ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

MINING PATIENTS' DATA FOR EFFECTIVE TUBERCULOSIS DIAGNOSIS: THE CASE OF MENELIK II HOSPITAL

ASIA NESREDIN

JUNE 2012

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

MINING PATIENTS' DATA FOR EFFECTIVE TUBERCULOSIS DIAGNOSIS: THE CASE OF MENELIK II HOSPITAL

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Information Science

By

ASIA NESREDIN

JUNE 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

MINING PATIENTS' DATA FOR EFFECTIVE TUBERCULOSIS DIAGNOSIS: THE CASE OF MENELIK II HOSPITAL

By

ASIA NESREDIN

Name and Signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____

DEDICATION

I would like to dedicate this thesis to my father and my mother, Ato Nesredin Seid and W/o Zemzem Tajedin.

ACKNOWLEDGMENT

The first and the most special thanks go to the almighty god Allah, thank you for giving me favors in your sight. All I am and all I have, it is because of you Alhamdulillah.

I gratefully acknowledge Mekelle University and I would also like to acknowledge Addis Ababa University for providing me the necessary benefits for my study. I am also very thankful to my instructors and all staff members of Information Science Department.

I wish to express my deepest gratitude to my advisor Dr. Rahel Bekele. I cannot have sufficient words to thank you for your dedicated and highly qualified support, swift feedback, friendliness, patience and understanding. Your advice and encouragement gave me confidence to think critically. Specially for effective reading of my paper. Without your close guidance, it would have been difficult to accomplish this work.

I gratefully appreciate the support and encouragement of my colleagues from Menelik II hospital TB clinic case team Department and the Addis Ababa Health Bureau. I owe my sincere gratitude to my colleague Sister Asnakech Tizazu for her encouragement and support during the fieldwork. I would love to thank Sister Meymuna Yesuf for her skilful contribution through encoding the manual data and giving the general information about the problem domain. I wish to thank Dr. Alemu for facilitating the data collection process.

I owe sincere gratitude to my father Nesredin and my mother Zemzem for raising me with love and encouragement to live a purpose-driven life. I wish to thank my whole family specially my brothers Jemal and Seyd for their love and constant moral support. I am deeply indebted to my uncle Beshir, for his financial and moral support. I earnestly appreciate the splendid care and sincere love expressed by mom to my daughter until I finish my study.

My mom, I can never thank you enough for your love, care and understanding. You are the source of inspiration and the reason of success in my life. My little daughter Selma I missed you so much as you missed me during my long absence. Thank you for your love and endurance. I love you so much.

Finally, I would like to thank my friends and others who participated in this study.

LIST OF ABBREVIATIONS

ARFF:	Attribute Relation File Format
ART:	Antiretroviral Therapy
CRISP-DM:	Cross Industry Standard Process for Data Mining
CSV:	Comma Separated Values
DM:	Data Mining
DOT's:	Direct Observation Therapy
HIV:	Human Immunodeficiency Virus
HSDP:	Health Sector Development Plan
KDD:	Knowledge Discovery in Data Bases
SEMMA:	Sample Explore Modify Model Assess
TB:	Tuberculosis
VCT:	Voluntary Counseling and Testing
WEKA:	Waikato Environment for Knowledge Analysis
WHO:	World Health Organization

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

TB is a common and deadly infectious disease that can occur at any age. The incidence of TB has more than doubled in Africa during the last two decades. Ethiopia is also ranked 8th among the 22 countries with the highest TB burden in the world.

On the other hand, the advances in computing and information storage have provided vast amounts of data. However, the challenge has been to extract knowledge from this raw data. This has lead to new methods and techniques such as data mining that can bridge the knowledge gap.

Data mining can be used to model health care problems. This research aimed to apply data mining techniques to patients' data to establish meaningful relationships or patterns for effective TB diagnosis. In general, 7069 data sets were extracted from Menelik II hospital. The data set contains patients' detail information. The research establishes whether patients' data are classified using various data mining techniques for predicting purpose.

The research specifically look at the use of clustering algorithm followed by classification for a data mining approach to help identify patients patterns and speed up the process of TB diagnosis system. The study has tried to apply k-means clustering with some enhancements to aid in the process of segmenting the dataset into TB-positive and TB-negative. The resulting cluster is then used for developing the classification model. Classification was employed in the study to identify patterns and predict the occurrence of TB. The classification task was made using J48 decision tree and Naïve Bayes classification algorithms and different experimentations was conducted. The model developed for predicting purpose has an accuracy of 85.93%. The discovered knowledge from the J48 decision tree is presented by traversing the tree for the ease of understanding.

The outcome of the research can have many benefits, to the organization especially for TB diagnosis activities.

CHAPTER ONE
INTRODUCTION

1.1. Background

The healthcare industry has experienced a proliferation of innovations aimed at enhancing life expectancy, quality of life, diagnostic and treatment options, as well as the efficiency and cost effectiveness of the healthcare system (Vincent et al, 2010). Moreover, primary health care plays a central role in health care systems worldwide. It can offer families cost-effective services close to their home. Particularly in developing countries community health centers usually offer a broad range of services, including prenatal care, immunizations, treatment of childhood illnesses, treatment of malaria, tuberculosis and other common infectious diseases, and other basic medical care.

Ethiopia experiences a heavy burden of disease mainly attributed to communicable infectious diseases and nutritional deficiencies. Shortage and high turnover of human resource; inadequacy of essential drugs and supplies have also contributed to the burden (Abdi et al, 2010).

In this regard, Ethiopia's main health problems are said to be communicable diseases caused by poor sanitation and malnutrition (Richard, 2009). These problems are exacerbated by the shortage of trained manpower and health facilities. Hence, the main goal of the country is to have a health care system that gives a comprehensive and integrated primary health care at the community level emphasizing on disease preventive aspects without neglecting the curative aspects of medicine. Naturally, this means building a wide-reaching system that focuses on communicative diseases, such as HIV, Tuberculosis, and Malaria; and maternal and childcare health issues such as immunization and reproductive health. Dissemination of Information on health, hygiene and nutrition are part of the envisioned health system called Health Sector Development Plan (HSDP). To this end, there has been encouraging improvements in the coverage and utilization of the health service over the periods of implementation of HSDP.

Tuberculosis (TB) is one of those infectious diseases caused by bacteria whose scientific name is mycobacterium. It was first isolated in 1882 by a German physician named Robert Koch. Many years ago, this disease was referred to us consumption because without effective diagnosis and treatment, these patients often would waste away. Today, of course, tuberculosis usually can be treated successfully with antibiotics but there is a problem with diagnosis of the disease. This is because TB can remain in an inactive (dormant) state for years without causing symptoms or spreading to other people. When the immune system of a patient with dormant TB weakens, the TB can became active (reactivate) and cause infection in the lungs or other parts of the body.

TB is a common and deadly infectious disease that can occur at any age. It mostly affects the lungs, but can also affect other organs, including the central nervous system (Abdi et al, 2009). TB is a leading cause of morbidity and mortality in adults worldwide, killing more than 1.5 million people every year (Elamy et al, 2010). Prolonged delay of such patients to treatment may lead to more advanced disease, high mortality, and enhance continual transmission in the community. Diagnostic delay reflects patient delays in seeking health care, health care providers delay in making prompt and correct diagnosis and initiation of treatment.

The incidence of TB has more than doubled in Africa during the last two decades. Ethiopia is ranked 8th among the 22 countries with the highest TB burden in the world (WHO, 2003/2004). This unprecedented increase in TB is attributable to a number of factors, one of the most important being the large number of infectious TB patients who remain undetected and untreated, thereby maintaining the cycle of TB transmission (Abdi et al, 2010). The ability of TB control programs to contain this growing number of undetected TB patients is constrained by factors that deter TB patients from seeking prompt medical care.

The international standard for tuberculosis control is the World Health Organization's Direct Observation of Therapy (DOT) strategy that aims to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic tuberculosis patients who come at health care facilities (WHO, 2010).

1.2. Statement of the problem

The underlying research problem that necessitated this research is the existence of high death rate of TB at a national level. Menelik II hospital is a hospital which have large amount of TB patients than the other hospitals in Addis Ababa. Although a good treatment of the patients is periodically carried out in the hospital, there is a challenge on the diagnosis of the disease. As described by the experts of the hospital, some of the patients' result of laboratory does not indicate clearly the bacteria for TB. Hence, by assuming the patients' disease can be TB, they started the therapy for the disease. After some weeks it may be discovered that it is wrongly diagnosed. This leads to delay the control program of the disease and because of such kind of problems lots of patients die.

Although, there are lots of data gathered in manual databases of the organization, due to lack of computerized databases and appropriate data analysis tools these data are not practically used to alleviate the problems faced by health-care professionals, planners and policy makers to identify major determinant factors for effective diagnosis of TB and in order to plan and implement TB control programme strategies to reduce death in Ethiopia.

TB is a great problem in most low income countries; it is the single most frequent cause of death in individuals aged fifteen to forty-nine years (Temurtas et al, 2010). TB continues to be a major public health concern. The devastating impact of TB on vulnerable populations is also driven by its deadly synergy with HIV. HIV infection compounds the problems of accurate diagnosis as well as adequate treatment. TB causes more rapid deterioration of the immune systems of people with HIV or AIDS, and they are 100 times more likely to have active TB during their lifetime than people who are HIV-negative (Elamy et al, 2010). People with advanced HIV infection can have active TB that is smear-negative or without typical chest radiography features, which means that co-existent TB infections remain untreated.

On the other hand, the amount of data stored in medical databases increases exponentially with time. The technological advancement resulted in the management of huge computerized data acquisition and storage of data bases are contains hidden knowledge that can be important and useful for decision making. It is impossible and time consuming to unravel this knowledge.

Moreover, improper conclusions ultimately affect decision making. Consequently, a need to use more efficient techniques and to have or provide knowledge in a comprehensive form as well as to arrive at better results has developed from both the owner and users of the data bases. This has lead to the exploration of a new field of research called data mining. Data mining refers to computer-aided pattern discovery of previously unknown interrelationships and recurrences across seemingly unrelated attributes in order to predict actions, behaviors and outcomes (Madan, 2006). Data mining, in fact, helps to identify patterns and relationships in the data.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns), clustering (finding and visually documenting groups of previously unknown facts) and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities) (Asha et al, 2009).

However, in developing countries there is lack of the most elementary data in databases. Those available are often not sufficient quality to provide health care planners with required information on levels and trends of TB diagnosis. In Ethiopia, the practical challenge for health care providers, planners, and policy makers working in primary health care prevention and control activities is lack of timely and reliable health information on the health states of defined population groups.

The problem in this research emanates from the fact that there is lack of knowledge among primary health care workers which has become obstacle to address TB diagnosis effectively especially in developing countries like Ethiopia. Therefore, building capacity and enhancing universal access to rapid and accurate laboratory diagnostics are necessary to control TB and HIV-TB co-infections in resource-limited countries.

Different substantial researches have been done on TB diagnosis abroad, such as Sebban et al (2001), demonstrate how Data Mining methods have been applied to the evolutionary genetics and molecular epidemiology of TB through specification of a technique that reduced some of the

experimental constraints, improving the expert's knowledge of unknown patterns. Another research were conducted on the title "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification" by Asha et al, the methodology they used is based on clustering and classification that classifies TB into two categories, Pulmonary Tuberculosis(PTB) and retroviral PTB(RPTB) that is those with Human Immunodeficiency Virus (HIV) infection and the model they proposed can help for doctors in the diagnosis and for categorization purpose. Moreover, Tolmie, (1997); Demssie, (2002); Abdi et al, (2009); Elamy et al, (2010) also studied factors that motivate tuberculosis transmission. Finally, they come up with a conclusion of delay in seeking health care, low case detection rate, poor public awareness about TB, poor quality in diagnostic procedures, and lack of proper knowledge by community health workers to be the major challenges of TB controlling program in many developing countries.

However, the problem is all those previous studies were conducted by using a very small proportion of the database. Besides, in those studies, data analysis was conducted by using simple statistical techniques (such as regression and verification techniques). Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world databases, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases (Plate et al, 1997).

Thus, in this study the researcher investigated how tuberculosis can be diagnosed using the background history of the patients data available in Menelik II hospital by applying the new computerized methods of data mining technology and to the best of the present researcher, there is no work done to alleviate such problems by building a model for tuberculosis diagnosis. Thus, the purpose of this study is to review current policies of local tuberculosis control programmes for the diagnosis of tuberculosis and describes a data-mining approach that uses patients' data to analyze the relationship among different variables and the tuberculosis diagnostic category registered for each patient. Hence, this research attempts to answer the following questions:

- ✓ Which DM algorithm can be more suitable for the purpose of identifying the behavior of patients for effective diagnosis?
- ✓ What are the patterns that predict whether a given patient have Tuberculosis or not?
- ✓ Which DM algorithm can be more suitable for the purpose of identifying/predicting the futurity of the patients?

1.3. Objective of the study

The general objective of the present work is to design a model using data mining techniques for effective tuberculosis diagnosis. Specific objectives include:

- ✓ To build data mining model for Tuberculosis diagnosis
- ✓ To evaluate the performance of the model
- ✓ To identify limitations with such a model
- ✓ To report findings of the result
- ✓ To draw recommendations for such a model

1.4. Ethical considerations

- ➢ The research has nothing to do with the personal identifiers (like name and address) of individual about whom the data is collected and hence there is no problem of privacy and the confidentiality of individuals
- ➢ The research is purely dedicated to academic purpose (Masters Thesis for the Partial Fulfillment of M.Sc. Degree in Information Science)
- ➢ Ethical clearance was also obtained from Addis Ababa Health Bureau
- ➢ The research is purely for public benefit in general and TB patients under study to improve their health in particular

1.5. Scope and limitation

The study is intended to design a model for effective tuberculosis diagnosis and focuses on Menelik II Hospital.

The main limitation of this study was the availability of the data in manual format. To encode the data in an MS-Excel format it took a lot of time. This research didn't include all of the data available in the hospital because of time and financial problems for encoding the data. Moreover, the problem of getting health related data for data mining researches is immense even though the ethical considerations are made.

1.6. Significance of the study

The developed model can have lots of advantages for the health workers and other related individuals. Some of them are as follows:

- The proposed approach can help doctors in their diagnosis decisions and also in their treatment planning procedures for different categories
- The model can give an advice for those health care workers to have knowledge of tuberculosis diagnosis and treatment
- Diagnosis of Extra-pulmonary TB is very difficult. Hence, the model can help doctors for effective diagnosis.
- To assist health care planners, policy makers, and decision makers as a decision support aid in planning and implementing health intervention programs aimed at improving patients treatment in the hospital
- The developed database can be used as a baseline for the hospital especially for TB clinic case team workers to encode their future data concerning the TB patients detail information
- The output of this study can also be an input for further research in this and other related areas in the context of our country
- This study can give hands on experience for the researcher for understanding studies in the future

1.7. Organization of the research

This study is organized into six chapters. The first chapter briefly describes background to the problem area, and states the problem, objective of the study, scope and significance of the output of the research.

The second chapter provides a description of Tuberculosis in Ethiopia. This includes description of TB, types of TB, overview about DOT's program, current practice of the organization for TB diagnosis and its treatment.

The third chapter deals with literature review about DM technology, methods/techniques and algorithms, the different methodologies and tasks of DM, and its application in health care industry.

The fourth chapter provides discussions about the experimental design of the research. The overall methodologies that were used to undertake this research work is discussed in this chapter. This includes the data sources, data understanding and data preprocessing phases.

The fifth chapter provides a detailed discussion about the experimentation part of this study. Evaluation of the discovered knowledge is also discussed at this stage.

Finally, the sixth chapter is about the conclusion and recommendations for future work.

CHAPTER TWO
TUBERCULOSIS IN ETHIOPIA

Ethiopia's health care system is among the least developed in Sub-Saharan Africa and is not, at present, able to effectively cope with the significant health problems facing the country. Communicable diseases are the primary illnesses. Acute respiratory infections such as tuberculosis, upper respiratory infections, and malaria are the Ministry of Health's priority health problems. These afflictions accounted for 17 percent of deaths and 24 percent of hospital admissions in 1994 and 1995 E.C. Poor sanitation, malnutrition, and a shortage of health facilities are some of the causes of communicable diseases.

2.1. Overview of Tuberculosis

Tuberculosis (TB) is a chronic bacterial infection caused by Mycobacterium tuberculosis. It is spread through the air and usually infects the lungs, although other organs and parts of the body can be involved as well. Most people who are infected harbor the tuberculosis bacterium without symptoms. This is known as latent tuberculosis. If the body's resistance is low because of aging, malnutrition, infections such as HIV, or other reasons, the bacteria may break out of hiding and cause active tuberculosis.

According to World Health Organization (WHO, 2008) estimates, each year, eight million people worldwide develop active tuberculosis and nearly two million die. One in 10 people who are infected with tuberculosis may develop active TB at some time in their lives. The risk of developing the active disease is greatest in the first year after infection, but active disease often does not occur until many years later.

2.1.1. Types of tuberculosis

Different scholars classify tuberculosis in different ways, based on the symptoms and the disease treats. The medical community divides tuberculosis in to two categories called pulmonary and extra-pulmonary, which together causes twelve distinct types of tuberculosis. Pulmonary tuberculosis is responsible for five of these and extra-pulmonary the remaining seven (Stanley and Swierzewski, 2011).

Pulmonary tuberculosis: this type of TB includes:

- **Primary Tuberculosis Pneumonia:** This uncommon type of TB presents as pneumonia and is very infectious. Patients have a high fever and productive cough. It occurs most often in extremely young children and the elderly. It is also seen in patients with immune-suppression, such as HIV-infected and AIDS patients.
- **Tuberculosis Pleurisy:** This usually develops soon after initial infection. A granuloma located at the edge of the lung ruptures into the pleural space, the space between the lungs and the chest wall. Usually, a couple of tablespoons of fluid can be found in the pleural space. Once the bacteria invade the space, the amount of fluid increases dramatically and compress the lung, causing shortness of breath (dyspnoea) and sharp chest pain that worsens with a deep breath (pleurisy). A chest x-ray shows significant amounts of fluid. Mild- or low-grade fever commonly is present. Tuberculosis pleurisy generally resolves without treatment; however, two-thirds of patients with tuberculosis pleurisy develop active pulmonary TB within 5 years.
- **Cavitary TB:** Cavitary TB involves the upper lobes of the lung. The bacteria cause progressive lung destruction by forming cavities, or enlarged air spaces. This type of TB occurs in reactivation disease. The upper lobes of the lung are affected because they are highly oxygenated (an environment in which M. tuberculosis thrives). Cavitary TB can, rarely, occur soon after primary infection. Symptoms include productive cough, night sweats, fever, weight loss, and weakness. There may be hemoptysis (coughing up blood). Patients with cavitary TB are highly contagious.
- **Miliary TB:** Miliary TB is disseminated TB. "Miliary" describes the appearance on chest x-ray of very small nodules throughout the lungs that look like millet seeds. Miliary TB can occur shortly after primary infection. The patient becomes acutely ill with high fever and is in danger of dying. The disease also may lead to chronic illness and slow decline. Symptoms may include fever, night sweats, and weight loss. It can be difficult to use the initial chest x-ray may be normal. Patients who are immunosuppressed and children who have been exposed to the bacteria are at high risk for developing miliary TB.

- **Laryngeal TB:** TB can infect the larynx, or the vocal chord area. It is extremely infectious.

Extra-pulmonary Tuberculosis

This type of tuberculosis occurs primarily in immune-compromised patients. These include:

- **Lymph Node Disease:** Lymph nodes contain macrophages that capture the bacteria. Any lymph node can harbor uncontrolled replication of bacteria, causing the lymph node to become enlarged. The infection can develop a fistula (passageway) from the lymph node to the skin.
- **Tuberculosis Peritonitis:** M. tuberculosis can involve the outer linings of the intestines and the linings inside the abdominal wall, producing increased fluid, as in tuberculosis pleuritis. Increased fluid leads to abdominal distention and pain. Patients are moderately ill and have fever.
- **Tuberculosis Pericarditis:** The membrane surrounding the heart (the pericardium) is affected in this condition. This causes the space between the pericardium and the heart to fill with fluid, impeding the heart's ability to fill with blood and beat efficiently.
- **Osteal Tuberculosis:** Infection of any bone can occur, but one of the most common sites is the spine. Spinal infection can lead to compression fractures and deformity of the back.
- **Renal Tuberculosis:** This can cause asymptomatic pyuria (white blood cells in the urine) and can spread to the reproductive organs and affect reproduction. In men, inflammation of the epididymis may occur.
- **Adrenal Tuberculosis:** TB of the adrenal glands can lead to adrenal insufficiency. Adrenal insufficiency is the inability to increase steroid production in times of stress, causing weakness and collapse.
- **TB Meningitis:** M. tuberculosis can infect the meninges (the main membrane surrounding the brain and spinal cord). This can be devastating, leading to permanent impairment and death. TB can be difficult to discern from a brain tumor because it may present as a focal mass in the brain with focal neurological signs. Headache and sleepiness are typical symptoms. The patient may appear to have had a stroke.

According to Stanley and Swierzewski (2011), Tuberculosis also categorized as active TB and inactive TB:

- **Active Tuberculosis:** Active TB means the bacteria are active in the body. The immune system is unable to stop these bacteria from causing illness. People with active TB in their lungs can pass the bacteria on to anyone they come into close contact with. When a person with active TB coughs, sneezes or spits, people nearby may breathe in the tuberculosis bacteria and become infected.

- **Inactive Tuberculosis:** Inactive TB infection is also called latent TB. If a person has latent TB, it means their body has been able to successfully fight the bacteria and stop them from causing illness. People who have latent TB do not feel sick, do not have symptoms and cannot spread tuberculosis. In the People who have HIV, the inactive TB may become active TB if their immune system becomes weakened.

2.1.2. Causes for Tuberculosis

All cases of TB are passed from person to person via droplets. When someone with TB infection coughs, sneezes, or talks, tiny droplets of saliva or mucus are expelled into the air, which can be inhaled by another person. Once infectious particles reach the alveoli (small saclike structures in the air spaces in the lungs), another cell, called the macrophage, engulfs the TB bacteria. Then the bacteria are transmitted to the lymphatic system and bloodstream and spread to other organs occurs. The bacteria further multiply in organs that have high oxygen pressures, such as the upper lobes of the lungs, the kidneys, bone marrow, and meninges: the membrane-like coverings of the brain and spinal cord. When the bacteria cause clinically detectable disease, the person can have TB. People who have inhaled the TB bacteria, but in whom the disease is controlled are referred to as infected. Their immune system has walled off the organism in an inflammatory focus known as a granuloma. They have no symptoms, frequently have a positive skin test for TB, yet cannot transmit the disease to others.

Tuberculosis is a serious health problem in its own right but it is also the most likely cause of death for HIV positive people. Like HIV, tuberculosis has had an uneven impact around the world (WHO, 2010).

Risk factors of Tuberculosis are:

- Aging
- Alcoholism
- Crowded living conditions
- Diseases that weaken the immune system
- Health care workers
- HIV infection
- Homelessness
- Low socioeconomic status
- Malnutrition
- Migration from a country with a high number of cases
- Nursing Homes
- Unhealthy Immune System
- Use of drugs for Arthritis

2.1.3. Symptoms of Tuberculosis

Most people who are infected with M. tuberculosis harbour the bacterium without symptoms (known as latent tuberculosis), but some will develop active tuberculosis. In other cases, the bacteria die off. WHO (2010) identified that, a positive TB skin test and old scars on a chest x-ray may provide the only evidence that a person was ever infected with tuberculosis.

The primary stage of the tuberculosis may be symptom-free, or the individual may experience a flu-like illness. According to Asha et al (2011) the main signs and symptoms of Tuberculosis include:

- A cough lasting for more than 2-3 weeks
- Chest Pain
- Discolored or bloody sputum
- Night sweats
- Severe Headache
- Slight fever
- Weight loss

Latent Tuberculosis Symptoms

In most people who breathe in the tuberculosis bacteria and become infected, the body is able to fight the bacteria to stop them from growing. The bacteria become inactive, but they remain alive in the body and can become active later. This is called latent tuberculosis. People with latent tuberculosis:

- Have no symptoms of tuberculosis
- Don't feel sick
- Can't spread TB to others
- Usually have a positive TB skin test (PPD test) reaction.

Some people with latent tuberculosis can develop active tuberculosis if they do not receive treatment.

2.1.4. Current Practice of the organization for Tuberculosis Diagnosis

A complete medical diagnosis for TB includes a review of complete medical history, physical examination of patient, a tuberculin skin test, a chest X-ray, and microbiologic smears and cultures. Tuberculosis can be diagnosed by many ways:

- **X-ray:** Diagnosis of tuberculosis in the lungs may be made using an X-ray. This is the most common diagnostic test that leads to the suspicion of infection. The main problems with X-ray are poor film quality, low specificity, and difficulties with interpretation.

- **The Mantoux skin test**: also known as a tuberculin skin test (TST). This test helps identify people infected with M. tuberculosis but who have no symptoms. A doctor must read the test. This test can often indicate disease when there is none (false positive). Also, it can show no disease when in fact have TB (false negative).

- **QuantiFERON-TB Gold test:** This is a blood test that is an aid in the diagnosis of TB. This test can help detect active and latent tuberculosis. The body responds to the presence of the tuberculosis bacteria. By special techniques, the patient's blood is incubated with proteins from TB bacteria. If the bacteria is in the patient, the immune cells in the blood sample respond to these proteins with the production of a substance called interferon-gamma (IFN-gamma). This substance is detected by the test. If someone had a prior BCG

vaccination and a positive skin test due to this, the QuantiFERON-TB Gold test will not detect any IFN-gamma.

- **Sputum testing:** Sample of sputum is test in laboratory to diagnose the TB. If sputum is available, or can be induced, a lab test may give a positive result in up to 30% of people with active disease.

2.1.5. Treatment

Treatment takes that long because the disease organisms grow very slowly and, unfortunately, also die very slowly. (Mycobacterium tuberculosis is a very slow-growing organism and may take up to six weeks to grow in a culture media.)

- Doctors use multiple drugs to reduce the likelihood of resistant organisms emerging.
- Often the drugs will be changed or chosen based on the laboratory results.
- If doctors doubt that the patient is taking the medicine, Prescribing doses twice a week helps assure compliance.
- The most common cause of treatment failure is people's failure to comply with the medical regimen. This may lead to the emergence of drug-resistant organisms. The patient must take the medications as directed, even if he/she are feeling better.
- Another important aspect of tuberculosis treatment is public health. This is an area of community health for which mandated treatment can occur. In some cases, the local health department will supervise administration of the medication for the entire course of therapy.
- Active TB disease can almost always be cured with a combination of antibiotics. Avoid nutrient deficiencies and imbalances.
- If someone is believed to have been in contact with another person who has TB, preventive antibiotic treatment may have to be given.
- Streptomycin, a drug that is given by injection, may be used as well, particularly when the disease is extensive
- The patient should take his/her pills under the guidance of someone who can supervise the therapy. The approach is called DOTS (Directly Observed Treatment, Short Course).
- Surgery on the lungs may be indicated to help cure TB when medication has failed

- If the patient develops any side effects from medications such as itching, change in color of skin, tiredness, or excessive fatigue other treatment must be given.

2.1.6. Tuberculosis Preventions

Preventive measures include strict standards for ventilation, air filtration etc.

- A BCG vaccine, is available and has been of some benefit in preventing TB
- Covering of mouth by a mask is helpful in prevention from TB
- Exercise regularly to keep the immune system
- Get adequate amounts of sleep
- Get tested regularly. Experts advise getting a skin test annually
- Keep the immune system healthy.
- The WHO recommends that HIV positive people who have latent TB should be offered isoniazid preventive therapy as needed.

2.2. Directly Observed Therapy Program

2.2.1. An Overview

In a directly observed therapy program, patients with tuberculosis meet with a healthcare worker every day or several times a week. During these meetings, the healthcare worker will monitor the patient as he or she takes the tuberculosis medication. The healthcare worker will also monitor the patient for signs of side effects from the medication. The directly observed therapy program ensures that the patient will complete the course of treatment until he or she is cured of tuberculosis (Marisa et al, 2011).

The best way to remember to take the tuberculosis medication is to get involved with a directly observed therapy program. As part of this program, the patient will meet with a healthcare worker every day or several times a week.

2.2.2. Benefits of a Directly Observed Therapy Program

A directly observed therapy program helps in several ways. The healthcare worker can help the patient remember to take the medicine and complete the tuberculosis treatment. This means the patient will get well as soon as possible. With a directly observed therapy program, the patient may need to take medicine only 2 or 3 times each week instead of every day.

The healthcare worker will make sure that the medicine is working as it should. This person will also watch for side effects and answer questions rise from the patient about tuberculosis.

CHAPTER THREE

DATA MINING AND KNOWLEDGE DISCOVERY

The amount of data stored in databases increases exponentially with time. As a consequence, the manual analysis of this data is complex and prone to errors. When the amount of data to be analyzed exploded in the mid-1990s, knowledge discovery emerged as an important analytical tool. The process of extracting useful knowledge from volumes of data is known as knowledge discovery in databases (Fayyad, 1996). Knowledge discovery's major objective is to identify valid, novel, potentially useful, and understandable patterns of data. Knowledge discovery is supported by three technologies: massive data collection, powerful multiprocessor computers, and data mining (Turban, 2005).

There is some confusion about the terms Data Mining and KDD. Often these two terms are used interchangeably. Many authors agree in that KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in this process. According to Fayyad, Shapiro & Smyth (1996), KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data whereas Data mining is the application of specific algorithms for extracting patterns from data.

Data mining is one step at the core of the knowledge discovery process, dealing with the extraction of patterns and relationships from large amount of data. It is a promising interdisciplinary area of research shared by several fields such as database systems, machine learning, intelligent information systems, statistics, data warehousing, and knowledge acquisition in expert systems (Lin & Cercone, 1997). It currently relies heavily on known techniques from statistics, Artificial Intelligence, and machine learning, the three roots of data mining. These techniques are used together to study data and find previously hidden trends or patterns within it.

3.1. OVERVIEW OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database, and mining a mountain for a vein of valuable ore. Data mining can generate new business opportunities by providing automated prediction of trends and behaviors, and discovery of previously unknown patterns.

One of the strengths of data mining, as opposed to more traditional statistical methods, is that it is not necessarily to know exactly what we are looking for before we start. Data mining uses powerful analytic tools to quickly and thoroughly explore mountains of data and pull out valuable and usable information. The primary use of data mining is to find something new in the data to discover a new piece of information that no one knew previously. This is data-driven or bottom-up approach because we start with the data and then build theories based on discovered patterns or trends.

DM requires massive collection of data to generate valuable information (Han & Kamber 2006). The data can range from simple numerical figures and text documents, to more complex information such as spatial data, multimedia data and hyper text documents. Deshpande and Thakare (2010) indicated that the data retrieval is simply not enough to take advantage of data. 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, data bases, 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 for decision-making.

A typical data mining process includes data acquisition, data pre-processing, model building and model validation (Deshpande & Thakare, 2010).

3.1.1. Data acquisition

The first step in data mining is to select the types of data to be used. Although a target data set has been created for discovery in some applications, DM can be performed on a set of variables or data samples in a larger database called training set to create and model while holding back some of the data sets which are called test dataset for latter validation of the model.

3.1.2. Data-preprocessing

Once the target data is selected, the data is then pre-processed for cleaning and transforming to improve the effectiveness of discovery. During this step, researchers remove the noise or outlier if necessary and decide on strategies for dealing with missing data fields. Then data is transformed to reduce the number of variables by converting one type of data to another such as numeric ones into categorical or deriving new attributes.

3.1.3. Building model

The third step of data mining refers to a series of activities such as deciding on the type of data mining operations, selecting the data mining algorithms and mining the data. First, the type of the data mining operation such as, classification, regression, clustering, association rule discovery, segmentation and deviation detection must be chosen. Based on the operations chosen for the application, an appropriate data mining technique is then selected based on the nature of the knowledge to be mined. The next step is selecting a particular algorithm within the data mining technique chosen. Choosing a data mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

3.1.4. Interpretation and model evaluation

The fourth step of data mining process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understanding terms by users. In the interpretation of results, the researcher determines and resolves potential conflicts with previously known or decides redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal.

3.2. DATA MINING TASKS

The data mining tasks are of different types depending on the use of data mining result (Hand et al, 2001). Predictive modeling, descriptive modeling, exploratory data analysis, patterns and rules discovery, and retrieval by content are some of the data mining tasks.

3.2.1. Predictive modeling

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, regression, prediction etc. are some examples of predictive modeling. As Tan et al (2009) indicated 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. Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable.

3.2.1.1. Classification

In supervised learning, classification refers to the mapping of data items into one of the pre-defined classes. In the development of data mining tools that use statistical approaches, one of the critical tasks is to create a classification model, known as a classifier, which will predict the class of some entities or patterns based on the values of the input attributes. Choosing the right classifier is a critical step in the pattern recognition process. A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known techniques that are used in data mining include decision trees, logistic regression, neural networks, NaïveBayes and nearest neighbor approach(Rashmi, 2010).

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 (Two Craws Corporation, 1999). 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 (Two Crows Corporation, 1999). 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 (Berry & Linoff, 2000).

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.

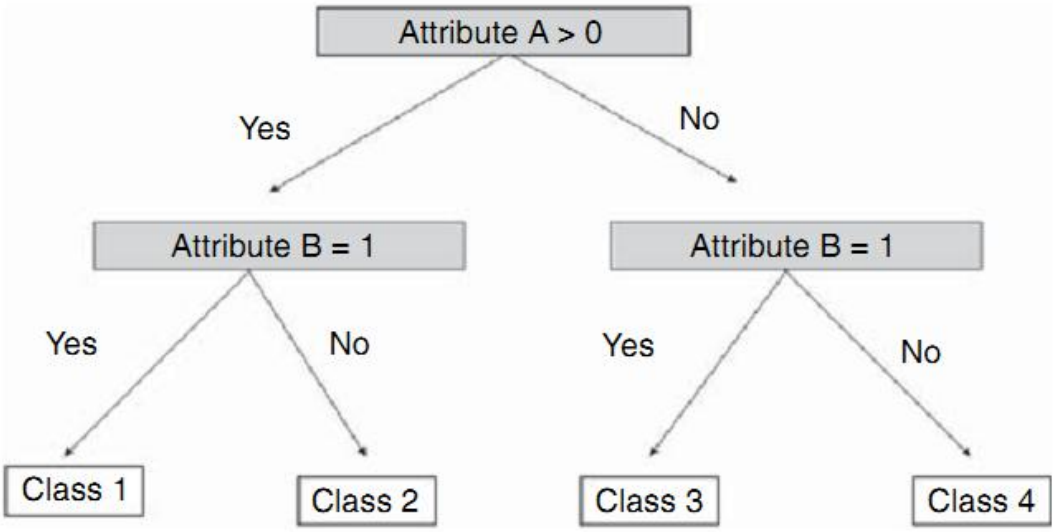

Figure3. 1 a simple decision tree

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.

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 articular feature of a class is unrelated to the presence (or absence) of any other feature (Bhargavi & Jyothi, 2009).

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.

$$Prob(B/A) = Prob(A/B)\, Prob(B)/Prob(A) \quad \ldots\ldots 3.1$$

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.

3.2.2. Descriptive modeling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined (Deshpande & Thakare, 2010). It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables. Clustering, association rule discovery, sequence discovery etc. are some of the examples. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone (Han & kamber, 2006). The association rule finds the association between different attributes. Association rule mining is a two-step process: finding all frequent item sets, generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in the data. This sequence can be used to understand the trend.

3.3.2.1. Clustering

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.

Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements.

Clustering techniques are very useful in data mining because of the speed, reliability, and consistency with which they can organize a large amount of data into distinct groupings. Despite the availability of a vast collection of clustering algorithms in the literature, they are based on two popular approaches: hierarchical clustering and partitioning clustering. The former, which is the most frequently used technique, organizes data in a nested sequence of groups that can be displayed in a tree-like structure, or dendrogram. Partitioning clustering constructs various partitions and then evaluates them by some criteria. K-means clustering and expectation maximization are the two methods of partitioning clustering.

As discussed earlier, clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (similar to loss in data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters.

Clustering techniques fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be variable to be predicted, thus no distinction is being made between independent and dependent variables (Rashmi, 2010).

According to Rashmi (2010), clustering techniques are used for combining observed instances into clusters (groups) which satisfy two main criteria:

1. Each group or cluster is homogenous; instances that belong to the same group are similar to each other.

2. Each group or cluster should be different from other clusters, that is instances that belong to one cluster should be different from the instances of other clusters.

Depending on the clustering technique, clusters can be expressed in different ways:

- Identified clusters may be exclusive: any instance belongs to only one cluster.
- They may be overlapping: an instance may belong to several clusters.
- They may be probabilistic: whereby an instance belongs to each cluster with a certain probability.
- Clusters might have hierarchical structure: having crude division of instances at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

The k-means Algorithm

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields (Rashmi, 2010).

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. The following shows how the K-means algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

Choosing the Number of Clusters

One of the main disadvantages to k-means is the fact of identifying the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if there are a group of people that were easily clustered based upon gender, calling the k-means algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and called the k-means algorithm with k=20, the results might be too generalized to be effective.

3.2.3. Pattern or Association rule discovery

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database (kotsiantisn & Kanellopoulos, 2006). The problem is usually decomposed into two sub problems: one is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimum confidence.

3.3. THE DATA MINING MODELS

There are different data mining process model standards. The six phase Cios et al model, the KDD(knowledge discovery in databases) process, CRISP-DM(CRoss Industry Standard Process for Data Mining) and SEMMA(Sample Explore Modify Model Asses), are some of the models used in different data mining projects.

3.3.1. Cios et al model

This model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps (Cios & Kurgan, 2005).

1. **Understanding of the problem domain:** in this step one works closely with domain experts to define the problem and determine the research goal, identify key people and learn about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the data mining goals, and include initial selection of data mining tools.

2. **Understanding of the data:** this step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, verification of the usefulness of the data is needed in respect to the data mining goal. Data needs to be checked for completeness, redundancy, missing value, etc.

3. **Preparation of the data:** this is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for data mining tools of step 4, is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. the cleaned data can be, further processed by feature selection and extraction algorithms(to reduce dimensionality), and by derivation of new attributes(say by discretization). The result would be new data records, meeting specific input requirements for the planned to be used data mining tools.

4. **Data mining:** this is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, pre-processing techniques, machine learning etc. This step involves the use of several data mining tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen data mining tools; the generated data model is verified by using testing procedures.

5. **Evaluation of the discovered knowledge:** this step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire data mining process may be revisited to identify which alternative actions could have been taken to improve the results.

6. **Using the discovered knowledge:** this step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

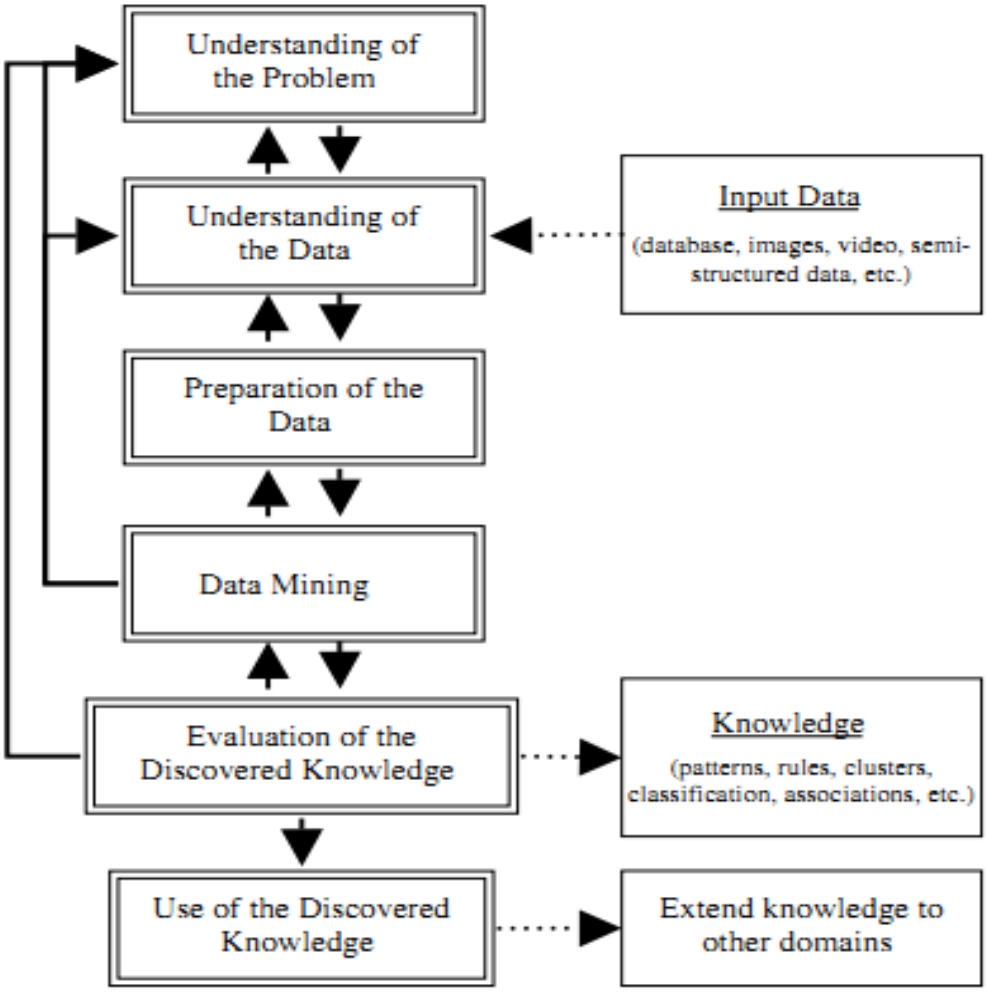

Figure3. 2 Cios et al. model

3.3.2. Knowledge discovery in database (KDD) process

KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database as Azevedo & Santos (2008) described. Generally there are five stages:

1. **Data 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. **Data Pre-processing** - this stage consists on the target data cleaning and pre-processing in order to obtain consistent data.

3. **Data 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 DM objective (usually, prediction).

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

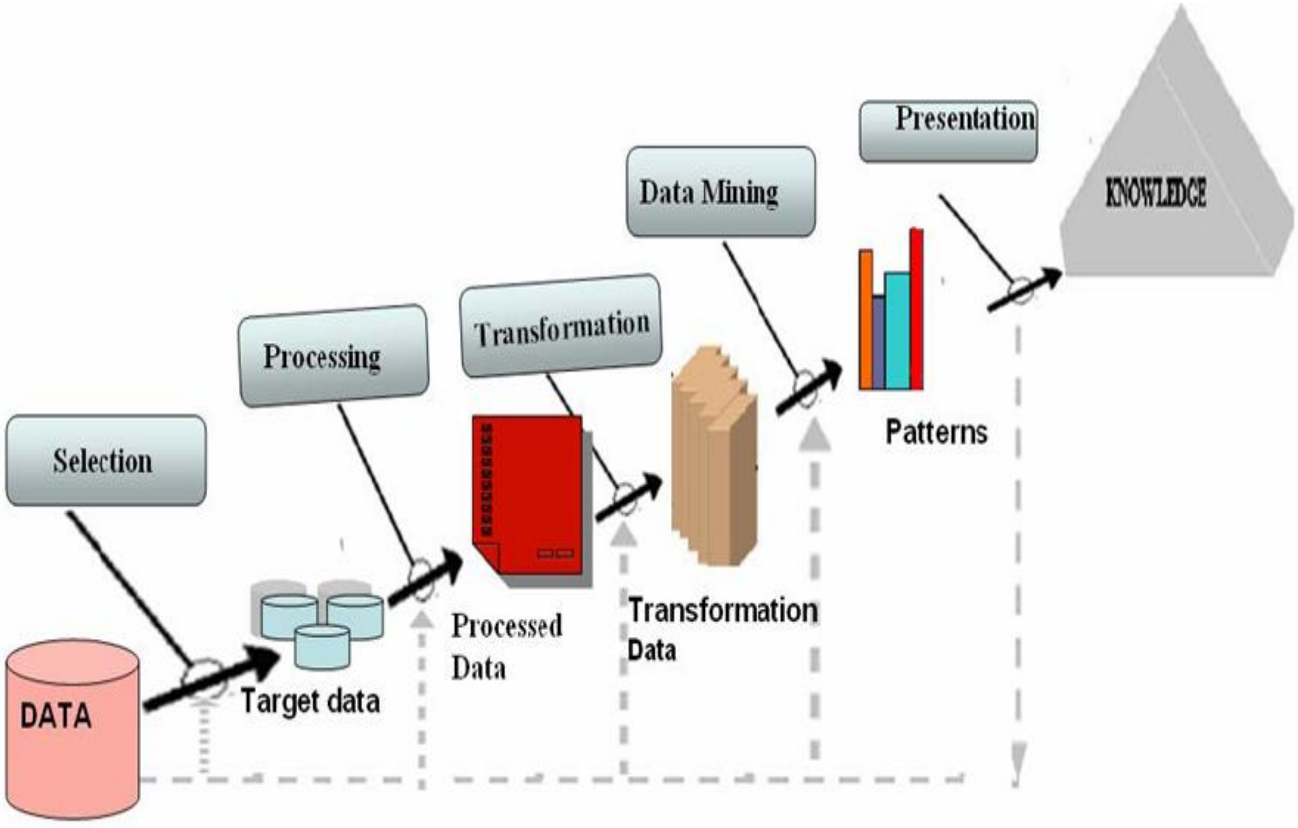

Figure3. 3 the KDD process

3.3.3. The CRISP-DM process

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 (Chapman et al, 2000; Azevedo & Santos, 2008).

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 are 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.

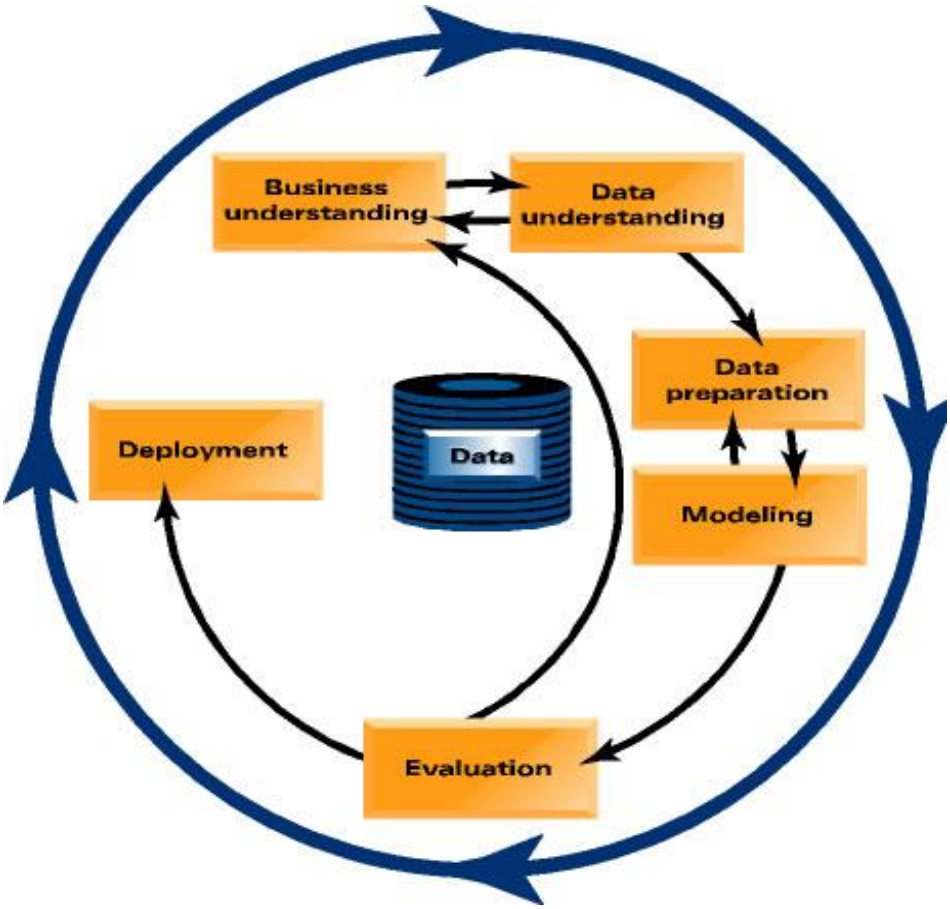

Figure3. 4 the CRISP-DM process

3.4. APPLICATION OF DATA MINING

The benefits of using Data Mining are numerous and the ever-increasing, newly developed applications of technologically enhanced information systems guarantee the establishment of Data Mining techniques as a very powerful and valuable tool for a wide variety of users in different fields. Today, though the primary application of Data Mining in the financial and marketing sectors, Data Mining usage has recently been expanded to other fields such as Medicine, Biology, Genetics and Biomedical Sciences in general (Eirini et al, 2005).

Accordingly as Seifert (2004) stated that, data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales.

3.4.1. Data mining in health care

As in other sector of economy, the health care industry has experienced many changes in information technology over the years. Advances in hardware, software, and networks have offered benefits, such as reduced cost and time of data processing and increased potential for profits, as well as new challenges particularly in the areas of increased competition. Health care industry can make better use of modern data mining technologies to develop more accurate and better performing models that are generated in less time than with previously known techniques. By generating better, extensively tested models, health firms can more accurately address issues such as moral hazard in underwriting and the adverse selection in customer satisfaction. The researcher has tried to find documents, which have been made so far on the application of data mining technology in support of various activities within the health care industry.

In this regard, data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective diagnosis and treatments and patients receive better and more affordable healthcare services (Chye & Tan, 2005). The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making.

The use of Data Mining algorithms in Medicine might well be one of the most interesting aspects of "computer" application in the field. Pattern search algorithms can search through vast databanks of patient information, providing new insights into conundrums that routinely trouble experts of the biomedical profession. A possible successful application of Data Mining may be in tuberculosis diagnosis.

3.4.2. Tuberculosis Diagnosis

Tuberculosis has been a major killer disease for several years. It is estimated that around 1.5 million people die each year from tuberculosis; and in 2009 figures indicate that approximately 8.8 million people developed the disease (WHO, 2010). The international standard for tuberculosis control is the World Health Organization's DOT strategy that aims to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic tuberculosis patients who present at health care facilities. As discussed earlier, the treatment is based on the strict supervision of medicines intake. The supervision is possible thanks to the availability of an information system that records the individual patient data. These data can be used at the facility level to monitor treatment outcomes, at the district level to identify local problems as they arise, at provincial or national level to ensure consistently high-quality tuberculosis control across geographical areas (WHO, 2011).

Identification of individuals latently infected and effective diagnosis are important parts of tuberculosis control. The DOT's strategy recommends identification of infectious tuberculosis cases by microscopic examination of sputum smears. However, this function requires a strong laboratory network and high-quality sputum smear microscopy. In children, the diagnosis of pulmonary tuberculosis is difficult because collection of sufficient sputum for smear microscopy and culture is difficult. The HIV epidemic has led to huge rises in incidence of tuberculosis in the worst affected countries, with disproportionate increases in smear negative pulmonary tuberculosis in children and adults (Getahun, 2007). Additionally, the use of chest radiography for diagnosis of pulmonary tuberculosis can be compromised by poor film quality, low specificity, and difficulties with interpretation (WHO, 2010).

Physicians are concerned about the poor specificity of current methods. In particular, there is a need to analyze tuberculosis diagnosis. In addition, the notification rate of relapsed cases is slightly increasing hence the researcher interested in finding patterns that can explain this trend.

CHAPTER FOUR

EXPERIMENT DESIGN

As discussed in Chapter3 Section 3.3 in data mining there are different kinds of standard methodologies used for modeling purpose, such as Cross-Industry Standard Process for Data Mining (CRISP-DM), Fayyad et al, Cios et al, and so on. Generally, in order to achieve the objectives of this study the researcher has used the KDD(Knowledge Discovery in Data bases) process model developed by Cios et al (2000). It was developed based on the CRISP-DM model by adopting it to academic research. This process has been chosen since it is a hybrid of both for academic and industrial purpose. According to Cios et al (2007), the Cios et al model has lots of advantages when compared it with other methodologies. 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.

4.1. Data Source

The main source of the data used to undertake this research was patients' real data taken from Menelik II hospital. Menelik II hospital is a public hospital which is found in Addis Ababa city. The main reason to select this hospital is, there are lots of data concerning the problem domain when compared with other hospitals. This can help for the learner to learn more and to give a better performance model. However, the data was in a hard copy format with 7069 records and 26 variables. A sample data is presented in Appendix2. As the medical officer described, this is because of the shortage of computer-skilled health workers and there are no database administrators available in the hospital. Hence, the researcher first encoded all the data in an

Excel format. After the data was encoded, the entire dataset is put in one file having many records. Each record corresponds to most relevant information of one patient. Initial queries by doctor as symptoms and some required test details of patients have been considered as main attributes. Next, Pre-processing techniques was applied to make it appropriate for mining purpose. This process was conducted through discussion with the domain experts as explained in section 4.3.

4.2. Data understanding

As discussed in chapter1. Section1.2, the problem domain was clearly defined. After understanding the problem to be addressed, the next step was analyzing and understanding the available data. The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of the available data (Cios et al, 2007). The original attributes and their description was presented in Appendix1.

4.3. Data preparation

The data preparation phase covers all activities to construct the final data set from the initial raw data. Tasks like data cleaning, record and attribute selection as well as transformation of data using discritization method were included.

4.3.1. Data cleaning

Data cleaning refers to the pre-processing of data in order to remove or reduce noise and the treatment of missing values. It is the process of ensuring that all values in a dataset are consistent and correctly recorded. To do so all the data which are available on the database was cleaned to the same format. As a result the data were prepared for data analysis.

The researcher makes use of the MS-Excel application for cleaning the data. In this subsection, different data cleaning tasks were carried out.

- **Missing value:** Medical data by nature has lots of missing values. Therefore, in the dataset collected for this research work there were missing values especially in the dependant variables “HIV_test_performed” and “HIV_test_result”. To handle those missing values, the researcher tries to fill it with correspondence between the dependant variables (i.e. if one variable have value the other dependant variable’s value is also

known). This is done with the help of the domain experts so that all the missing values were filled with the appropriate value.

- **Detecting noisy data and outliers:** a database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. In this research, the data objects do not have any outlier since the values of each column fits the database's general behavior.

4.3.2. Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. For example, smoothing techniques including binning, regression and attribute construction are the most used ones.

From the dataset the "AGE" and "WEIGHT" attributes were descritized (binned) to reduce the distinct values of the attributes so that it will suit the mining tool and to obtain meaningful patterns. Data descritization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results. A concept hierarchy for a given numerical attribute defines a descritization of the attribute (Han & Kamber, 2006). Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as the numerical values for the attribute "AGE") and high-level concepts (small, middle, and old).

Therefore, the researcher performed descritization on the attributes "AGE" and "WEIGHT" using a binning method. The attribute "AGE" is binned into three levels (small, middle and large) where as the attribute "WEIGHT" is binned to low, medium and high.

4.3.3. Numerousity Reduction (Attribute Selection)

As discussed in section4.1, the original table was contained 26 attributes. From this a total of 13 attributes were selected for the research based on their relevance and pre-processing activities of

the problem. There were 13 attributes which were excluded in the preliminary data observation like 'Medical record number', 'unit TB number', 'name of patient', 'address of patient', 'name of contact person', 'address of contact person', 'laboratory number', 'drug', 'month', 'days of month', 'CPT started date', 'Enrolled in HIV care', 'ART started', since they are no more important for the mining purpose. 'Medical record number', 'unit TB number' and 'laboratory number' have no important to know the research output. The other four attributes namely 'name of patient', 'address of patient', 'name of contact person' and 'address of contact person' have values such as name, phone-number, house-number, kebele and so on. Therefore, those values can make the model complicated and not understandable. Accordingly, the 'drug' attribute has only one value and is not necessary for mining activity. The researcher then removed the whole column since it is meaningless to use this attribute which have similar values throughout the records. 'CPT started date', 'Enrolled in HIV care', and 'ART started date' were removed since they do not hold values for all patients but patients with HIV- reactive only. Hence, the decision was made to remove those columns from the data set used for model building. Table 4.1 shows final attributes used for model building and their description.

SN.	ATTRIBUTES	DESCRIPTION	DATA TYPE
1	Sex	Patient's sex	Nominal
2	Age	Patient's age in years	Numeric
3	Weight	Weight of the patient	Numeric
4	Year	The day treatment started	Numeric
5	HIV performed	Patient tested for HIV	Nominal
6	HIV test results	HIV test result of the patient	Nominal
7	Headache	Whether the patient have headache	Nominal
8	Cough	Cough for about 2 weeks	Nominal
9	Chest pain	Some pain around the chest	Nominal
10	Bloody sputum	Sputum mixed with a blood	Nominal
11	Fever	An expected increase in temperature	Nominal
12	Weight loss	Whether the patient reduced in weight	Nominal
13	Night sweats	Whether the patient have sweats	Nominal

Table4. 1 Final selected attributes and their Description

Ranked Attributes: the final selected attributes by the Weka and their rank is shown as follows:

Value	Attribute-No	Attributes	Rank
0.12863861	7	Headache	1
0.06981118	11	Fever	2
0.05742172	8	Cough	3
0.01159513	4	Year	4
0.00806487	13	Night-Sweats	5
0.00098714	12	Weight-Loss	6
0.00039984	3	Weight	7
0.00026696	6	HIV-Test-Result	8
0.00024894	10	Bloody-Sputum	9
0.00014473	1	Sex	10
0.00012574	2	Age	11
0.00006743	9	Chest-Pain	12
0.00000486	5	HIV-Performed	13

Selected attributes: 7,11,8,4,13,12,3,6,10,1,2,9,5 : 13

4.4. Data mining

In this phase, appropriate techniques of data mining were applied to the data set available. Typically, there are several techniques for the same data mining problem. Data mining techniques: clustering and classification were applied to the dataset available.

After the data was cleaned and prepared, it was analyzed using a data mining tool. There are varieties of tools available for data mining such as the knowledge Studio, Weka, xlminer, and others. Among those tools, Weka is selected since the whole suite of Weka is written in java, so it can be run on any platform. In addition to this, the package has three different interfaces: a command line interface, an Explorer GUI interface which allows for preparation, transformation and modeling algorithms on a dataset, and an Experimenter GUI interface which allows to run different algorithms in batch and to compare the results.

Although the choice of data mining techniques for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are neural networks, decision trees and NaïveBayes. As it is indicated previously, for the purpose of this research work the researcher experimented the potential applicability of data mining technology in developing a model that predicts

the occurrence of TB in patients. To this end the researcher has employed and tested the applicability of decision tree and NaïveBayes techniques to the problem domain. Weka software was employed to build both decision tree and NaïveBayes models. This software partitions the dataset prepared for analysis into training and test facts where training facts are used to train and build the models and test facts are used to test the performance of the model. By default the software automatically sets aside 10% of the prepared dataset for testing purposes. Besides for validation purpose the researcher splits the dataset 75% for training and the remaining 25% for testing. 75% of the original data was selected for training purpose since the classifier learns more from large amount of data and increases its performance. The test data (25%) was selected from the original data using Simple Random Sampling technique.

In this research, numerous models were built by using the Weka3.7.5 software, and the proposed models was tested using test sets of data. Besides, the validity and performance of the model was tested to check its efficiency and effectiveness. Finally, the confusion matrix was used to evaluate the accuracy and performance of the model built with the decision tree algorithm.

CHAPTER FIVE

EXPERIMENTATION

This chapter presents steps and procedures followed during the experimentations. The main objective of this research is, discovering patterns for predicting patients' whether they have TB or not within the patients' dataset. Having this purpose in mind, the model-building phase in the DM process of this investigation is carried out following two-step process i.e. clustering then classification DM approach. This is because some of the available dataset didn't incorporate the target class for this study. The clustering sub phase has been then conducted using the K-means algorithm for segmenting the data into the target classes of TB-positive and TB-negative. After that classification was applied to predict the occurrence of TB for each patient. Both the clustering and the classification tasks were applied to the training dataset. As described earlier, these techniques were implemented using Weka 3.7.5 DM tool.

List of Abbreviated Attributes		**List of Abbreviated values**	
Abbreviated Attributes	**Description**	**Abbreviated values**	**Description**
S	Sex	F	Female
A	Age	M	Male
W	Weight	S	Small
YR	Year	ML	Middle
HP	HIV Performed	O	Old
HTR	HIV Test Result	H	High
HA	HeadAche	MM	Medium
C	Cough	L	Low
CP	Chest Pain	Y1	2009
BS	Bloody Sputum	Y2	2010
F	Fever	Y3	2011
WL	Weight Loss	Y4	2012

NS	Night Sweats	NR	Not Reactive
		R	Reactive
		Y	Yes
		N	No

Table5. 1 list of abbreviated attributes and their value

5.1. Clustering modeling

There have been four experimentations done for the clustering modeling. These experimentations were analyzed and compared to each other in terms of different measurements such as Number of iterations, with in cluster sum of squared errors and the judgment of the expert. The models were also compared with regard to the patterns /knowledge discovered.

As it was seen in chapter 4, section4.4 the method of validation is decided to be full training set splitting at 75% (5303) of the data set for training and allocating the rest 25%(1766) testing data set.

Parameters	Description	Usage
K	The number of clusters	To set the number of clusters to be created
Distance function	The method for calculating the distance	To find the similarity and dissimilarity between clusters
Seed value	Defines the number of data tubles the cluster must start with	To set the random number of seed to be used
Cluster distribution	The number of instances segmented to each cluster	To know the distribution percentage of the whole data

Table5. 2 parameters and their description for clustering modeling

Experimentation I

The first experimentation was done for K=2, with default seed value and default distance function. All of the final selected attributes and the instances for the training set (5303) records were used as an input for the experimentation. In order to cluster the records based on their values the model was trained by using the default values of the algorithm.

The following table exhibits the parameter values and the segmentation of the first experiment. As can be shown in the table the distribution of the dataset for each cluster is presented.

K	**Distance function**	**Seed value**	**Cluster distribution**	
2	EuclideanDistance	10	**C1**	3184(60%)
			C2	2119(40%)

Table5. 3 Default Parameter Values and Cluster Distribution for the First Experimentation

As we can see from Table 5.3, the first experimentation is conducted with the default values of the K-Means algorithm. The following table shows the result of the first experimentation.

Cluster No	**Distribution of instances in %**	**Attribute Names**												
		S	A	W	YR	HP	HTR	HA	C	CP	BS	F	WL	NS
1	3184(60%)	F	ML	MM	Y2	Y	R	Y	N	N	N	N	N	Y
2	2119(40%)	M	ML	MM	Y1	Y	NR	N	N	N	Y	Y	N	Y

Table5. 4 Clustering Result of the First Experiment

Table 5.4 shows the result of the first experiment with the average values of the attributes for each segment. The following table exhibits the description for each segment of values.

Cluster No	Description
1	Female, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result reactive, with headache, no cough, no chest-pain, no bloody-sputum, no fever, no weight-loss, no night-sweats
2	Male, middle age, medium weight, the year of 2010, HIV-performed, HIV-test-result not-reactive, no headache, no cough, no chest-pain, with bloody-sputum, with fever, no weight-loss, with night-sweats

Table5. 5 Cluster Summary of the First Experiment

As the expert told that, the main determinant factors for a patient to have a TB is, the value for the symptoms have to be 'yes' specially for the symptoms: Cough, Chest-pain, Bloody-sputum, Weight-loss and Night-sweats . And for a patient to be TB-negative, those attribute values have to be 'no' value. Moreover, in the current practice of the organization, the experts use those symptoms for identifying patients for TB-positive and TB-negative.

As can be seen in Table 5.5, all of the symptoms of TB in cluster1 have 'no' values. This means cluster1 is considered as the class for TB-negative. Accordingly, the second cluster cluster2 with the determinant attribute values have a value of 'yes' and the remaining as 'no' is considered to be as TB-positive.

Experimentation II

The second experiment is conducted with a default K value, a default distance function and seed= 50.

K	Distance function	Seed value	Cluster distribution	
2	EuclideanDistance	50	**C1**	3095 (58%)
			C2	2208 (42%)

Table5. 6 Parameters of the Second Experiment with Seed=50 and Other Default Values

Table 5.6 depicts the parameters and the cluster distribution of the second experiment. The following table shows the result of the model by the second experiment.

Cluster No	**Distribution of instances in %**	**Attribute Names**												
		S	A	W	YR	HP	HTR	HA	C	CP	BS	F	WL	NS
1	3095 (58%)	F	ML	MM	Y2	Y	NR	N	N	N	N	N	N	Y
2	2208 (42%)	F	ML	MM	Y1	Y	NR	Y	N	N	Y	Y	N	Y

Table5. 7 Clustering Result of the Second Experiment

Cluster No	**Description**
1	Female, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result not-reactive, no headache, no cough, no chest-pain, no bloody-sputum, no fever, no weight-loss, with night-sweats
2	female, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result not-reactive, with headache, no cough, no chest-pain, with bloody-sputum, with fever, no weight-loss, with night-sweats

Table5. 8 Cluster Summary of the Second Experiment

As can be seen from Table 5.8, cluster1 is considered as TB-negative since almost all of the values of the attributes contains ‘no’ except for the attribute value ‘Night-sweats’ which is ‘yes’. The second segment for cluster2 is about TB-positive because the most determinant attributes for a patient to have TB contains a value ‘yes’.

In this experiment one can understand that the most determinant factors for a TB to happen are not the detail information of the patients but the symptoms that the patient have. As can be seen in Table5.8, in both clusters, cluster1 and cluster2, the other attributes other than the symptoms

have the same value. Hence, this indicated that the result of the model is matched the experts judgment.

Experimentation III

The third experiment of the clustering modeling is with seed=500 and the other parameters have a default value. Table 5.9 shows the parameters used for this experiment and the segmentation of each clusters.

K	**Distance function**	**Seed value**	**Cluster distribution**	
2	EuclideanDistance	500	**C1**	2985(56%)
			C2	2318(44%)

Table5. 9 Parameter Values of the Third Experiment with Seed=500 and other Default Values

The following table presents the result of the third experiment and the segmentation of the two clusters.

Cluster No	**Distribution of instances in %**	**Attribute Names**												
		S	A	W	YR	HP	HTR	HA	C	CP	BS	F	WL	NS
1	2985(56%)	M	ML	MM	Y1	Y	NR	N	N	N	Y	Y	N	Y
2	2318(44%)	F	ML	MM	Y3	Y	R	Y	Y	Y	Y	Y	Y	N

Table5. 10 Clustering Result of the Third Experiment

Cluster No	Description
1	Male, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result not-reactive, no headache, no cough, no chest-pain, with bloody-sputum, with fever, no weight-loss, with night-sweats
2	female, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result reactive, with headache, with cough, with chest-pain, with bloody-sputum, with fever, with weight-loss, no night-sweats

Table5. 11 Cluster Summary of the Third Experiment

As can be seen from the above table i.e. Table 5.11, the first segment cluster1 contains 'yes' for some of the attributes and 'no' for the remaining. The second cluster cluster2 have 'yes' value for all of the attributes (symptoms) of TB. So, cluster2 is considered for the cluster of TB-positive and the first segment for TB-negative.

Compared to the previous experiments, in this experiment a good model were built. This is because the second cluster: cluster2 have 'yes' value for all of the symptoms of TB. In addition to this, the HIV-test-result has also 'reactive' value. As discussed in Chapter1, this leads to the HIV and TB co-existence.

Experimentation IV

The last experiment was taken by changing the distancefunction value to ManhattanDistance, and seed=1000. The cluster distribution for each segment shows in Table 5.12.

K	Distance function	Seed value	Cluster distribution	
2	ManhattanDistance	1000	**C1**	3029(57%)
			C2	2274(43%)

Table5. 12 Parameter Values of the Fourth Experiment with Distancefunction =Manhattandistance and Seed=1000

Table 5.13 shows the result of the final experimentation with ManhattanDistance and the cluster segmentation of the attributes for cluster1 and cluster2.

Cluster No	**Distribution of instances in %**	**Attribute Names**												
		S	A	W	YR	HP	HTR	HA	C	CP	BS	F	WL	NS
1	3029(57%)	F	ML	MM	Y1	Y	NR	Y	N	N	Y	Y	N	Y
2	2274(43%)	M	ML	MM	Y2	Y	NR	N	N	N	N	N	N	Y

Table5. 13 Clustering Result of the Fourth Experiment

Cluster No	**Description**
1	Female, middle age, medium weight, the year of 2009, HIV-performed, HIV-test-result not-reactive, with headache, no cough, with chest-pain, with bloody-sputum, with fever, no weight-loss, with night-sweats
2	Male, middle age, medium weight, the year of 2010, HIV-performed, HIV-test-result not-reactive, no headache, no cough, no chest-pain, no bloody-sputum, no fever, no weight-loss, with night-sweats

Table5. 14 Clustering Result of the Fourth Experiment

As can be seen from Table 5.14, the first segmentation is about TB-positive. This is because the attributes for the symptoms of TB have ‘yes’ value. The second segmentation is for TB-negative, since almost all of the values of the attributes of the symptoms have ‘no’ values.

This is the last experiment of the clustering model since it is not better than the previous experiments when compared the measurements and the judgment of the experts. Table 5.15 summarizes the entire summary of the criteria taken for choosing the best clustering model.

Choosing the best clustering model

The following table presents the values of Number of iteration, Within cluster sum of squared errors and Time taken to build each model of the four experimentations.

Experimentation No	Number of iteration	Within cluster sum of squared errors	Time taken
I	4	2643	0.25
II	3	2628	0.23
III	2	2596	0.06
IV	4	2536	0.1

Table5. 15 Comparison between Clustering Models

As can be seen from Table 5.15, the third experiment has a small number of iterations compared to the other experimentations. The last experimentation took smallest time of all the experimentations to build the model and have also minimum value of within cluster sum of squared errors.

The best cluster model has a value of minimum number of iteration, smallest value of within cluster sum of squared errors and minimum time to build the model. In this regard, the third experiment with minimum number of iteration, better value of within cluster sum of squared errors and time taken is selected as the final model for the classification model. In addition to this, the model built using the third experimentation was clustered the dataset by fulfilling the qualification given by the experts. This means, the discovered knowledge from the developed clustering model is essential for segmenting the patients' data into TB-positive and TB-negative. Hence, it is chosen by the experts as a best model for the next step (classification).

5.2. Classification modeling using J48 decision tree

Analysis of the decision tree models are made in terms of detailed accuracy of the classifier on the training dataset as tested on the tested data based on a confusion matrix of each model result. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Confusion matrix shows four important numerical quantities namely: True-Positive, False-Positive, False-Negative and True-Negative as shown below.

	Predicted class	
Actual class	Yes	No
Yes	TP: True-Positive	FN: False-Negative
No	FP: False-Positive	TN: True-Negative

Table5. 16 The Confusion Matrix

The True-positives (TP) and True-negatives(TN) are correct classifications. A False- positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A False-negative(FN) occurs when the outcome is incorrectly predicted as no when it is actually yes.

Once the clustering model is developed, the next step of this study was developing the predictive model using the classification techniques. As can be seen in the forgoing discussion, the resulted clustering model identified the segments of the patients' data in to high intra-similarity and low inter-similarity. Since the developed model does not classify a new instance in to a certain segment the development of the classification model was essential. The list of the parameters for J48 decision tree and their description is presented as follows as taken from Weka manual (2008).

Option	Description
binarySplits	Whether to use binary splits on nominal attributes when building the trees
confidenceFactor	The confidence factor used for pruning(smaller values incur more pruning)
debug	If set to true, classifier may output additional info to the console
minNumObj	The minimum number of instances per leaf
numFolds	Determines the amount of data used for reduced-error pruning, one fold is used for pruning and the rest for growing the tree
reducedErrorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning
saveInstanceData	Whether to save the training data for visualization
Seed	The seed used for randomizing the data when reduced-error pruning is used
subtreeRaising	Whether to consider the sub-tree raising operation when pruning
unpruned	Whether pruning is performed
useLaplace	Whether counts at leaves are smoothed based on Laplace

Table5. 17 Description of J48 Decision Tree Parameter Options in Weka

For starting the classification experimentation, J48 decision tree and Naïve Bayes were selected. There were four experiments to be done for decision tree classification and two experiments for NaïveBayes. These experimentations were going to be experimented and analyzed and compared to each other in terms of different performance matrix values, accuracies, number of leaves, the size of trees generated and execution time. The models were also compared with the discovered knowledge and the judgment of the experts.

Experimentation I: J48 decision tree with 10-fold cross-validation and default parameters

In this experiment, J48 decision tree with 10-fold cross-validation was applied. The result of this experiment is presented below.

Test mode: 10-fold cross-validation

Number of Leaves: 49

Size of the tree: 84

Time taken to build model: 0.21 seconds

=== 10-fold cross-validation ===

=== Summary ===

Correctly Classified Instances	4528	85.39%
Incorrectly Classified Instances	775	14.61%

The base for calculating correctly classified instances and incorrectly classified instances is the confusion matrix. The confusion matrix of the class which is a base for calculating accuracy measures and performance is presented below.

Actual	**Predicted**		**Total**
	Cluster1	Cluster2	
Cluster1	2122	366	2488
Cluster2	409	2406	2815
	2531	2772	5303

Table5. 18 Confusion Matrix of 10-fold cross-validation with the Default Parameters

The number of True-positives in this confusion matrix is 2122 records. Those records which were predicted as ‘cluster1’ class by the classifier and also happened in ‘cluster1’ actually are True-positives. The number of the records (2406) which were classified to the ‘cluster2’ class by the classifier and they are actually in ‘cluster2’ are True negative. The sum of the True positive

and True-negative is gives us correctly classified instances. The total number of records which were correctly classified to ‘cluster1’ and ‘cluster2’ classes of the TB-result of the patients’ was 4528(85.39%). 775(14.61%) of the records are miss classified.

As can be seen from Table 5.18, the accuracy of the model is moderate i.e. 85.39%. The number of leaves and the size of the tree are 49 and 84 respectively.

In this experiment, the researcher attempted to modify the default values of the parameters so as to minimize the number of the leaves and the size of the tree and to increase the accuracy of the model. With this objective in mind, the minNumobj(minimum number of objects in a leaf) parameter was tried with a value 5, 10, 15 and 20. But the result is not that much improved when compared it with the default parameter (i.e. 2). This is because if the value of minNumobj increases, the number of the leaves and size of the tree also decreases but the accuracy and the performance of the model decrease as well. Hence, the default value of minNumobj=2 was better than the other experiments with the changed values.

Experimentation II: J48 decision tree with percentage split

The second experiment of the J48 decision tree was conducted with percentage split 75% for training and 25% for testing. The first line reports the split point for training and testing dataset.

Test mode: split 75% train, remainder test

Number of Leaves: 49

Size of the tree: 84

Time taken to build model: 0.03 seconds

=== Evaluation on test split===

=== Summary ===

Correctly Classified Instances	1129	85.14%
Incorrectly Classified Instances	197	14.86%

Actual	Predicted		Total
	Cluster1	Cluster2	
Cluster1	558	89	647
Cluster2	108	571	679
	666	660	1326

Table5. 19 Confusion Matrix of Percentage Split with the Default Parameters

As can be seen in Table 5.19, the percentage split with 75% training set and remaining 25% for testing purpose was applied. In the above confusion matrix, the number of correctly classified instances is 1129(85.14%) out of 1326 records. This means that the number of records which are correctly classified to both the 'cluster1' and 'cluster2' classes of the TB-result class of the patients while they are actually in those classes. This hasn't improved the accuracy of the model as shown in the confusion matrix than the previous experiment. Compared to the performance, accuracy measures and other matters with the first experiment, it is poorer than the first experiment with an accuracy of 85.14%.

The Number of leaves and the size of the tree have a value of 49 and 84 respectively. This experiment has the same value of Number of leaves and the size of the tree with the first experiment. Therefore, J48 decision tree with 10-fold cross-validation and J48 decision tree with 75% percentage split can equally perform on the patients' data set.

Experimentation III: J48 decision tree with reduced attributes

Some of the attributes were excluded in this experiment. Reduced attributes which were not appeared in this experiment were those attributes which were on the last four ranks in Chapter4 section 4.3.3 Attribute Selection method. These attributes were 'Age', 'Sex', 'Chest-pain' and 'HIV-performed'. So this experiment was going on without these attributes together with J48 decision tree default parameters. The result of the third experiment of the classification modeling looks like as follows.

Test mode: 10-fold cross validation

Number of Leaves: 29

Size of the tree: 50

Time taken to build model: 0.03 seconds

=== Stratified cross validation ===

=== Summary ===

Correctly Classified Instances	4557	85.93 %
Incorrectly Classified Instances	746	14.06 %

Actual	**Predicted**		**Total**
	Cluster1	Cluster2	
Cluster1	2122	366	2488
Cluster2	380	2435	2815
	2502	2801	5303

Table5. 20 Confusion Matrix of 10-fold Cross-Validation with Reduced Attributes

In this experiment, J48 decision tree with reduced number of attributes and default parameters were applied. As can be seen in Table5.20, correctly classified instances are 4557(85.93%) out of the whole dataset 5303. This means the accuracy of this model is high compared to the previous experimentations.

The J48 decision tree with full training set with reduced attributes in this experiment has generated a tree with 29 Number of leaves and 50 Size of the tree. To compare the number of leaves and the size of the tree with the previous three experiments, here both number of the leaves and size of the tree decreases. From this we can understand that the experimentations with all of the final selected attributes have low performance than the experiment with reduced number of attributes. Accordingly, its ability in correctly classifying records in to both ‘cluster1’ and ‘cluster2’ classes are increased to 85.93%.

A trial also made by reducing the number of instances to 1050 and 2500, the accuracy of the model decreases to 82.42%. In addition to this the number of the leaves and the number of the tree also becomes increased much more and complex to understand. So that a conclusion is made if the number of the records increases the performance of the model also increase.

The model developed from this experiment was validated using the separated test set 25% (1766) and the performance of the model increases to 86.07%. This registers high performance and is better when compared it with the previous experimentations.

Experimentation IV: J48 decision tree 10-fold cross-validation with unpruned tree

The last experiment for the J48 decision tree is built with 10-fold cross-validation and unpruned tree with all attributes. The result of this experiment is presented below. The researcher compared the result of this model to the previous models in terms of all matters pertaining to the performance, Number of leaves and Size of the tree after the result of the model.

Test mode: 10-fold cross-validation

Number of Leaves: 350

Size of the tree: 593

Time taken to build model: 0.03 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	4422	84.39%
Incorrectly Classified Instances	881	16.61%

Actual	**Predicted**		**Total**
	Cluster1	Cluster2	
Cluster1	2084	404	2488
Cluster2	477	2338	2815
	2561	2742	5303

Table5. 21 Confusion Matrix of 10-fold Cross-Validation with Unpruned Tree

This experiment is taken place for in case some improvement might occur due to the change in the parameter unpruned to 'True'. The accuracy of the model reduces to 84.39% compared to the previous experimentations. In terms of Number of leaves and size of the tree, this experiment with unpruned tree increased the value than the previous experiments with 350 Number of leaves and 593 Size of the tree. This indicates the complexity of the tree increased as can be observed in the Number of leaves and Size of the tree. The tree structure is not understandable. That means it has many leaf nodes as well as it is very lengthy. Compared the performance, accuracy with the previous experiments, the previous experiments has much better than this experiment.

This is the last experiment for the J48 decision tree since there is no improvement in the performance of the model. This means the previous models were better than this experiment.

5.3 Classification modeling using Naïve Bayes model building

The second data mining technique employed for the classification modeling to build the classification model is the naïve Bayes. In order to build the model the clustered data set were used as input to the naïve Bayes classifier algorithm.

For the purpose of classification model with naïve Bayes algorithm, the following two experimentations were under taken. The experiments were conducted by changing the default parameters of the classifier. These experiments were presented and discussed in the next sub-sections. A comparison and analysis is made for each of them.

Experimentation I: Naïve Bayes with default parameters

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Time taken to build model: 0.02 seconds

=== Summary ===

Correctly Classified Instances	4413	83.21 %
Incorrectly Classified Instances	890	16.78 %

Actual	Predicted		Total
	Cluster1	Cluster2	
Cluster1	2073	415	2488
Cluster2	475	2340	2815
	2548	2755	5303

Table5. 22 Confusion Matrix of NaïveBayes 10-fold Cross-Validation with the Default Parameters

In this experiment Naïve Bayes classifier with default parameters were applied to the clustered data set. As can be seen in the confusion matrix of this model, correctly classified instances by the classifier are 4413(83.21%) of the total training data set 5303. This shows that the accuracy is reduced when compared it with the models created by the J48 decision tree.

To make some improvements another experiment were conducted by splitting the clustered data set using percentage split option of the classifier algorithm. The following experiment presents the result of the model and some descriptions.

Experimentation II: Naïve Bayes with percentage split

This experiment was going using Naïve Bayes with default percentage split (66%). The result of this experiment and the analysis is presented below.

Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	1484	82.30 %
Incorrectly Classified Instances	319	17.70 %

Actual	Predicted		Total
	Cluster1	Cluster2	
Cluster1	710	157	867
Cluster2	162	774	936
	872	931	1803

Table5. 23 Confusion Matrix of Naïve Bayes Percentage Split with the Default Parameters

The second experiment of the naïve Bayes classifier was conducted by splitting the dataset in to 66% training and 34% for testing the model. From 1803 number of instances, 1484(82.30%) were correctly classified and 319(17.70%) instances were misclassified by the classifier. The accuracy of the model built by this experiment is poorer than the previous models build by the J48 decision tree and Naïve Bayes. Hence, since there is not an improvement in the accuracy of the model this experiment was the last experiment for the classification modeling.

Comparison of J48 decision tree and Naïve Bayes models

A summary to the models built by the J48 decision tree and the Naïve Bayes is presented below.

Experiment No	Type of classifier	Number of leaves	Size of tree	Correctly classified instances	Incorrectly classified instances
I	J48 decision tree	49	84	85.39%	14.61%
II	J48 decision tree	49	84	85.14%	14.86%
III	J48 decision tree	29	50	85.93%	14.06%
IV	J48 decision tree	350	593	84.39%	16.61%
I	Naïve Bayes	-	-	83.21%	16.78%
II	Naïve Bayes	-	-	82.30%	17.70%

Table5. 24 Summary of the J48 Decision Tree and Naïve Bayes Models

As can be seen from Table5.24, the first experiment and the second experiment of the J48 decision tree have the same value for both the number of leaves and size of the tree. This means the model built by the two experiments was almost similar even if it differs in accuracy. The

accuracy of the J48 classifier to build the third experiment is high than the remaining classifiers. So, the J48 decision tree with reduced number of attributes is effective classification technique among remaining classification techniques. Accordingly, the accuracy of the third experiment with 85.93% is better than all the experiments. When compared all of the experiments with the available criteria the third experiment for the J48 decision tree with 29 Number of leaves and 50 size of tree were chosen. Here, the models which were built by these experiments were evaluated by the domain experts and the third experiment was chosen for TB prediction purpose.

5.4 Evaluation of the discovered knowledge

At this stage in the data mining task a model was built to have high quality from a data analysis perspective. Besides, it is important to thoroughly evaluate the model and review the steps executed to construct the model and to be certain that it achieves the business objectives. The developed model was evaluated using the test set data (1766) prepared for evaluation purpose. At the end of this phase, a decision on the use of the data mining results is reached. This is performed based on the domain expert's advice and the parameters set and the researcher's personal judgment.

It is good to see the meaning of the patterns generated by decision tree. As shown in Appendix3, the number of instances for each label is given at the leave as name of the majority class followed by number of instances for the majority class/ number of the minority class in brackets. It is possible to calculate the likelihood predictability of the majority class from the numbers of instances. The following are few of the patterns which were discovered between the attributes. Those patterns have also got an acceptance by the domain experts as consulted informally.

Rule1: If Headache=yes and Fever=yes and Cough=Yes Then TB-Result=Cluster2 (TB-positive)

Rule2: If Headache=no and Cough=yes and Fever=yes Then TB-Result=Cluster2 (TB-Positive)

Rule3: If Headache=yes and Fever=yes and Cough=No and Year=Y3 and weight=Low and Bloody-sputum=yes Then TB-Result=Cluster2 (TB-positive)

Rule4: If Headache=no and Cough=no and Fever=no and Year=Y3 and Bloody-sputum=yes and Weight-loss=yes and HIV-Test-Result=Reactive Then TB-Result=Cluster2 (TB-Positive)

Rule5: If Headache=no and Cough=no and Fever=no and Year=Y2 and Bloody-sputum=yes and Weight-loss=yes and HIV-Test-Result=Not-reactive Then TB-Result=Cluster1 (TB-negative)

Rule6: If Headache=yes and Fever=yes and Cough=no and Year=Y2 and Weight=Low and HIV-Test-Result=Not-reactive Then TB-Result=Cluster1 (TB-negative)

Rule7: If Headache=no and Cough=no and Fever=yes and Year=Y1 and Night-Sweats=no and Bloody-Sputum=no and Then TB-Result=Cluster1 (TB-negative)

Rule8: If Headache=yes and Fever=yes and Cough=yes and Year=Y3 and Weigh=Medium and Weight-Loss=yes and Then TB-Result=Cluster2 (TB-positive)

Rule9: If Headache=no and Fever=yes Cough=yes and Night-Sweats=no and Year=4 Then TB-Result=Cluster2 (TB-positive)

A set of rules are generated from the developed J48 decision tree classification model. All the selected attributes were used to build the decision tree. From these, the generated decision tree has shown that Headache is the most determinant variable, which is the top splitting variable of the model. One of the generated rule shows this fact that even the patient doesn't have cough it has a TB-positive. This indicated that the type of the TB can be extra-pulmonary TB. This decision tree has also shown that HIV-test-result variable can used in the process of decision-making. If the patient's HIV-test-result is reactive then most of the generated rules show that the patient can have TB (TB-positive). Generally, if the patient has Headache, Fever, Cough, Bloody-sputum and Night-sweats, the decision tree indicated that the patient is a victim of TB (TB-positive).

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

The effective use of information and technology is crucial for health care organizations to stay competitive in today's complex, evolving environment. The challenges faced when trying to make sense of large, diverse, and often complex data source are considerable. In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficacy of patient care. Data mining can be used to help predict future patient behavior and to improve diagnosis and treatment programs.

An overview of the KDD process and basic data-mining methods were discussed. Given the broad spectrum of data-mining methods and algorithms, the overview is inevitably limited in scope: There are many data-mining techniques, particularly specialized methods for particular types of data and domain. Although various algorithms and applications might appear quite different on the surface, it is not uncommon to find that they share many common components.

Data mining, extracting meaningful patterns and rules from large quantities of data, is clearly useful in any field where there are large quantities of data and something worth learning. In this respect, health care industry is a potential area for data mining. It is filled with lots of data.

In this regard one of the areas where huge amount of data found is in medical data where different tuberculosis patients data are recorded and stored for long period of time. Therefore, application of data mining tools is needed to convert such data to useful information and knowledge. Moreover, the availability of DOT'S records gives an opportunity to use Data Mining techniques such as demographic clustering, classification and association rule discover.

Tuberculosis is an important health concern as it is also associated with HIV. Retrospective studies of tuberculosis suggest that active tuberculosis accelerates the progression of HIV infection. In this paper, an efficient hybrid model for the prediction of tuberculosis were proposed. This paper presented a data mining technique for diagnosis of TB based on the

clustering and classification model building techniques. K-means clustering is combined with different classifiers to improve the accuracy in the prediction of TB. This approach not only helps doctors in diagnosis but also to consider various other features involved within each class in planning their treatments.

In this research, an attempt was made to assess the potential applicability of data mining technology in support TB diagnosis activity in Menelik II hospital. This experimental research, which employed the commonly used methodological approach in data mining researches, made use of two predictive modeling techniques, decision tree and Naive Bayes, to address the problem. Defining the target class was one of the Challenging tasks in this research. The initial data collected from the organization was not pre-defined. So before moving to the successive stages of the study, the data had to be segmented using clustering approach to two classes namely: TB-positive and TB-negative.

Thus, a model to classify patients data were built. Two basic tasks made in model building are decision tree and Naive Bayes. Various experiments were made iteratively by making adjustments on the modeling parameters in both tasks to come up with meaningful results. Accordingly, the better decision tree selected as a working model generates meaningful rules that would assign new patients records to the classes. Moreover, the classification accuracy of the decision tree was so convincing, that among the 5303 data, 85.93% of them were correctly classified. The misclassification of the remaining records is mainly attributed to the problem associated with the definition of the target variable.

Compared to the result of decision tree, the Naive Bayes performs slightly low. This performance difference by no means shows the weakness of the predictive capability of Naive Bayes. In addition, the comparison of the results of the decision tree and Naive Bayes models showed an interesting pattern.

To conclude, results from the study have shown that the problem in TB diagnosis, could be leveraged using data mining techniques.

6.2. Recommendations

The following recommendations are forwarded based on the findings of the experiment.

- Even though the research is done for academic achievements the research output would give the hospital's workers in identifying the main factors for the TB diagnosis purpose. This can help for solving problems rising during diagnosis and treatment.

- The developed data base can be used as a base line for the hospital specially the TB case team, to encode the detail information of the patients' for future use.

- The researcher implemented only two classes of data mining techniques i.e. clustering and classification. But, those data mining techniques which were not experimented by this study might reveal important patterns which were related to the effective TB-diagnosis. And this might increase the performance of the model.

- In the study a model for predicting the accuracy of TB patients were built. But if knowledge based system were added to it, it becomes a better advisory system for the health workers of the organization.

- Adding knowledge based system with the developed model helps for identifying the type of TB. Since there are about 12 distinct types of TB as discussed in this research.

- In this study the scope was limited to the Menelik II hospital, if a huge amount of data is considered from a number of hospitals the performance of the model might increase.

- Pattern identification in data mining is the most difficult task. It is recommended that if some visualization techniques like graphically or logically representation of the useful knowledge and translating them in to understanding terms by users.

- This research was undertaken for TB diagnosis; hence it is recommended that a data mining concept might apply for diagnosis of other types of diseases in our country as well.

REFERENCES

Abdi A., Gunnar B., and Fekadu A., 2009, 'Pastoralism and Delay in Diagnosis of TB in Ethiopia'.

Abdi A., Mette S., Fekadu A., and Gunnar A., 2010, 'Barriers to Tuberculosis Care: A Qualitative Study Among Somali Pastoralists in Ethiopia'.

Asha T., S. Natarajan, K.N.B. Murthy, 2011, 'Effective Classification Algorithms to Predict the Accuracy of Tuberculosis: A Machine Learning Approach', (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 7, pp. 89-94.

Asha T., S. Natarajan, K.N.B. Murthy, 2011, 'A Data Mining Approach to the Diagnosis of Tuberculosis by cascading Clustering and Classification'.

Aynalem A., 2008, 'The Medical Geography of Ethiopia'.

Azevado A. and Santos F., 2008, 'KDD, SEMMA AND CRISP-DM: A Parallel Overview ', IADIS European Conference Data Mining, Portugal, pp. 182-185.

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

Bhargavi P. and Jyothi S., 2009, 'Applying Naïve Bayes Data Mining Technique for Classification of Agricultural Land Soils', (IJCSNS) International Journal of Computer Science and Network Security, Vol. 9, No. 8.

Delphine S., Haileyesus G., Reuben G., Christian L. and Matteo Z., 2010, 'Priority Research Questions for TB/HIV in HIV-Prevalent and Resource-Limitted Setings', (WHO) World Health Organization.

Demissie M., 2002, 'Challenges of Tuberculosis Control in Ethiopia'.

Deshpande S. and Thakare V., 2010, 'Data Mining System and Applications: A Review', (IJDPS) International Journal of Distributed and Parallel Systems, Vol. 1, No. 1, Pp. 32-44.

Elamy A., Mandal M., Far B., Basu A., Cheng I. and Long, R., 2010, 'An Intelligent Cad System For Automated Detection Of Pulmonary Tuberculosis On Chest Radiograph and Ct Thorax: A Road Map'.

Eirini P., Kotsioni I., Linos A., 2005, 'Data Mining: A New Technique in Medical Research'.

Fayyad U., Grgory P. and Padhraic S., 1997, 'From Data Mining to Knowledge Discovery in Data Bases', American Association for Artificial Intelligence, pp. 37-54.

George D. and Andriana P., 2000, 'Machine Learning in Medical Applications'.

Han J. and Kamber M., 2006, 'Data Mining: Concepts and Techniques'.

Hian C. and Tan G., 2005, 'Data Mining Applications in Health Care', Journal of Health Care Information Management, Vol. 19, No. 2, pp. 64-72.

Khabaza T., 2010, ' Nine Laws of Data Mining'.

Madan L., 2006, 'Data Mining: A Competitive Tool in the Banking and Retail industries', The charter Accountant, pp. 588-594.

Marisa A., Sonia U. and Pablo A., 2011, 'Mining Tuberculosis Data'.

Marry K., and Obe S., 2004, 'Application of Data Mining Techniques to Health Care Data', (SHEA) The Society for Health Care Epidemiology of America, Vol. 25, No. 8, pp. 690-695.

Melli G., Osmar R. and Kitts B., 2007, 'Introduction to the Special Issue on Successful Real World Data Mining Applications', (SIGKDD) Explorations, Vol. 8, No. 1, pp. 1-2.

Michael J. and David C., 2007, 'Data Mining and Clinical Decision Support Systems'.

Ministry Of Health, 'Health and Health Related Indicators. Planning and Programming Department', 2007, Moh.

Mirjana b. and dijana c., 2008, 'Data mining usage in health care management: literature survey and decision tree application', vol. 5, no. 1, pp. 57-64.

Padmapriadarsini C., Narendran G. and Soumya S., 2011, 'Diagnosis and Treatment of Tuberculosis in HIV Co-Infected Patients', Indian Journal of Medical Research, Vol. 134, No. 6, pp. 850-865.

Rashmi, 2010, 'Clustering in Data Mining'.

Ravichandra r., 2003, 'data mining and clustering techniques'.

Richard G., 2009, 'Reviewing Ethiopia's Health System Development', JMAJ, International Medical Community, Vol. 52, No. 4, Pp. 279-286.

Sebban M., Mokrousov I., Rastogi N., Sola C., 2001, 'A Data Mining Approach To Spacer Oligonucleotid Typing Of Mycobacterium Tuberculosis', Bioinformatics, vol. 18, pp. 235-242.

Seifert W., 2004, 'Data Mining: An Overview', Analysis In Information Science And Technology Policy Resources Science And Industry Division.

Shah S., Demissie M., Lambert L., Ahmed J., Leulseged S., Kebede T., Melaku Z., Mengistu Y., Lemma E., Charles D., Wuhib T. and Lisa J., 2009, 'Intensified Tuberculosis Case Finding Among HIV-Infected Persons From A Voluntary Counseling and Testing Center In Addis Ababa, Ethiopia', Vol. 50, No. 5, pp. 537-545.

Simon J., Neel G., Wafaa E., and Gerald F., 2005, 'Tuberculosis and HIV: Operational Challenges Facing Collaboration And Integration'.

Srinivas K., Kavihta B. And Govrd A., 2010, 'Application of Data Mining Techniques In Health Care and Prediction of Heart Attacks', (IJCSE) International Journal on Computer Science and Engineering, Vol. 2, No. 2, pp. 250-255.

Tariku A., 2011, 'Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company'.

Teklu U., 2010, 'Application of Data Mining Techniques on Antiretroviral Therapy (ART): The Case of Adama And Asella Hospitals'.

Tollman S., Doherty J. and Mulligan J., 2007, 'General Primary Health Care in Disease Control Priorities in Developing Countries'.

Two Crows Corporation, 1999, 'Introduction to Data Mining and Knowledge Discovery', Two Crows Corporation.

Vincent H., 2007, 'Developing a Consumer Health Informatics Decision Support System Using Formal Concept Analysis'.

APPENDICES

Appendix1. The original attributes and their description

SN.	ATTRIBUTES	DESCRIPION
1	Medical record number	Unique individual identifier used on medical information folder
2	Unit TB number	TB unit identification number
3	Name of patient	The patient's name
4	Address of the patient	Address of patient
5	Sex	Patient's sex
6	Age	Patient's age in years
7	Name of contact person	Name of the contact person
8	Address of contact person	Address of the contact person
9	Lab.No.	Laboratory number for the sputum test
10	Weight	Weight of the patient
11	Drug	The drug therapy used
12	Year	The day treatment started
13	Month	The month treatment started
14	Days of month	Each day the patient receives DOT's treatment
15	HIV performed	Client tested for HIV
16	HIV test results	HIV test result of the patient
17	CPT started date	The date CPT started
18	Enrolled in HIV care	The date the patient enrolled in HIV care
19	ART started	The date the patient started ART
20	Headache	Whether the patient have headache
21	Cough	Cough for about 2 weeks
22	Chest pain	Some pain around the chest
23	Bloody sputum	Sputum mixed with a blood
24	Fever	An expected increase in temperature
25	Weight loss	Whether the patient Reduced in weight
26	Night sweats	Whether the patient have sweats

Appendix2. Sample values of the final selected attributes

Sex	Age	Weight	Year	HIV-Performed	HIV-Test-Result	Headache	Cough	Chest-Pain	Bloody-Sputum	Fever	Weight-Loss	Night-Sweats
male	SMALL	MEDIUM	2009	yes	nr	no	no	no	yes	yes	yes	no
female	OLD	HIGH	2012	yes	nr	yes	no	no	yes	yes	no	yes
male	MIDDLE	MEDIUM	2009	no	refused	yes	no	no	yes	no	no	no
female	MIDDLE	MEDIUM	2009	yes	nr	yes	yes	no	no	yes	yes	yes
female	MIDDLE	HIGH	2009	yes	r	yes	yes	no	no	yes	no	yes
female	MIDDLE	MEDIUM	2012	yes	r	yes	no	no	no	yes	no	no
female	SMALL	MEDIUM	2009	yes	nr	yes	no	no	no	yes	no	no
female	MIDDLE	LOW	2009	yes	r	yes	yes	yes	yes	yes	yes	yes
female	MIDDLE	LOW	2009	no	refused	no	no	no	no	yes	yes	no
male	MIDDLE	MEDIUM	2010	yes	r	no	no	no	no	yes	no	no
male	MIDDLE	MEDIUM	2009	yes	nr	no	no	no	no	yes	yes	no
male	MIDDLE	HIGH	2009	yes	nr	no	yes	no	yes	yes	no	yes

Appendix3. A sample decision tree generated from the J48 decision tree

A. Experimentation I: J48 decision tree with 10-fold cross-validation

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: patient_data_clustered
Instances: 5303
Attributes: 14
SEX
AGE
WEIGHT
YEAR
HIV_PERFORMED
HIV_TEST_RESULT
HEADACHE
COUGH
CHEST_PAIN
BLOODY_SPUTUM
FEVER
WEIGHT_LOSS
NIGHT_SWEATS
class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

HEADACHE = yes
| FEVER = yes
| | COUGH = yes
| | | NIGHT_SWEATS = yes: cluster1 (377.0/28.0)
| | | NIGHT_SWEATS = no: cluster0 (2.0)
| | COUGH = no
| | | YEAR = Y1: cluster0 (339.0/89.0)
| | | YEAR = Y2
| | | | CHEST_PAIN = yes
| | | | | WEIGHT = LOW: cluster0 (3.0)
| | | | | WEIGHT = MEDIUM: cluster1 (8.0/1.0)
| | | | | WEIGHT = HIGH: cluster0 (3.0/1.0)
| | | | CHEST_PAIN = no: cluster1 (99.0/6.0)
| | | YEAR = Y3
| | | | WEIGHT = LOW
| | | | | HIV_TEST_RESULT = r: cluster1 (12.0/3.0)
| | | | | HIV_TEST_RESULT = nr
| | | | | | SEX = male: cluster0 (6.0/1.0)
| | | | | | SEX = female
| | | | | | | BLOODY_SPUTUM = yes: cluster1 (3.0)
| | | | | | | BLOODY_SPUTUM = no
| | | | | | | | NIGHT_SWEATS = yes: cluster0 (4.0)

B. Experimentation II: J48 decision tree with percentage split

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     patient_data_clustered
Instances:    5303
Attributes:   14
              SEX
              AGE
              WEIGHT
              YEAR
              HIV_PERFORMED
              HIV_TEST_RESULT
              HEADACHE
              COUGH
              CHEST_PAIN
              BLOODY_SPUTUM
              FEVER
              WEIGHT_LOSS
              NIGHT_SWEATS
              class
Test mode:    split 75.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

HEADACHE = yes
|  FEVER = yes
|  |  COUGH = yes
|  |  |  NIGHT_SWEATS = yes: cluster1 (377.0/28.0)
|  |  |  NIGHT_SWEATS = no: cluster0 (2.0)
|  |  COUGH = no
|  |  |  YEAR = Y1: cluster0 (339.0/89.0)
|  |  |  YEAR = Y2
|  |  |  |  CHEST_PAIN = yes
|  |  |  |  |  WEIGHT = LOW: cluster0 (3.0)
|  |  |  |  |  WEIGHT = MEDIUM: cluster1 (8.0/1.0)
|  |  |  |  |  WEIGHT = HIGH: cluster0 (3.0/1.0)
|  |  |  |  CHEST_PAIN = no: cluster1 (99.0/6.0)
|  |  |  YEAR = Y3
|  |  |  |  WEIGHT = LOW
|  |  |  |  |  HIV_TEST_RESULT = r: cluster1 (12.0/3.0)
|  |  |  |  |  HIV_TEST_RESULT = nr
|  |  |  |  |  |  SEX = male: cluster0 (6.0/1.0)
|  |  |  |  |  |  SEX = female
|  |  |  |  |  |  |  BLOODY_SPUTUM = yes: cluster1 (3.0)
|  |  |  |  |  |  |  BLOODY_SPUTUM = no
|  |  |  |  |  |  |  |  NIGHT_SWEATS = yes: cluster0 (4.0)
|  |  |  |  |  |  |  |  NIGHT_SWEATS = no: cluster1 (9.0/2.0)
|  |  |  |  |  HIV_TEST_RESULT = refused: cluster0 (1.0)
```

C. Experimentation III: J48 decision tree with reduced attributes

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     patient_data_clustered
Instances:    5303
Attributes:   10
              WEIGHT
              YEAR
              HIV_TEST_RESULT
              HEADACHE
              COUGH
              BLOODY_SPUTUM
              FEVER
              WEIGHT_LOSS
              NIGHT_SWEATS
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

HEADACHE = yes
|   FEVER = yes
|   |   COUGH = yes: cluster1 (379.0/30.0)
|   |   COUGH = no
|   |   |   YEAR = Y1: cluster0 (339.0/89.0)
|   |   |   YEAR = Y2: cluster1 (113.0/12.0)
|   |   |   YEAR = Y3
|   |   |   |   WEIGHT = LOW
|   |   |   |   |   BLOODY_SPUTUM = yes: cluster1 (10.0/2.0)
|   |   |   |   |   BLOODY_SPUTUM = no
|   |   |   |   |   |   NIGHT_SWEATS = yes: cluster0 (9.0/1.0)
|   |   |   |   |   |   NIGHT_SWEATS = no: cluster1 (16.0/5.0)
|   |   |   |   WEIGHT = MEDIUM: cluster1 (141.0/54.0)
|   |   |   |   WEIGHT = HIGH: cluster0 (65.0/23.0)
|   |   |   YEAR = Y4: cluster0 (62.0)
|   FEVER = no: cluster0 (1423.0/156.0)
HEADACHE = no
|   COUGH = yes
|   |   FEVER = yes
|   |   |   NIGHT_SWEATS = yes: cluster1 (371.0/30.0)
|   |   |   NIGHT_SWEATS = no: cluster0 (6.0/1.0)
|   |   FEVER = no: cluster1 (886.0/1.0)
|   COUGH = no
|   |   FEVER = yes
|   |   |   NIGHT_SWEATS = yes: cluster1 (424.0/61.0)
|   |   |   NIGHT_SWEATS = no
|   |   |   |   BLOODY_SPUTUM = yes: cluster0 (9.0)
|   |   |   |   BLOODY_SPUTUM = no: cluster1 (200.0/55.0)
```

D. Experimentation IV: J48 decision tree with unpruned tree

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -U -M 2
Relation:     patient_data_clustered
Instances:    5303
Attributes:   14
              SEX
              AGE
              WEIGHT
              YEAR
              HIV_PERFORMED
              HIV_TEST_RESULT
              HEADACHE
              COUGH
              CHEST_PAIN
              BLOODY_SPUTUM
              FEVER
              WEIGHT_LOSS
              NIGHT_SWEATS
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

HEADACHE = yes
|  FEVER = yes
|  |  COUGH = yes
|  |  |  NIGHT_SWEATS = yes
|  |  |  |  YEAR = Y1
|  |  |  |  |  WEIGHT_LOSS = yes
|  |  |  |  |  |  HIV_TEST_RESULT = r: cluster1 (5.0/1.0)
|  |  |  |  |  |  HIV_TEST_RESULT = nr: cluster0 (5.0/2.0)
|  |  |  |  |  |  HIV_TEST_RESULT = refused: cluster1 (1.0)
|  |  |  |  |  WEIGHT_LOSS = no
|  |  |  |  |  |  SEX = male
|  |  |  |  |  |  |  WEIGHT = LOW: cluster1 (3.0/1.0)
|  |  |  |  |  |  |  WEIGHT = MEDIUM
|  |  |  |  |  |  |  |  HIV_PERFORMED = yes
|  |  |  |  |  |  |  |  |  HIV_TEST_RESULT = r: cluster1 (6.0/1.0)
|  |  |  |  |  |  |  |  |  HIV_TEST_RESULT = nr
|  |  |  |  |  |  |  |  |  |  CHEST_PAIN = yes: cluster1 (3.0)
|  |  |  |  |  |  |  |  |  |  CHEST_PAIN = no: cluster0 (5.0/2.0)
|  |  |  |  |  |  |  |  |  HIV_TEST_RESULT = refused: cluster1 (0.0)
|  |  |  |  |  |  |  |  HIV_PERFORMED = no: cluster1 (4.0)
|  |  |  |  |  |  |  WEIGHT = HIGH: cluster1 (8.0)
|  |  |  |  |  |  SEX = female: cluster1 (55.0/2.0)
|  |  |  |  YEAR = Y2
|  |  |  |  |  WEIGHT_LOSS = yes: cluster1 (5.0)
```

DECLARATION

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

_____

Date

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

_____

Date