**Table of Contents**

CHAPTER ONE: [INTRODUCTION 1](#_bookmark5)

* 1. [Background 1](#_bookmark6)
  2. [Statement of the Problem 3](#_bookmark7)
  3. [Objective 5](#_bookmark8)
     1. [General Objective 5](#_bookmark9)
     2. [Specific Objectives 5](#_bookmark10)
  4. [Research Methodology 5](#_bookmark11)
     1. [Research Design 5](#_bookmark12)

[Tools 9](#_bookmark13)

* 1. [Scope and Limitation of the Study 10](#_bookmark14)
  2. [Significance of the Study 10](#_bookmark15)
  3. [Dissemination of Results 11](#_bookmark16)

CHAPTER TWO[: LITERATURE REVIEW 13](#_bookmark18)

* 1. [Overview of Data Mining 13](#_bookmark19)
  2. [Developmental History of Data Mining 14](#_bookmark20)
  3. [Knowledge Discovery in a Database and Data Mining 15](#_bookmark21)
  4. [Data Mining Process Models 17](#_bookmark22)
     1. [The KDD Process Model 18](#_bookmark23)
  5. [Data Mining Tasks 22](#_bookmark24)
     1. [Classification 22](#_bookmark25)

[Clustering .........25](#_bookmark26)

* + 1. [Association Rule Discovery 27](#_bookmark27)
  1. [Application of Data Mining in the Health Sector 28](#_bookmark28)
  2. [Related Works 37](#_bookmark35)

CHAPTER THREE[: ALGORITHMS USED FOR MODEL BUILDING AND PERFORMANCE MEASURES 39](#_bookmark36)

* 1. [Decision Trees 39](#_bookmark37)
     1. [J48 Classifier Algorithm 40](#_bookmark38)
  2. [Neural Networks 42](#_bookmark39)
     1. [Multilayer Perception 44](#_bookmark40)
  3. [Bayesian Classifier 45](#_bookmark41)
     1. [Naïve Bayes 45](#_bookmark42)
  4. [Performance Measures 47](#_bookmark43)
     1. [10-Fold Cross Validation 48](#_bookmark44)
     2. [Confusion Matrix 49](#_bookmark45)
     3. [Area under the ROC Curve 51](#_bookmark46)

CHAPTER FOUR: [BUSINESS AND DATA UNDERSTANDING 54](#_bookmark47)

* 1. [Business Process Description 54](#_bookmark48)
  2. [Heart Diseases Diagnosis 55](#_bookmark49)
     1. [Echocardiography 56](#_bookmark50)
  3. [Medical Problem Definition 58](#_bookmark51)
     1. [Defining Medical Goals 58](#_bookmark52)
     2. [Defining Data Mining Goals 59](#_bookmark53)
  4. [Selecting and Creating the Target Dataset 59](#_bookmark54)
     1. [Data Understanding 59](#_bookmark55)
     2. [Creating the Database 60](#_bookmark56)
     3. [Description of the Dataset 60](#_bookmark57)

CHAPTER FIVE: [EXPERIMENTATION 72](#_bookmark69)

* 1. [Experimental Setup 72](#_bookmark70)
  2. [Model Building Using J48 Decision Tree 72](#_bookmark71)
  3. [Model Building Using Naïve Bayes Classifier 76](#_bookmark72)
  4. [Model Building Using Neural Network 78](#_bookmark73)
  5. [Discussion 80](#_bookmark74)
     1. [Effect of Decision Tree Pruning 80](#_bookmark75)
     2. [Effect of Attribute Selection 81](#_bookmark76)
     3. [Model Comparison 82](#_bookmark77)
     4. [Specific Rule Extraction 85](#_bookmark78)
     5. [Performance of the Selected Model 87](#_bookmark79)

CHAPTER SIX: [CONCLUSION AND RECOMMENDATIONS 89](#_bookmark80)

* 1. [Conclusion 89](#_bookmark81)
  2. [Recommendations 90](#_bookmark82)

[REFERENCES 92](#_bookmark83)

# LIST OF FIGURES

#### Title Page

Figure 2.1: The KDD Process for Knowledge Discovery 18

Figure 2.2: Between-cluster variation and Within-cluster variation 26

Figure 3.1: A simple Decision Tree 39

Figure 3.2: A Neural Network Architecture 43

Figure 3.3: A simple Confusion Matrix 49

Figure 3.4: A Sample ROC curves 52

Figure 4.1: Images from echocardiography examination 57

Figure 4.2: Distribution of heart disease among Gender 61

Figure 4.3: Distribution of Heart Disease among different age groups 62

Figure 5.1: Boxplot for numeric attributes in the dataset to identify outliers 67

Figure 6.1: Effect of Attribute Selection on Classification Accuracy 81

Figure 6.2: Performance Comparisons of the Models 83

# LIST OF TABLES

**Title Page**

Table 3.1: Performance measure of ROC Area 53

Table 4.1: Attributes and their description 61

Table 4.2: Numerical Attributes Data Summary 63

Table 5.1: Attributes with missing values 66

Table 5.2: After Discretizing the Age Attribute 69

Table 5.3: Encoding mechanisms Used for Dimensionality Reduction 70

Table 6.1: Confusion Matrixes for Experiment 1 73

Table 6.2: Detailed Performance Measures for Experiment 1 73

Table 6.3: Confusion Matrixes for Experiment 2 75

Table 6.4: Detailed Performance Measures for Experiment 2 75

Table 6.5: Confusion Matrixes for Experiment 3 77

Table 6.6: Detailed Performance Measures for Experiment 3 77

Table 6.7: Confusion Matrixes for Experiment 4 79

Table 6.8: Detailed Performance Measures for Experiment 4 79

Table 6.9: Performance Summary of the Models 82

Table 6.10: Selected Misclassified Cases 88

# ABSTRACT

**Background:** Most countries face high and increasing rates of heart disease or Cardiovascular Disease. Even though, modern medicine is generating huge amount of data every day, little has been done to use this available data to solve the challenges that face a successful interpretation of echocardiography examination results.

**Objective:** To design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography.

**Methodology:** Knowledge Discovery in Database (KDD) methodology consisting of nine iterative and interactive steps was adopted to extract significant patterns from a dataset containing 7,339 echocardiography examination reports of patients. The data used for this study was collected by International Cardiovascular Hospital from October, 2008 to March, 2011.

**Results:** The findings of this study revealed all the models built from J48 Decision Tree classifier, Naïve Bayes classifier and Neural Network have high classification accuracy and are generally comparable in predicting heart disease cases. However, comparison that is based on True Positive Rate suggests that the J48 model performs slightly better in predicting heart disease with classification accuracy of 95.56%.

**Conclusion:** This study showed that data mining techniques can be used efficiently to model and predict heart disease cases. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease.

**Keywords:** KDD, Data Mining, Decision Tree, Neural Network, Bayesian classifier, Heart Disease

## Background

# CHAPTER ONE INTRODUCTION

Heart disease or cardiovascular disease is the class of diseases that involve the heart or blood vessels (arteries and veins). Today most countries face high and increasing rates of heart disease and it has become a leading cause of debilitation and death worldwide in men and women over age sixty-five and today in many countries heart disease is viewed as a "second epidemic," replacing infectious diseases as the leading cause of death (Gale Nutrition Encyclopedia, 2011).

Traditionally, heart disease was thought to be the problem of developed countries, but now it is becoming a headache for developing countries too and it is especially devastating for the developing countries since they do not have adequate health care. As noted by Office on Women’s Health (2006) heart disease was considered to be a man’s problem, but now it is recognized as number one killer of women, just as it is of men.

Making a diagnosis of heart disease includes taking a complete medical evaluation and history and physical examination and early diagnosis of heart disease can help reduce the rate of mortality (Thaksin University, 2006). One of the best ways to diagnose a heart disease is by using echocardiography. Echocardiography, or echo, is a painless test that uses sound waves to create pictures of the heart. The test gives information about the size and shape of the heart and how well the heart chambers and valves are working. Echo also can be done to detect heart problems in infants and children.

The test also can identify areas of heart muscle that aren't contracting normally due to poor blood flow or injury from a previous heart attack. In addition, a type of echo called Doppler ultrasound shows how well blood flows through the chambers and valves of the heart (Joel and Robert, 1976).

As described by National Heart Lung and Blood Institute (2008) echo can detect possible blood clots inside the heart, fluid buildup in the pericardium (the sac around the heart), and problems with the aorta (aorta is the main artery that carries oxygen-rich blood from the heart to the body). But, the interpretation of echo recordings remains a challenge since there are no available precise

rules that are deduced from databases. The analysis of Echo data by experts is time consuming and this is in concomitant with the shortage of experts possessing knowledge on the analysis of Echo data. Thus, methods to automate the interpretation of Echo recordings by minimizing human efforts are important for diagnosis of heart disease in patients. In order to solve this and many other problems in the health sector related to disease diagnosis, one has to come up with a way to extract hidden information from enormous datasets that are collected in the past. Data mining can be a solution by generating rules from those enormous datasets which can be used in echo readings.

Data mining is a crucial step in discovery of knowledge from large datasets. In recent years, Data mining has found its significant hold in every field including health care. Mining process is more than the data analysis which includes classification, clustering, and association rule discovery. It also spans other disciplines like Data Warehousing, Statistics, Machine learning and Artificial Intelligence (Larose, 2005).

Data mining is predicted to be one of the most revolutionary developments of the century, according to the online technology magazine ZDNET News (Rachel, 2001). In fact, The Technology Review Ten, (2001) chose data mining as one of 10 emerging technologies that will change the world. Effectively analyzing information from customers, partners, and suppliers has become important to more companies. Furthermore, many companies have implemented a data warehouse strategy and are now starting to look at what they can do with all that data (Larose, 2005).

Data mining can be a useful tool in the health sector and healthcare. Organizations that perform data mining are better positioned to meet their long-term needs, Benko and Wilson (2003) argue that data can be a great asset to healthcare organizations, but they have to be first transformed into information.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. In recent years new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets.

Kangwanariyakul *et al.*, (2010), Patil and Kumaraswamy, (2009) have tried to apply data mining techniques in the diagnosis of heart disease. Different classification methods such as Neural Networks and Decision Trees were applied to predict the presence of heart disease and to identify the most significant factor which contributes for the cause of the disease, while association rule discovery was used to identify the effect of diet, lifestyle, and environment on the outcome of the disease. Clustering algorithms like the k-means algorithm were used on heart disease data warehouse which contains screening clinical data of patients to identify instances which are more relevant to heart attack. The results showed a bright future of data mining in the diagnosis of heart disease.

## Statement of the Problem

WHO, (2011) reported Cardiovascular Diseases (CVDs) are the number one cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.1 million people died from CVDs in 2004, representing 29% of all global deaths, of these deaths, an estimated 7.2 million were due to coronary heart disease which is one of the most common types of heart disease and 5.7 million were due to stroke.

Low- and middle-income countries are disproportionally affected, 82% of CVD deaths take place in low- and middle-income countries and occur almost equally in men and women. By 2030, almost 23.6 million people will die from CVDs, mainly from heart disease and stroke. These are projected to remain the single leading causes of death. The largest percentage increase will occur in the Eastern Mediterranean Region. The largest increase in number of deaths will occur in the South-East Asia Region due to change in lifestyle, work culture and food habits. Hence, more careful and efficient methods of cardiac diseases and periodic examination are of high importance (WHO, 2011).

Another report by WHO shows that Cardiovascular Disease and Ischemic heart disease together account 6% of total deaths in Ethiopia for all ages, which makes them the 7th and 8th deadliest diseases in Ethiopia and persons dying from heart disease are expected to grow drastically partly as a result of increasing longevity, urbanization, lifestyle changes, work culture changes and food habits changes (WHO, 2006).

In order to decrease mortality from heart diseases there should be a fast and effective detection method especially, in developing countries like Ethiopia where there is a shortage of specialists and wrongly diagnosed cases are high. Data mining can be a convenient tool to assist physicians in detecting the disease by obtaining knowledge and information regarding the disease from patient’s data.

After discussion with a domain expert, who is a consultant cardiologist at International Cardiovascular Hospital, the researcher was able to understand that early diagnosis and treatment of heart disease is vital to preventing serious, even life threatening complications such as cardiac arrest and death, but due to mistakes made during the analysis of the measurements that are taken during the echocardiography examination diagnosis of heart disease may be overlooked or delayed. The interpretation of echo recordings remains a challenge since there are no available precise rules that are deduced from databases.

As Patil and Kumaraswamy, (2009) noted like other diseases detection of heart diseases is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option.

Deciding on doctors’ intuition and experience rather than on the knowledge-rich data hidden in the database leads to unwanted biases, errors and excessive medical costs which affect the quality of service provided to patients. Wu *et al*. (2002) proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, and decrease unwanted practice variation. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

Data mining have shown a promising result in prediction of heart disease. It is widely applied for prediction or classification of different types of heart disease. For example, different data mining techniques were applied for prediction of ischemic heart disease and diagnosis of coronary artery disease (Tsipouras and Fotiadis, 2008; Kangwanariyakul et al., 2010). These successful studies which are conducted abroad have motivated this study to tackle the underlying problem that exists in our country related to heart disease diagnosis because, to the best of the researcher’s

knowledge there are no researches conducted locally regarding heart disease diagnosis using data mining techniques.

The purpose of this study is, therefore to apply data mining techniques for extracting hidden patterns, which are significant to heart diseases, from a data collected by International Cardiovascular Hospital.

## Objective

### General Objective

The general objective of this study is to design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography.

### Specific Objectives

* + - * To identify key patterns or features from the dataset.
      * To Identify and select attributes that are more relevant in relation to heart disease diagnosis.
      * To compare Decision Tree, Neural Network and Bayesian Classifiers in predicting heart disease cases.
      * To interpret and analyze the results of the selected model with the help of domain expert.

## Research Methodology

### Research Design

In this study, to develop a prediction model that can predict heart disease cases based on measurements taken from transthoracic echocardiography examination, the researcher has used the Knowledge Discovery in Database (KDD) methodology. As described by Fayyad *et al*. (1996), a given data mining project has a life cycle consisting of nine steps.

The steps are iterative and interactive, with many decisions made by the user. Note that the process is iterative at each step, meaning that moving back to previous steps may be required.

The process starts with determining the KDD goals, and ends with the implementation of the discovered knowledge.

This methodology is selected for this specific study because of three reasons. First, the KDD methodology is best suited for academic researches. Second, as an outsider for the domain using KDD methodology reduces the skill required for the knowledge discovery. Third, the KDD methodology is also independent from any tools and techniques, so one can follow any desired technique during the study. So, based on the KDD methodology, for this specific data mining project, the following nine steps were undertaken:

#### Understanding the Application Domain

To define the problem and determine medical goals, the researcher has worked closely with the hospital’s Echo Department head, who is also a consultant cardiologist at the hospital. The discussions with the domain expert have helped the researcher to learn about the problem and to know about current solutions to those problems. Since the knowledge gained from the domain expert is a high-level description of the problem from the medical point of view, a literature review was carried out and relevant works related to data mining and heart disease have been reviewed to have more knowledge about the domain. Furthermore, a real time observation of the system was performed to understand the business process of the hospital.

A key sub goal in this step is determination of data mining goals and their success criteria. The goals are obtained by translating medical goals into data mining goals

#### Selecting and Creating the Target Dataset

The researcher used data collected by international cardiovascular Hospital which contains transthoracic echocardiography report of 7,708 patients from the year 2008 to the first quarter of year 2011. The report contains different measurements that are taken during the echocardiography examination, including information on 20 variables. In an effort to reduce the number of variables, the researcher turned to a domain expert for assistance. The expert selected 15 of the most important variables for inclusion in the dataset.

As the hospital keeps the record of each patient in a separate Microsoft Word file, in order to integrate the data it was needed to create a database with variables of interest and record the

values of each variable into the new database. After recording, the new database now contains 7,339 instances each instance resembling a single file.

#### Preprocessing and Cleansing

The selected data was checked for noise, inconsistency and missing values using distribution frequency while outlier detection was done using box plots. Noises and inconsistencies identified in the dataset were corrected manually, while missing values were replaced with the most probable value determined with regression and outliers were replaced with the mean value of the attributes. All the data cleaning was performed after addressing issues and requirements of the tools selected for the preprocessing phase.

#### Data Transformation

At this stage after consulting with the domain expert a few transformations were implemented on the dataset to make the data more suitable for the data mining algorithms.

To reduce the dimensionality of the dataset encoding mechanisms were used. These methods were applied so as to obtain a reduced or compressed representation of the original data. In order to be able to reconstruct the original data from the compressed data without any loss of information a lossless reduction method was adopted.

To convert continuous valued variables to discrete values discretization method was used. This method was applied to reduce distinct values of continuous valued variables by allowing to have limited numbers of labels to represent the original variables.

The other data transformation performed was attribute selection. Attribute selection was necessary to reduce the number of features a classification algorithm has to examine and reduce errors from irrelevant features. The researcher has used the best first search method to select the best attributes from 15 attributes that were available.

#### Choosing the appropriate Data Mining Task

In this step the major task was to decide on what type of data modeling technique to use. And selecting an appropriate modeling technique depends on the goal of the study, and as the goal of

this study is to detect heart disease using data mining techniques a classification technique was adopted to develop predictive models.

#### Choosing the Data Mining Algorithm

To select the best algorithms the performance of the algorithms, the data mining goal set on the first step, the structure of the data available and the algorithm’s application in the medical field were taken into consideration.

After thoroughly checking the available algorithms in Weka machine learning software the algorithms Decision Tree, Neural Network and Bayesian Classifier were selected for this study.

#### Employing Data Mining Algorithm

To employ the selected classification algorithms four experiments were designed and the experiments were conducted on a full training dataset containing 7,339 instances. In all of the experiments two scenarios were considered, one containing all 15 attributes and the other only 8 selected attributes. 10-Fold Cross Validation was adopted for randomly sampling the training and test data sets. The Weka 3.6.4 machine learning software was used for these purposes.

#### Evaluation

In this step, models were evaluated to assess the degree to which the models met the data mining goals which were identified in the first step. The steps involved includes understanding the results, checking if the new information is novel and interesting, medical interpretation of the results, and checking their impact and the project goal. The evaluation for the algorithms was based on the classification accuracy, Area under the ROC curve and confusion matrix table. The approved models were kept and for those disapproved models the entire knowledge discovery was revised to identify failures and misleading steps.

#### Using the Discovered Knowledge

This is the last step in the knowledge discovery process and the success of this step determines the effectiveness of the entire KDD process.

After all the steps including selecting the target data, preparing data for analysis, applying data mining techniques on the data and evaluating the performance of the models are carried out a final report was written to summarize the project outcome. The final report was offered for the hospital to incorporate the discovered knowledge with the heart disease diagnosis method the hospital is already using.

### Tools

To select a data mining tool that best suits the objective of this study some factors were needed to be taken into consideration. The major factors include:

* + - * The specific analysis which needs to be performed, in this case classification.
      * The algorithms supported by the data mining tool.
      * The format of the data available.
      * The platform where the data mining tool runs on.
      * The budget for tools, although the price tag is not a good indicator of data mining tool quality.

After taking the above factors into consideration and comparing different data mining softwares Weka 3.6.4 machine learning software and SPSS version 16 were selected for this study.

Weka 3.6.4 machine learning software was selected because of three principal reasons, First, it is open source, which not only means that it can be obtained free, but more importantly it is maintainable, and modifiable, without depending on the commitment, health, or longevity of any particular institution or company. Second, it provides a wealth of state-of-the-art machine learning algorithms that can be deployed on any given problem. Third, it is fully implemented in Java and runs on almost any platform.

#### Weka Machine Learning Software

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software, developed at the University of Waikato, New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems—and even on a personal digital assistant (Witten and Frank, 2005).

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It includes virtually all the popular algorithms. It is designed so that you can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand (Witten and Frank, 2005).

#### SPSS

SPSS (Statistical Package for the Social Sciences) is a is a package of programs for manipulating, analyzing, and presenting data that provides a powerful statistical-analysis and data-management system using descriptive menus and simple point-and-click interface.

## Scope and Limitation of the Study

This study focuses on using data mining tools and techniques specifically classification techniques in developing a predictive model that is capable of predicting heart disease cases in general (i.e. specific heart disease types could not be identified using the prediction model developed by this study). To this end, data collected from Echocardiography examination by International Cardiovascular Hospital from the year 2008 to 2011 was used.

Due to the time available to complete the study, the researcher was forced to use data from a single source: International Cardiovascular Hospital found in Addis Ababa, Ethiopia, because of this the findings of this study cannot be generalized beyond the cases studied.

## Significance of the Study

It is envisioned that the results of this study will reduce medical errors, enhance patient safety and reduce mortality rate from heart disease. In addition, other medical institutions both private and public can use the result of this study for their medical decisions related to heart disease diagnosis. Furthermore, Researchers from the medical science and IT fields can use the result of

this research as an input to their study. It will also help them to come up with better solutions to the problems facing heart disease diagnosis using echocardiography.

Finally, the public will get a proper medical care if the result of this study is used along with the existing system.

## Dissemination of Results

Every effort will be made to disseminate the results of the study through the following ways:

* Presentation for the school of public Health and School of Information Science
* Publishing in different journals
* Presentation on different conferences/workshops
* Putting the hardcopy in the libraries of concerned organizations so that interested readers can get access to the research output to be used for decision support, take action or to use it as a base for further research in the area;

## Organization of the Thesis

The thesis is organized into seven chapters. The first chapter is an introductory chapter which gives a brief introduction about the study. It also describes the problem that constitutes the basis of the paper, objectives of the study, the research methodology, significance and scope of the study.

Chapter two focuses on reviewing relevant literatures on data mining and heart disease that are necessary to understand the methods and terms that are introduced in the parts of the thesis. Finally, a few researches which address problems similar to this study are reviewed.

The third chapter of this thesis is devoted on describing different algorithms and performance measures that are employed during model building and performance evaluation.

Chapter four is concerned with understanding the application domain and the dataset selected for the study. Furthermore, medical goals, data mining goals and dataset description are presented in this chapter.

Chapter Five deals with the experimental part of the study. In this chapter experiments that are performed on the preprocessed data are discussed with their corresponding interpretations. In addition, the performances of the models are presented.

The six and final chapter of this thesis contains a summary of the study. This chapter sums up the methods that are used, the experimentation results and draws general conclusions. Afterwards recommendations are forwarded and ideas for future work are presented.

# CHAPTER TWO LITERATURE REVIEW

## Overview of Data Mining

Due to a wide availability of huge amount of data and a need to convert this available huge amount of data to useful information necessitates the use of data mining techniques. Data Mining and KDD have become popular in recent years. The popularity of data mining and KDD shouldn’t be a surprise since the size of the data collections that are available are far too large to be examined manually and even the methods for automatic data analysis based on classical statistics and machine learning often face problems when processing large, dynamic data collections consisting of complex objects.

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories become “data tombs”—data archives that are seldom visited. Consequently, important decisions are often made based not only on the information-rich data stored in data repositories, but also on a decision maker’s intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In addition, consider expert system technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. The widening gap between data and information calls for a systematic development of data mining tools that will turn data tombs into “golden nuggets” of knowledge (Han and Kamber, 2006).

As noted by Fayyad *et al.* (1996) historically the notation of finding useful patterns in data have been given a variety of names including data mining, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing but recently the

terms data mining and KDD are dominating in the Management Information Science (MIS) communities and database fields.

KDD is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex datasets. Data Mining is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction (Oden and Lior, 2005).

## Developmental History of Data Mining

Just a few short years ago, few people had even heard of the term data mining. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s.

As Han and Kamber (2006) described the building blocks of today’s data mining techniques date back to 1960s where database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. Since the 1970s database systems has progressed from early hierarchical and network database systems to the development of relational database systems (where data are stored in relational table structures), data modeling tools, and indexing and accessing methods. In addition, users gained convenient and flexible data access through query languages, user interfaces, optimized query processing, and transaction management.

Starting from the mid-1980s database technologies adopted relational technology and there were an upsurge of research and development activities on new and powerful database systems. Issues related to the distribution, diversification, and sharing of data have been studied extensively. Heterogeneous database systems and Internet-based global information systems such as the World Wide Web (WWW) have also emerged and play a vital role in the information industry (Han and Kamber, 2006).

The confluence of disciplines like Artificial Intelligence, Information Retrieval, statistics, and database systems plus the availability of fast microcomputers opened up a world of possibilities

for retrieving and analyzing data. During this time new programming languages were developed and new computing techniques were developed (Dunham, 2003).

The 1990s saw the development of database warehouses, a term used to describe a large database (composed of a single schema), created from the consolidation of operational and transactional database data. Along with the development of data warehouses came online analytical processing, decision support systems, data scrubbing/staging (transformation), and association rule algorithms. During the same period, data mining changed from being an interesting new technology to becoming part of standard business practice. This occurred because the cost of computer disk storage went down, processing power went up, and the benefits of data mining became more apparent. Businesses began using “data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers” (Two Crows Corporation, 2005).

## Knowledge Discovery in a Database and Data Mining

Fayyad *et al*. (1996) defined KDD as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. According to this definition, data is a set of facts that is somehow accessible in electronic form. The term “patterns” indicates models and regularities which can be observed within the data. Patterns have to be valid, i.e. they should be true on new data with some degree of certainty.

KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, Artificial Intelligence, knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large datasets. The data mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data mining step of the KDD process. KDD focuses on the overall process of knowledge discovery from data, including how the data are stored and accessed, how algorithms can be scaled to massive datasets still run efficiently, how results can be interpreted and visualized, and how the overall man- machine interaction can usefully be modeled and supported (Sciences Applied, 2010).

A driving force behind KDD is the database field. Indeed, the problem of effective data manipulation when data cannot fit in the main memory is of fundamental importance to KDD. Database techniques for gaining efficient data access, grouping and ordering operations when accessing data, and optimizing queries constitute the basics for scaling algorithms to larger datasets. Most data mining algorithms from statistics, pattern recognition, and machine learning assume data are in the main memory and pay no attention to how the algorithm breaks down if only limited views of the data are possible. A related field evolving from databases is data warehousing, which refers to the popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. Data warehousing helps set the stage for KDD in two important ways (Sciences Applied, 2010).

KDD and data mining are often used interchangeably in some literatures, according to Chen *et al*. (1997), data mining, which is also referred to as knowledge discovery in databases (KDD), is defined as a process of extracting nontrivial, implicit, previously unknown and potentially useful information (such as knowledge, rules, constraints, regularities) from data in databases.

Fayyad *et al*. (1996) defined data mining as a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. According to this definition Data mining is the step that is concerned with the actual extraction of knowledge from data. To emphasize the necessity that data mining algorithms need to process large amounts of data, the desired patterns has to be found under acceptable computational efficiency limitations.

As Hand *et al*. (2001) defined it “Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”. The definition above refers to "observational data," as opposed to "experimental data." Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. This means that the objectives of the data mining exercise play no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as "secondary" data analysis.

According to Connolly *et al*. (1999) Data mining is “a process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions”.

The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. In this regard, Thearling (2000) notes that data mining tools can answer business questions that traditionally were too time-consuming to resolve. These tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

To summarize, data mining is a way to find previously unknown, valid patterns and relationships from huge amount of data represented in qualitative, textual or multimedia formats by applying different data analysis tools and also most of the time the datasets are collected for other purposes.

## Data Mining Process Models

SEMMA, CRISP-DM and KDD are popular data mining process models that are used in data mining projects.

SEMMA (Sample, Explore, Modify, Model, Assess) was developed by the SAS Institute. It refers to the process of conducting a data mining project. The model considers a cycle with 5 stages for the process. The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals.

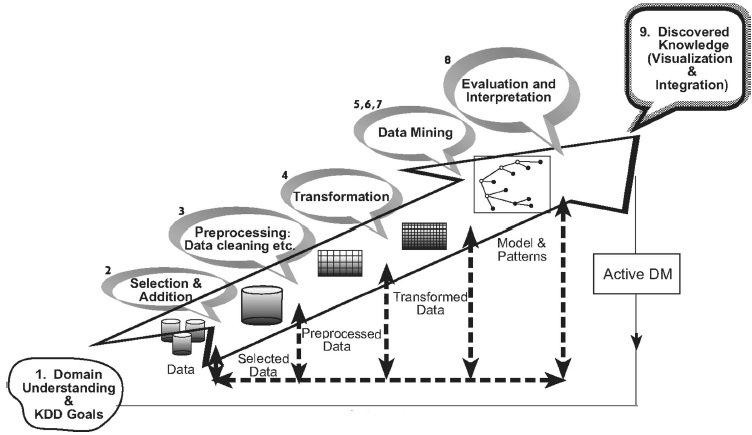
CRISP-DM (Cross-Industry Standard Process for Data Mining) was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR (National Cash Register), the model consists a cycle that comprises six stages. CRISP-DM is extremely complete and documented. All his stages are duly organized, structured and defined, allowing that a project could be easily understood or revised (Larose, 2005).

KDD process, as presented in Fayyad *et al*. (1996), is the process of using data mining 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. The KDD process model is adopted for this study, therefore

### The KDD Process Model

According to Fayyad *et al.* (1996) The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. Each step attempts to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

The KDD process model proposed by Fayyad *et al*. (1996) consists of nine steps that are executed in iterative mode. Each subsequent step is initiated upon successful completion of the previous steps, and requires the result generated by the previous step as its input, and moving back to previous steps may be required. The process starts by developing an understanding about the application domain and ends with using the discovered knowledge as shown on Figure 2.1 (Oden and Lior, 2005)

**Figure 2.1: The KDD Process for Discovering Knowledge**

#### Understanding the Application Domain

This is the initial step where experts who are in charge of the data mining research project need to understand and accordingly define the goals of the end-user and the environment in which the knowledge discovery process will take place including relevant prior knowledge. As the KDD process proceeds, there may be even a revision of this step. After understanding the KDD goals, creating and preprocessing of the target data proceeds next.

#### Creating a Target Dataset

Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one dataset, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From this respect, the more attributes are considered, the better. On the other hand, to collect, organize and operate complex data repositories is expensive and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. This starts with the best available dataset and later expands and observes the effect in terms of knowledge discovery and modeling.

#### Data Cleaning and Preprocessing

In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. There are many methods, from doing nothing to becoming the major part (in terms of time consumed) of a KDD project in certain projects. It may involve complex statistical methods or using a Data Mining algorithm in this context. For example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors. In any case,

studying the aspects is important and often revealing by itself, regarding enterprise information systems.

#### Data Transformation

In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step can be crucial for the success of the entire KDD project, and it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. In marketing, we may need to consider effects beyond our control as well as efforts and temporal issues (such as studying the effect of advertising accumulation). However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed.

#### Choosing the appropriate Data Mining Task

In this step we decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta- learning for the particular set of available data.

#### Choosing the Data Mining Algorithm

Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers). For example, in considering precision versus understandability, the former is better with neural networks, while

the latter is better with decision trees. For each strategy of meta-learning there are several possibilities of how it can be accomplished. Meta-learning focuses on explaining what causes a Data Mining algorithm to be successful or not in a particular problem. Thus, this approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate. Each algorithm has parameters and tactics of learning (such as ten-fold cross-validation or another division for training and testing).

#### Employing the Data Mining Algorithm

Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm’s control parameters, such as the minimum number of instances in a single leaf of a decision tree.

#### Interpreting and Evaluating Mined Patterns

In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Here we consider the preprocessing steps with respect to their effect on the Data Mining algorithm results. This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage.

The last step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining:

#### Using the Discovered Knowledge

We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we may make changes to the system and measure the effects. Actually the success of this step determines the effectiveness of the entire KDD process. There are many challenges in this step, such as losing the “laboratory conditions” under which we have operated. For instance, the knowledge was discovered from a certain static snapshot (usually sample) of the data, but now the data becomes dynamic.

Data structures may change (certain attributes become unavailable), and the data domain may be modified such as, an attribute may have a value that was not assumed before.

## Data Mining Tasks

As Han and Kamber (2006) stated data mining tasks are used to specify the kind of patterns to be found in data mining process. In general, data mining tasks can be classiﬁed into two categories: **Predictive** and **Descriptive**.

A Predictive model makes a prediction about values of data using known results found from different data and its goal is to Identify strong links between variables of a data table (columns) while the, examples include providing a diagnosis for a medical patient on the basis of a set of test results, estimating the probability that customers will buy product A given a list of other products they have purchased, or predicting the value of something six months from now, given current and past values of the index (Hand *et al*., 2001).

Descriptive model identifies patterns or relationships in data. It simply summarizes data in convenient ways or in ways that will lead to increased understanding of the way things work. For example, a physician might be interested in discovering the influence of climate among malaria patients by grouping patients in different climate zones, or may be a physician might want to cluster contraceptive users on their education background.

The major difference between the two models is that, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. In contrast, a predictive model has the specific objective of allowing us to predict the value of some target characteristic of an object on the basis of observed values of other characteristics of the object.

Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rule Discovery, and Sequence analysis.

### Classification

According to Han and Kamber (2006) Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able

to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

As stated by Larose (2005) examples of classification tasks include the following:

* Determining whether a particular credit card transaction is fraudulent
* Assessing whether a mortgage application is a good or bad credit risk
* Diagnosing whether a particular disease is present or absent
* Determining whether a will was written by the actual deceased, or fraudulently by someone else
* Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat

Some of the major tools used for constructing a classification model include Decision Trees, Artificial Neural Networks and Bayesian Classifier.

#### Decision Tree

Berry and Linoff (2004) defined decision tree as “a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another “.

A record enters the tree at the root node. The root node applies a test to determine which child node the record will encounter next. There are different algorithms for choosing the initial test, but the goal is always the same: to choose the test that best discriminates among the target classes. This process is repeated until the record arrives at a leaf node. All the records that end up at a given leaf of the tree are classified the same way. There is a unique path from the root to each leaf. That path is an expression of the rule used to classify the records (Berry and Linoff, 2004).

Different leaves may make the same classification, although each leaf makes that classification for a different reason. For example, in a tree that classifies fruits and vegetables by color, the leaves for apple, tomato, and cherry might all predict “red,” albeit with varying degrees of confidence since there are likely to be examples of green apples, yellow tomatoes, and black cherries as well (Berry and Linoff, 2004).

Han and Kamber (2006) noted that the construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy for linearly separable problems.

Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology (Kantardzic, 2003).

#### Artificial Neural Networks

As noted by Larose (2005) the inspiration for neural networks was the recognition that complex learning systems in human brains consisted of closely interconnected sets of neurons. Although a particular neuron may be relatively simple in structure, dense networks of interconnected neurons could perform complex learning tasks such as classification and pattern recognition.

According to Hand *et al*. (2001) Artificial Neural Networks (ANNs) are one of a class of highly parameterized statistical models that have attracted considerable attention in recent years. The fact that ANNs are highly parameterized makes them very flexible, so that they can accurately model relatively small irregularities in functions.

One of the advantages of using neural networks is that they are quite robust with respect to noisy data. Because the network contains many nodes (artificial neurons), with weights assigned to each connection, the network can learn to work around these uninformative (or even erroneous) examples in the dataset. However, unlike decision trees, which produce intuitive rules that are understandable to non-specialists, neural networks are relatively opaque to human interpretation. Also, neural networks usually require longer training times than decision trees, often extending into several hours.

#### Bayesian classifiers

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is

based on Bayes’ theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve” (Han and Kamber, 2006).

Various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains. In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data (Han and Kamber, 2006).

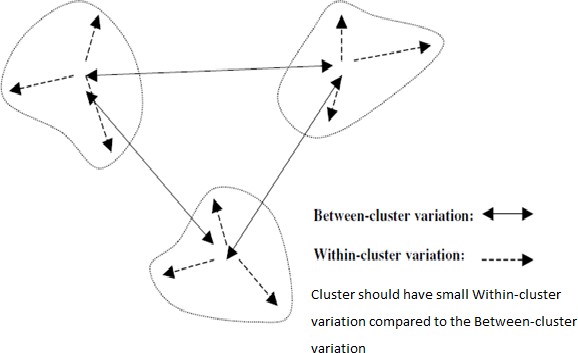
### Clustering

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived (Larose, 2005).

As noted by Larose (2005) examples of clustering tasks include the following:

* Target marketing of a niche product for a small-capitalization business that does not have a large marketing budget.
* For accounting auditing purposes, to segmentize financial behavior into benign and suspicious categories.
* As a dimension-reduction tool when the dataset has hundreds of attributes.
* For gene expression clustering, where very large quantities of genes may exhibit similar behavior.

Clustering is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks. Due to the enormous size of many present day databases, it is often helpful to apply clustering analysis first, to reduce the search space for the downstream algorithms (Larose, 2005).



**Figure 2.2: Between-cluster variation and Within-cluster variation**

Some of the major clustering algorithms include hierarchical clustering and K-means clustering. In hierarchical clustering, a tree like cluster structure (dendrogram) is created through recursive partitioning (divisive methods) or combining (agglomerative) of existing clusters. Agglomerative clustering methods initialize each observation to be a tiny cluster of its own. Then, in succeeding steps, the two closest clusters are aggregated into a new combined cluster. In this way, the number of clusters in the dataset is reduced by one at each step. Eventually, all records are combined into a single huge cluster. Divisive clustering methods begin with all the records in one big cluster, with the most dissimilar records being split off recursively, into a separate cluster, until each record represents its own cluster (Larose, 2005).

The K-means algorithm is one of the most commonly used partitioning clustering algorithms. The “K” in its name refers to the fact that the algorithm looks for a fixed number of clusters which are defined in terms of proximity of data points to each other. The technique is illustrated using two-dimensional diagrams. In practice the algorithm is usually handling many more than two independent variables. This means that instead of points corresponding to two-element

vectors (x1, x2), the points correspond to n-element vectors (x1, x2, . . . xn). The procedure itself is unchanged (Berry and Linoff, 2004).

Descriptions of K-means and related algorithms gloss over the selection of K. But since, in many cases, there is no a priori reason to select a particular value, there is really an outermost loop to these algorithms that occurs during analysis rather than in the computer program.

Even though, K-means clustering algorithm is the simplest and most commonly used algorithm it is very sensitive to noise and outlier data points, because a small number of such data can substantially influence the mean value (Kantardzic, 2003).

### Association Rule Discovery

The association task for data mining is the job of finding which attributes “go together.” Most prevalent in the business world, where it is known as affinity analysis or market basket analysis, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Association rules are of the form “If antecedent, then consequent,” together with a measure of the support and confidence associated with the rule.

According to Han and Kamber (2006) confidence and support are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. Which is the probability P( ) where indicates that a transaction contains both X and Y, that is, the union of itemsets X and Y. Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability ( | ), that is, the probability that a transaction containing X also contains Y. More formally, support and confidence are defined as

( ) ( ) ( )

( ) ( | ) ( )

Association differs from classification in two ways: they can “predict” any attribute, not just the class, and they can predict more than one attribute’s value at a time. Because of this there are far

more association rules than classification rules, and the challenge is to avoid being swamped by them (Witten and Frank, 2005).

As noted by Larose (2005) examples of association tasks in business and research include the following:

* Investigating the proportion of subscribers to a company’s cell phone plan that respond positively to an offer of a service upgrade
* Examining the proportion of children whose parents read to them who are themselves good readers
* Predicting degradation in telecommunications networks
* Finding out which items in a supermarket are purchased together and which items are never purchased together
* Determining the proportion of cases in which a new drug will exhibit dangerous side effects

One of major association rule discovery algorithms is Apriori algorithm. Apriori is a classic algorithm for learning association rules. Apriori uses breadth-first search and a tree structure to count candidate itemsets efficiently. It generates candidate itemsets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. This algorithm has emerged as one of the best association rule mining algorithms. It also serves as the base algorithm for most parallel algorithms (Larose, 2005).

## Application of Data Mining in the Health Sector

Without a question the health sector has more need for data mining today. There are several arguments that support the use of data mining in the health sector. Kob and Tan (2005) argue that “data mining is becoming popular in the health care, if not essential”. To back their argument they stated that healthcare insurers use data mining to reduce the existence of medical insurance fraud and abuse. As they have stated use of data mining applications for fraud detection is prevalent in the commercial world, for example, in the detection of fraudulent credit card

transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection.

Another factor for the need of data mining in health sector is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. As Biafore, (1999) pointed out using data mining to analyze such huge amounts of data has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data.

There are also other factors that boost data mining popularity in the health sector one of them is non-invasive diagnosis and decision support. Some diagnostic and laboratory procedures are invasive, costly and painful to patients. An example of this is conducting a biopsy in women to detect cervical cancer. Thangavel *et al.* (2006) used the K-means clustering algorithm to analyze cervical cancer patients and found that clustering found better predictive results than existing medical opinion. They found a set of interesting attributes that could be used by doctors as additional support on whether or not to recommend a biopsy for a patient suspected of having the cervical cancer.

## Related Works

Numerous works related to heart disease diagnosis using data mining techniques have motivated this study.

A model Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network was proposed by Palaniappan and Awang, (2008), they used a CRISP-DM methodology to build the mining models on a dataset obtained from the Cleveland Heart Disease database. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease.

Another study conducted by Guru *et al*. (2007) experimented on a sample database of patients’ records. The Neural Network is tested and trained with 13 input variables such as Age, Blood Pressure, Angiography’s report and the like. The supervised network has been recommended for diagnosis of heart diseases. Training was carried out with the aid of back propagation algorithm. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of probable diseases that the patient is vulnerable to. The success rate for imprecise inputs to retrieve the desired output is closest to 100%.

The problem of identifying constrained association rules for heart disease prediction was studied by Carlos, (2004). The assessed dataset encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably

besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries.

Kangwanariyakul *et al*. (2010) presented a predictive model for the Ischemic Heart Disease (IHD); they applied Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN) and the support vector machine (SVM) to develop classification models for identifying IHD patients on a data obtained from measurements of cardiac magnetic field at 36 locations (6 × 6 matrices) above the torso. The result shows that BPNN and BNN gave the highest classification accuracy of 78.43 %, while RBF kernel SVM gave the lowest classification accuracy of 60.78 %. BNN presented the best sensitivity of 96.55

% and RBF kernel SVM displayed the lowest sensitivity of 41.38 %. Both polynomial kernel SVM and RBF kernel SVM presented the minimum and maximum specificity of 45.45 % and

86.36 %, respectively.

After reviewing the above literatures the researcher was motivated to work on a classification model that is sought to predict heart disease cases based on patterns generated from International Cardiovascular Hospital database. This study is different in two ways from the studies that are presented above, the first one is the data that is used for this study is collected from transthoracic echocardiography report of patients and the second one is the classification models are developed on a much larger dataset. As far as the knowledge of the researcher is concerned this study will be the first of its kind in Ethiopia that applied data mining to predict heart disease.

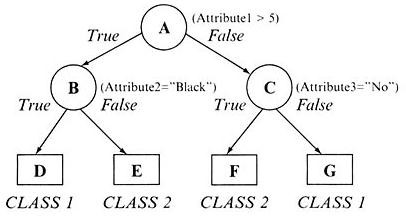
# CHAPTER THREE ALGORITHMS USED FOR MODEL BUILDING AND PERFORMANCE

**MEASURES**

This study aims at applying classification techniques for heart disease detection. An attempt was made to construct prediction model using Decision Tree, Neural Network and Bayesian Classifier. After constructing the models, performance of each models were evaluated, and also their performances were compared to each other. In this section the algorithms used to build the models and matrices used for performance measure and comparison are discussed in detail.

## Decision Trees

Han and Kamber (2006) defined decision tree as a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.



**Figure 3.1: A simple Decision Tree**

As the construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery they have become popular. Decision trees can handle high dimensional data.

Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand.

Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology (Han and Kamber, 2006).

### J48 Classifier Algorithm

During the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). This work expanded on earlier work on concept learning systems, Quinlan later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared.

C4.5 algorithm is an improvement of IDE3 algorithm. It is based on Hunt’s algorithm and also like IDE3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. Unlike IDE3, C4.5 accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too-much details in the training data set. Like IDE3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute (Quinlan, 1993).

The C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. Suppose that we have a variable X whose k possible values have probabilities P1, P2, . . . , Pk the smallest number of bits, on average per symbol, needed to transmit a stream of symbols representing the values of X observed called the entropy of X and is defined as

For an event with probability p, the average amount of information in bits required to transmit the result is(p). For example, the result of a fair coin toss, with probability 0.5, can be transmitted using(0.5) = 1 bit, which is a zero or 1, depending on the result of the toss. For

variables with several outcomes, we simply use a weighted sum of the (Pj)’s, with weights equal to the outcome probabilities, resulting in the formula 3.1.

C4.5 uses this concept of entropy as follows. Suppose that we have a candidate split S, which partitions the training dataset T into several subsets, T1, T2, . . . , Tk . The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

Where Pi represents the proportion of records in subset i. We may then define our information gain to be gain(S) = H(T) − Hs(T), that is, the increase in information produced by partitioning the training data T according to this candidate split S. At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain(S).

J48 implements Quinlan’s C4.5 algorithm for generating a pruned or unpruned C4.5 decision tree. The decision trees generated by J48 can be used for classification (Quinlan, 1993).

A feature of the J48 algorithm is that it “prunes” leaves that do not contribute greatly to the predictive accuracy of the tree. This creates smaller trees, which may be more resistant to over fitting

Han and Kamber stated that, when a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data. Such methods typically use statistical measures to remove the least reliable branches.

According to Han and Kamber (2006), there are two common approaches to tree pruning: prepruning and postpruning. In the prepruning approach, a tree is “pruned” by halting its construction early (e.g., by deciding not to further split or partition the subset of training tuples at a given node). Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.

When constructing a tree, measures such as statistical significance, information gain, Gini index, and so on can be used to assess the goodness of a split. If partitioning the tuples at a node would

result in a split that falls below a pre-specified threshold, then further partitioning of the given subset is halted. There are difficulties, however, in choosing an appropriate threshold. High thresholds could result in oversimplified trees, whereas low thresholds could result in very little simplification.

The second and more common approach is post pruning, which removes sub trees from a “fully grown” tree. A sub tree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the sub tree being replaced.

As described by Quinlan (1993) J48 Decision tree classifier follows the following simple algorithm. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class.

J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation.

## Bayesian Classifier

Han and Kamber (2006) stated that Bayesian classifiers are statistical classifiers that are based on Bayes’ theorem. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

Bayes’ theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. Let X be a data tuple. In Bayesian terms, X is considered “evidence.” As usual, it is described by measurements made on a set of n attributes. Let H be some hypothesis, such that the data tuple X belongs to a specified class C. For classification problems, we want to determine ( | ) the probability that the hypothesis H holds given the “evidence” or observed data tuple X. In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

The probabilities P(H) and P(X) may be estimated from the given data. Bayes’ theorem is useful in that it provides a way of calculating the posterior probability, from P(H) and P(X) as shown in equation 3.3.

### Naïve Bayes

The Naïve Bayes classifier is based on the Bayes rule of conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other.

Various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains. In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data (Han and Kamber, 2006).

As described by Han and Kamber (2006) the naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2, . . . , xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2, . . . , An.

Suppose that there are m classes, C1, C2, . . . , Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if

As P(X) is constant for all classes, only ( | ) ( ) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = . . . = P(Cm), and we would therefore maximize P(X│Ci). Otherwise, we maximize ( | ) ( ). Note that the class prior probabilities may be estimated by number of training tuples of class Ci in D.

Given datasets with many attributes, it would be extremely computationally expensive to compute . In order to reduce computation in evaluating, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).

## Performance Measures

In order to minimize the bias associated with the random sampling of the training and test data samples k-Fold Cross Validation was adopted. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2, : : : , Dk, each of approximately equal size. Training and testing is performed k times (Han and Kamber, 2006).

As Witten and Frank (2005) stated, extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up. Although these arguments are by no means conclusive, and debate continues to rage in machine learning and data mining circles about what is the best scheme for evaluation, 10-fold cross-validation has become the

standard method in practical terms. Tests have also shown that the use of stratification improves results slightly. Thus the standard evaluation technique in situations where only limited data is available is stratified 10-fold cross-validation.

### 10-Fold Cross Validation

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold.

According to Olson and Delen (2008) 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation. In fact, in data mining community, for methods-comparison studies with relatively smaller datasets, k-fold type of experimentation methods are recommended. In essence, the main advantage of 10-fold (or any number of folds) cross validation is to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample.

The cross validation estimate of the overall accuracy of a model is calculated by simply averaging the 10 individual accuracy measures

Where CVA stands for cross validation accuracy and A is the accuracy measure (e.g., it-rate, sensitivity, specificity, etc.) of each folds.

There are three steps to perform 10-Fold Cross Validation:

**Step 1:** The complete dataset is randomly divided into 10 disjoint subsets (i.e., folds) with each containing approximately the same number of records. Sampling is stratified by the class labels to ensure that the proportional representation of the classes is roughly the same as those in the original dataset.

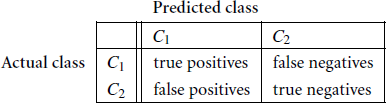
**Step 2:** For each fold, a classifier is constructed using all records except the ones in the current fold. Then the classifier is tested on the current fold to obtain a cross-validation estimate of its error rate. The result is recorded.

**Step 3:** After repeating the step 2 for all 10 folds, the ten cross validation estimates are averaged to provide the aggregated classification accuracy estimate of each model type.

### Confusion Matrix

In classification problems, the primary source of performance measurements is a confusion matrix (Coincidence matrix, classification matrix or a contingency table). Given m classes, a confusion matrix is a table of at least size m by m (Olson and Delen, 2008).

If the instance is positive and it is classified as positive, it is counted as a true positive (TP); if it is classified as negative, it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as a true negative (TN); if it is classified as positive, it is counted as a false positive (FP). These terms are useful when analyzing a classifier’s ability.



**Figure 3.3: A simple Confusion Matrix**

Figure 3.3 shows a confusion matrix for a two-class classification problem. The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. The equations of most commonly used metrics that can be calculated from the coincidence matrix are discussed below.

The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples. Other performance measures, such as recall (sensitivity), specificity and F-measure are also used for calculating other aggregated performance measures (e.g., area under the ROC curves).

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

Han and Kamber (2006) gave an example to emphasize on using other measures as an alternative to the accuracy measure. ” Suppose that you have trained a classifier to classify medical data tuples as either “cancer” or “not cancer.” an accuracy rate of, say, 90% may make the classifier seem quite accurate, but what if only, say, 3–4% of the training tuples are actually “cancer”? Clearly, an accuracy rate of 90% may not be acceptable—the classifier could be correctly labeling only the “not cancer” tuples, for instance. Instead, we would like to be able to access how well the classifier can recognize “cancer” tuples (the positive tuples) and how well it can recognize “not cancer” tuples (the negative tuples).

The sensitivity and specificity measures can be used, respectively, for this purpose. Sensitivity is also referred to as the true positive (recognition or Recall) rate (that is, the proportion of positive tuples that are correctly identified), while specificity is the true negative rate (that is, the proportion of negative tuples that are correctly identified). We may use precision to access the percentage of tuples labeled as “cancer” that actually are “cancer” tuples. These measures are defined as

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

F-measure (also known as F1 or F-score) is a measure of test’s accuracy. It considers both the Precision and the Recall of the test to compute the score. It can be interpreted as a weighted average of the Precision and the Recall, where 1 is its best value and 0 its worst. The F-Measure

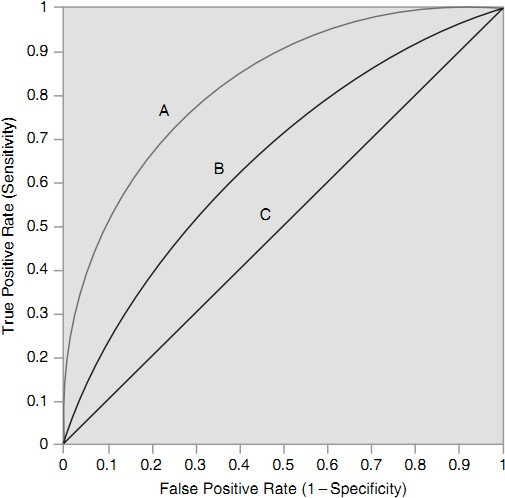
### Area under the ROC Curve

A Receiver Operating Characteristics (ROC) curve is a technique for visualizing, organizing and selecting classifiers based on their performance. In essence, it is another performance evaluation technique for classification models and also useful tool for comparing two or more classification models. ROC curves have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers. The use of ROC analysis has been extended into visualizing and analyzing the behavior of diagnostic systems. Recently, the medical decision making community has developed an extensive literature on the use of ROC curves as one of the primary methods for diagnostic testing (Olson and Delen, 2008).

As Han and Kamber (2006) described ROC curve shows the trade-off between the true positive rate or sensitivity (proportion of positive tuples that are correctly identified) and the false positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model. That is, given a two-class problem, it allows us to visualize the trade-off between the rate at which the model can accurately recognize ‘yes’ cases versus the rate at which it mistakenly identifies ‘no’ cases as ‘yes’ for different “portions” of the test set. Any increase in the true positive rate occurs at the cost of an increase in the false-positive rate. The area under the ROC curve is a measure of the accuracy of the model.

In order to plot an ROC curve for a given classification model, true positive (TP) rate is plotted on the Y axis and false positive (FP) rate is plotted on the X axis (see Figure 3.6). As Han and Kamber (2006) described the process of drawing Roc curve we start at the bottom left-hand corner (where the true positive rate and false-positive rate are both 0), we check the actual class label of the tuple at the top of the list. If we have a true positive (that is, a positive tuple that was correctly classified), then on the ROC curve, we move up and plot a point. If, instead, the tuple really belongs to the ‘no’ class, we have a false positive. On the ROC curve, we move right and

plot a point. This process is repeated for each of the test tuples, each time moving up on the curve for a true positive or toward the right for a false positive.



**Figure 3.4: A Sample ROC curves**

As Olson and Delen (2008) noted to compare classifiers or to judge the fitness of a single classifier one may want to reduce the ROC measures to a single scalar value representing the expected performance. A common method to perform such task is to calculate the area under the ROC curve, Area under the Curve (AUC) is a portion of the area of the unit square, and its value will always be between 0 and 1.0. A perfect accuracy gets a value of 1.0. The diagonal line y = x represents the strategy of randomly guessing a class. For example, if a classifier randomly guesses the positive class half the time (much like flipping a coin), it can be expected to get half the positives and half the negatives correct; this yields the point (0:5; 0:5) in ROC space, which in turn translates into area under the ROC curve value of 0.5. No classifier that has any classification power should have an AUC less than 0.5.

For example, In Figure 3.4 classification performances of three classifiers (A, B and C) are shown in a single ROC graph. Since the AUC is the commonly used metric for performance comparison of prediction models, one can easily tell that the best performing classifier (out of the three that is being compared to each other) is A, followed by B. The classifier C is not shoving any predictive power; staying at the same level as random chance.

**Table 3.1: Performance measure of ROC Area**

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

To summarize, three algorithms namely J48 Classifier, Multilayer Perception and Naïve Bayes were used for model building and 10-fold cross validation, Predictive accuracy, TP Rate, TN Rate, Precision, and F-Measure, are six measures used for the evaluation of classification and prediction methods while Predictive accuracy, TP Rate, TN Rate and ROC Area were used to compare the models.

# CHAPTER FOUR

**BUSINESS AND DATA UNDERSTANDING**

This study is based on a data collected by the International Cardiovascular Hospital. So, to understand the medical domain, the researcher has worked closely with the hospital’s administration and Echo department. In addition, real time observation of the business process was performed to gain an insight how the hospital functions.

International Cardiovascular Hospital is one of the two privately owned cardiac hospitals in Addis Ababa. The hospital is established with the aim of serving the needs of the local community related to cardiac diseases by providing the highest quality health care to patients and their families. International Cardiovascular Hospital aspires to transform the future of healthcare, through filling the gap created by the limited ability of government owned hospitals to cope up with the growing cardiovascular disease.

To understand the medical problem domain, the first step was identifying the current procedures that are used to diagnose heart disease, after the procedures are identified the researcher continued to define the problems which occur during the diagnosis of the disease. Finally medical goals were determined and data mining goals were set to identify and prepare data required for the study.

## Business Process Description

Basically there are three types of patients who come to the hospital: new patients, referred patients and patients who had examination at the hospital at least once. The process of medical examination varies depending on the patient type, thus the process of medical examinations are discussed separately for each type of patient.

If the patient is new he/she directly requests for the service at the reception desk and the receptionist fills all the required fields on the patients’ medical chart and after completing the registration the receptionist will issue an index card and a receipt upon payment for the patient and will send the patient’s medical chart to the nurses’ station.

After taking vital signs from patients the nurses will send the patient to a waiting room and the patient’s medical chart to the physician. When the physician is ready to examine the patient one of the nurses will take the patient to examination room and the physician diagnosis the patient, depending on the symptoms, then the physician may advices and discharges the patient if he/she is disease free or otherwise may request laboratory tests (including echocardiography examination), after the patient undergoes a laboratory test the result will be sent to the physician from the laboratory directly.

Next, the physician examines the results and may prescribe drugs, might admit the patient at the hospital for further treatment or the physician may discharge the patient with follow up instructions if the patient is stable enough to go home. The physician might prescribe drugs, admit the patient or refer the patient even without any laboratory test.

Referred patients come to the hospital for two reasons; the first is for only laboratory tests’ including ECG, Echocardiography and stress test, the second is for specialized medical examination. Patients that come for laboratory tests don’t need to go through registration and medical examination like new patients all they have to do is pay the fee required for the laboratory test, undergo the test and take the result with them to the health institution that referred them, but patients that are referred for specialized medical examination might need to go through all the steps like new patients.

If the patient had an examination at the hospital at least once and if he/she came back to the hospital either for a follow up or without any appointment, the receptionist will search the patient’s medical card using the unique card no, and will send the chart which contains the patient’s medical history to concerned physician.

## Heart Diseases Diagnosis

There are many methods that can be used to diagnose for possible heart disease. Usually the recognition of the disease starts with identifying the patient’s history, habits such as smoking, life style and by preforming the physical examination including blood pressure, blood chemistries, ECG, and Echocardiography. The healthcare provider will try to choose the testing modality that will best provide the diagnosis.

To understand the medical problem related to heart disease diagnosis using echocardiography test the focus will be on echocardiography, processes of the echocardiography examination and different types of echocardiography tests.

### Echocardiography

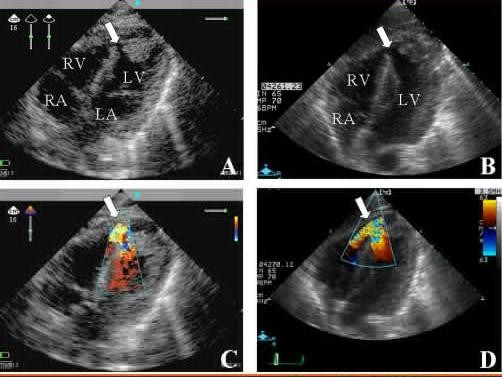
Echocardiography is one of the most widely used diagnostic tests for heart disease. The term echocardiography refers to a group of tests that utilize ultrasound to examine the heart and record information in the form of echoes, i.e. reflected sonic. In addition to providing single-dimension images, known as M-mode echo that allows accurate measurement of the heart chambers, the echocardiogram also offers far more sophisticated and advanced imaging. This is known as two- dimensional (2-D) Echo and is capable of displaying a cross-sectional "slice" of the beating heart, including the chambers, valves and the major blood vessels that exit from the left and right ventricle (Joel and Robert, 1976).

A standard echocardiography or Transthoracic Echocardiography which is the most common type of echocardiogram examination generally lasts between 15–30 minutes. The patient lies bare-chested on an examination table. A special gel is spread over the chest to help the transducer make good contact and slide smoothly over the skin. The transducer, also called a probe, is a small handheld device at the end of a flexible cable.

The transducer, essentially a modified microphone, is placed against the chest and directs ultrasound waves into the chest. Some of the waves get echoed (or reflected) back to the transducer. Since different tissues and blood reflect ultrasound waves differently, these sound waves can be translated into a meaningful image of the heart that can be displayed on a monitor or recorded on paper or tape (See Figure 4.1).

Finally, the physician reviews the examination prior to completion of the final report. The patient does not feel the sound waves, and the entire procedure is painless (Lee *et al.*, 2001).

Echocardiography is usually performed in the cardiology department at a hospital, but may also be performed in a cardiologist's office or an outpatient imaging center. Because the ultrasound scanners used to perform echocardiography are portable (handheld) or mobile, echocardiography can be performed in the hospital's emergency department or at the bedside of patients who cannot be transported to the cardiology department.



**Figure 4.1: Images from echocardiography examination**

Occasionally, variations of the echocardiography test are used. This includes Doppler echocardiography and Stress echocardiography.

**Doppler echocardiography:** Doppler is a special part of the ultrasound examination that assesses blood flow (direction and velocity). In contrast, the M-mode and 2-D Echo evaluates the size, thickness and movement of heart structures (chambers, valves, etc.).

During the Doppler examination, the ultrasound beams will evaluate the flow of blood as it makes it way though and out of the heart. This information is presented visually on the monitor (as color images or gray scale tracings and also as a series of audible signals with a swishing or pulsating sound).

**Stress echocardiography:** or exercise echo, is an echocardiogram performed during exercise, when the heart muscle must work harder to supply blood to the body. This allows physicians to detect heart problems that might not be evident when the body is at rest and needs less blood. For patients who are unable to exercise, certain drugs can be used to mimic the effects of exercise by dilating the blood vessels and making the heart beat faster. During the examination the

sonographer can take measurements and, using the ultrasound scanner's computer, make calculations, including measuring blood flow speed.

Most ultrasound scanners are equipped with videotape recorders or digital imaging/archiving devices to record the real-time examination, and with medical image printers to print out hard copies of still images.

## Medical Problem Definition

After discussing the current issues related to heart disease diagnosis with the hospital’s Echo Department head, who is a consultant cardiologist the researcher was able to understand that interpretation of echo recordings remains a challenge since there are no available precise rules that are deduced from databases. The rules they are following now are gained through experience and professional learning.

There are limited number of specialists and a high number of patients, because of this the hospital is forced to use junior cardiologists which lead to inconsistencies and errors, currently there is only one specialist who does all the echocardiography examinations facing 400 up to 700 examinations per month in addition to his regular duties.

These inconsistencies and errors arise from being solely dependent on experience rather than linking experience with other useful tools like rules that can be generated from existing data. Thus, there is a need for a tool to assist in the diagnosis process that is based on rules that are generated from the data collected by the hospital.

### Defining Medical Goals

To solve the problems that exist in the current system, the following medical goals were set. The medical goals set for this study are:

* + - * To create a tool that could be used in heart disease diagnostic process by assisting cardiologists to interpret echo recordings.
      * To make the interpretation procedure easier, more consistent, and efficient based on the rules that are reformulated by the system.

### Defining Data Mining Goals

The data mining goals are translations of the medical goals, here the goals are set towards the technical part of the solution.

The main data mining goals are

* + - * Given patients’ echo result for each attribute, classify patients into two categories; those who are diagnosed with heart disease and those who are free from the disease.
      * To identify key features or patterns from the dataset.
      * To Identify and select attributes that are significant in relation to the predictable state – heart disease.

To achieve these data mining goals it was required to use different classification algorithms such as Neural Network, Decision Trees and Bayesian classifiers.

## Selecting and Creating the Target Dataset

After establishing data mining goals and the project plan, the next step was selecting and creating a target dataset which is suitable for the study. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one dataset, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data.

### Data Understanding

Data collected from patients include personal information, medical history, and laboratory test results. These data is usually generated and stored in the context of making a medical decision, for the purpose of augmenting choices about further testing and/or treatment and also for future access when needed.

#### 4.4.1.1 Selecting the Target Dataset

The first stage of this step is to select the related data from many available datasets to correctly describe the given medical task. The transthoracic echocardiography report of 7,708 patients from the October, 2008 to March, 2011 with a size of 300 Megabytes was selected as a target

dataset for this study and the hospital kindly provided the data. The report contains different measurements that are taken during the echocardiography examination, including information on 20 variables.

### Creating the Database

As the hospital keeps the record of each patient in a separate Microsoft Word file, in order to integrate the data it was needed to create a database with variables of interest and record the values of each variable into the new database.

As discussed above the original dataset contains 7,708 records, but some of the Microsoft Word files were unusable due to presence of errors that have been introduced unintentionally and again there were duplicated files on a single patient.

After removing these redundant and corrupted files by deleting from the original dataset the researcher was left with 7,339 files. Before creating the new database the researcher turned to a domain expert to select the most valuable attributes which should be included in the new database and the domain expert selected 15 of the most important variables and removed the remaining 5 because of ethical issues.

Finally, the new database was created using Microsoft Access 2010 and the values corresponding to the 15 selected attributes were recorded into the new database. After creating the new database, it now contains 7,339 instances each instance mapping a single file.

### Description of the Dataset

Each record in the dataset corresponds to a single patient’s exam results which are collected during the echocardiography examination. The information measured by the echo from a number of views around the patient is used to construct the report.

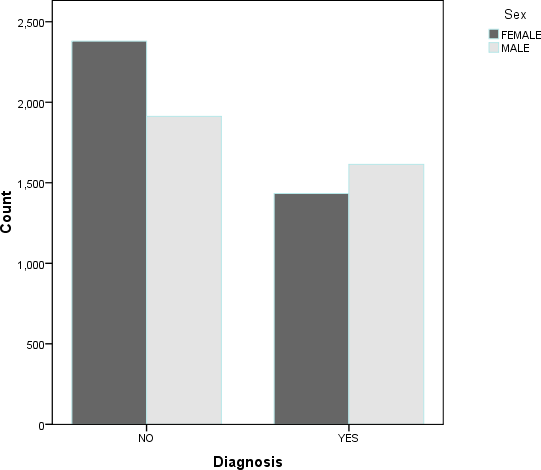
The variables in the dataset include Age, Sex, Aortic root – diameter, Left atrium: (sys) diameter, Left ventricle in: diastole, Left ventricle in systole, Posterior wall of LV, Interventricular septum, LV- ejection fraction, Main Pulmonary Artery diameter, Pericardial effusion, TR Velocity, Em/Am velocity ratio, Rhythm, and Diagnosis. The attributes and their description are presented in Table 4.1.

**Table 4.1: Attributes and their description**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Attribute** | **Description** | **Type** |
| 1 | Age | Age of the patient in years | Numeric |
| 2 | Sex | Sex of the patient (Male / Female) | Nominal |
| 3 | Aortic root – diameter | Size of Aortic root – diameter in mm | Numeric |
| 4 | Left atrium: (systole) diameter | Size of Left atrium: (sys) diameter in mm | Numeric |
| 5 | Left ventricle in: diastole | Left ventricle in: diastole in mm | Numeric |
| 6 | Left ventricle in systole | Size of Left ventricle in systole in mm | Numeric |
| 7 | Posterior wall of LV | Size of Posterior wall of LV in mm | Numeric |
| 8 | Interventricular septum | Size of in Interventricular septum in mm | Numeric |
| 9 | LV- ejection fraction | Fraction of blood pumped out of ventricles with each  heart beat in percentage | Numeric |
| 10 | Main Pulmonary Artery diameter | Size of Main Pulmonary Artery diameter in cm | Numeric |
| 11 | Pericardial effusion | Presence of an abnormal amount and/or character of fluid in the pericardial space | Ordinal |
| 12 | TR Velocity | Tricuspid Regurgitation Velocity in cm/sec | Numeric |
| 13 | Em/Am velocity ratio | The ratio between myocardial early and atrial peak  velocities | Numeric |
| 14 | Rhythm | Type of the heart rhythm observed | Nominal |
| 15 | Diagnosis | Does the patient has a heart disease (Yes or No) | Nominal |

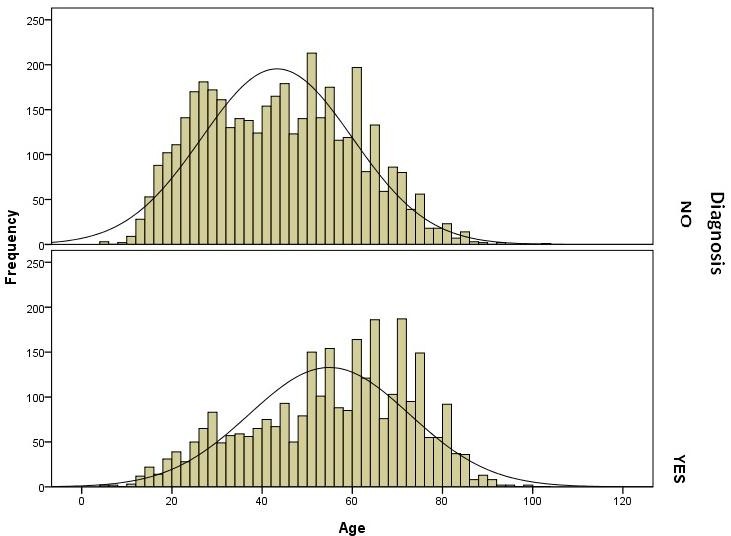
#### Descriptive Data Visualization

The dataset includes 3,813 (51.96%) females and 3,526 (48.04%) males age ranging from 4 years to 102 years. Out of the total 7,339 cases included in this study 3,047 (41.52%) were diagnosed with the disease while the remaining 4,292 (58.48%) were free from the disease.



**Figure 4.2: Distribution of heart disease among Gender**

Figure 4.2 shows that there is a gender gap related to diagnosis. Men who took the test are more likely to have the disease than women who took the test. Out of 3,526 males who were examined 1,614 (45.77%) were diagnosed with the disease while out of 3,813 females who were examined 1,433 (37.85%) were diagnosed with the disease.

The distribution of the disease is more over normal throughout different age groups (See Figure 4.3). However, from all the patients who took the test, patients whose ages were above 60 were slightly susceptible to the disease while patients whose age were less than 40 were free from the disease compared to those who are above 60.

**Figure 4.3: Distribution of Heart Disease among different age groups**

To summarize attributes with numerical data type measures of centrality and measure of spread were done. To measure centrality, mean (arithmetic average) was used and to measure how the data is spread standard deviation and range (minimum value and maximum value) were used. The result is presented in Table 4.2.

From all the measures of dispersion listed on Table 4.2 standard deviation has a more statistical significance, thus standard deviation is used as a major tool for measure of variation. The age of the patient’s included in the dataset has a standard deviation value of 18.128 suggesting there is fairly unique values for age and each patient’s age is not too close to the average while the

standard deviation of Main Pulmonary Artery diameter which is 0.62 suggests every patient is pretty close to average.

**Table 4.2: Numerical Attributes Data Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **Attribute** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| 1 | Age | 48.184 | 18.128 | 4 | 102 |
| 2 | Aortic root – diameter | 28.833 | 4.073 | 15 | 76 |
| 3 | Left atrium: (systole) diameter | 36.234 | 7.667 | 10 | 87 |
| 4 | Left ventricle in: diastole | 47.489 | 6.91 | 20 | 83 |
| 5 | Left ventricle in systole | 30.819 | 6.738 | 12 | 71 |
| 6 | Posterior wall of LV | 10.095 | 1.709 | 6 | 30 |
| 7 | Interventricular septum | 9.952 | 1.795 | 4 | 25 |
| 8 | LV- ejection fraction | 53.091 | 8.588 | 10 | 74 |
| 9 | Main Pulmonary Artery diameter | 1.874 | 0.62 | 1 | 50 |
| 10 | TR Velocity | 1.665 | 0.959 | 0 | 7 |

# CHAPTER FIVE EXPERIMENTATION

As the goal of this study is to detect heart disease using data mining techniques a classification technique was adopted to develop a predictive model. The models were built with three different supervised machine learning algorithms i.e. Decision Tree Classification Algorithm, Bayesian Classifier and Neural Network using Weka 3.6.4 machine learning software.

## Experimental Setup

Four experiments were conducted for this study and for all experiments two scenarios were considered, one containing all the 15 attributes and the other containing 8 selected attributes. With four experiments and eight different scenarios a total of eight models were developed.

The experiments were conducted on a full training dataset containing 7,339 instances and 10- Fold Cross Validation was adopted for randomly sampling the training and test sets. While performing the experiments all parameters were set to their default setting for each algorithm except for J48 classifier where the parameter “Unpruned” which had a default value “False” was changed to “True” for the first experiment to observe the performance of J48 unpruned tree.

The performances of the models in this study were evaluated using the standard metrics of accuracy, precision, recall and F-measure which were calculated using the predictive classification table, known as Confusion Matrix. ROC area was also used to compare the performances of the classifiers.

## Model Building Using J48 Decision Tree

Two experiments were conducted using J48 decision tree classifier. These two experiments were designed to investigate:

* + - The effect of attribute selection on classification accuracy as well as model complexity on both unpruned and pruned J48 Decision Tree Classifiers.
    - The effect of tree pruning methods on classification accuracy, Decision Tree size and model complexity when building a J48 decision tree model.

#### Experiment 1

The first experiment was designed to evaluate the performance of a J48 classifier Unpruned tree in predicting heart disease and to investigate the effect of attribute selection on the performance of the model. In this experiment two scenarios were considered, one containing all 15 attributes and the other containing the selected 8 attributes.

On the first scenario the algorithm was run on a full training set containing 7,339 instances with 15 attributes. It took 0.89 second to build the model and the model generated a tree with a size of 473 and 323 leaves.

On the second scenario the algorithm was run on a full training set containing 7,339 instances with selected 8 attributes. It took 0.36 second to build the model and the model generated smaller and less complex tree with a size of 126 and 71 leaves making it less complex and faster than the experiment conducted on all attributes.

**Table 6.1: Confusion Matrixes for Experiment 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Confusion Matrix** | | |
| **J48 unpruned with all attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,841 | 206 | **Yes** |
| 213 | 4,079 | **No** |
| **J48 unpruned with Selected attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,875 | 172 | **Yes** |
| 157 | 4,135 | **No** |

**Table 6.2 Detailed Performance Measures for Experiment 1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **TP Rate** | **TN Rate** | **Precision** | **F-Measure** | **ROC Area** |
| **J48 unpruned with all attributes** | 94.29% | 0.932 | 0.95 | 0.943 | 0.943 | 0.942 |
| **J48 unpruned with Selected attributes** | 95.52% | 0.944 | 0.963 | 0.955 | 0.955 | 0.965 |

The model built with J48 unpruned tree with all attributes correctly classified (predicted the correct outcome) 6,920 (94.29%) instances while 419 (5.71%) of the instances were classified incorrectly. The overall accuracy rate of the model is highly successful, but it is worth to consider the TP Rate (Sensitivity), that is, patients which have heart disease and that are correctly identified and TN Rate (Specificity), that is, patients which do not have heart disease and that are correctly identified, to see the performance of the model for each class.

The model correctly identified 2,841 patients out of 3,047 patients who had heart disease and the remaining 206 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.932. The model is better in identifying negative cases as the TN Rate of the model is 0.95 by correctly identifying 4,079 patients out of 4,292 patients who didn’t had heart disease and the remaining 213 were identified to have the disease while they actually didn’t.

Regarding to Precision score of the model, 93% of patients labeled as belonging to class Yes does indeed belong to class Yes while 95.2% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 94.3% it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.943 it can be concluded that the Precision and the Recall of the model are significantly balanced.

The second model built with J48 unpruned tree with selected 8 attributes correctly classified 7,010 (95.52%) instances while 337 (4.48%) of the instances were classified incorrectly. Like the experiment the overall accuracy rate of the model is highly successful, in fact it is better compared to J48 unpruned tree implemented on all attributes.

The model correctly identified 2,872 patients out of 3047 patients who had heart disease and the remaining 172 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.944. The model is better in identifying negative cases as the TN Rate of the model is 0.963 by correctly identifying 4,135 patients out of 4,292 patients who didn’t had heart disease and the remaining 157 were identified to have the disease while they actually didn’t.

Regarding to Precision score of the model, 94.8% of patients labeled as belonging to class Yes does indeed belong to class Yes while 96% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 95.5% it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.955 it can be concluded that the Precision and the Recall of the model are significantly balanced.

The results of this experiment indicated that a J48 unpruned decision tree algorithm is highly capable in predicting heart disease cases. Furthermore, the results showed the impact of attribute selection on classification accuracy, Decision tree size and model complexity.

**Experiment 2**

The second experiment was designed to investigate:

* + - The performance of a J48 classifier pruned tree in predicting heart disease
    - The effect of attribute selection on the performance of a J48 classifier pruned tree model.
    - The effect of tree pruning methods when building a J48 decision tree model,

Like Experiment 1 in this experiment two scenarios were considered, one containing all 15 attributes and the other containing the selected 8 attributes. On the first scenario the algorithm was run on a full training set containing 7,339 instances with 15 attributes. It took 1.05 second to build the model and the model generated smaller and less complex tree with a size of 104 and 63 leaves.

On the second scenario the algorithm was run on a full training set containing 7,339 instances with only 8 selected attributes. It took 0.41 second to build the model and the model generated smaller and less complex tree with a size of 93 and 52 leaves making it the least complex model built from J48 classifier.

**Table 6.3: Confusion Matrixes for Experiment 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Confusion Matrix** | | |
| **J48 pruned with all attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,872 | 175 | **Yes** |
| 162 | 4,130 | **No** |
| **J48 pruned with selected attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,892 | 155 | **Yes** |
| 171 | 4,121 | **No** |

**Table 6.4 Detailed Performance Measures for Experiment 2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **TP Rate** | **TN Rate** | **Precision** | **F-Measure** | **ROC Area** |
| **J48 pruned with all attributes** | 95.41% | 0.943 | 0.962 | 0.954 | 0.954 | 0.964 |
| **J48 pruned with selected attributes** | 95.56% | 0.949 | 0.96 | 0.956 | 0.955 | 0.965 |

The predictive model built from a J48 classifier pruned tree with all attributes correctly classified 7,002 (95.41%) instances while 337 (4.59%) of the instances were classified incorrectly. Like the experiment 1 the overall accuracy rate of the model is highly successful, in fact it is better compared to J48 unpruned tree.

The model correctly identified 2,872 patients out of 3047 patients who had heart disease and the remaining 175 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.943. The model is better in identifying negative cases as the TN Rate of the model is 0.962 by correctly identifying 4,130 patients out of 4,292 patients who didn’t had heart disease and the remaining 162 were identified to have the disease while they actually didn’t.

Regarding to Precision score of the model, 94.7% of patients labeled as belonging to class Yes does indeed belong to class Yes while 95.9% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 95.4% it is a very successful model in retrieving relevant values for each class. F-Measure value of 0.954 suggests that the Precision and the Recall of the model are significantly balanced.

The predictive model built from a J48 classifier pruned tree with 8 selected attributes correctly classified 7,013 (95.56%) instances while 326 (4.44%) of the instances were classified incorrectly. The model scored a highly successful overall accuracy rate.

The model correctly identified 2,892 patients out of 3047 patients who had heart disease and the remaining 155 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.949. The model is better in identifying negative cases as the TN Rate of the model is 0.96 by correctly identifying 4,121 patients out of 4,292 patients who didn’t had heart disease and the remaining 171 were identified to have the disease while they actually didn’t.

Regarding to Precision score of the model, 94.4% of patients labeled as belonging to class Yes does indeed belong to class Yes while 96.4% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 95.6% it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.956 it can be concluded that the Precision and the Recall of the model are significantly balanced.

## Model Building Using Naïve Bayes Classifier

The third experiment was designed to evaluate the performance of Naïve Bayes Classifier in predicting heart disease. In this experiment two scenarios were considered, one containing all 15

attributes and the other containing the selected 8 attributes. The intention here is to investigate the effect of attribute selection on the performance of the models.

#### Experiment 3

On the first scenario the algorithm was run on a full training set containing 7,339 instances with 15 attributes and the execution time of the model is 0.11 second. On the second scenario the algorithm was run on a full training set containing 7,339 instances with selected 8 attributes and the execution time of the model is 0.05 second.

**Table 6.5: Confusion Matrixes for Experiment 3**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Confusion Matrix** | | |
| **Naïve Bayes with all attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,635 | 412 | **Yes** |
| 178 | 4,114 | **N0** |
| **Naïve Bayes with selected attributes** | **Yes (Predicted)** | **No (Predicted)** | **Actual** |
| 2,654 | 393 | **Yes** |
| 163 | 4,129 | **No** |

**Table 6.6 Detailed Performance Measures for Experiment 3**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **TP Rate** | **TN Rate** | **Precision** | **F-Measure** | **ROC Area** |
| **Naïve Bayes with all attributes** | 91.96% | 0.865 | 0.959 | 0.92 | 0.919 | 0.97 |
| **Naïve Bayes with selected attributes** | 92.42% | 0.871 | 0.962 | 0.925 | 0.924 | 0.972 |

The first Naïve Bayes model built on all 15 attributes correctly classified 6,749 (91.96%) instances while 590 (8.04%) of the instances were classified incorrectly. This has a moderately successful overall accuracy rate. The model correctly identified 2,635 patients out of 3047 patients who had heart disease and the remaining 412 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.943. The model is better in identifying negative cases as the TN Rate of the model is 0.962 by correctly identifying 4,130 patients out of 4,292 patients who didn’t had heart disease and the remaining 178 were identified to have the disease while they actually didn’t.

About the Precision score of the model, 93.7% of patients labeled as belonging to class Yes does indeed belong to class Yes while 90.9% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 92% it is a very successful model in retrieving

relevant values for each class. With F-Measure value of 0.899 it can be concluded that the Precision and the Recall of the model are significantly balanced.

The second Naïve Bayes model built on 8 selected attributes correctly classified 6,783 (92.42%) instances while 556 (7.58%) of the instances were classified incorrectly. The overall classification accuracy of this model is better than the similar experiment conducted on all attributes, but still it’s moderately successful. The model correctly identified 2,654 patients out of 3047 patients who had heart disease and the remaining 343 were identified incorrectly to be free from the disease while they actually had the disease. This result gave the model a TP Rate of 0.871. The model is better in identifying negative cases as the TN Rate of the model is 0.962 by correctly identifying 4,129 patients out of 4,292 patients who didn’t had heart disease and the remaining 163 were identified to have the disease while they actually didn’t.

Regarding to Precision score of the model, 94.2% of patients labeled as belonging to class Yes does indeed belong to class Yes while 91.3% of patients labeled as belonging to class No does indeed belong to class No. With an average precision of 92.5% it is a very successful model in retrieving relevant values for each class. With F-Measure value of 0.924 it can be concluded that the Precision and the Recall of the model are significantly balanced.

Here, Naïve Bayes model performed better on the selected attributes. The classification accuracy increased to 92.42% from 91.96%. And also, the execution time decreased by half compared to the model built on all 15 attributes.

# CHAPTER SEVEN CONCLUSION

## 

In this study, the aim was to design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography.

Data collected by International Cardiovascular Hospital from the year 2008 to 2011 containing 7,339 instances was selected and preprocessed for this study. The models were built on the preprocessed Transthoracic Echocardiography dataset with three different supervised machine learning algorithms i.e. J48 Classifier, Naïve Bayes and Multilayer Perception using Weka 3.6.4 machine learning software.

The performances of the models were evaluated using the standard metrics of accuracy, precision, recall and F-measure. 10-Fold Cross Validation was adopted for randomly sampling the training and test data samples. All eight models performed well in predicting heart disease cases. The most effective model to predict patients with heart disease appears to be a J48 classifier implemented on selected attributes with a classification accuracy of 95.56%.

Three data mining goals were defined based on the medical problems. The goals were evaluated against the selected model and the selected model built with J48 Decision Tree Algorithm successfully met all the three data mining goals.

Significant rules that are useful for predicting the presence of heart disease were extracted from the dataset. The domain expert confirmed that most of the rules generated are important in interpretation of echocardiography examinations.

From a total of 15 attributes that were available, 8 attributes that are highly relevant in predicting heart disease from Transthoracic Echocardiography dataset were selected.

Heart disease is a fatal disease by its nature and misdiagnosis of this disease can cause serious, even life threatening complications such as cardiac arrest and death. The best model selected for

predicting heart disease could not exceed a classification accuracy of 95.56% and still much remains to fill the gap of 4.44% misclassified cases.

This study showed that data mining techniques can be used efficiently to model and predict heart disease cases. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis.

# REFERENCES

Benjamin, M. (2006). *Introduction to Heart Disease*. Available at [http://www.mentalhelp.net/poc/view\_doc.php?type=doc&id=4496](http://www.mentalhelp.net/poc/view_doc.php?type=doc&amp;id=4496) (Accessed 10 January 2011)

Benko, A. and Wilson, B. (2003). *Online Decision Support Gives Plans An Edge*. Managed Healthcare Executive

Berry J.A. Michael and Linoff S. Gordon (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. Second Edition. Wiley Publishing, Inc., Indianapolis, Indiana

Biafore, S. (1999). *Predictive Solutions Bring More Power To Decision Makers.* Health Management Technology, 20(10): 12-14.

Braunwald, E., Douglas, P. Zipes, Peter, L., Robert, B. (1988). *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Third Edition, Harcourt Brace Jovanovich Inc., New York

Carlos, O. (2004), *Improving Heart Disease Prediction Using Constrained Association Rules*, Seminar Presentation at University of Tokyo

Chakrabarti, S., Earl, C., Frank, E., Ralf Hartmut, G., Han, J., Xia Jiang, Kamber, M., Sam S. Lightstone and others (2009).*Data Mining Know It All*. Morgan Kaufmann Publishers, Elsevier Inc., Burlington

Chen, M.S., Han, J., and Yu, P. S. (1997). *Data Mining: An Overview from Database Perspective*. National Taiwan University, Taipei, Taiwan.

Cindy H. (2008). *5 Common Types of Heart Disease*. EzineArticles.com. Available at <http://ezinearticles.com/?5->•Common-•Types-•of-•Heart-•Disease&id=1073496 (March 03

January 2011)

Connolly, T., Begg, C. and Strachan, A. (1999). *Database Systems: A Practical Approach to Design, Implementation and Management.* Second Edition. Addison-Wesley, New York. David L. Olson and Dursun, D. (2008). *Advanced Data Mining Techniques*. Springer-Verlag

Berlin Heidelberg

Dunham, M.H. (2003). *Data Mining Introductory and Advanced Topics*. Pearson Education, Inc., Upper Saddle River, New Jersey

# 