effective drug supply and management and monitor and evaluate performance and impact.

2. **Address TB/HIV, MDR-TB and the needs of poor and vulnerable population**

Scale up collaborative TB/HIV activities, scale up prevention and management of MDR-TB and Address the needs of TB contacts and of poor and vulnerable population.

3. **Contribute to health system strengthening based on primary health care**

Help improve health policies, human resource development, financing, supplies, service delivery and information, Strengthen infection control in health services, congregate settings and household, Upgrade laboratory networks and implement practical approach to lung health and Adapt successful approaches from other field and sectors, and foster actions on the social determinants of health.

4. **Engage all care providers**

Involve all public, voluntary and corporate and private providers through public-private mix (PPM) approaches and Promote use the International Standard for TB care.

5. **Empower people with TB and communities through partnership**

Pursue advocacy, communication and social mobilization, foster community participation in TB care, prevention and health promotion and Promote use of patients' charter for Tuberculosis Care.

6. **Enable and promote research**

Conduct program-based operational research and Advocate for and participate in research to develop new diagnostics, drugs and vaccines.

CHAPTER THREE

DATA MINING TECHNOLOGY

This chapter presents review of literatures on data mining technology. It begins with a short introduction of data mining and knowledge discovery in database (KDD). The chapter continues to discuss the major tasks of data mining and the different algorithms with their performance measurement. The two popular data mining models: KDP and CRISP with their mix (Hybrid) also discussed. The chapter ends up with the presentation of the connection between statistics and data mining.

3.1 DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASE

Due to the advent of information technologies, organizations collect huge amount of data in their databases which in turn results in data explosion. This massive collection of data is beyond the capacity of human experts to analyze the data and gain knowledge for decision making purposes. Statistics also could not perform the analysis work on such large amount of data [15].

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year. The ability to use these data to extract useful information for quality healthcare is crucial [16].

Therefore, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD) [17].

Deshpande et al [18] states:

> To take complete advantage of data accumulated in databases; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data.
>
> With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'.

3.2 WHAT IS DATA MINING?

Data Mining, also popularly known as **Knowledge Discovery in Databases** (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 3.1) shows data mining as a step in an iterative knowledge discovery process.

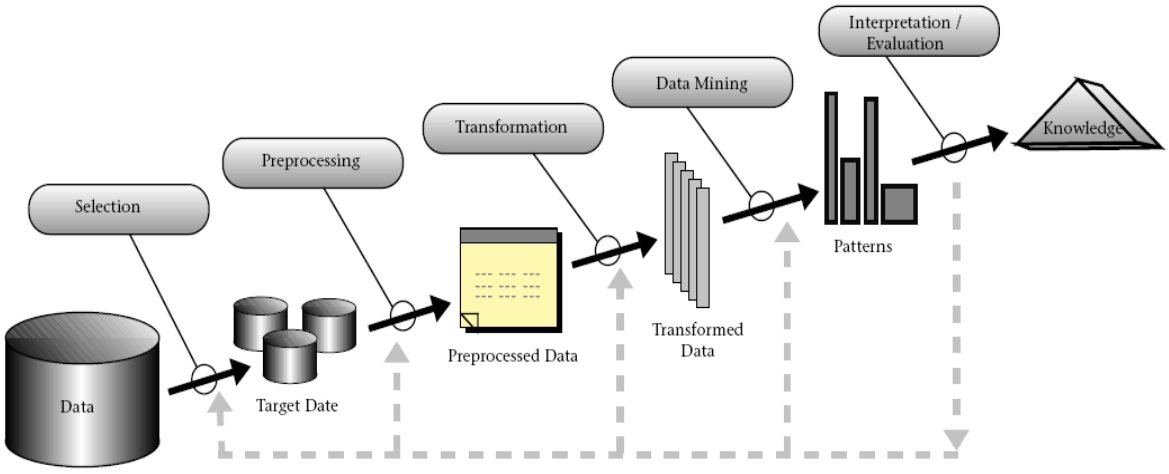


Figure 3.1 The place of data mining in the knowledge discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

The iterative process consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. [19]

3.3 DATA MINING TASKS AND ALGORITHMS

The knowledge discovery goals are defined by the intended use of the system. We can distinguish two types of goals: (1) verification and (2) discovery. See figure 3.1

With *verification,* the system is limited to verifying the user's hypothesis. With *discovery,* the system autonomously finds new patterns. We further subdivide the discovery goal into *prediction,* where the system finds patterns for predicting the future behavior of some entities, and *description,* where the system finds patterns for presentation to a user in a human-understandable form. In this research, the current researcher is primarily concerned with discovery-oriented data mining. [17]

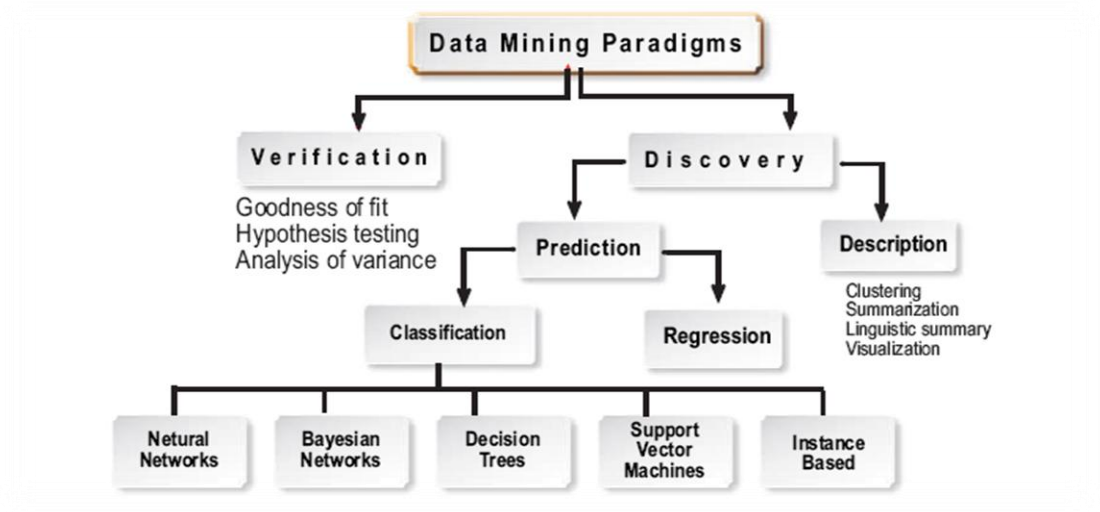


Figure 3.2 Data Mining Taxonomy

3.3.1 Descriptive Methods

It is a method that identifies the patterns or relationships in data and explores the properties of the data examined. Clustering and Association rule are the two popular descriptive methods [18].

3.3.1.1 Association Rule Discovery

The association task for data mining is the job of finding which attributes "go together." Most prevalent in the business world, where it is known as affinity analysis or market basket analysis, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Association rules are of the form "If antecedent, then consequent," together with a measure of the support and confidence associated with the rule. For example, a particular supermarket may find that of the 1000 customers shopping on a Thursday night, 200 bought diapers, and of those 200 who bought diapers, 50 bought beer. Thus, the association rule would be "If buy diapers, then buy beer" with a support of 200/1000 = 20% and a confidence of 50/200 = 25% [20].

3.3.1.2 Clustering

According to Rashmi as cited by Nesredin [21], Clustering refers to situations where the goal is to classify a diverse collection of unlabeled data into different groups based on different features in a data set. Clustering, also known as cluster analysis or unsupervised classification, is a general term to describe methodologies that are designed to find natural groupings or clusters based on measured or perceived similarities among the items in the clusters using a multidimensional data set.

There is no need to identify the groupings desired or the features that should be used to classify the data set. In addition, clustering offers a generalized description of each cluster, resulting in better understanding of the data set's characteristics and providing a starting point for exploring further relationships.

3.3.2 Prediction Methods

It makes prediction about unknown data values by using the known values. Classification and Regression are the two popular methods of prediction

Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state [16].

3.3.2.1 Classification

Classification is a task that occurs frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes. The term 'mutually exhaustive and exclusive' simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all [22].

For example, assigning people or objects to one of a number of categories:

- A patient who is more likely to get cured and died
- Customers who are likely to buy or not buy a particular product in a supermarket

- People who are at high, medium or low risk of acquiring a certain illness
- People who closely resemble, slightly resemble or do not resemble someone seen committing a crime
- People who are at high, medium or low risk of a car accident in the next 12 months
- The likelihood of rain the next day for a weather forecast (very likely, likely, unlikely, very unlikely).

3.3.2.1.1 Classification Algorithms

Decision tree

Data Mining uses machine-learning methods using decision trees to classify objects based on the dependent variable. There are two main types of decision trees [23]. Decision trees, which are used to predict categorical variables, are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called regression trees. Classification trees label records and assign them to the proper class. Classification trees can also provide the confidence that the classification is correct. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. Regression trees, on the other hand, estimate the value of a target variable that takes on numeric values. When a tree model is applied to data, each record flows through the tree along a path determined by a series of tests until the record reaches a leaf or terminal node of the tree.

There it is given a class label based on the class of the records that reached that node in the training set or, in the case of regression trees, assigned a value based on the mean (or some other mathematical function) of the values that reached that leaf node in the training set.

Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make prediction [20]. Various decision tree algorithms such as CHAID (Chi-squared Automatic Interaction Detection), C4.5/5.0, CART (Classification and Regression Trees), J48 and any with less familiar acronyms, produce trees that differ from

one another in the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over-fitting [24].

Today's data mining software tools allow the user to choose among several splitting criteria and pruning rules, and to control parameters such as minimum node size and maximum tree depth allowing one to approximate any of these algorithms. Figure3.1. shows how decision tree works.

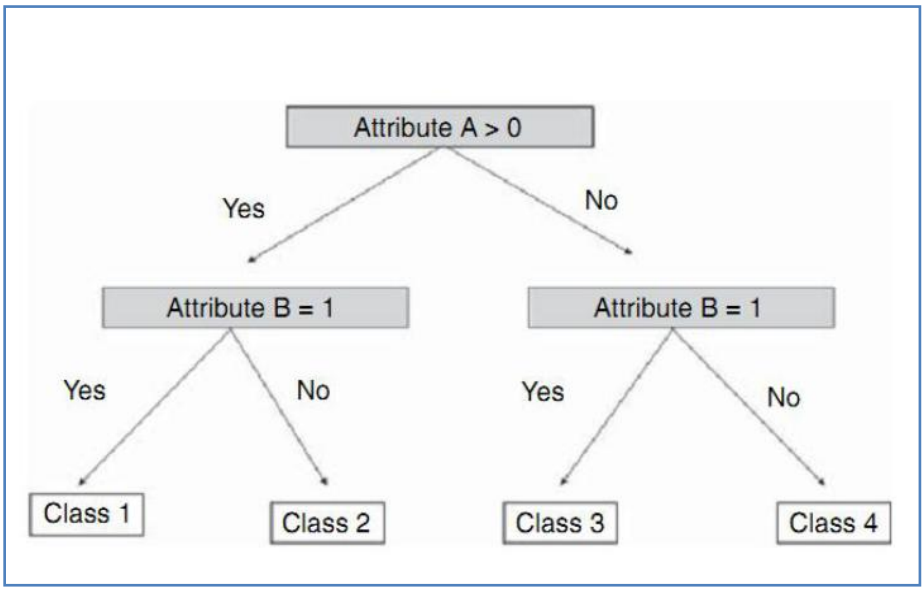


Figure 3.3 A simple decision tree

RJ-48 Algorithm

Decision tree algorithms are based on a divide-and-conquer approach to the classification problem. They work from the top down, seeking at each stage an attribute to split on that best separates the classes, then recursively processing the sub problems that result from the split. This strategy generates a decision tree, which if necessary can be converted into a set of classification rules—although if it is to produce effective rules, the conversion is not trivial.

Algorithm: Generate decision tree.
Input: Sets of training dataset (D), Attribute list ,Attribute selection method;
Output: A decision tree.
Method:
(1) Create a node ***N***;
(2) If tuples in ***D*** are all of the same class, ***C*** then
(3) Return ***N*** as a leaf node labeled with the class ***C***;
(4) If ***attribute list*** is empty then
(5) Return ***N*** as a leaf node labeled with the majority class in ***D***; // *majority voting*
(6) Apply Attribute selection method (***D***, ***attribute list***) to find the "best" ***splitting criterion***;
(7) Label node *N* with ***splitting criterion***;
(8) If ***splitting attribute*** is discrete-valued and multiway splits allowed then // *not restricted to binary trees*
(9) ***Attribute list*** ← *attribute list- splitting_ attribute*; // *remove splitting attribute*
(10) For each outcome ***j*** of ***splitting criterion*** // *partition the tuples and grow subtrees for each partition*
(11) Let ***Dj*** be the set of data tuples in ***D*** satisfying outcome ***j***; // *a partition*
(12) If ***Dj*** is empty then
(13) attach a leaf labeled with the majority class in ***D*** to node ***N***;
(14) Else attach the node returned by **Generate decision tree** (***Dj***, ***attribute list***) to node ***N***;
endfor
(15) Return ***N***;

Figure 3.4 decision tree algorithm

Trees and Rules

Decision tree methods are often chosen for their ability to generate understandable rules. It is certainly true that for any particular classified record, it is easy to simply trace the path from the root to the leaf where that record landed in order to generate the rule that led to the classification, and most decision tree tools have this capability. Many software products can output a tree as a list of rules in different format, including SQL code, pseudo code, or pseudo-English. However, since every split in a decision tree is a test on a single variable, decision trees can never discover rules that involve a relationship between variables. It is up to the miner to add derived variables to express relationships that are likely to be important.

PART Algorithm

PART algorithm combines the divide-and-conquer strategy (the top-down approach) for decision tree construction with the separate-and-conquer approach for rule learning.

The separate-and-conquer strategy first builds a rule and then removes those instances that the rule covers. These consecutive activities continue recursively for the remaining instances until none are left which generates sets of rules called ‘decision lists’ or ordered set of rules. On the other hand, in the partial decision tree, a pruned decision tree is built for part of the training instances, the leaf with the largest coverage is made into a rule, and the tree is discarded. Using partial decision trees in conjunction with the separate-and-conquer methodology adds flexibility and speed. A partial decision tree is an ordinary decision tree that contains branches to undefined sub trees. During the generation of such a tree, construction and pruning operations are integrated in order to find a “stable” sub tree that cannot be simplified further. Once this sub tree has been found, tree building ceases and a single rule is read off [29].

The rule sets that PART produces are as accurate as those generated by C4.5 and more accurate than other fast rule-induction methods. However, its main advantage over other schemes is simplicity. The close similarity in accuracy with C4.5 is due to the use of the C4.5 algorithm itself for building the partial decision tree whose "best" leaves are later converted into a rule [29].

Support Vector Machine

Explained by Huang et al (2002), and cited by Thomas, until 1992 Support Vector Machine (SVM) were largely unnoticed due to widespread belief in the statistical and/or machine learning community, despite being theoretically appealing. They were taken seriously only when excellent results achieved in numeral recognition, computer vision, and text categorization; today SVM show better results than Neural network comparable outcome and other statistical models [43].

SVM is a promising new method for the classification of both linear and nonlinear data; it uses a non linear mapping to transform the original training data into a higher dimension [7].

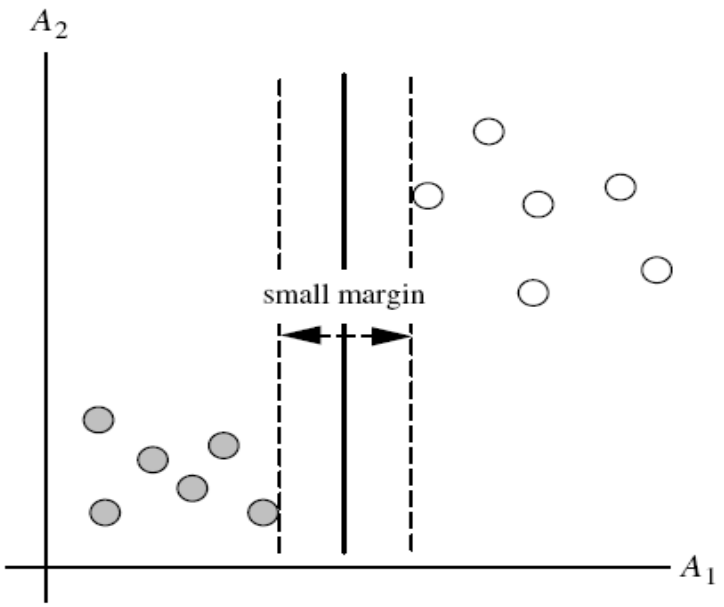


Figure 3.5 the 2-D training data are linearly separable

To explain SVM; if it is a two-class problem where the classes are linearly separable, an algorithm is implemented to finds a special kind of linear models. Let $\mathrm{xi} \in \mathrm{R_{n,}}(\mathrm{i} = 1,2,3..\mathrm{m})$ represents the vectors and$\mathrm{y_i} \in \{1, -1\}$. The term $\mathrm{f(xi)}$can be represented by a linear function of the form by $\mathrm{y_i} = \mathrm{f(x)}$

$$f(xi) = (w.x) + bf(xi)$$

Where W is a weight vector namely, $\mathrm{w} = \{\mathrm{w1}, \mathrm{w2}, \mathrm{w3} \ldots \mathrm{w_n}\}$and b is a scalar, often referred to as bias [29]. There is infinite number of hyperplane / separating lines that could be drawn for classifying the two-classes [27]. To find the optimal linear model or Hyperplane (n dimensions) that will have the minimum classification error on previously unseen tuples, SVM search for maximum marginal hyperplane [29].

Maximum marginal hyperplane is the one that gives the greatest separation between the classes [29]. Figure 3.2 depict that hyperplanes that can correctly classify all of the given data tuples.

But larger margin are likely to be more accurate at classifying future data tuples than the hyperplane with the smaller margin. During learning phases, SVM searches for the hyperplane

with largest margin, which is the maximum marginal hyperplane (MMH) [27]. When dealing with the MMH, this distance is the shortest distance from the MMH to the closest training tuple of either class [27]. A hyperplane separating the two classes' decision boundary may be written as

$$x = b + \sum \alpha_i y_i a(i) \, . \, a$$

Here, y_i is the class value of training instance$a(i)$; while b and $\propto_i$are numeric parameters that have to be determined by the learning algorithms. $a(i)$ and a represent the vectors. The vector a represent the test instances and $a(i)$ are the training instances.

Naïve Bayes

Naïve Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (Naïve) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of particular feature of a class is unrelated to the presence (or absence) of any other feature [25].

The Naïve Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

$$prob(B/A) = \frac{prob(A/B)prob(B)}{prob(A)}$$

In probability theory Bayes theorem shows how one conditional probability (such as the probability of a hypothesis given observed evidence) depends on its inverse (in this case, the probability of that evidence given the hypothesis). In more technical terms, the theorem expresses the posterior probability (i.e. after evidence B is observed) of a hypothesis A in terms

of the prior probabilities of A and B, and the probability of B given A. It implies that evidence has a stronger confirming effect if it was more unlikely before being observed.

Naïve Bayes Algorithm

As cited by Wei at al [44] Naïve Bayes (NB) is a machine-learning method that has been used for over 50 years in biomedical informatics. It is very efficient computationally and has often been shown to perform classification surprisingly well, even when compared to much more complex methods.

Naive Bayesian classifier uses the Bayes' rule to compute the probability of each possible value of the target attribute given the instance, assuming the input attributes are conditionally independent given the target attribute i.e. class conditional independence. Due to the fact that this method is based on the simplistic, and rather unrealistic assumption that the causes are conditionally independent given the effect, this method is well known as Naive Bayes [29, 27]. But despite the disparaging name, Naive Bayes works very well particularly when combined with some attribute selection procedure is applied to eliminate redundant (nonindependent attributes) [29].

According to Han and Kamber [27], the naive Bayesian classifier works as follows:

1. Let D be a training set of instances and their associated class labels. As usual, each instance is represented by an n-dimensional attribute vector, $x = (x_1 + x_2, \dots x_n$…depicting n measurements made on the instance from n attributes, respectively, $A_1 + A_2 , \dots , A_n$
2. Suppose that there are m classes, C_1, C_2, … , C_m. Given an instance, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier predicts that instance x belongs to the class c_j if and only if

$$p(c_j/x) > p(c_j/x) \;\; for\; 1 \leq j \leq m; j \neq i$$

Thus probability is obtained for$p(c_j/x)$). The class c_j for which $p(c_j/x)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$p(c_j/x) = \frac{p(x/c_i)\, p(c_i)}{p(x)}$$

3. As $p(x)$is constant for all classes, only $p(x/c_j)p(c_j)$needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, (C_m), $p(c_1 = p(c_2 = \cdots = p(c_m)$, and we would therefore maximize $p(x/c_j)$.Otherwise, we maximize $p(x/c_j)\, p(c_j)$. Note that the class prior probabilities may be estimated by $p(cj) = |c_{j,D}|/|D|$, where $c_{j.D}$ is the number of training instances of class c_j in D.
4. Given datasets with many attributes, it would be extremely computationally expensive to compute$p(x/c_j)$. In order to reduce computation in evaluating$p(x/c_j)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the instance (i.e., that there is no dependence relationships among the attributes). Thus,

$$p(x/c_j) = \prod_{k=1}^{n} p(x_k/c_j)$$
$$= p(x_1/c_1) * (p\left(\frac{x_2}{c_1}\right) * \ldots * p(x_n/c_1)$$

We can easily estimate the probabilities $p(x_1/c_j), p(x_2/c_j) \ldots, p(x_n/c_j)$ from the training instances. Recall that here x_k refers to the value of attribute $A_k$for instance x.

6. In order to predict the class label of $x, p(x/c_j)\, p(c_j)$ is evaluated for each class c_j. The classifier predicts that the class label of instance x is the class c_j if and only if

$$p(x/c_j)p(c_j) > p(x/c_j)p(c_j)\ for\ 1 \le j \le m, j \ne 1$$

In other words, the predicted class label is the class C_i for which $P(X/C_i)\ P(C_i)$ is the maximum.

The Naive Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which can be used to calculate the probability of each of the possible classifications in turn. Having done this, the class with the largest value will be selected as the class of the new instance [22].

3.3.2.2 Regression

Description by Two Crows Corporation as cited in Beshah [26], Regression on the other hand used to deal with non-discrete that means continuous variable. Regression is similar to classification, except that the label is not discrete.

For example, predicting salary or the price of stock is a regression, whereas predicting whether the salary is in a given range or whether a stock will go up or down is a classification task. It uses existing values to forecast what other values will be.

Although at a simplest senses regression uses standard statistical techniques such as linear regression, because of the complex nature of the real world problems, more complex techniques like decision tree and neural networks may be necessary to forecast future values.

3.3.2.3 CLASSIFIER ACCURACY MEASURES

Using the same dataset to derive a classifier or predictor and then to estimate the accuracy of the resulting learned model results in misleading overoptimistic estimates due to over specialization of the learning algorithm to the data. Instead, accuracy is better measured on a test set consisting of class-labeled instances that were not used to train the model. Then, the classifier is applied on the test set and the number of instances that were assigned to their actual classes and the number of instances that were assigned to different class by the classifier are counted, a process whose result is effectively represented by confusion matrix [27].

Confusion Matrix

Confusion matrix is a useful tool for analyzing how well a learned model can recognize instances of different classes. A confusion matrix for two mutually exclusive classes is shown in the Table 3.1. If there are m classes, a confusion matrix will be a table of size m by m. An entry, $CM_{i,j}$ indicates the number of instances of class i that were labeled by the learned model as class j. For a learned model to have good accuracy, ideally most of the instances would be represented along the diagonal of the confusion matrix, from entry $CM_{1,1}$ to entry $CM_{m,m}$, with the rest of the entries being close to zero [27,28].

Let C_1 be actual class label is positive, and C_2 actual class label is negative. P+ is predicted class label is positive, and P- is predicted class label is negative. Then the matrix shown in figure 2.5 provides four entries as a result of combination of the actual class label and class label provided to instances by the classifier [27, 29].

Table 3.1 A confusion matrix for two mutually exclusive classes

Predicted class

		P+	P-
Actual	C1 +	True Positive (TP)	False Negative (FN)
	C2 -	False Positive (FP)	True Negative (TN)

The number of true positives and true negatives respectively refers to the positive and negative instances that were correctly labeled by the learned model. The number of false positives is the negative instances that were incorrectly labeled. Similarly, false negatives are the positive instances that were incorrectly labeled [27, 29]. The number of false positive and false negative are summed up to give number of errors and help calculate the learned models/classifiers error rate represented by the equation below in [29]

$$Error\ rate = \frac{Number\ of\ errors}{Number\ of\ instances} = \frac{FP + FN}{TP + TN + FP + FN}$$

On the other hand, the overall success rate is the number of correct classifications divided by the total number of classifications [29] which is commonly called as accuracy of a classifier.

The accuracy of a classifier on a given test set indicates the percentage of test set instances that are correctly classified by the classifier [27].

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

A relationship between the above two equations can be constructed as Accuracy + Error rate = 1 which ultimately leads to the conclusion that the one can be obtained by subtracting the value of the other from 1. For example, Error rate =1 - accuracy [30].

Learned model performance measures that ignore correctly predicted negative instances give additional information about the ability of the model more than the information obtained from the models predictive accuracy. Measures used for the purpose discussed here are precision, recall, and F measure [28].

Precision indicates the percentage of instances classified as positives by the learned model and that are actually positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall shows the percentage of actual positives which the learned model has classified as positives.

$$Recall = \frac{TP}{TP + FN}$$

Another measure called the F measure combines precision and recall with the formula [29]

$$\text{F } meausre = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * recall * Precion}{recall + precision} = \frac{2 * TP}{2 * TP + FP + FN}$$

Sensitivity and specificity

With slight difference, these four notations in figure 2.3 are used in medicine and health care for the purpose of characterizing the performance of diagnostic tests. For example, if a certain diagnostic test shows a positive test result for people with a disease, this is referred to as sensitivity. On the other hand, specificity refers to the proportion of people without disease who have a negative test result, which is 1 - FP. Sensitivity is also referred to as the true positive recognition rate (that is, the proportion of positive instances that are correctly identified), while specificity is the true negative rate (that is, the proportion of negative instances that are correctly identified) [28-30].

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specifity = \frac{TN}{TN + FP}$$

ROC curve

According to Altman and Bland cited in Tefera [7] the critical step before any data mining model can be used in routine clinical practice is to compare its performance with equivalent statistical methods like sensitivity and specificity. ROC (receiver operating characteristics) curves that originated from signal detection theory has added more value to these two measures by creating trade-off [31]. AUC (Area Under Curve) is a measure of the area under the ROC curve [32].

ROC curve is a two-dimensional graph to select possibly optimal models based on the TP rate and FP rate. It also represents trade-off between benefits (TP) and costs (FP). In the ROC curve, the sensitivity (TP) rate is represented on the Y-axis and the specificity (FP) rate on the X-axis. Each prediction result one instance of a confusion matrix represents one point in the ROC space. Several points on a ROC graph should be noted. The lower left point (0, 0) represents that the classifier labeled all instances out of their actual class. The upper right point (1, 1) is the case where all instances are classified in their actual class.

The point (0, 1) represents perfect classification and the line y = x defines the strategy of randomly guessing the class. In order to assess the overall performance of a classifier, the fraction of the total area that falls under the ROC curve is considered. AUC varies between 0 and 1. Larger AUC values indicate generally better classifier performance [30].

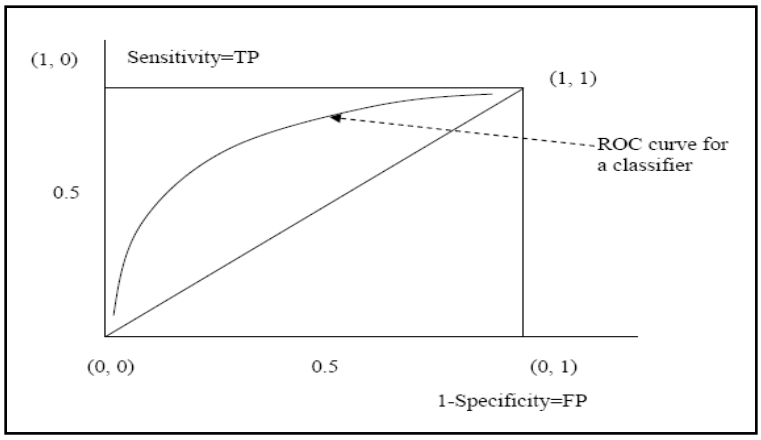


Figure 3.6 A ROC curve for a particular classifier

3.4 KNOWLEDGE DISCOVERY PROCESS MODELS (KDP)

The KDP model consists of a set of processing steps to be followed by practitioners when executing a knowledge discovery project. The model describes procedures that are performed in each of its steps. It is primarily used to provide a roadmap to follow while planning and executing a project, this in turn results in cost and time savings, better understanding, and acceptance of the results of such projects [30].

Of the five known KDP models emerged from academia and industry, the most popular ones are discussed below.

3.4.1 The Fayyad et al. Knowledge Discovery Process (KDP)Model

The KDP process, as presented in (Fayyad et al, 1996) is the process of using DM methods to extract what is deemed (considered, regarded as, estimated as, viewed) knowledge according to the specification of *measures and thresholds*, using a database along with any required preprocessing, sub sampling, and transformation of the database.

There are considered five stages, presented in figure 3:

1. **Selection** – This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

2. **Pre processing** – This stage consists on the target data cleaning and pre processing in order to obtain consistent data.

3. **Transformation** – This stage consists on the transformation of the data using dimensionality reduction or transformation methods.

4. **Data Mining** – This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)

5. **Interpretation/Evaluation** – This stage consists on the interpretation and evaluation of the mined patterns.

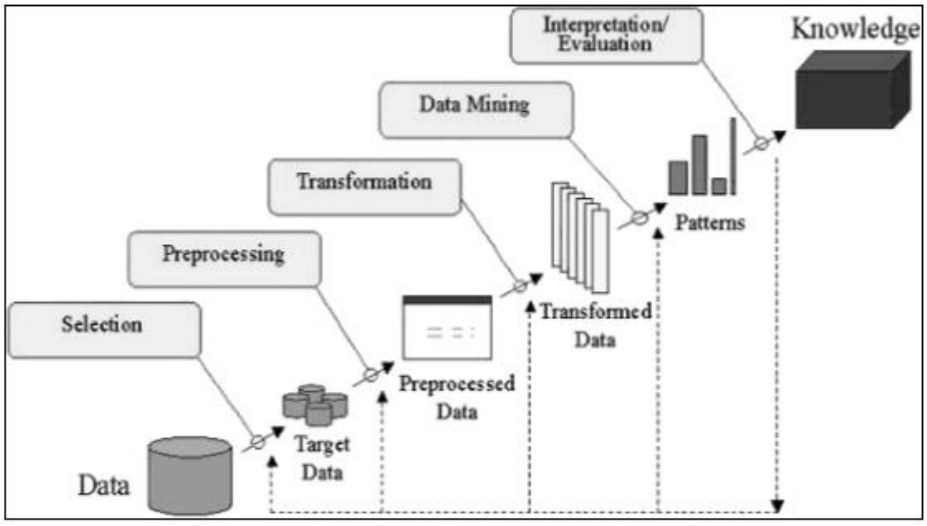


Figure 3.7 Knowledge Discovery Process (KDP)

3.4.2 CRoss-Industry Standard Process (CRISP) KDP Model

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a data mining project compromises a multi-step, iterative process. It consists on a cycle that comprises six stages [26].

1. Business understanding- this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

2. Data understanding- the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data preparation- the data preparation phase covers all activities to construct the final dataset from the initial raw data.

4. Modeling- in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation- at this stage the model (or models) obtained is more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives.

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

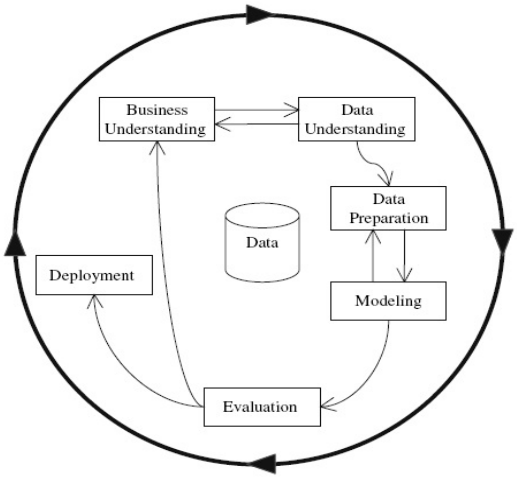


Figure 3.8 The CRISP-DM KD process model (source: http://www.crisp-dm.org/).

3.4.3 The Cios et al Hybrid KDP Model

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include

- providing more general, research-oriented description of the steps,
- introducing a data mining step instead of the modeling step,
- introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

A description of the six steps follows

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, project 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 derivation of new attributes (say, 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. Data mining. Here the data miner uses various DM 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. [26]

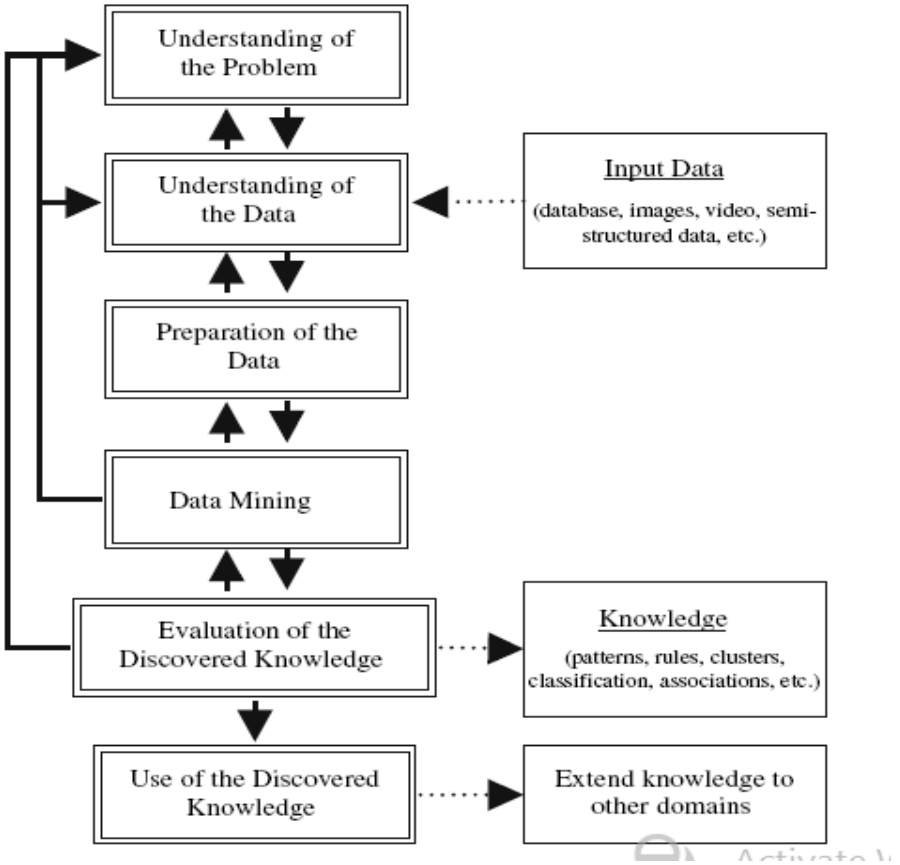


Figure 3.9 The six-step KDP model

3.5 DATA MINING APPLICATION

In today's data rich but useful information poor situation, data mining has several applications if the methods and algorithms are used with care and intelligently.

Some examples of applications (potential or actual) are [22]:

- predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care.
- a supermarket chain mines its customer transactions data to optimise targeting of high value customers
- a credit card company can use its data warehouse of customer transactions for fraud detection

- a major hotel chain can use survey databases to identify attributes of a ‘high-value’ prospect
- predicting the probability of default for consumer loan applications by improving the ability to predict bad loans

3.6 DATA MINING AND STATISTICAL METHODS

There is a marked difference between statistics and data mining, comparing both fields of study from the size and type of data is discussed in this brief part of the research.

The connection between data mining and statistics is seen differently from different angles. One is the view of computer scientists or data base experts and the other is that of Statistician.

For the later, data mining is statistics plus more but for the former professionals data mining is methodologies are developed outside the field of statistics.

Statistics considers a few hundred records as large but today's modern database contains records of size gigabytes or terabyte. Analyzing such large amount of data is beyond statistics and a new method is required to acquire useful information [15].

In addition to the large size of data, large set of records with outliers and missing values cannot be handled with standard statistical methods, in that checks the records from the source. This is impossible for large datasets [15].

Therefore, statistics, especially as taught in most statistics texts, might be described as being characterized by data sets which are small and clean, which permit straightforward answers via intensive analysis of single data sets, which are static, which were sampled in an iid manner, which were often collected to answer the particular problem being addressed, and which are solely numeric. None of these apply in the data mining context.

Finally, classical statistics deals solely with numeric data. Increasingly nowadays, databases contain data of other kinds. Four obvious examples are image data, audio data, text data, and geographical data [15], [33].

CHAPTER FOUR

RELATED WORKS

In this section, the application of data mining system in health care is briefly discussed and different works which are related with treatment prediction outcome is presented. Finally, TB treatment outcome prediction is discussed.

4.1 DATA MINING TECHNIQUES IN PREDICTING TREATMENT OUTCOME

In medical and health care areas, due to recordkeeping regulations and due to the availability of computers, a large amount of data is becoming available. Even though, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules [34].

According to Abdullah [35] The applications of data mining can be found in many areas such as evaluating risks of financial investment, detection of credit card fraud, patient diagnosis etc. Data mining can be applied to health care effort as well [2].

Data Mining Technology provides a user-oriented approach to novel and hidden pattern in the data. The discovered knowledge can also be used by the medical practitioner to reduce the adverse effect of drugs, point to less costly treatment alternatives, and predicting treatment outcome.

Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care [36].

Data mining has been used in a number of treatment datasets such as surgical outcome prediction, heart attack prediction, prevention of mental illness, hypertension control, tuberculosis diagnosis, etc...

4.2 DATA MINING TREATMENT OUTCOMES PREDICTION

4.2.1 The Process of Treatment Prediction

The process of treatment outcome prediction is lucidly described in a research dissertation, *Data mining in diagnostic charts and treatment outcome prediction for Vision Restoration Therapy on treatment outcome prediction.*

"In medical applications, treatments were administered by the physician to support the patient's recovery. If the treatment is time consuming and if it does not have the same result for all patients with a specific disease, it is desirable to use a prediction model telling whether the regimen is helpful or not. A simple manual approach to build such a prognostic system would require the following steps:

1. Collect as much data as possible about patients: the therapeutic regimen and the result after treatment.
2. To make the prognosis for a *new patient* P, just look into the set of collected samples and find a patient P0 (*treated patient*)with identical properties (with respect to age, the history of the patient's diseases, the regimen and some other criteria which were collected).
3. Return the treatment outcome from P0 as prognosis for P (new patient).

This approach lacks some very important aspects. If the database is huge (containing more than 10,000 cases), it is tedious to go through the database manually and finding the correct match. Furthermore, each patient is unique; no two patients are identical. This example shows that computers are helpful to support the physician who wants to make a prognosis for a new patient. Data mining tools were designed to efficiently access large databases and can be used to find cases which best match the properties of the patient P0." [37]

As it is described above finding a pattern (knowledge) in a dataset is extremely difficult, machine language algorithms and data mining techniques are the remedies proposed by experts. Though, this does not mean that the blind use of the algorithms and the techniques bring about successful results.

4.2.2 Data Mining Treatment Prediction

A study conducted on patients with urinary fistula, surgical repair outcome is investigated using predictive data mining techniques with the objective of coming up with a model that helps us to predict treatment outcome [7].

The study showed that the WEKA rule based algorithm, PART outperforms decision tree algorithm and regression analysis. The performance of the model was analyzed using the ROC area and resulted in 0.742 [7].

According to a research conducted at Center stone, the largest community based mental health provider in the United States, predicting treatment outcome based on a data extracted from EMR, achieved a 70% success rate in predicting treatment outcomes using data mining methods[38].

The research used a number of algorithms (23 different algorithms) and the model is also evaluated using multiple performance metrics such as accuracy, ROC analysis, AUC, TP rate, FP rate. From the finding achieved, clinical decisions could be supported by data mining technologies [38].

In as study conducted in Saudi Arabia 2005, a data mining regression analysis using non communicable diseases, the data sets for different age groups in case of blood pressure treatment for hypertension for Male using different modes have been studied. The Oracle data miner predicts the best mode of treatment such as drug, diet, weight reduction, smoke cessation and exercise for each group by which one can analyze the appropriate treatment [35].

A study by Asli [39], using imbalanced in-vitro fertilization (IVF) dataset, for predicting implantation outcome revealed the same performance by adjusting the decision threshold and under sampling as minority oversampling technique. The study made use Naive Bayes algorithm and produce significantly better predictive performance. It also described the problem of imbalanced dataset and the best performance measurement used in such situation. ROC analysis is proposed for the problem.

In a research conducted in the university of Manchester, on predicting the treatment outcome of tuberculosis-supervised and unsupervised learning, using seventeen features of dataset Iranian TB patients such as sex, weight, nationality, area of residency, current stay in prison, low body weight, TB type, treatment category, length of disease, TB case type, recent TB infection, diabetic or HIV positive, and social risk factors like history of imprisonment, IV drug usage, and unprotected sex decision tree gave the best prediction accuracy (74.21%) compared with other methods. The data mining algorithms used for the purpose includes logistic regression and support vector machine were shown least performance [40].

4.3 SUMMARY OF LITERATURE REVIEW AND RELATED WORKS

As it is pointed out in chapter 2, TB is one among the major public health problems and Ethiopian is one of the high burden counties. However, with the expansion of DOT and HEWs and community volunteers, case finding and treatment of tuberculosis has been successful. The Stop TB strategy is also adopted by Ethiopia and basically it is expansion of DOT and other strategies.

In order to assist the effort of tackling the TB problem, be able to predict the treatment outcome using data mining techniques supports the endeavor.

From the above description of related works, TB treatment outcome prediction model was proposed in a research conducted in Manchester University. But the features/ attributes selected for the purpose is different from the dataset found in health facilities found in Ethiopia.

In addition to that, the performance measure used for the model was not simple accuracy measurement but in this study a combination of accuracy measurement and ROC area are used.

Among the major five outcomes indicated in table 2.1, the dataset contains only three outcomes that is two favorable outcomes (Treatment completed and Cure) and one unfavorable outcome (Death). Treatment Failure and Defaulter are not found.

The problem is be attributed to (1) failure to register treatment failure by the clinicians (2) there are only 2% failure in treatment, as it is reported by FMoH (3) Defaulter are not also found.

Finally, this research is the first attempt in this country to prepare a predictive model for TB treatment outcome prediction using the TB dataset found in the health facilities.

CHAPTER FIVE

DATA UNDERSTANDING AND PREPARATION OF THE DATA

As it was indicated in the first chapter, the knowledge discovery process is guided by the six steps Cios et al hybrid model. The advantage and the steps the model follows were discussed in brief in chapter 3. The steps of the model guide this chapter and the subsequent one.

5.1 BUSINESS UNDERSTANDING

5.1.1 Overview

Data mining starts by understanding the business or problem domain in order to gain the business knowledge. The method used to understand the TB treatment and its potential outcome, as the other parts of the study, are literatures and experts in the domain.

First, different literatures are used to gain a through/complete understanding of the domain and the primary source referred to meet the goal is "GUIDELINES FOR CLINICAL AND PROGRAMMATIC MANAGEMENT OF TB, LEPROSY AND TB/HIV IN ETHIOPIA, FIFTH EDITION APRIL, 2012 Addis Ababa". In addition to the guideline, WHO sources are used extensively.

The researcher did not solely depend on literatures but also consulted experts and professionals in the area so that the literatures would be better understood. TB focal persons, Clinicians, Nurses and TB advisors were some of the people contributed to this step of the project – understanding of the problem domain.

5.1.2 Understanding TB Diagnosis and Treatment

The very first step of TB diagnosis begins with the identification of suspects, who are coughing for two weeks or more and have other symptoms such as low grade fever, night sweating and weight loss. This could be done with the help of HEWs and community volunteers. The suspects referred to a diagnostic health facility for sputum smear examination and further clinical evaluation.

There are different diagnostic methods to confirm the existence of the bacilli in the suspects but the mainstay of diagnostic methods for TB in Ethiopia is the Bacteriological Methods called Direct Light Smear Microscopy /conventional microscopy/. It is a method that uses sputum specimens and examines the specimen in two consecutive days with a certain defined procedure.

The procedure followed to confirm a suspect has TB or not is a patient use two sputum smears (one sputum positive is enough for HIV positive patients) or culture positive for mycobacterium tuberculosis.

Definite case of tuberculosis is also defined as a patient with Mycobacterium tuberculosis complex identified from a clinical specimen, either by culture or a newer method such as molecular line probe assay and for EPTB proven by one culture-positive specimen from an extra-pulmonary site or histo-pathological evidence from a biopsy.

Treatment of TB also depends on whether the patient has never had treatment or had a relapse or is returning after default or failure of prior treatment.

Clinically confirmed TB patients should receive adequate chemotherapy in the right combination and correct dosage regularly for a sufficient period of time. First line drugs for the treatment of TB in Ethiopia include: Rifampicin(R) Ethambutol (E) Isoniazid (H) Pyrazinamide (Z) and Streptomycin(S).

The Chemotherapy has two phases:

Intensive (initial) phase: This phase consists of treatment with combination of four drugs for the first 8 weeks for new cases, and with combination of five drugs for the first eight weeks followed by four drugs for the next four weeks for re-treatment cases.

Continuation Phase: This phase immediately follows the intensive phase and is important to ensure cure or completion of treatment. It is necessary in order to avoid relapse after completion of treatment.

There are also special cases in treatment of TB. One such case is HIV Patients on Anti-retroviral: TB patients with HIV infection or HIV/AIDS may experience a temporary worsening of symptoms and signs after starting TB treatment.

In TB patients infected with HIV, treatment with anti-retroviral (ARV) may interact with treatment of TB, reducing the efficacy of anti-retroviral and of anti-TB drugs while increasing the risk of drug toxicity.

The issue at the final stage of TB treatment effort is who does monitor the treatment. The best solution for this is DOT (Directly Observed Treatment) strategy.

This strategy is for a health worker or a community TB treatment supporter to watch each patient swallow every single dose of the drugs. Directly observed treatment can take place in a hospital, health center or health post, the patient's workplace, or at the home of the patient.

DOT ensures that all anti-TB drugs are swallowed. DOT is supposed to build supportive relationship between patient and health worker or community TB treatment supporter. A good relationship enables the patient to discuss any question or fear about the disease and treatment.

The TB patient treatment has six defined outcomes i.e. Treatment completed (no proof of negative specimen), Cure (final sputum result is smear or culture negative), Failure, Defaulter, Death and Transferred Out.

5.2 DATA UNDERSTANDING

Data mining requires collecting great amount of data (available in data warehouses or databases) to achieve the intended objective. Both literatures and experts in the area are used to understand the record of DOT TB registry book and also special effort is devoted to understand the 12 variables selected for the data mining prediction.

5.2.1 Data Acquisition and Attribute Subset Selection

Though, a computer based database is not available in the facilities where the dataset is located, the researcher encoded the data from the paper registry book with the help of a human data encoder. The records selected for the encoding had all class information.

Prior to the encoding of the data, as indicated in the table below, the researcher selected 12 out of 24 attributes (see Appendix F) through discussion with domain expert and confirmed using literatures as indicated in related works section 4.3.2. The 12 variables were selected over the others for the data mining work believed to be more relevant for treatment outcome prediction. These dataset were collected from 5 Health Centers of Addis Ababa City government and merged in to one file, using Excel Program, giving a total of 6,320 records.

Table 5.1 Attributes selected for data mining task

No	Attribute Name	Description	Data Type
Demographic Characteristics			
1.	Sex	Sex of the patient	Nominal
2.	Age	Age of the patient in years	Numerical
3.	Weight	Weight of the patient in Kilograms	Numerical
Clinical Features			
4.	Smear Result	Sputum examination result	Nominal
5.	Patient Category	Type of Patient as new, defaulter	Nominal
6.	TB Type	TB Type as PNeg,Ppos, EP	Nominal
7.	HIV Test Result	HIV status of a patient	Nominal
8.	CPT	Cotrimoxazol Preventive Therapy	Nominal
9.	ART	Anti-Retroviral Drug for HIV patients	Nominal
10	Second Month Sputum Result	Sputum result of the patient after 2 months	Nominal
11.	Second Month Weight	Weight in Kilograms after two months	Numerical
Treatment Outcome			
12.	Outcome		Nominal

5.2.2 Data Integration

The data needed for the mining task is collected from 5 health centers. The TB registry format is a standard for all health centers throughout the country; the values entered for the variables are all the same, so that integration of the dataset from the health centers was not a problem.

To make the dataset suitable for data mining task, pre-processing work is done. Such as data cleaning, missing value, etc. The detail of the process is indicated in the next section.

5.3 DATA PREPROCESSING

Before feeding data to data miner we have to make sure the quality of data. The dataset used for the project contains missing values, inconsistencies and noisy data. In order to use the data for the intended purpose filling missing values, correcting inconsistency has been made.

Noise in the data is defined as a value that is a random error or variance in a measured feature [42]. Depending on the amount in the data, it can be a substantial problem that can jeopardize the knowledge discovery process.

Exploratory data analysis precedes the effort of data preprocessing in that it exposes the problems found in the dataset.

5.3.1 Exploratory data analysis

In this section efforts were made to present the description of the selected attribute together with the exploratory data analysis performed with the use of frequency tables. The attribute's description, data type, unit of measure and list of values or range of values are described. With the use of frequency tables, the exploratory data analysis is performed to detect bad data i.e. attributes with the missing values and wrong entries or noises and inconsistency in values of attributes. The frequency tables for the selected attributes show the original distribution of values of attributes in instances of the dataset before any preprocessing is done on the dataset.

Sex: It is *nominal* value that has two values which is Male or Female. The number of male is approximately equal to female patients. There are only 15 instances are missing and 6 instances of noise are found in the dataset out of 6332 instances, which amounts 0.24% and 0.09% respectively. Table 5.2 shows the statistical summary of this variable.

Table 5.2 Statistical Summary of Sex

Sex		**Frequency**	**Percent**
Valid	Male	3153	49.79
	Female	3158	49.87
Missing Values		15	0.24
Noise		6	0.09
Total		6332	100

Age: It is a *continuous* valued attribute. Missing values and noises in the dataset were found to be 24 (0.38%) and 4(0.06%) respectively. Table 5.3 describes summary of age in the dataset.

Table 5.3 Statistical Summary of Age

Age		**Frequency**	**Percent**
Valid	1-120	6304	99.56
Missing Values		24	0.38
Noise		4	0.06
Total		6332	100

Smear Result: variable that classifies patients based on smear test as Positive, Negative and Extra-pulmonary. It is *nominal* variable that assumes values such as Positive (P), Negative (N) and Extra-pulmonary (EP). There were 4,510 (71.23%) missing values and there were very few errors (0.06%). Table 5.4 shows summary of smear result.

Table 5.4 Statistical Summary of Smear Result

Smear Result		**Frequency**	**Percent**
Valid	N	1015	16.01
	P	761	12.02
	EP	42	0.66
Missing Values		4510	71.23
Noise		4	0.06
Total		6332	100

Weight: It is *continuous* variable. The value is taken at the health facility after the patient had been suspected having TB. All the values in the data were correct but had 231 (3.65%) missing values. Table 4.4 shows the summary of weight measurements of patients.

Table 5.5 Statistical Summary of Weight

Weight		Frequency	Percent
Valid		6101	96.35
Missing Values		231	3.65
Noise		0	0
Total		6332	100

Patient Category: this variable tells the patient type as New, Relapse, Failed, Default, Transferred and Other. It is a *nominal* variable assumes values such as New (N), Relapse (R), Failure (F), Defaulter (D), T (Transferred out), and O (other). Table 5.6 summarize patients category. The data set reveals few missing values and errors with a magnitude of 97 (1.53%) and 6 (0.09%) respectively.

Table 5.6 Statistical Summary of Patient Category

Patient Category		Frequency	Percent
Valid	N	5388	85.1
	R	196	3.09
	F	12	0.19
	D	16	0.25
	T	220	3.47
	O	397	6.27
Missing Values		97	1.53
Noise		6	0.09
Total		6332	100

TB Type: It is also *nominal* attribute that assumes three values Pulmonary Positive (P/Pos), Pulmonary Negative P/Neg and Extra-pulmonary (EP). Only 31 (0.49%) missing values and 24 (0.38%) noises are found under this attribute. Table 5.7 summarizes attribute values in the dataset.

Table 5.7 Statistical Summary of TB Type

TB Type		Frequency	Percent
Valid	P/Pos	1562	24.67
	P/Neg	2673	42.21
	EP	2042	32.25
Missing Values		31	0.49
Noise		24	0.38
Total		6332	100

HIV Test Result: this attribute indicates HIV status of TB patient. This attribute also assumes valid nominal values such as R and NR for Reactive and Non-Reactive respectively. The fields with missing values are only 361 (5.70%) of the total number of instances. Noise is insignificant amounts to 0.06% of the total dataset. Table 5.8 shows summary of this attribute value in the dataset.

Table 5.8 Statistical Summary of HIV Test Result

HIV Test Result		Frequency	Percent
Valid	R	1953	30.84
	NR	4014	63.39
Missing Values		361	5.70
Noise		4	0.06
Total		6332	100

CPT: stands for Cotrimoxazol Preventive Therapy. The attribute indicates the drug HIV patients take to guard themselves against other opportunistic infections. It is a binary *nominal* attribute assumes values as YES or NO. The frequency of missing value in the dataset amounts 4241 (66.97%) of the total instances while the errors are only 0.13 percent. Table 4.8 shows the statistical summary of the CPT value in the dataset.

Table 5.9 Statistical summary of CPT

CPT		Frequency	Percent
Valid	YES	1622	25.62
	NO	461	7.28
Missing Values		4241	66.97
Noise		8	0.13
Total		6332	100

ART: stands for Anti Retroviral Therapy. It is a drug for HIV positive patients. It assumes binary *nominal* attribute. It takes a value YES or NO for those HIV Reactive patients who takes the drug or not respectively. There were 4,431 (69.97%) missing values and there were only 2 errors. Table 4.9 show the summary of ART value in the dataset.

Table 5.10 Statistical Summary of ART

ART		Frequency	Percent
Valid	YES	1010	15.95
	NO	889	14.04
Missing Values		4431	69.97
Noise		2	0.03
Total		6332	100

Second Month Sputum Result: This variable indicates to see the patient convert sputum after two months treatment. It is a binary *nominal* attributes that assumes two values as YES or NO. Noises are not identified under this attribute and the missing values amount to 76.55% of the total instances. Table 5.11 shows statistical summary of second month sputum.

Table 5.11 statistical summary of second month sputum result

Second Month Sputum Result		**Frequency**	**Percent**
Valid	YES	49	0.77
	NO	1436	22.68
Missing Values		4847	76.55
Noise		0	0
Total		6332	100

Second Month Weight: The weight measurement taken after the patient completed treatment administered for two months. It is a *continuous* variable. There is 1 error but significant amount of missing values existed that was 1296 (20.47%).

Table 5.12 statistical summary of second month weight

Second Month Weight		**Frequency**	**Percent**
Valid		5035	79.52
Missing Values		1296	20.47
Noise		1	0.01
Total		6332	100

Outcome: This is the dependent variable the researcher wants to predict in this data mining project work. It is a *nominal* variable that assumes three values TC stands for *Treatment Completed*; C stands for *Cure*; the above two are outcomes represent treatment success. D stands for *Death*. The data set contained only 0.20% missing values and had no error at all. Table 4.12 shows statistical summary of outcome variable in the dataset.

Table 5.13 statistical summary of outcome

Outcome		**Frequency**	**Percent**
Valid	TC	4868	76.87
	C	1135	17.92
	D	316	4.99
Missing Values		13	0.20
Noise		0	0
Total		6332	100

5.3.2 Data Cleaning

Real world data is full of flaws. They may contain missing values, noises and inconsistencies. These problems are hard to see to clean in huge dataset but with the help of statistical software tools, the difficult task of cleaning the data has become easier.

Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies [22]. The next sections clearly discuss the tasks done in cleaning the dataset in order to gain a “clean” data for the mining algorithms.

5.3.2.1 Managing Missing Values

Many datasets are plagued by the problem of missing values. The missing values problem may happen for various reasons such as incomplete manual data entry, incorrect measurements, and equipment errors [30].

In some domains (as in medicine), it is common to encounter data with a large percentage of missing values, even over 50% of all values [22].The dataset also reflects this fact and large number of missing values are observed as indicated.(Table 5.14)

There are two ways in dealing with missing values, one is removal of the missing data and the other is filling in the missing feature [30]. Both methods are employed to handle the missing value problem in the dataset. Except for the outcome class value, it is the later used in dealing with the problem.

As most data preprocessing literatures discusses, the method used to fill values depends on the type of attributes used in the project, for nominal attributes the most frequent values (mode) is used; for continuous variables arithmetic mean is used. But unintelligent application of both methods shocks the validity of the solution.

In the previous phase of data understanding using EDA, the number of missing values for each attribute is obtained. The strategy to handle each problem is explained briefly as follows:

Sex the number of male approximately equal to female patients, the replacing by the most frequent strategy does not work. Since the number of missing values are few (15 instances out of 6332), half of the missing values are replaced by male and the other half by female. (Table 5.2)

Age is simply replaced by mean, in this case the distribution of age is analyzed using WEKA and age reveals a normal distribution.

Replacing the missing values may likely lead to an error unless they are used with care []. In the case of missing values of patient's **weight**, before taking the average value of all the instances, group the patients according to age and take the mean value for that age group. Weight and **Second Month Weight** is replaced by mean by taking care of the age group of the patients.

The missing values of S**mear result** can be successfully resolved by comparing with *TB type* attribute (Ex. Pulmonary Negative (PNeg) has Negative Smear result, if Negative smear result is missing it can be replaced using TB Type which is Pulmonary Negative). Since the missing values of **TB type** attributes are quite minimal and replaced by most frequent value in turn solves the problem of smear result.

Patient Category is replaced by the most frequent value (mode). If values are distributed evenly, the validity of this approach is questionable. Buy as it is clearly shown in Table 5.6, Patient Category New (N) represents 85% and replacing the missing values with New (N) gives statistical sense.

HIV Test Result: Only patients with HIV Positive test result was registered in the dataset as Positive (R); for those patients who are HIV Negative, the field was left unfilled, thus filled with Negative (NR).

Most of the missing value happens simply because there are some attributes that are not applicable (eg. HIV negative patients do not take CPT and ART medication). Missing values of **CPT, ART, and Second Month Sputum** are replaced by Not Applicable (NA).The high number of missing values in these features is attributed to such a reason.

Finally, the **outcome** class with missing value is totally discarded because it is the dependent variable (statistically speaking) that we want to predict. Table 5.14 reveals the missing values problems with the various strategies to deal with them.

Table 5.14 summarizes missing values and how it is corrected.

No	Attributes	Missing Value	Handling Mechanism
1	Sex	0.24%	Replaced by most frequent value
2	Age	0.38%	Replaced by mean
3	Weight	3.65%	Replaced by mean
4	Smear Result	71.23%	Replaced by cross checking with TB type
5	Category	1.53%	Replaced by most frequent value
6	TB Type	0.45%	Replaced by most frequent value
7	HIV Test Result	5.70%	Replaced by Negative (NR)
8	CPT	66.97%	Replaced by NA[1]
9	ART	69.97%	Replaced by NA
10	Second Month Sputum Result	76.55%	Replaced by NA
11	Second Month Weight	0.01%	Replaced by mean
12	Outcome	0.20%	Removal of instances

5.3.2.2 Handling Noise

There were very few errors found in the dataset and the errors are corrected manually and the rest filled by the most frequent values. Though few are the errors, efforts are shown to understand and manage the problem.

As indicted in the below table, out of the 12 attributes selected, only 9 of them had revealed noise in the data.

[1] NA: Not Applicable

Table 5.15 summarizes noisy value and how it is handled.

No	Attributes	Noisy data	Handling Mechanism
1	Sex	N, E	It is typo, N and E is corrected by M and F respectively
2	Age	0, 299,422,440	0 is discarded, typing a character twice and corrected as 29, 42 40
3	Smear Result	EPTB, PN, F, M	It is typo, replaced by EP and N. F and M are replaced by the most frequent
4	Patient Category	M, P, PN, NPN	It is typo. M is replaced by N, NPN with N and the replaced by the most frequent
5	TB Type	P,E,N,ER,NR,F,R,Y	PP, EP,PN are manually corrected and the rest replaced by the most frequent
6	HIV Test Result	PN,PP,EP	Replaced by the most frequent
7	CPT	N, NR,O	Replaced by the most frequent
8	ART	57, N	Replaced by the most frequent
9	Second Month Weight	565	Corrected manually by 56.5 and checked with the weight when the patient first comes to facility

5.3.2.3 Resolving Inconsistencies

Since the data is encoded manually, much inconsistency was not seen in the data and those identified was due to data entry during encoding were manually managed. The inconsistency is due to data entry during encoding.

5.4 PREPROCESSED FINAL DATASET

Even though, the data was collected for the intended objectives discussed in the first chapter, the data was plagued with problems such as missing values, inconsistencies and errors. The problem of inconsistencies and errors were very few and handled manually.

The most apparent problem was missing values but the problems were dealt in discussion with domain experts. For instance, if a patient is HIV negative, their status is not filled and in that case Negative (NR) is filled for the feature. There were similar problem in the dataset and successfully handled. In the case of the outcome class, only 13 records were without class value and they were discarded. The final processed dataset contains 6319 instances with 12 variables including the class variable.

CHAPTER SIX

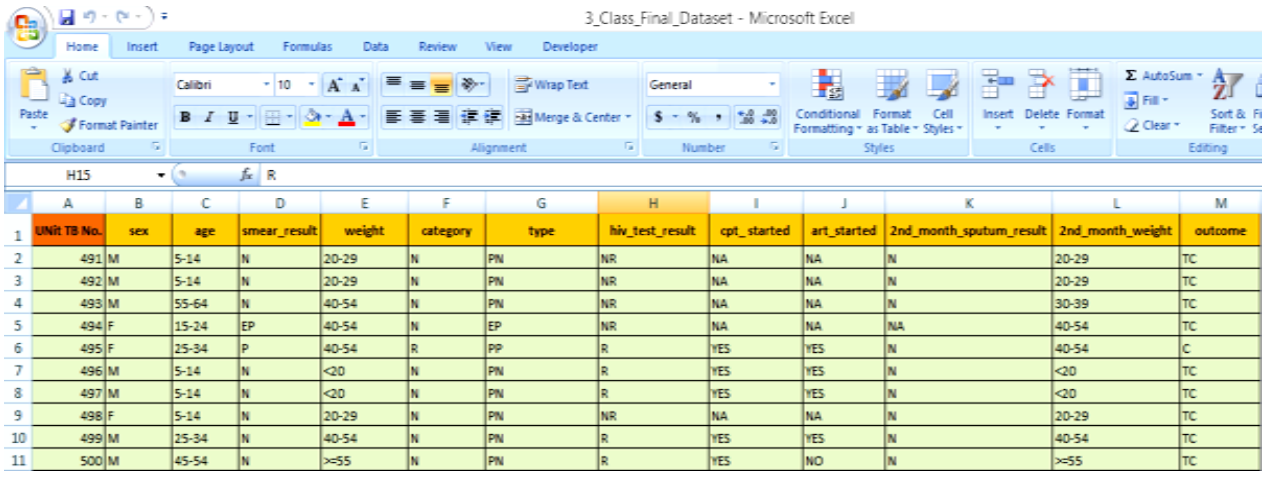
EXPERMENTATION AND EVALUATION OF THE DISCOVERED KNOWLEDGE

After preprocessing and preparation of the dataset the next step was the mining task using selected algorithms. The three selected algorithms to prepare a predictive model are J48, Naïve Bayes, SMO and PART. In this chapter successive experimentation with their performance measurement is provided and the rules generated by the best algorithm are selected and discussed with the help of an expert for its interestingness. Finally, prototype of user interface provided.

6.1 EXPERIMENTAL SETUP

The TB dataset collected from the five health centers in Excel format was converted to Comma Separated Value (CSV) file format so as to make it ready for WEKA. The dataset contained 6320 instances and a total of 12 attributes including class attribute which is an “outcome”.

The 12 attributes that were selected include: "sex", "age", "smear result", "weight", "category", "TB type", "HIV test result", "CPT started", "ART started", "second month sputum", "second month weight" and the class attribute "outcome" (Table 6.1).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	UNit TB No.	sex	age	smear_result	weight	category	type	hiv_test_result	cpt_started	art_started	2nd_month_sputum_result	2nd_month_weight	outcome
2	491	M	5-14	N	20-29	N	PN	NR	NA	NA	N	20-29	TC
3	492	M	5-14	N	20-29	N	PN	NR	NA	NA	N	20-29	TC
4	493	M	55-64	N	40-54	N	PN	NR	NA	NA	N	30-39	TC
5	494	F	15-24	EP	40-54	N	EP	NR	NA	NA	NA	40-54	TC
6	495	F	25-34	P	40-54	R	PP	R	YES	YES	N	40-54	C
7	496	M	5-14	N	<20	N	PN	R	YES	YES	N	<20	TC
8	497	M	5-14	N	<20	N	PN	R	YES	YES	N	<20	TC
9	498	F	5-14	N	20-29	N	PN	NR	NA	NA	N	20-29	TC
10	499	M	25-34	N	40-54	N	PN	R	YES	YES	N	40-54	TC
11	500	M	45-54	N	>=55	N	PN	R	YES	NO	N	>=55	TC

Figure 6.1 Example dataset after preprocessing and preparation

6.2 DATA MINING TOOL AND PARAMETER SELECTION

To develop these models, a suitable environment is required. A number of machine learning algorithms including the above mentioned classifiers are well known implementations in the freely available code package WEKA (Waikato Environment for Knowledge Analysis). This package is developed at the University of Waikato in New Zealand.

The system provides a uniform interface to a number of different learning tasks such as classification, regression, clustering, associated rules and visualization. The algorithms can either be employed directly to a dataset or called from other JAVA code. Moreover, the environment is capable of pre-processing and post-processing to evaluate the result of a learning scheme on any offered dataset.

The dataset was imported to WEKA 3.6.9, stable version, selected for the data mining task as shown below in figure 6.1.

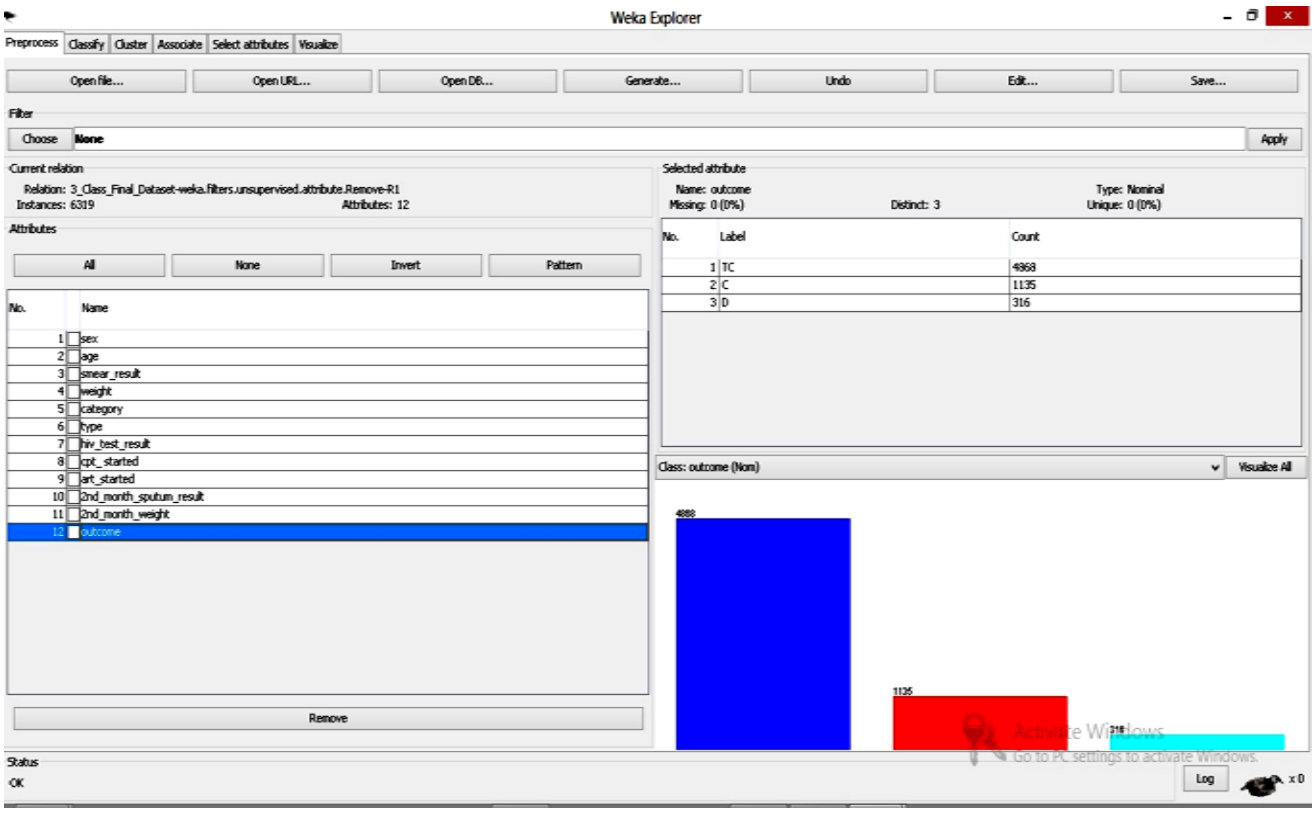


Figure 6.2 Data loading

In order to get statistically meaningful results, the default number of iterations is 10. In case of 10-fold cross-validation this means 100 calls of one classifier with training data and tested against test data. All the other parameters are left in their default, except for tree pruning, in the case of J-48 algorithm, is used.

6.3 POST-PROCESSING OF THE DATASET

Understanding the dataset also includes looking in to the distribution of outcome class carefully since it determines the choosing of the right predictive performance indicators. One major issue is the problem of imbalanced outcome class.

6.3.1 The Problem of Imbalanced Dataset and Performance Metric

A dataset is imbalanced if the classes are not approximately equally represented [41]. The TB dataset used for the mining task in this project is represented approximately on the order of 100 (Successful Treatment Class) to 5 (Death Class).

Performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced. [41]

In this project, high rate of correct detection of the minority class (Death Class) is required and small errors in the majority class in order to achieve this. Simple predictive accuracy is clearly not appropriate in such situations.

Rather, the Receiver Operating Characteristic (ROC) curve is a standard technique for summarizing classifier performance over a range of tradeoffs between true positive and false positive errors [41].

ROC curves can be thought of as representing the family of best decision boundaries for relative costs of TP and FP. On an ROC curve the X-axis represents %FP = FP/(TN+FP) and the Y-axis represents %TP = TP/(TP+FN). The ideal point on the ROC curve would be (0,100), that is all positive examples are classified correctly and no negative examples are misclassified as

positive. The area under the curve (AUC) is an accepted performance metric for an ROC curve [41].

As it is pointed above, the dataset analyzed represents an imbalanced distribution of Treatment Success (Treatment Completed plus Cure) and Death. The number of instances of Treatment Success dominates that of Death. Such imbalance leads to bias towards the majority class. In order to deal with this problem, SMOTE (Synthetic Minority Over-sampling Technique), a technique that helps us to adjust the distribution of minority class instances artificially, was used.

6.4 CLASSIFICATION MODELING USING J48 DECISION TREE

The decision tree used in WEKA is termed as J48 which is a modification of the C4.5 algorithm. The outcome variable takes three values i.e. Treatment completed, Cure and Death. The number of death in the dataset is very small in that it shows class imbalance as it is clearly seen in the bar graph below in Figure 6.1 and this imbalance leads to incorrect classification by the decision tree algorithms. The strategy to mitigate this problem is applying SMOTE.

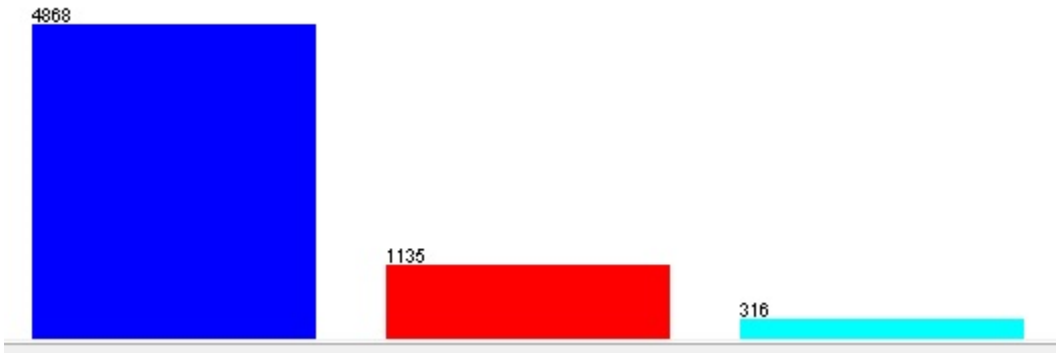


Figure 6.3 Outcome class (TC= 4868, C=1135 and D=315) before SMOTE is applied

6.4.1 J48 Experiment I

The first experiment was done with default parameters and without the application of SMOTE. In this case, though SMOTE was not applied to the dataset the accuracy seems good. In imbalanced dataset such as this, accuracy is not a good measure but ROC area with accuracy is the performance measure used to compare experiment findings. The next two experiments were done with the application of SMOTE. Table 6.3

Table6.2 Experimentation with J48 using default parameters before SMOTE

Experiment No	Accuracy	WTP Rate	WFP Rate	WROC Area
1	87.3714%	0.874	0.179	0.848

6.4.2 J48 Experiment II

Experiment two was done with default parameters but applying successive SMOTE to rebalance the class. As the histogram in figure 6.2 reveled, with 300% SMOTE, the minority class is re-balanced but after 300% SMOTE the previously minority class (Death) became the majority class.

Table6.3 Experimentation with J48 using default parameters after SMOTE 300%

Experiment No	SMOTE	Accuracy	WTP Rate	WFP Rate	WROC Area
1	100%	83.2102%	0.832	0.247	0.794
2	200%	79.9642%	0.8	0.222	0.840
3	300%	81.1473%	0.811	0.16	0.863

Though, the accuracy decreased from 87% before SMOTE to 81.15% after 300% SMOTE, ROC area slightly increased from 0.848 to 0.863.

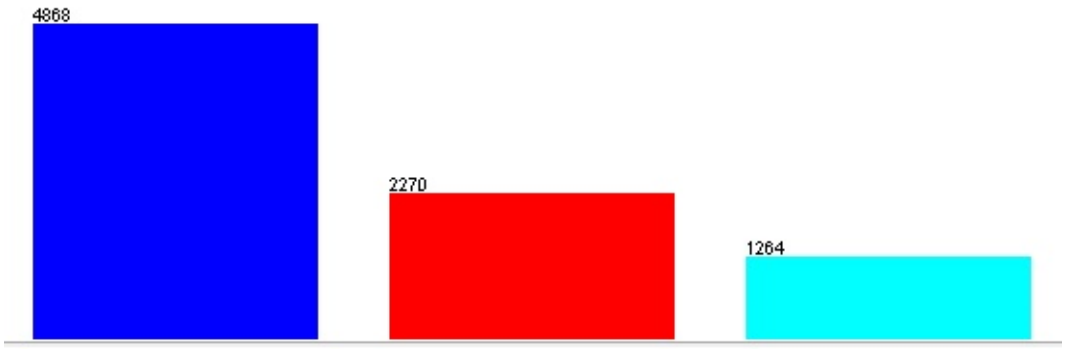


Figure 6.4 Outcome Class after 300% SMOTE

6.4.3 J48 Experiment III

The third experiment was done by setting unprunned parameter value to “True” after SMOTE 300%. The tree generated will represent unprunned decision tree. The last two experiments were done by setting the unprunned parameter value to “False”.

Table6.4 Experimentation with J48 using unprunned parameter true after SMOTE 300%

Experiment No	SMOTE	Accuracy	WTP Rate	WFP Rate	WROC Area
1	300%	81.0997 %	0.811	0.142	0.888

Accuracy remained the almost the same as unprunned parameter value false but the ROC area increased from 0.863 to 0.888.

6.5 CLASSIFICATION MODELING USING NAÏVE BAYES METHOD

The Naïve Bayes classification reads a set of examples from the training set and uses the Bayes theorem to estimate the probabilities of all classifications. For each instance, the classification with the highest probability is chosen as the prediction class.

6.5.1 Naïve Bayes Experiment I

The application of Naïve Bayes technique with 10-fold cross validation as in evaluating the prediction model based on the correctly classified instances, the model has produced 87.0866 percent accuracy rate and ROC Area 0.874 (see Table 6.5).

Table6.5 Experimentation with Naïve Bayes default parameters

Experiment No	Accuracy	WTP Rate	WFP Rate	WROC Area
1	87.0866 %	0.871	0.18	0.874

6.5.2 Naïve Bayes Experiment II

Using a successive SMOTE (100%-400%), the accuracy as well as ROC area declined dramatically (see Table 6.6).

Table6.6 Experimentation with Naïve Bayes using default parameters and 100%-300% SMOTE

Experiment No	SMOTE	Accuracy	WTP Rate	WFP Rate	WROC Area
1	100%	81.3263 %	0.813	0.237	0.837
2	200%	72.5609 %	0.726	0.248	0.796
3	300%	75.6963%	0.757	0.176	0.845

These two experiments in the above tables describe the performance of the algorithm in the given dataset. First, the algorithm performs well in both measurement using default parameters and as the data is in the unbalanced state. The successive attempt to balance the data resulted in low performance.

6.6 CLASSIFICATION MODELING USING SMO ALGORITHM

The experiments for SVM was conducted with same dataset size (N= 6319) as the same as for others algorithms and test options was also the default 10 fold cross-validation pertained.

6.6.1 SMO Experiment I

The result below shows (Table 6.7) that SMO without rebalancing the dataset reflects an accuracy of 87.3714 % and ROC area 0.848 respectively. This finding clearly shows that the performance of SMO is almost the same as Naïve Bayes.

Table 6.7 Experimentation with SMO default parameters

Experiment No	Accuracy	WTP Rate	WFP Rate	WROC Area
1	87.3714 %	0.874	0.179	0.848

6.6.2 SMO Experiment II

In order to get a better result, this time rebalancing is done successively to see how the algorithm performs. But at SMOTE 200%, both the accuracy and ROC area exhibits a better performance compared to the firs and the third experiment at 100% and 300% SMOTE.

Table 6.8 Experimentation with SMO using default parameters and 100%-300% SMOTE

Experiment No	SMOTE	Accuracy	WTP Rate	WFP Rate	WROC Area
1	100%	78.5527 %	0.786	0.212	0.798
2	200%	83.2102 %	0.832	0.247	0.794
3	300%	75.9736 %	0.76	0.311	0.728

The two experiments conducted above using SMO reveal that performance is better without rebalancing the dataset artificially.

6.7 CLASSIFICATION MODELING WITH PART ALGORITHM

According to Witten and Frank there are two industrial-strength rule induction algorithms. But the one that works by repeatedly building partial decision trees and extracting rules from them (i.e. PART) is preferred to and used in this research because of its simplicity and its ability to achieve the same level of performance with others [29].

6.7.1 PART Experiment I

As it was used in J48, SMO and Naïve Bayes algorithms, an experiment with PART was also done with default parameters without any effort of re-balancing the dataset (Table 6.7).

Table 6.9 Experimentation with PART using default parameters without SMOTE

Experiment No	Accuracy	WTP Rate	WFP Rate	WROC Area
1	86.4377 %	0.864	0.222	0.866

6.7.2 PART Experiment II

PART experiment with successive SMOTE resulted in an accuracy of 81.314 and ROC area 0.892. Table 6.10 shows the result of PART algorithm with ROC increases steadily.

Table 6.10 Experimentation with PART using default parameters and 100%-300% SMOTE

Experiment No	SMOTE	Accuracy	WTPR	WFPR	WROC
1	100%	83.1047 %	0.831	0.237	0.853
2	200%	79.7991 %	0.798	0.208	0.864
3	300%	81.314 %	0.813	0.145	0.892

Comparing the two experiments, the first experiment shows a better accuracy but comparison based on ROC area is reveals that experiment two came up with a better result.

6.8 PERFORMANCE COMPARISON of NAÏVE BAYES,J-48, SMO and PART

Comparison of J48, Naïve Bayes, SMO and PART indicated that PART algorithm believed to work well in unbalanced dataset with an accuracy rate of 81.314 and ROC area 0.892.

Table 6.11 Summary of J-48, Naïve Bayes, SMO and PART

Exp No	Algorithms	SMOTE	Accuracy	WTPR	WFPR	WROC
1	J48	300%	81.1473%	0.811	0.16	0.863
2	Naïve Bayes	NO	87.0866 %	0.871	0.18	0.874
3	SMO	NO	87.3714%	0.874	0.179	0.848
4	PART	300%	81.314 %	0.813	0.145	0.892

A closer look at the four algorithms reveled something interesting. J48 and PART algorithms produced almost the same accuracy. It is not only a better accuracy rate but also a better performance in the area under the ROC except for a slight difference in the later. From the two algorithms PART has better performance.

The other two algorithms Naive Bayes and SMO, as the above two algorithms, have also generated comparable accuracy rate and ROC area for a small increase in the former. Naive Bayes performs better than SMO.

Bayes and PART are the two candidate algorithms in this project work. Due to a better ROC area the former is taken though it is less in accuracy rate. Therefore, PART is the algorithm of choice to generate rules and for further analysis for their interestingness in the next section.

6.9 EVALUATION OF THE CLASSIFICATION RULES

The rules generated using PART algorithms are given as follows. The selections of the rules are based on objective measurements given in parenthesis at the end of each rules. Then the interestingness of the knowledge discovered is confirmed with experts. Among the 54 rules generated only 12 is selected. Out of the 12 rules only 5 rules were found to be interesting for the experts.

6.8.1 Rules selected for Treatment completed

Rule #1 IF smear_result = EP AND age = 15-24 AND hiv_test_result= NR THEN TC (518.0/31.0)

Rule #2 IF smear_result = EP AND age = 25-34 AND 2^{nd}_month_weight = >=55 AND category = N THEN TC (223.0/3.0)

Rule #3 IF smear_result = N AND weight = >=55 AND 2^{nd}_month_weight = >=55: THEN TC (543.0/51.0)

Rule #4 IF smear_result = EP AND category = N AND age = 25-34 AND hiv_test_result = NR: THEN TC (252.0/42.0)

Rule #5 IF smear_result = N AND age = 15-24 AND weight = 40-54 AND ART_started = NA: TNEN TC (355.0/58.0)

Discussion

Among the five rules selected for Treatment completed outcome, **Rules #4** is accepted by an expert. New cases of EPTB in a given age range (25-34) with HIV status negative has a high chance of completing treatment. Most EPTB patients are HIV positive and do not complete their treatment due to various reasons. Adults in the age range specified above and HIV negative are the ones who are completing treatment.

6.8.2 Rules selected for Cure

Rule #1 IF smear_result = P AND hiv_test_result = NR AND sex = F AND age = 15-24 AND 2nd_month_weight = 40-54: THEN C (328.0/42.0)

Rule #2 IF smear_result = P AND hiv_test_result = NR AND sex = F AND category = N AND age = 25-34: THEN C (238.0/24.0)

Rule #3 IF smear_result = P AND hiv_test_result = NR AND age = 35-44 AND category = N 2nd_month_sputum_result = N AND weight = 40-54: THEN C (79.0/16.0)

Discussion

As stated by a TB expert at FMoH, there is "High cure rate among New smear positive TB cases that are HIV negative as compared to HIV positives. Young females will also have better chance of cure." Thus, **Rule #2** gives sense.

The same expert who confirmed the interestingness of Rule #2 also accepted **Rule #3**. A new patient with second month smear negative and HIV test result negative, the chance of got cured is high.

6.8.3 Rules selected for Death

Rule #1 IF smear_result = EP AND sex = M AND weight = >=55 AND 2nd_month_weight = >=55 AND age = 55-64 AND art_started = NA: THEN D (42.0/11.0)

Rule #2 IF smear_result = EP AND sex = M AND category = O AND age = 25-34 AND hiv_test_result = R AND weight = 40-54: THEN D (30.0/5.0)

Rule #3 IF smear_result = N AND weight = >=55 AND age = 35-44 AND hiv_test_result = R AND category = N: THEN D (63.0/17.0)

Rule #4 IF smear_result = N AND weight = 40-54 AND age = 25-34 AND 2nd_month_weight = 40-54 AND sex = M: D (85.0/28.0)

Discussion

Among the four rules selected for Death Outcome, Rule #2 and Rule #3 are accepted by two experts.

Rule #2: A patient with EPTB, a case difficult for diagnosis, male (less adherence compared to female patients), patient category of “other” (previously treated and with new form of TB that smear negative or EPTB), HIV positive in a given age (25-34) and weight (40-54) range results in Death.

Rule #3: A new patient with Smear Result negative but Positive Serostatus, HIV Positive in a given age and weigh range likely to die. Older, new cases of smear negative TB patients who are HIV positive tend to die due to poor treatment compliance since they believed that they are free from TB (exert opinion)

6.9 PROTOTYPE DEVELOPMENT

Due to the move to smart phones applications, the rules, generated using PART algorithm, are imbedded in to mobile program using Sun Java. The ease of use of smart phones and their accessibility made them a choice of computing device today.

As indicated in the diagrams below (figure 6.3), the user fill up the form with the demographic features and the selected clinical features and may predict the outcome of the treatment with an accuracy of more than 81%.

Figure 6.5 Prototype of DM TB Outcome Prediction

CHAPTER SEVEN

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

7.1 SUMMARY

Useful information for decision makings are buried in the accumulated mountain of data over the years in the health sector. This massive collection remains as they are without much benefit to the sector.

Health informatics is a field of study that endeavors to assist health care related decisions through the application of theories and tools from information science and computer science. One of the efforts in health informatics is to use machine learning algorithms and data mining techniques to uncover the useful information in the data.

The objective of this research was to prepare a predictive model for TB treatment outcome that assists clinical decisions in connection with TB treatment. For this purpose data is collected from health facilities that provide tuberculosis treatment in their clinics. Having pre-processed the data, the data was fed in to data mining tools with the selected algorithms based on their performance discussed in different literatures. In addition, the algorithms selected are tested in bio-medical fields where much missing features and imbalanced class problem is apparent.

The data mining task used is classification with which we can successfully predict the outcome of treatment and three classifier algorithms such as J-48, Naïve Bayes, SMO and PART are used. A serious of experiments was done with the three algorithms through manipulation of some parameters and using data re-balancing techniques such as SMOTE.

7.2 CONCLUSIONS

The results of the experiments were compared using the two performance metrics accuracy and area under ROC. PART outperforms the decision tree algorithm and Naïve Bayes and found to be the best learner and best algorithm in a dataset such as tuberculosis.

The results of PART were accuracy and area under ROC 81.31 and 0.89 respectively. From the rules selected by the experts for their interestingness, the following attributes were important for the prediction of treatment outcome: Age, Sex, Initial Weight with Second Month Weight, Patient Category, TB Type and HIV Result.

7.3 RECOMMENDATIONS

Based on the findings of the research the following recommendations have been made to different levels of the health system.

Policy level:

- Data mining requires a huge collection of data in a database or data warehouse to apply data mining; such accumulated data is found in health facilities in Addis Ababa but it is in a registry book. The attempt to store data in electronic repositories had began in the country and some hospitals in Addis Ababa began the work. Such effort should also be extended to health centers in the city.

Program level:

- The features recorded in the TB registry book only reflect demographic characteristics and clinical attributes. The outcome of TB treatment is not only dependent on those features but socio-economic factors such as income level and education has also a significant influence and should be included.
- In relation to the above recommendation, TB is only associated with HIV result and treatment but chronic diseases such as diabetes mellitus and hypertension have also an impact on the prognosis of the patients.

Facility level:

- The TB registry book contains many missing values for the attributes especially Smear Result, CPT, ART, Second Month Sputum Result and others. To avoid this problem the use of training and supervision need to be considered.

Researchers:

- Due to the limitation of time and resources, the study was conducted only with 6319 records and four algorithms. However, building a model with more data and additional algorithms help to optimize the model.
- The dataset collected from the facilities only reflect three classes namely Treatment completed, Cure and Death. The other outcomes such as failure and defaulter are not found in the dataset. In order to predict the outcome of TB treatment in full magnitude with additional dataset and outcomes, further efforts should be exerted in collecting data that reflect all the TB treatment outcomes.
- Preparing a complete knowledge based system based on the rules generated and found to be interesting.

REFERENCES

[1] Introduction to Health Informatics, available at:

http://www.pitt.edu/~super4/36011-37001/36991.ppt, Accessed on January 1, 2013

[2] "Data Mining in Health Informatics", Available at http://yavar.naddaf.name/downloads/Data%20Mining%20in%20Health%20Informatics.pdf, Accessed on December 20, 2012

[3] Health Informatics, available at: http://www.nlm.nih.gov/hsrinfo/informatics.html Access on December 22, 2012

[4] T. Asha, S. Natarajan and K. N. B. Murthy, Data Mining Techniques in the Diagnosis of Tuberculosis, Bangalore Institute of Technology, Department of Information Science and Engineering, Bangalore Institute of Technology, India

[5] Federal Ministry of Health, Guidelines for Clinical and Programmatic Management of TB, Leprosy and TB/HIV in Ethiopia, April 2012 Addis Ababa

[6] Tewodros Mengesha, Electronic Solutions for Ethiopian Health Sector Electronic Medical Record (EMR) System, autumn 2011, Oulu University of Applied Sciences

[7] Minale Tefera, Application of Data Mining Techniques to Predict Urinary Fistula Surgical Repair Outcome , Faculty of Informatics, Addis Ababa University June 2012

[8] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia, College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Al-Kharj, Kingdom of Saudi Arabia

[9] Millennium Development Goals (MDGs), available at http://www.who.int/mediacentre/factsheets/fs290/en/ Accessed on May 13, 2013

[10] Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC; 2007.

[11] Tuberculosis, available at: http://www.cdc.gov/tb/topic/basics/default.htm Accessed on May 14, 2013

[12] Directly Observed Therapy, available at: http://www.harlemtbcenter.org/dot_program.htm Accessed on May 14, 2013

[13] World Health Organization, Treatment of tuberculosis Guidelines 4th Edition, 2010

[14] Tuberculosis Treatment Outcomes, available at: http://www.atsjournals.org/doi/full/10.1164/rccm.200401-095OC#abstract,Accessed on May 14, 2013

[15] David J. Hand (1998). Data Mining: Statistics and More?, The American Statistician, Vol. 52, No. 2 pp. 113

[16] Arun George Eapen., (2004) Application of Data mining in Medical Applications. MSc, University of Waterloo

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996).From Data Mining to Knowledge Discovery in Databases, Magazine Volume 17 p.37 [Online version] [Viewed on May 10, 2013]

[18] Deshpande, S. P. and Thakare, V. M., (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS), Vol.1, No.1, p. 32 [Online version] [viewed on: 4/12/2013]

[19] Osmar R. Zaïane, (1999). Principles of Knowledge Discovery in Databases, University of Alberta

[20] Larose Daniel T. Discovering Knowledge in Data - An Introduction to Data Mining. New Jersey, USA: John Wiley & Sons Inc; 2005.

[21] Asia Nesredin (2012). Mining Patients' Data for Effective Tuberculosis Diagnosis: The Case of Menelik Ii Hospital. MSc. Addis Ababa University

[22] Bramer Max. Principles of Data Mining. London. Springer-Verlag Limited; 2007

[23] Two Crows Corporation, 1999, Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation.

[24] Berry M. and Linoff G., 2004, Data Mining Techniques for Marketing, Sales and Customer Relationship Management.

[25] Bhargavi P. and Jyothi S., Applying Naïve Bayes Data Mining Technique for Classification of Agricultural Land Soils, (IJCSNS) International Journal of Computer Science and Network Security. 2009, 9(8).

[26] Tibebe Beshah (2005). Application of Data Mining Technology to Support Road Traffic Accident Severity Analysis at Addis Ababa Traffic Office. MSc. Thesis, Addis Ababa University

[27] Han Jiawei, Kamber Micheline. Data Mining: Concepts and Techniques. New York. USA: Morgan Kaufmann Publishers; 2001.

[28] Weiss Sholom M, Zhang Tong. Performance Analysis and Evaluation. In: Ye Nong, Editor. The handbook of data mining. New Jersey. USA: Lawrence Erlbaum Associates Inc; 2003.

[29] Witten Ian H, Frank Eibe. Data Mining: Practical Machine Learning Tools and Techniques. Second edition. USA: Elsevier inc; 2005.

[30] Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC; 2007.

[31] Sumathi S, Sivanandam SN. Introduction to Data: Mining and its Applications. Berlin, German: Springer-Verlag inc; 2006.

[32] Bath Peter A. Data Mining in Health and Medical Information. In: Blaise Cronin, editor. Annual review of Information Science and Technology. Vol 38. USA: Information Today Inc; 2004.

[33] Jerome H. Friedman (2000) Data Mining and Statistis: What is the Connection. Department of Statistics and Stanford Linear Accelerator Center, Stanford University

[34] Arun George Eapen (2004). Application of Data Mining in Medical Application. MSc Thesis. Univesity of Waterloo

[35] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia, College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Al-Kharj, Kingdom of Saudi Arabia

[36] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan (2010). "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks". IJCSE Vol. 02, No. 02 pp. 250

[37] Rudolf Kruse, Bernhard A. Sabel, Christian Borgelt, Data mining in diagnostic charts and treatment outcome prediction for Vision Restoration Therapy , March 2008, Otto-von-Guericke University Faculty of Computer Science

[38] Casey Bennett, Thomas W. Doub, Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice, Proceedings of the 6th International Conference on Data Mining. pp. 313-318

[39] Asli Uyar , Ayse Bener, H. Nadir Ciray, Mustafa Bahceci (2009). Predicting Implantation Outcome from Imbalanced IVF Dataset. Proceedings of the World Congress on Engineering and Computer Science Vol II

[40] Sharareh RostamNiakanKalhori (2011). An Integrated Supervised and Unsupervised Learning Approach to Predict the Outcome of Tuberculosis Treatment Course. PhD Desertation, University of Manchester

[41] Chawla, Bowyer, Hall, Kegelmeyer (2002) SMOTE: Synthetic Minority Over-sampling TEchinque.Journal of Arteficial Intelligence Research, Volume 16, 321-357

[42] Chakrabarti,Cox, Frank, Güting, Han, Jiang, Kamber, Lightstone, Nadeau, Neapolitan, Pyle, Refaat,Schneider, Teorey, Witten . Data Mining Know It All. USA: ELSEVIER; 2009

[43] Te-Ming Huang, Vojislav Kecman, IvicaKopriva. (2006). Kernel Based Algorithms for Mining Huge Data Sets. Springer-Verlag Berlin Heidelberg

[45] Warner H, Toronto A, Veasy L, et al. A mathematical approach to medical diagnosis. Application to congenital heart diseases. JAMA 1961;177:177e83.

[46] Wei Wei, Shyam Visweswaran, Gregory F Cooper. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.

APPENDICES

APPENDIX A: J-48 OUTPUT

Classifier output

```
Time taken to build model: 0.11 seconds

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

Correctly Classified Instances        6818               81.1473 %
Incorrectly Classified Instances      1584               18.8527 %
Kappa statistic                          0.6586
Mean absolute error                      0.1974
Root mean squared error                  0.3197
Relative absolute error                 52.0638 %
Root relative squared error             73.4227 %
Total Number of Instances             8402

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.881     0.224      0.844     0.881     0.862      0.859    TC
                 0.923     0.094      0.784     0.923     0.848      0.93     C
                 0.343     0.03       0.668     0.343     0.453      0.754    D
Weighted Avg.    0.811     0.16       0.801     0.811     0.797      0.863

=== Confusion Matrix ===

    a    b    c   <-- classified as
 4289  410  169 |    a = TC
  128 2096   46 |    b = C
  662  169  433 |    c = D
```

APPENDIX B: SMO OUTPUT

Classifier output

```
Time taken to build model: 52.98 seconds

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

Correctly Classified Instances        6600               78.5527 %
Incorrectly Classified Instances      1802               21.4473 %
Kappa statistic                          0.5919
Mean absolute error                      0.2799
Root mean squared error                  0.3629
Relative absolute error                 73.8179 %
Root relative squared error             83.3625 %
Total Number of Instances             8402

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.912     0.314      0.8       0.912     0.853      0.799    TC
                 0.951     0.113      0.757     0.951     0.843      0.919    C
                 0         0          0         0         0          0.577    D
Weighted Avg.    0.786     0.212      0.668     0.786     0.722      0.798

=== Confusion Matrix ===

    a    b    c   <-- classified as
 4442  426    0 |    a = TC
  112 2158    0 |    b = C
  996  268    0 |    c = D
```

APPENDIX C: NAÏVE BAYES OUTPUT

Classifier output

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances        5503               87.0866 %
Incorrectly Classified Instances       816               12.9134 %
Kappa statistic                          0.6585
Mean absolute error                      0.1187
Root mean squared error                  0.287
Relative absolute error                 47.8679 %
Root relative squared error             81.5379 %
Total Number of Instances             6319

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.909     0.212      0.935     0.909     0.922      0.873    TC
                 0.948     0.095      0.686     0.948     0.796      0.937    C
                 0         0.003      0         0         0          0.657    D
Weighted Avg.    0.871     0.18       0.844     0.871     0.853      0.874

=== Confusion Matrix ===

    a    b    c   <-- classified as
 4427  426   15 |    a = TC
   58 1076    1 |    b = C
  249   67    0 |    c = D
```

APPENDIX D: PART OUTPUT

Classifier output

```
Time taken to build model: 0.47 seconds

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

Correctly Classified Instances        6832               81.314  %
Incorrectly Classified Instances      1570               18.686  %
Kappa statistic                          0.6678
Mean absolute error                      0.1816
Root mean squared error                  0.3107
Relative absolute error                 47.8971 %
Root relative squared error             71.3699 %
Total Number of Instances             8402

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.199      0.857     0.864     0.861      0.883    TC
                 0.904     0.082      0.804     0.904     0.851      0.944    C
                 0.454     0.051      0.61      0.454     0.521      0.835    D
Weighted Avg.    0.813     0.145      0.805     0.813     0.807      0.892

=== Confusion Matrix ===

    a    b    c   <-- classified as
 4207  370  291 |    a = TC
  143 2051   76 |    b = C
  559  131  574 |    c = D
```

APPENDIX E: RULES GENRARTED USING PART

TREATMENT COMPLETED = 19 Rules			
smear_result = EP AND age = 45-54 AND art_started = NO: TC (26.0)	smear_result = EP AND age = 45-54 AND art_started = NA: TC (101.0/5.0)	smear_result = EP AND age = 5-14 AND art_started = NA: TC (102.0/6.0)	smear_result = EP AND age = 15-24 AND hiv_test_result = NR: TC (518.0/31.0)
smear_result = EP AND age = 35-44 AND 2nd_month_weight = >=55: TC (156.0/3.0)	smear_result = EP AND age = 25-34 AND 2nd_month_weight = >=55 AND category = N: TC (223.0/3.0)	smear_result = N AND hiv_test_result = NR AND weight = 30-39 AND category = N: TC (135.0/6.0)	smear_result = N AND weight = >=55 AND 2nd_month_weight = >=55: TC (543.0/51.0)
smear_result = EP AND art_started = YES AND weight = 40-54 AND category = N: TC (124.0/17.0)	smear_result = EP AND category = N AND age = 25-34 AND hiv_test_result = NR: TC (252.0/42.0)	smear_result = N AND age = 15-24 AND weight = 40-54 AND art_started = NA: TC (355.0/58.0)	smear_result = N AND age = 25-34 AND category = N AND art_started = YES AND weight = 40-54: TC (126.0/15.0)
smear_result = N AND age = 55-64 AND 2nd_month_weight = 40-54: TC (69.0/2.0)	smear_result = N AND weight = 40-54 AND sex = M AND cpt_ started = YES AND art_started = NO AND 2nd_month_weight = 40-54 AND age = 35-44 AND category = N: TC (29.0/5.0)	smear_result = N AND weight = 40-54 AND sex = M AND cpt_ started = NA AND category = N: TC (188.0/56.0)	smear_result = N AND weight = 40-54 AND art_started = YES: TC (109.0/25.0)
smear_result = EP AND hiv_test_result = R AND cpt_ started = YES AND 2nd_month_weight = 40-54 AND age = 35-44 AND weight = 40-54: TC (31.0/7.0)	smear_result = EP AND cpt_ started = NO AND weight = 40-54: TC (14.0/2.0)		
Cure =14 Rules			
smear_result = P AND hiv_test_result = NR AND age = 25-34 AND weight = 40-54 AND category = N: C (243.0/43.0)	smear_result = P AND hiv_test_result = NR AND sex = F AND category = N AND age = 15-24: C (101.0/14.0)	smear_result = P AND hiv_test_result = NR AND age = 35-44 AND category = N AND 2nd_month_sputum_result = N AND weight = 40-54: C (79.0/16.0)	2nd_month_weight = >=55 AND age = 15-24 AND category = N AND art_started = NA: C (132.0/28.0)

smear_result = P AND 2nd_month_weight = 20-29 AND age = 5-14 AND weight = 20-29: C (18.0/4.0)	2nd_month_weight = >=55 AND age = 35-44 AND weight = >=55 AND sex = M: C (82.0/16.0)	2nd_month_weight = 40-54 AND age = 15-24 AND category = N AND smear_result = P: C (236.0/48.0)	2nd_month_weight = 40-54 AND sex = F AND category = N AND smear_result = P AND age = 25-34: C (101.0/20.0)
smear_result = P AND art_started = NA AND sex = F AND age = 45-54: C (52.0/6.0)	category = N AND sex = F AND 2nd_month_weight = 40-54 AND age = 35-44: C (64.0/12.0)	category = N AND sex = M AND art_started = NA: C (172.0/72.0)	category = N AND age = 15-24 AND smear_result = P AND art_started = NO: C (11.0/1.0)
art_started = NO AND age = 45-54 AND category = N AND cpt_ started = YES: C (26.0/3.0)	cpt_ started = NA AND age = 45-54 AND 2nd_month_weight = 40-54: C (13.0/2.0)		
Death = 21 Rules			
smear_result = EP AND sex = M AND cpt_ started = NA AND age = >=65 AND 2nd_month_weight = 40-54: D (29.0/11.0)	smear_result = N AND weight = 30-39 AND cpt_ started = NA AND sex = F: D (12.0/4.0)	smear_result = N AND weight = 30-39 AND category = N AND 2nd_month_weight = >=55 AND cpt_ started = YES: D (11.0/1.0)	smear_result = N AND weight = >=55 AND age = 35-44 AND hiv_test_result = R AND category = N: D (63.0/17.0)
smear_result = N AND weight = 40-54 AND sex = F AND category = N AND 2nd_month_weight = >=55 AND age = 45-54: D (66.0/16.0)	smear_result = EP AND age = >=65 AND weight = >=55: D (20.0/2.0)	smear_result = N AND sex = F AND category = N AND age = >=65 AND 2nd_month_weight = 40-54: D (37.0/18.0)	smear_result = EP AND age = 35-44 AND 2nd_month_weight = 40-54 AND weight = 30-39: D (17.0/3.0)
smear_result = EP AND cpt_ started = YES AND age = 15-24 AND sex = F AND weight = 40-54: D (21.0/5.0)	smear_result = EP AND cpt_ started = YES AND sex = M AND age = 25-34 AND art_started = NO: D (28.0/13.0)	art_started = NO AND age = 25-34 AND sex = F: D (30.0/3.0)	category = N AND art_started = NO AND sex = M: D (58.0/21.0)
cpt_ started = YES AND age = 45-54 AND category = N: D (29.0/8.0)			

APPENDIX F: TB REGISTER

Left Hand Page

Sex (M/F)	Name of contact person	Smear result	Category N.R.F.D.T.O	Intensive phase		Treatment started (DD/MM/YY)	Write the month	Intensive phase treatment monitoring chart
Age	Address contact person (Woreda, Kebele, HNo)	Lab. no. / Weight	P/Pos, P/Neg or EP	Drug	Dose			Days: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41)

Right Hand Page

UNIT TB REGISTER

HIV test performed (√)	HIV test result (R or NR or I)	CPT started (DD/MM/YY)	Enrolled in HIV care (DD/MM/YY)	ART started (DD/MM/YY)	Sputum results, lab.name, serial nr.& wt			Continuation phase		Continuation phase treatment monitoring chart weekly attendance**														Write the date (DD/MM/YY) that treatment was stopped in appropriate column:						Remarks
					Months*					Month:																				
					2nd	5th	6th	Drug	Dose	Ham	Neh	Pag	Mes	Tik	Hid	Tah	Tir	Yek	Meg	Mia	Gin	Sen	Cured	Treatment Completed	Died	Failure	Defaulted	Transferred out	
(43)	(44)	(45)	(46)	(47)	(48)	(49)	(50)	(51)	(52)	(53)	(54)	(55)	(56)	(57)	(58)	(59)	(60)	(61)	(62)	(63)	(64)	(65)	(66)	(67)	(68)	(69)	(70)	(71)	(72)