Prototype Development of a Novel Heart Disease Risk Evaluation Tool Using Data Mining Analysis

Mai Mohammed Abbas Shouman

B.Sc. in Information Systems and Technology

M.Sc. in Information Systems

Faculty of Computer Science and Informatics, Zagazig University, Egypt

UNSW
THE UNIVERSITY OF NEW SOUTH WALES

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (Computer Science) from the School of Engineering and Information Technology (SEIT), the University of New South Wales (UNSW) @ Canberra.

March 2014

[This Page is Left Blank Intentionally]

Abstract

In the last decade, heart disease has been the leading cause of death all over the world. However, it is among the most preventable and controllable diseases. The World Health Organization reported that early detection of heart disease reduces progression to severe and costly illness and complications. Early detection of heart disease patients helps in recovering the patients' health and decreasing the mortality rate from heart disease.

Although heart disease can be detected by several tests, such as electrocardiogram, stress tests, and cardiac angiogram, these tests are expensive and cannot be used as community-level screening tests. The Framingham Heart Disease Risk Evaluation Tool and the Australian Absolute Cardiovascular Risk Calculator are two common heart disease risk evaluation screening tests. However, both tests need prior blood sample investigations, an invasive and relatively costly process, which reduces their usability in other than medical settings.

Motivated by the increasing mortality rates of heart disease patients, researchers have been applying different data mining techniques in the diagnosis of heart disease. Research finds that the same data mining technique shows different results across different heart disease datasets indicating that there can be significant attributes for heart disease diagnosis. Furthermore, researchers suggest that hybrid data mining techniques show better performance in the diagnosis of heart disease patients.

This research seeks to help healthcare professionals in the early detection and risk evaluation of heart disease patients using data mining analysis. To achieve this objective, the significant attributes in the diagnosis of heart disease patients are identified using a benchmark dataset and a new larger dataset, the reliability of non-invasive attributes in the diagnosis of heart disease is investigated, the enhancement of applying hybrid data mining model to the non-invasive attributes is tested, and a heart disease expert system risk evaluation tool (HD - ESRET) using hybrid data mining model on non-invasive data attributes is constructed.

Although this research builds a low-cost heart disease expert system risk evaluation tool using a novel non-invasive data attributes combination, its usability testing among healthcare providers still needs further investigation.

[This Page is Left Blank Intentionally]

Keywords

Heart Disease Diagnosis, Data Mining, Decision Tree, Naïve Bayes, K-nearest Neighbour, Non-Invasive Attributes, Heart Disease Expert System Risk Evaluation Tool

[This Page is Left Blank Intentionally]

Acknowledgements

I would like to express my appreciation to all the individuals without whom the completion of this thesis would not be possible. First of all, my special thanks to my supervisor Dr. Tim Turner, for generously providing guidance on the technical aspects of this thesis, for continuously encouraging me and pushing me to my limit to complete my thesis, and for all of the patience and support he gave me and accepting me as a student. My appreciation likewise extends to my co-supervisor Dr. Rob Stocker, for his useful discussions, suggestions, comments and his valuable assistance.

I would like to express my grateful thanks to Professor Leonard Arnolda, director of cardiology department, Canberra Hospital, for his continuous help in getting access to Canberra Hospital data, helpful discussions and continuous support. My grateful thanks extend to Associate Professor Greg Kyle, head of pharmacy discipline, Faculty of Health, Canberra University, for his helpful discussions and valuable conversations in identifying the importance of non-invasive attributes in the risk evaluation of heart disease. I would like to express my thanks to Denise Russel for proof reading my thesis, providing me with useful comments in completing my thesis, and continuous help and support.

To my ever supportive partner in life, my lovely husband (Man3oomy), thank you for helping me in my life, patience and support. Thank you for your love, continuous understanding, endless motivation and rewarding, as well as your believing in me that I can finish my PhD thesis.

I would like to express my continuous love to my grandfather (Mahmoud) for his love and continuous motivation while I was a little girl. Although he left our life in 2005, I can still hear his words in my ear that I will be a doctor one day. I still remember when I was in my high school and you allowed me to study in your office using your special desk. Words cannot describe how your motivations pushed me to my top. I cannot find the words to express my deepest love and missing to you.

My deepest gratitude goes to my lovely mother, father, mother in law, and father in law for their continuous motivation, prayers for me and endless support. My grateful thanks extend to my lovely sisters (Mahytab, Maram and Menna), brothers in law (Wael, Moustafa and Sameh) and their families for their unflagging love and support throughout my life.

Individual acknowledgement to my friends for sharing their joyful moments with me and their moral support; Heba Zaki, Eman Samir, Noha Hamza, Hafsa Magdy, Yasmin Abdelraouf, Mayada Tharwat, Nour-el-Hoda Mohammed, Eman Freekh, Mona Said, Heba Nassar and their families.

Last but not least, I would like to thank all the staff of the University of New South Wales (UNSW) Canberra for providing me with a scholarship to study at this institution.

Table of Contents

List of Tables

[This Page is Left Blank Intentionally]

List of Figures

List of Equations

[This Page is Left Blank Intentionally]

Chapter 1

Introduction

1.1. Background

Heart disease has been the leading cause of death in the world over the past decade on different continents and countries regardless of their income (World Health Organization 2011b). The World Health Organization (WHO) reported that heart disease is the leading cause of death worldwide, causing 7.25 million death cases representing 12.8% of all deaths (World Health Organization 2013c). Different health organizations reported that heart and respiratory diseases are the main cause of death in different continents (Statistics South Africa 2008, ESCAP 2010, Australian Bureau of Statistics 2013, European Public Health Alliance 2013). Different countries have also reported that heart disease is a leading cause of death showing high mortality rates (World Bank Disease Control Priorities Project 2013, World Health Organization 2013b, Center for Disease Control and Prevention 2014).

Although heart diseases are among the most common chronic diseases causing a high rate of death all over the world, they have also been identified as among the most preventable and controllable diseases (Centers for Disease Control and Prevention 2013). Early detection and healthy behaviour are two critical issues in controlling heart disease. Early detection of heart disease patients can help in recovering patients' health and decreasing the mortality rate from heart disease (Centers for Disease Control and Prevention 2013). WHO reported that early detection and treatment reduce progression to severe and costly illness and complications of heart disease (World Health Organization 2010). Furthermore, the relative success of chronic disease treatments are dependent on the earliness of detection of those diseases (Paladugu and Shyu 2010). There is a vital need for accurate systematic tools that identify patients at high risk and provide information for early detection of heart disease (Paladugu and Shyu 2010).

Modern computer systems are gathering huge amounts of data every day through automatic recording systems in various sectors. Data mining technology appeared as a result of the evolution of database technology, information technology and storage devices (Obenshain 2004). Although data mining has been around for more

than two decades, its potential is only being realized recently. Data mining researchers have long been concerned with the application of tools to facilitate and improve data analysis on large and complex data sets. The current challenge is to make data mining and knowledge discovery systems applicable to a wider range of domains (Shillabeer and Roddick 2006). Data mining applications in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes (Porter and Green 2009), stroke (Panzarasa, Quaglini et al. 2010), cancer (Li L 2004), and heart disease (Das, Turkoglu et al. 2009). Motivated by the increasing mortality rates of heart disease, researchers have been using data mining techniques to help healthcare professionals in the diagnosis of heart disease patients (Parthiban and Subramanian 2007, Polat , Sahan et al. 2007, Tu, Shin et al. 2009, Rajkumar and Reena 2010, Lakshmi, Krishna et al. 2013).

1.2. Problem Description

Although heart disease is the leading cause of death all over the world (World Health Organization 2011b), early detection of heart disease patients can help in recovering patients' health and decreasing the mortality rate from heart disease (Centers for Disease Control and Prevention 2013). So there is a vital need for accurate and systematic tools that provide information for early detection of heart disease to identify those patients at high risk (Paladugu and Shyu 2010).

Community-level screening tests play an especially important role in the early detection of heart disease (Kotnik 2010). These community-level screening tests can be applied in pharmacies or public health clinics where non-medical healthcare professionals can screen the community at large for potential heart disease sufferers and refer those potential sufferers to more thorough screening tests. It can also be applied where there is limited availability of resources such as electrocardiogram, stress tests, and cardiac angiogram machines needed for the diagnosis of heart disease. Recent research focuses on discovering new specific, sensitive and cheap community-level screening tests (Kotnik 2010).

There are two famous heart disease risk evaluation screening tests commonly used in heart disease diagnosis: the Framingham Heart Disease Risk Evaluation Tool

(Framingham Heart Study 2013) and the Australian Absolute Cardiovascular Risk Calculator (National Heart Foundation of Australia 2009). Both of these two heart disease risk evaluation tests use a set of attributes such as age, sex, systolic blood pressure, total cholesterol, diabetes and smoking status to identify if a patient is at high, moderate or low risk of heart disease (National Heart Foundation of Australia 2009, Framingham Study. 2013). Although these two tests help in identifying patients at risk of heart disease, they need prior blood based investigation to identify the cholesterol and diabetes levels. Such tests are both invasive, requiring physical samples to be taken, and relatively expensive. The tests also pre-suppose medical facilities in which they can be undertaken. Early detection in a community setting needs lower cost tests that can be used by non-specialist healthcare advisers that is still reliable.

Hence, there is a need to simplify the heart disease risk evaluation tool attributes so that affordable detection strategies can be implemented (Bitton and Gaziano 2010). There is a need to find less costly tests and accurate systematic tools that can be used for community-level screening to identify patients at high risk of heart disease and provide information to enable early intervention (Paladugu and Shyu 2010).

Researchers have been using several data mining techniques to help healthcare professionals in the diagnosis of heart disease patients. Researchers have been comparing different data mining techniques' performance across different datasets to identify which data mining technique will provide better results in heart disease diagnosis (Palaniappan and Awang 2007, Tu, Shin et al. 2009, Rajkumar and Reena 2010, Abdullah and Rajalaxmi 2012). However, the results obtained on different datasets cannot easily be compared as different datasets have different data attributes. Also, the same data mining technique shows different results across different heart disease datasets. This shows that there are some data attributes that affect data mining techniques' performance in the diagnosis of heart disease patients. Although it is widely accepted that non-invasive attributes such as age, sex, blood pressure, smoking, and diabetes are major risk factors for developing heart disease, there are no previous investigations into the significance of applying data mining techniques in the diagnosis of heart disease patients using those non-invasive heart disease data attributes.

Recently researchers are suggesting that integrating more than one data mining technique in a hybrid model can enhance data mining techniques performance in the

diagnosis of heart disease patients (Parthiban and Subramanian 2007, Das, Turkoglu et al. 2009, Ozsen and Gunes 2009 , Soni, Ansari et al. 2011). However, the significance of the hybrid models on different data attribute combinations and their influence in the diagnosis of heart disease patients needs further investigation.

Although applying different data mining techniques to identify the significant one in the diagnosis of heart disease is critical, the ability to use the significant data mining technique in a community-level screening is of great benefit to the early detection of heart disease patients. Finding the most accurate data mining technique in the diagnosis of heart disease patients is helpful if the results can be used in a community-level screening test. So there is a need for data mining techniques analysis to build a data mining tool that can help healthcare professionals in the community-level screening of heart disease patients.

1.3. Motivation and Objectives

Motivated by the increasing mortality rates of heart disease all over the world and the fact that early detection helps in recovering patients' health and decreasing the mortality rate from heart disease, the main objective of this thesis is helping healthcare professionals in the early detection and risk evaluation of heart disease patients. To achieve this objective, this research poses the question:

Can data mining assist healthcare professionals in the early detection of heart disease in a community setting?

Although researchers have been applying different data mining techniques to help healthcare professionals in the diagnosis of heart disease patients, there is not a clear view of different data mining techniques' performance across different data attribute combinations in the diagnosis of heart disease patients. The main objective of this research in answering the above question is:

To provide healthcare professionals with a community-level screening tool for the early detection and risk evaluation of heart disease patients

To achieve this, the research is focussed on the following key questions:

1. Can significant attributes in the diagnosis of heart disease patients be identified?

2. Can applying data mining techniques on non-invasive attributes be usefully applied to the diagnosis of heart disease patients?
3. Can hybrid data mining techniques be usefully applied to enhance performance on non-invasive data attributes in the diagnosis of heart disease patients?
4. Can a reliable heart disease expert system risk evaluation tool, using non-invasive heart disease data attributes, be constructed?

1.4. Contribution to the Scientific Knowledge

The main contribution of this thesis is helping healthcare professionals in the risk evaluation of heart disease patients using data mining analysis, through:

Building a low-cost heart disease expert system risk evaluation tool using a novel non-invasive data attributes combination

This main contribution involves a set of sub-contributions including:

- **Identifying the significant attributes in the diagnosis of heart disease patients:** The significant attributes in the diagnosis of heart disease patients are identified by applying different data mining techniques over two different heart disease datasets (the benchmark Cleveland dataset and a larger Canberra dataset). The performance of different data mining techniques is tested over different attribute combinations to identify which attribute combination will provide reliable performance in the diagnosis of heart disease patients. It will also identify which data mining technique will provide stable performance across different attributes combinations.
- **Identifying the significance of non-invasive data attributes in the diagnosis of heart disease patients:** The significance of different single, combined and calculated non-invasive attribute combinations is tested across the Cleveland and Canberra heart disease datasets. The non-invasive attributes main importance is that they are low cost attributes; i.e. it costs very little or nothing to determine a value for that attribute when diagnosing a patient. The performance of applying Decision Tree data mining technique to identify which non-invasive attribute combination will show the best accuracy in the diagnosis of heart disease patients is investigated.

- **Investigating applying hybrid data mining model in the diagnosis of heart disease patients:** The significance of integrating K-Means clustering with different initial centroid selection methods with the Decision Tree data mining technique in the diagnosis of heart disease patients is investigated. The hybrid model is applied to both investigated datasets using all attributes and only non-invasive attribute combinations to identify which data mining technique combination will provide better performance in the diagnosis of heart disease patients.
- **Building a heart disease expert system risk evaluation tool using hybrid data mining model on non-invasive data attributes:** A heart disease expert system risk evaluation tool is built that uses the rules from a two cluster Outlier K-Means clustering Decision Tree applied to non-invasive data attributes to help healthcare professionals in the screening of heart disease patients. The expert system can act as a community-level screening test to identify if a patient is at high or low risk of heart disease.

1.5. Thesis Organization

The thesis is divided into seven chapters. This introductory chapter presents the background, problem description, motivation and objectives of this research, the contribution to the scientific knowledge, organization of this thesis as well as the publications resulting from this thesis. The remaining chapters are:

- Chapter 2: Technical Background and Literature Review
- Chapter 3: Applying Data Mining Techniques in Heart Disease Diagnosis
- Chapter 4: Non-Invasive Attributes Significance in Heart Disease Risk Evaluation
- Chapter 5: Integrating Clustering With Decision Tree in Heart Disease Diagnosis
- Chapter 6: Heart Disease Expert System Risk Evaluation Tool
- Chapter 7: Conclusions and Future Work

- Chapter 2 introduces the basic concepts of the topics related to this research. It starts with an overview of heart disease mortality rates followed by heart disease detection and prevention as well as the heart disease risk evaluation tools. Data mining as a step in knowledge discovery is discussed followed by an

explanation of data mining techniques and data mining technique performance evaluation. Afterwards, data mining applications in healthcare are discussed. Finally, applying data mining techniques in the diagnosis of heart disease patients and especially using the Cleveland benchmark dataset is reviewed.

- Chapter 3 presents an investigation to identify the significant attributes needed by data mining techniques in the diagnosis of heart disease patients. The investigation applies different data mining techniques over two different heart disease datasets i.e. Cleveland and Canberra to test if they provide reliable results over the two datasets. Then the investigation applies different data mining techniques over the different data attributes combinations of the two heart disease datasets to identify their significance in the diagnosis of heart disease patients.
- Chapter 4 presents an investigation to identify the significance of non-invasive attributes in the diagnosis of heart disease patients. The investigation applies the Decision Tree data mining technique to identify which non-invasive attributes combination will show the best performance in the diagnosis of heart disease patients. The investigation considers the effect of using single, combined and calculated non-invasive data attributes on both the Cleveland and Canberra datasets in the diagnosis of heart disease patients.
- Chapter 5 presents an investigation that applies hybrid data mining techniques to identify if the hybrid model will show better performance in the diagnosis of heart disease patients. It applies K-Means clustering as one of the most popular and well-known clustering techniques with different initial centroid selection methods and Decision Tree over the Cleveland and Canberra datasets using all attributes and non-invasive attribute combinations in the diagnosis of heart disease patients.
- Chapter 6 describes the building of a heart disease expert system risk evaluation tool using the results from two cluster Outlier K-Means clustering Decision Tree data mining analysis to help healthcare professionals in the diagnosis of heart disease patients. The heart disease expert system risk evaluation tool identifies the degree of risk of heart disease using non-invasive data attributes. This expert system risk evaluation tool can act as a community-level screening test that can identify the risk of heart disease as being high or low.

- Chapter 7 concludes the main findings of this thesis. It presents the research objectives, summarizes the research conclusions followed by discussion of the research limitations and future directions. Finally, the main contributions of this thesis are presented.

1.6. Publications Resulting from this Thesis

This section lists the publications resulting from this thesis. Several of the published papers have been cited since their publication.

- **Refereed journal papers:**

1. Mai Shouman, Tim Turner, Rob Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", at the International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.
2. Mai Shouman, Tim Turner, Rob Stocker," Integrating Clustering with Different Data Mining Techniques in the Diagnosis of Heart Disease", at the Journal of Computer Science and Engineering, Volume 20, Issue 1, August 2013.

- **Refereed conference papers:**

3. Mai Shouman, Tim Turner, Rob Stocker, "Integrating Naïve Bayes and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", at the International Conference of Data Mining & Knowledge Management Process (CKDP) Dubai, United Arab Emirates, December 2012.
4. Mai Shouman, Tim Turner, Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", at the International Conference on Data Mining (DMIN'12), LasVegas, USA, July 2012.
5. Mai Shouman, Tim Turner, Rob Stocker, "Using Data Mining Techniques In Heart Disease Diagnosis And Treatment" at the IEEE Japan-Egypt Conference on Electronics, Communications and Computers, Alexandria, Egypt, March 2012.
6. Mai Shouman, Tim Turner, Rob Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients", at the Ninth Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, December 2011.

- **Refereed posters:**

7. Mai Shouman, Tim Turner, Rob Stocker, "Applying data mining techniques in diagnosing heart disease patients" poster presentation at the Canberra Health Annual Research Meeting, Canberra, Australia, August 2012.

- **In-preparation journal papers:**

8. Mai Shouman, Tim Turner, Rob Stocker, "Investigating Non-Invasive Attributes Significance in the Risk Evaluation of Heart Disease Using Data Mining Techniques", To be submitted to Artificial Intelligence in Medicine in May 2014.
9. Mai Shouman, Tim Turner, Rob Stocker, "Building an Expert System for Heart Disease Risk Evaluation using Novel Non-Invasive attributes", To be submitted to Expert Systems With Applications in June 2014.

The following lists materials (or part) of the publications presented in the thesis:

- Chapter 2: publication [5]
- Chapter 3: publications [1,6,7]
- Chapter 4: publication [8]
- Chapter 5: publications [2,3,4]
- Chapter 6: publication [9]

[This Page is Left Blank Intentionally]

Chapter 2

Technical Background and Literature Review

2.1. Introduction

Over the last decade heart disease has been the leading cause of death in the world. Heart disease is a critical death cause in all low, middle, and high-income countries. The heart disease mortality cases are distributed almost equally for men and women (World Health Organization 2011a). The trends of heart disease mortality rates over ten years across twenty one different countries show that heart disease remains the main cause of death through the period (Tunstall-Pedoe, Kuulasmaa et al. 1999). It is obvious that this health challenge has significant impact in terms infrastructure and finance. A clear imperative is for early identification of potential victims to develop better health outcomes for people and governments.

Modern computer systems gather huge amounts of data every day through automatic recording systems in the health sector from which data mining can extract useful knowledge. This chapter provides an overview of heart disease, data mining, and current application of data mining techniques for heart disease diagnosis. This chapter provides an overview of heart disease, including its mortality rates, its prevention and detection, and its risk evaluation. Afterwards it describes data mining as a step in knowledge discovery, supervised and unsupervised data mining tasks, data mining techniques, and data mining techniques performance evaluation. Data mining applications in healthcare are discussed with a review of the application of data mining in heart disease diagnosis, and the use of single and hybrid data mining techniques in the Cleveland heart disease diagnosis dataset. Finally the chapter summary and conclusion is presented

2.2. Heart Disease Overview

Section 2.2 overviews heart disease mortality rates, understanding heart disease, heart disease prevention and detection, and heart disease risk evaluation tools.

2.2.1. Heart Disease Mortality Rates

As the leading cause of death in the world (World Health Organization 2011a), the World Health Organization (WHO) reported that heart disease mortality cases are distributed almost equally for men (3.8 million) and women (3.4 million) (World Health Organization 2011b). Heart disease caused 7.25 million deaths representing 12.8% of all the deaths (World Health Organization 2013c). Figure 2.1 (adapted from WHO) shows the heart disease mortality rates per 1000 around the world.

Different health organizations reported that heart and respiratory diseases are the main cause of death in different continents. The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2013). The Economic and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular diseases, cancers, diabetes and chronic respiratory diseases (ESCAP 2010). The Australian Bureau of Statistics reported that heart and circulatory system diseases are the leading cause of death in Australia, causing 33.7% of all deaths (Australian Bureau of Statistics 2013). Statistics of South Africa reported that heart and circulatory system diseases are the third leading cause of death in Africa (Statistics South Africa 2008). The Disease Control Priorities Project reported that non-communicable diseases such as cardiovascular heart disease, diabetes, cancer, and mental health disorders are causing significant illness, disability, and death in the Southern Cone of South America (World Bank Disease Control Priorities Project 2013).

Different countries have also reported that heart disease is a leading cause of death showing high mortality rates. For instance, in North America the Center for Disease Control and Prevention reported that heart disease is the leading cause of death and a major cause of disability in the United States of America causing almost 25% of all deaths (Center for Disease Control and Prevention 2014). Another example, in South America the Disease Control Priorities Project reported that in Argentina, cardiovascular heart disease is the biggest killer causing 18% of total deaths (World Bank Disease Control Priorities Project 2013). Similarly in Europe, WHO reported that heart disease is the leading cause of death in Spain (33% of all deaths) followed by cancer and other chronic diseases (World Health Organization 2013a). Also, in Africa,

WHO reported that heart disease is the leading cause of death in Egypt causing 42% of all deaths (World Health Organization 2013b). The percentage of people dying from heart disease is dependent on the culture lifestyle. (World Health Organization 2013c).

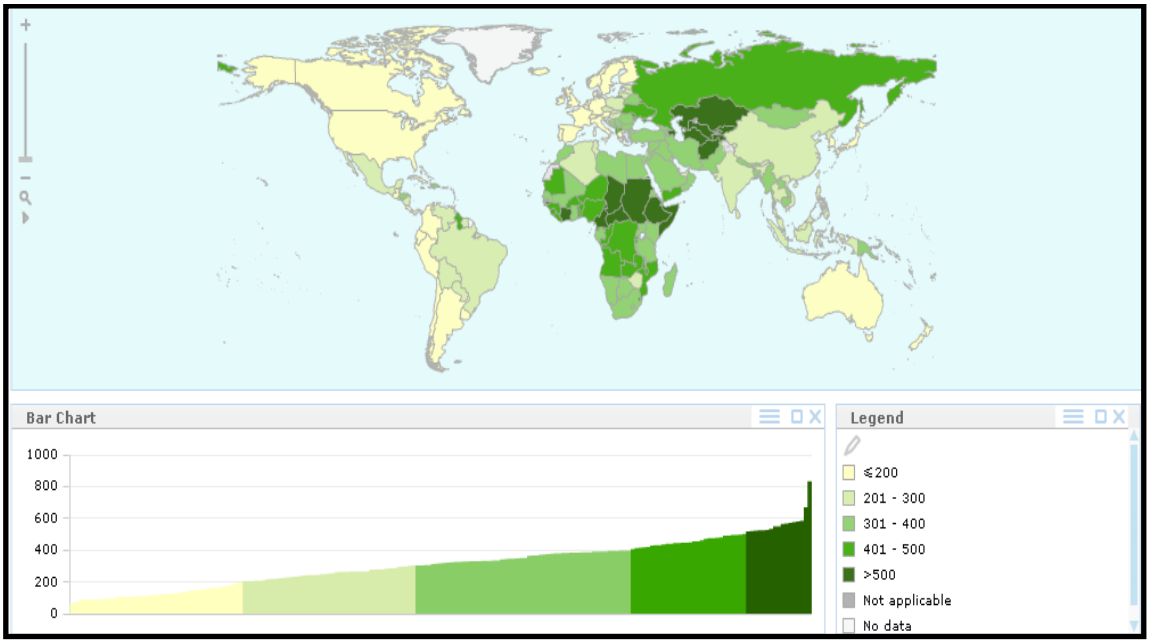

Figure 2.1: Mortality Rates around the World (adapted from WHO 2011)

From these health organizations' statistics, heart disease is confirmed as a leading cause of death in the world in different continents and countries regardless of their income.

2.2.2. Understanding Heart Disease

Section 2.2.1 describes heart disease as the leading cause of death. The mortality rates described require an understanding of heart disease. Symptoms differ from one heart disease to another, but generally include chest discomfort for a few minutes, pain in the jaw, neck, or back, weakness in the stomach, pain in the arms or shoulder, and shortness of breath (National Center for Chronic Disease Prevention and Health Promotion 2013). Several heart problems make up heart disease including coronary artery disease, heart failure and stroke (U.S. Department of Health and Human Services 2005). Section 2.2.2 discusses several heart diseases.

Coronary artery disease is a prevalent form of heart disease. It happens when the coronary arteries, which supply blood to the heart muscle, become hardened and narrowed due to the build-up of plaque on the heart arteries' inner walls (U.S. Department of Health and Human Services 2005). This building-up reduces the blood

flow to the heart over time (National Center for Chronic Disease Prevention and Health Promotion 2013). Plaque is an accumulation of cholesterol, fat, and other substances (U.S. Department of Health and Human Services 2005) and the resultant narrowing in the arteries usually causes heart attack (National Center for Chronic Disease Prevention and Health Promotion 2013).

Heart failure is a critical heart condition where the heart cannot pump enough blood to the body (National Center for Chronic Disease Prevention and Health Promotion 2013). It does not mean that the heart has stopped working, but that the body's need for blood and oxygen isn't being met (American Heart Association 2011). It occurs when excess fluid collects in the body as a result of heart weakness that leads to a build-up of fluid in the lungs, causing swelling of the feet, tiredness, weakness, and breathing difficulties (U.S. Department of Health and Human Services 2005). Too much alcohol, diabetes, high blood pressure, and previous heart attack can damage the heart muscle, leading to heart failure (U.S. Department of Health and Human Services 2005).

Stroke is another type of heart disease. It occurs because of a blood-clot leaving the heart and lodging in the arteries of the brain (American Heart Association 2013), thus causing a loss of blood supply to a part of the brain that then dies (U.S. Department of Health and Human Services 2005). Stroke can result in death or disability in walking or talking (U.S. Department of Health and Human Services 2005, American Heart Association 2013). One of the main risk factors for stroke is high blood pressure or hypertension. If stroke is detected and treated within an hour, this can help in preventing death or disability (American Heart Association 2013).

The World Health Organization reported that at least 80% of heart disease could be prevented (World Health Organization 2013c). The U.S Department of Health and Human Services reported that if different heart disease conditions are detected at the right time, sudden death can be prevented and patients can maintain a good healthy life (U.S. Department of Health and Human Services 2005). This clearly identifies the need for early detection in different heart disease conditions.

2.2.3. Heart Disease Prevention and Detection

Although heart disease is among the most common chronic diseases causing high rate of deaths all over the world, it has also been identified as among the most preventable and controllable diseases (Centers for Disease Control and Prevention 2013). At least 80%

of heart disease could be prevented by healthy diet, regular physical activity, and avoidance of tobacco products. These people dying from heart disease have one or more major risk factors that are influenced by lifestyle. (World Health Organization 2013c).

The main aim of disease detection and prevention is to avoid the disease and to interrupt the development of the disease. Prevention activities are performed at three levels: primary, secondary and tertiary prevention. Primary prevention is concerned with healthy people and how to reduce the risk factors that could result in a diseases's occurrence. Secondary prevention is concerned with the risk factors and early disease detection to increase the probability of successful medical treatments. Tertiary prevention is concerned with medical treatment of the disease and controlling the risk factors.

Healthy behaviour (primary prevention) and early detection (secondary prevention) are two critical elements in controlling heart disease. Healthy behaviour plays an important role in controlling the effects of heart diseases (American Heart Association 2011, Department of Health & Aging 2012, Centers for Disease Control and Prevention 2013). Early detection of heart disease patients can help in recovering patients' health and decreasing the mortality rate from heart disease (Centers for Disease Control and Prevention 2013). If proper preventative actions are not taken, then it is expected that the number of people dying due to heart disease will increase to reach 23.3 million by 2030 (Mathers CD 2006, World Health Organization 2011b).

The World Health Organization (2010) reported that early detection and treatment are aimed to reduce progression to severe and costly illness and complications of heart disease. The relative success of chronic disease treatments are dependent on the early detection of those diseases (Paladugu and Shyu 2010). There is a vital need for accurate and systematic tools that provide information for early detection of heart disease to identify those patients at high risk (Paladugu and Shyu 2010).

Regular medical checks (secondary prevention) are important in securing early detection and prevention of complications of heart disease (Department of Health & Aging 2012). It can be detected by several tests such as chest X-rays, coronary angiograms, electrocardiograms, and exercise stress tests (National Center for Chronic Disease Prevention and Health Promotion 2013). However, those tests are very costly

and require sophisticated equipment and a visit to a medical facility for heart disease detection.

Death rates from heart disease have decreased in North America and many western European countries. This decline has been due to early detection of heart disease and improved prevention regimes. In particular, reduced cigarette smoking among adults and lower average levels of blood pressure and blood cholesterol are critical factors for the prevention of heart disease. Unfortunately, it is expected that 82% of the future increase in heart disease mortality will occur in developing countries (World Health Organization 2013c). The economic circumstances of developing countries tend to limit the availability of the sophisticated equipment and medical facilities needed to meet the demand for heart disease detection. Approaches that allow early detection with less sophisticated equipment at a broader, community level may provide assistance in meeting that demand at lower cost. Community screening tests are one of the main opportunities of secondary prevention of heart disease (Kotnik 2010).

2.2.4. Heart Disease Risk Evaluation

There is a strong need worldwide for health maintenance services such as lifestyle education and screening tests in pharmacies to help enhancing patients' healthcare (Tang 2008). Community screening tests can help in early detection of heart disease and hence enhancing monitoring and managing patient's health. Recent researches has focused on discovering new specific, sensitive and cheap screening tests (Kotnik 2010). There is obviously a need for accurate, systematic tools that identify those patients at high risk and provide information for early intervention (Paladugu and Shyu 2010). Heart disease can be detected by several tests as electrocardiogram, stress tests, and cardiac angiograms. However these tests are expensive, require specialist equipment and therefore cannot be used as community screening tests. This clearly identifies a need to find less expensive tests that can be conducted as community screening.

It is widely accepted that age, gender, high blood pressure, smoking, cholesterol, and diabetes are the major risk factors for developing cardiovascular disease (Cupples and D'Agostino 1987). Heart disease risk factors also include obesity, left ventricular hypertrophy, and family history of heart disease (The Expert Panel 1994). The age and gender attributes are the most significant factors in the risk evaluation of heart disease (Cupples and D'Agostino 1987, The Expert Panel 1994). Moreover

obesity is a significant independent predictor of heart disease for different age groups (Hubert, Feinleib et al. 1983).

The Framingham Heart Disease Risk Evaluation (idiomatically, the Framingham test) uses those factors to identify the degree of risk for the patient. The researchers used extensive physical examinations and lifestyle interviews that they later analyzed for common patterns related to heart disease development (Framingham Heart Study 2013). The estimated heart disease event rates are used to quantify risk and to guide preventive care (D'Agostino, Vasan et al. 2008).

The Australian Absolute Cardiovascular Risk Calculator (idiomatically, the absolute risk calculator) has been developed for use by general practitioners, health workers, physicians, and health care professionals to help them in assessing the risk of cardiovascular disease in adults (National Heart Foundation of Australia 2009). It is developed from data extracted from large cohort studies and derived from the Framingham test study. The calculator evaluates the numerical probability of a cardiovascular event occurring within a five-year period. It reflects a person's overall risk of developing cardiovascular heart disease replacing the traditional method that considers various risk factors, such as high cholesterol or high blood pressure, in isolation. It is recommended for use in Australian primary care. The absolute risk calculator uses age, gender, systolic blood pressure, total cholesterol, diabetes and smoking status to identify if a patient is at high, moderate or low risk of heart disease (National Heart Foundation of Australia 2009).

Although the test and the absolute risk calculator help in identifying patients at risk of heart disease, there is not clearly known performance accuracy for them. Brindle, Emberson et al. established the predictive accuracy of the Framingham test for heart disease for twenty four towns in British population showing that 2.8% (95% confidence interval 2.4% to 3.2%) died from coronary heart disease compared with 4.1% predicted (Brindle, Emberson et al. 2003). In other research, Brindle, McConnachie et al. investigated the accuracy of the Framingham test for different socio-economic groups showing that the ratio of predicted to observed cardiovascular mortality rate was 0.56 (95% confidence interval [CI] = 0.52 to 0.60), a relative underestimation of 44% (Brindle, McConnachie et al. 2005).

The Framingham test and the absolute risk calculator use various scores that need blood tests prior to using the evaluation tool. These scores may be difficult to implement where there are limited resources available. Hence, there is a need to simplify the Framingham test attributes so that affordable detection strategies can be implemented (Bitton and Gaziano 2010). For instance, the Framingham test and the absolute risk calculator use total cholesterol for calculating the risk evaluation of heart disease. The total cholesterol is the sum of cholesterol in the blood. It needs a blood test to identify the total cholesterol level before being able to use either test (U.S department of health and human services 2005). Using the cholesterol, an invasive attribute, limits the ability of using the tests for the risk evaluation of heart disease.

The Framingham test and the absolute risk calculator use a set of invasive and non-invasive attributes in the risk evaluation heart disease patients. Although non-invasive attributes are easily known and low cost attributes, the use of only non-invasive attributes in the risk evaluation heart disease patients has not been investigated before. Moreover, if non-invasive attributes show significant performance in the risk evaluation of heart disease patients then this investigation would be of great benefit to the early detection of heart disease.

Techniques that fall into the field of data mining are being applied to consider exactly these kinds of questions. Section 2.3 looks at data mining to see how these techniques work and what assistance they can provide to heart disease detection.

2.3. Overview of Data Mining

Modern computer systems gather huge amounts of data every day through automatic recording systems in various sectors. Although there are huge amounts of stored data, the knowledge is buried within it (Bramer 2007), leading to the need for powerful data analysis tools. Researchers describe this state as being data rich and knowledge poor (Han and Kamber 2006), suggesting deeper analysis to extract useful knowledge. Such knowledge can lead to important discoveries in science; for example, it can enable scientists to predict weather and natural disasters, or make the difference between life and death of a patient (Han and Kamber 2006, Bramer 2007)

Although data mining has been used for more than two decades, its potential for detecting hidden knowledge is only recently being realized. Several researchers agreed that data mining lies at the interface of other research fields such as statistical analysis, database technology, pattern recognition, machine learning, data visualization, and expert systems (Fayyad 1997, Lee, Liao et al. 2000, Thuraisingham 2000, Obenshain 2004).The intersection between three main disciplines including databases, machine learning, and pattern recognition is shown in Figure 2.2.

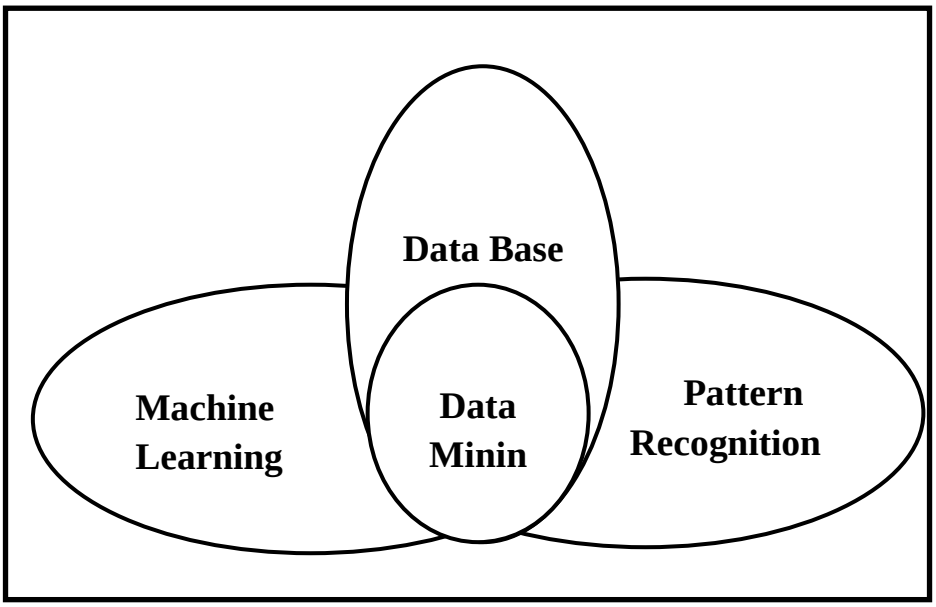

Figure 2.2: Data Mining and Intersecting Disciplines

Data mining can be defined from different prespectives. Fayyad (1997) describes data mining as the extraction of previously unknown, useful, and meaningful information from large amounts of data and defines data mining to be “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database” (Fayyad 1997). Scales and Embrechts (2002) define data mining as attempting to discover hidden patterns where these patterns are hidden and are difficult to detect with traditional statistical methods (Scales and Embrechts 2002). Han and Kamber (2006) agreed and added that data mining is an essential step in knowledge discovery (Han and Kamber 2006). Sitar-Taut, Zdrenghea et al. (2009) added that data mining is exploring very large datasets with the aim of seeking new patterns and relationships between variables and generalizing these relationships in a new model, formula, or decision tree (Sitar-Taut, Zdrenghea et al. 2009). These definitions of data mining allow us to conclude that data mining is the extraction of

useful knowledge from large amounts of data to identify hidden and unknown patterns, relationships and knowledge.

2.3.1. Data Mining as a Step in Knowledge Discovery

Data mining is one of the important steps in the knowledge discovery process to extract useful and meaningful knowledge and information. However, to extract knowledge from data, pre-processing steps are required. These steps involve data cleaning and integration, data selection and transformation, data mining, and data evaluation and presentation - see Figure 2. 3 (Han and Kamber 2006, Bramer 2007).

Data cleaning and integration is the first step in the knowledge discovery where noisy and irrelevant data are removed from the data source. Multiple data sources can be combined to form one data source. The second step is data selection and transformation where the data that are relevant to the knowledge discovery are selected and retrieved from the database. The selected data are then transformed into an appropriate form suitable for knowledge discovery. The third step is data mining which is responsible for extracting useful patterns and relationships from the data. Data mining techniques are used to explore data and retrieve meaningful knowledge. The fourth step is data evaluation and representation where the patterns and relationships are evaluated to identify the true and interesting ones. These patterns and relationships are formally presented (Zaïane 1999, Han and Kamber 2006, Bramer 2007).

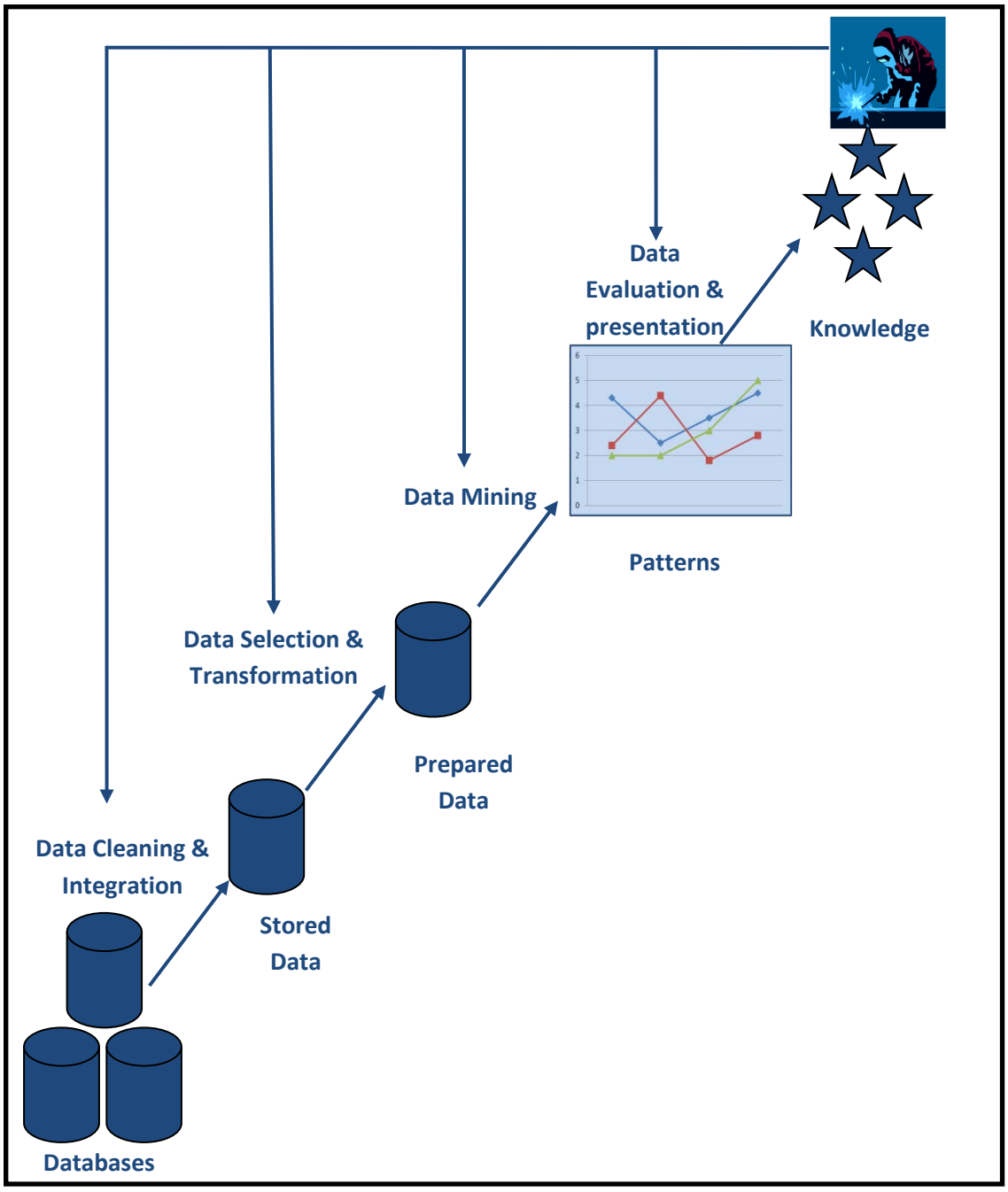

Figure 2.3: The Knowledge Discovery Steps

2.3.2. Supervised and Unsupervised Data Mining Tasks

The goal of data mining is to learn from data. There are two broad categories of data mining tasks: supervised and unsupervised learning as shown in Figure 2.4 (Matkovsky and Nauta 1998, Obenshain 2004, Maimon and Rokach 2010).

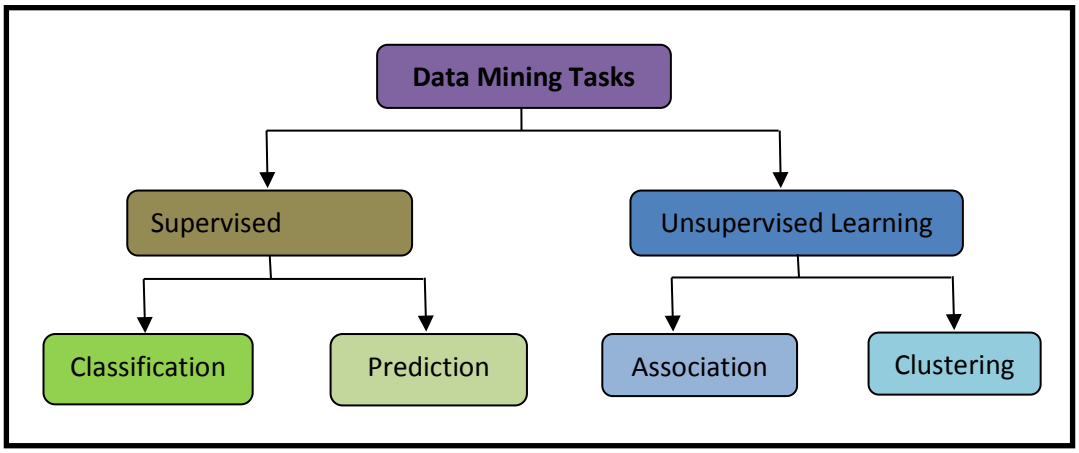

Figure 2.4: Data Mining Tasks and Techniques

In supervised learning, a training set is used to learn model parameters (Obenshain 2004). It is used when values of input variables are used to make predictions about another target variable with known values (Matkovsky and Nauta 1998). In unsupervised learning there is no training set used (Obenshain 2004). It refers to modelling the distribution of instances in a typical, high-dimensional input space (Maimon and Rokach 2010). Two types of supervised learning data mining techniques are classification and prediction. Two types of unsupervised learning data mining techniques are association and clustering as shown in Figure 2.4 (Han and Kamber 2006, Bramer 2007, Maimon and Rokach 2010).

Classification data mining applications are supervised learning data mining tasks that predict categorical (discrete and unordered) values of the target attribute given the input data. The target attribute is a set of pre-defined classes e.g. excellent, very good, good, bad, and very bad (Maimon and Rokach 2010). If the target attribute's values are two values (such as healthy and sick), this is called binary classification. If the target attribute has multiple possible values (for example, car, bicycle, person, bus and taxi then this is called multi-class classification (Bramer 2007). For example, a bank can use the customers' details such as age, gender, income, and job to categorize the home loan of those customers' applications to be safe or at risk. Such a classification is a binary classification, as the target attribute has two values, as shown in Figure 2.5. Similarly, a medical doctor can classify blood pressure patients' cases based on some attribute such as age, weight, smoking status, and systolic blood pressure to take treatment A, or treatment B, or treatment C. This classification is a multi-class classification as shown in Figure 2. 6.

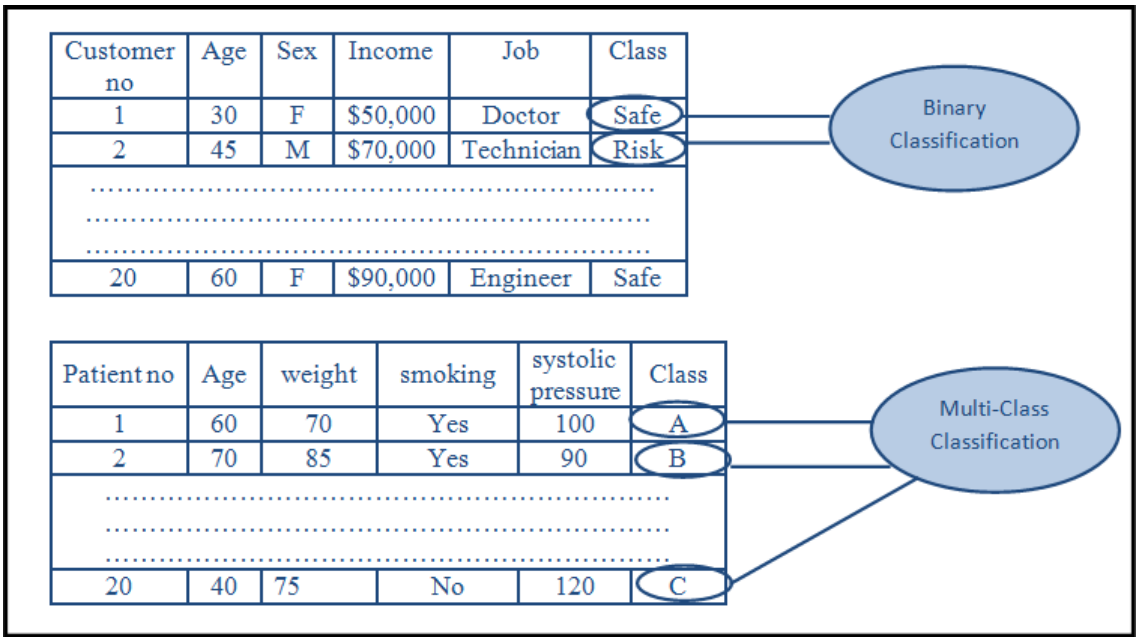

Figure 2.5: Binary and Multi-Class Classification Data Mining Example

Prediction data mining applications predict continuous valued functions of the target attribute given the input data. The target attribute is a continuous real value domain (Maimon and Rokach 2010). For example, a supermarket that wants to predict its profit in a subsequent month having recorded the profits for previous months. The profit value is a continuous function and its likely future values are estimated by prediction data mining (see Figure 2.6).

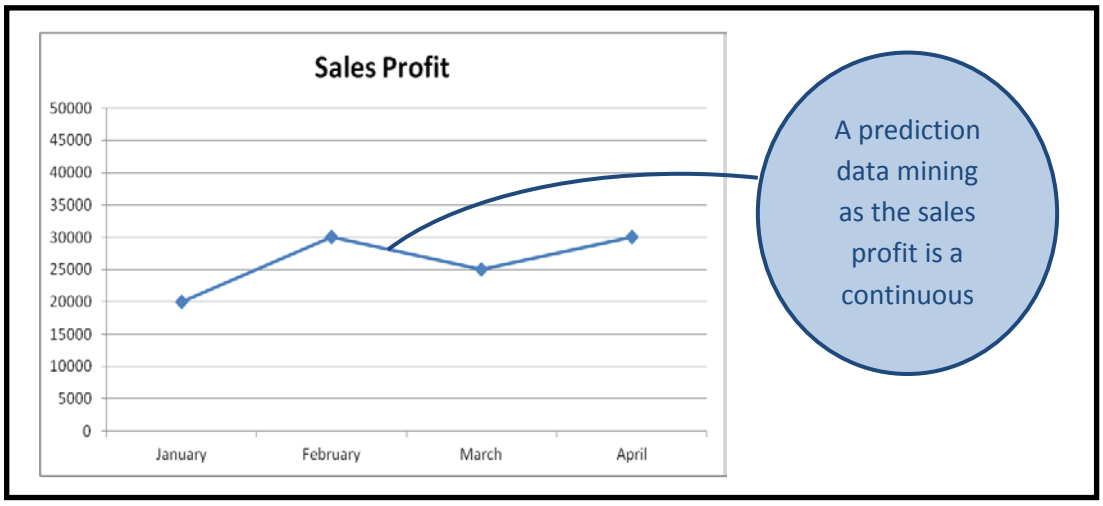

Figure 2.6: Prediction Data Mining Example

The association data mining application finds the relationship and rules that are associated with the values of some attributes of the variables (Han and Kamber 2006). Extracting rules and relationships from a dataset is called association rule mining

(Bramer 2007). Association rule mining allows overlapping between patterns to efficiently search the solution space. It enables users to extract all the patterns that satisfy some predefined constraints in the dataset (Gupta 2010). For example, market basket analysis extracts the association rules between re-occurring items in a customer's shopping list. In Figure 2.7, the customers that buy bread also buy eggs or milk. These rules help the supermarket management to arrange their products to help increase their sales and profit.

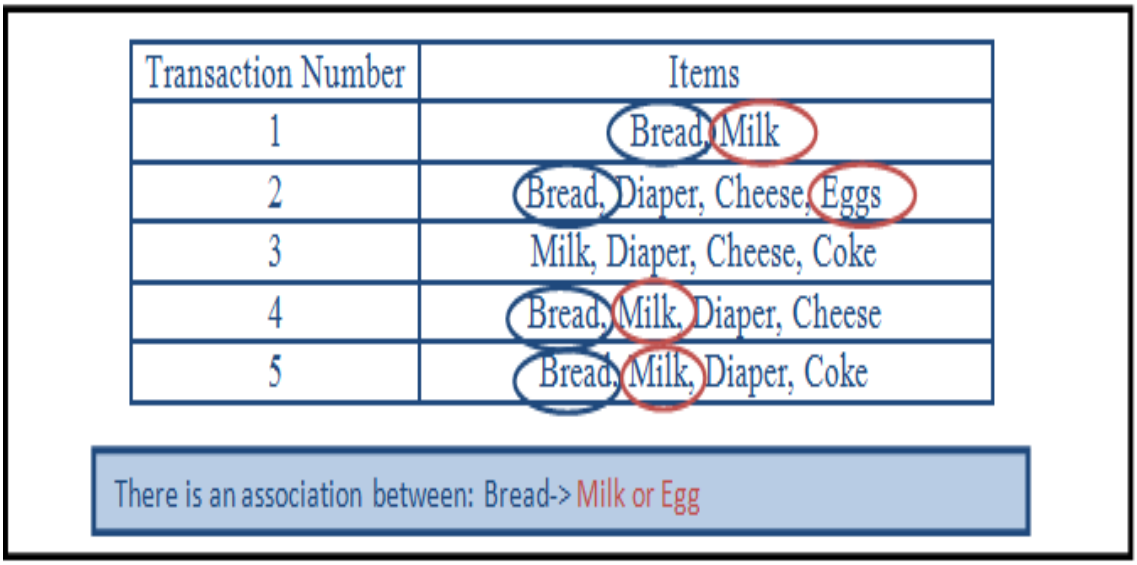

Figure 2.7: Association Rule Data Mining Example

The clustering data mining applications are unsupervised data mining tasks that group items with similar features together (Han and Kamber 2006, Bramer 2007). It combines objects with related or similar properties in one group and combines different or unrelated objects in other groups. The distance between the instances in the same cluster is called intra-clustering distance. The distance between different clusters is called inter-clustering distance (Tan, Steinbach et al. 2006). In applying clustering data mining it is important to identify the number of clusters into which we wish to group the data. The number of clusters is usually a small number between two and six but it could be larger in some applications (Bramer 2007). For example, an insurance company can cluster its customers based on their age, policy purchased or income into two clusters (see Figure 2. 8).

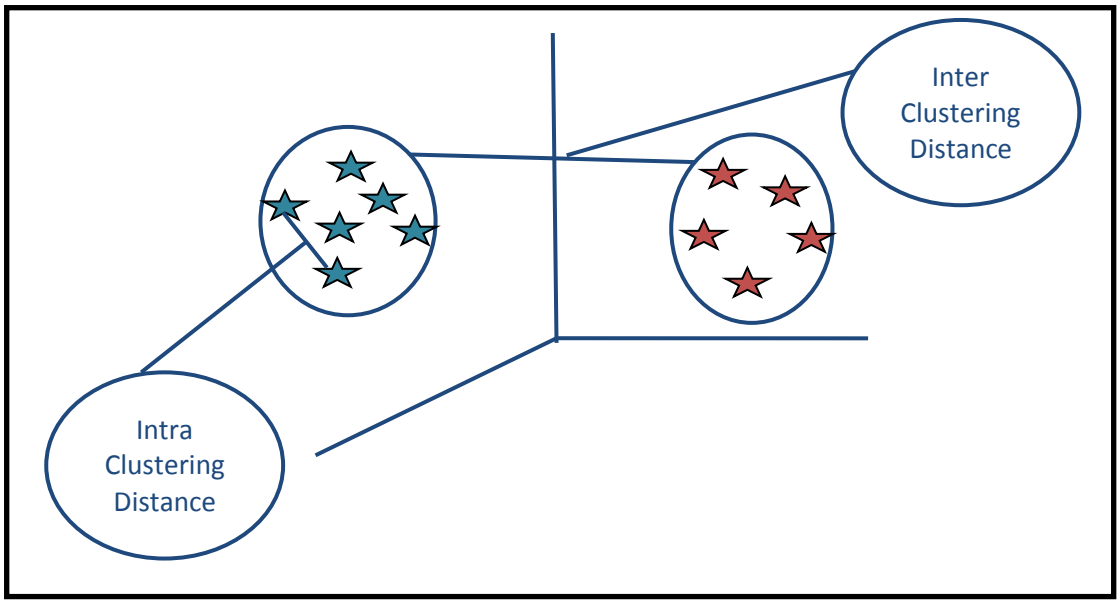

Figure 2.8: Clustering Data Mining Example

2.3.3. Data Mining Techniques

Various data mining techniques are applied depending on whether classification, prediction, association or clustering tasks are required. The focus of this thesis is the application of classification techniques that assist healthcare professionals to identify patients at risk of heart disease. Classification data mining techniques include: Decision Tree, Naïve Bayes, k-Nearest Neighbour, Support Vector Machine, and Neural Networks.

Decision Tree is a widely used data mining technique in classification problems due to its reliable performance and rules extraction ability (Bramer 2007). The Decision Tree creates a flow chart to represent a classification tree of the data. This tree is a sequence of simple questions that trace the path to the answer or the classification value (Moore, Jesse et al. 2001). The Decision Tree is based on top-down induction where decision tree rules are extracted by repeatedly splitting the values of the different attributes in the dataset. The Decision Tree uses the training data to extract decision tree rules then applies those rules to any instance in the testing data to classify it (Han and Kamber 2006). The extracted rules help in understanding how a testing instance is classified. There are different types of Decision Trees: information gain, gini index, and gain ratio (Han and Kamber 2006, Bramer 2007).

Naïve Bayes is one of the most successful and widely-used classification data mining techniques (Bramer 2007). It does not extract rules as does the Decision Tree

but it uses mathematical probability theory (Han and Kamber 2006). The name Naïve is based on class conditional independence. Class conditional independence assumes that the effect of an attribute value on a given class is independent of the values of the other attributes (Bramer 2007). Naïve Bayes techniques calculate the prior probability of the target attribute and the conditional probability of the remaining attributes. The prior and conditional probability is calculated for the training data. Then, for each testing instance in the testing dataset, the probability is calculated with each of the target attribute values. The target attribute value with the largest probability is then selected (Han and Kamber 2006, Bramer 2007).

K-Nearest Neighbour (KNN) is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time (Alpaydin 1997). KNN is different from Decision Tree that extracts rules and Naïve Bayes that uses probability. KNN uses learning by analogy, which compares a given testing instance with the training instances to find the similarity between them. This similarity is calculated by measuring the distance between the testing instance and all the training instances. The K-Nearest instances are used to calculate the target value of the testing instance (Han and Kamber 2006, Bramer 2007).

Support Vector Machine is a successful and accurate data mining technique for classification problems. It transforms the training data using non-linear mapping into a higher dimensional area (Han and Kamber 2006). It separates data into classes based on a hyper-plane with maximum margin. It searches for the optimal hyper-plane between classes to create the support vectors (Han and Kamber 2006, Sajda 2006).The Support Vector Machine is a highly accurate classification data mining technique. However, it suffers from slow processing when training with a large set of data (Han and Kamber 2006).

Neural Networks are another successful data mining technique used in classification problems. It is a set of input and output units with connections between them (Han and Kamber 2006). The connections between the inputs are weighted to provide the desired output. During the training phase, the network learns by adjusting the weights to be able to predict the correct class label of the output (Piction 2000). The advantage of Neural Networks is the ability to tolerate to noisy data and to classify instances on which they have not been trained (Han and Kamber 2006). The

disadvantage of Neural Networks is that is they are a 'black box' with limited representational power and only linear classifiers can be constructed (Krose and Smagt 1996).

Chapter 3 discusses the data mining techniques applied in this research.

2.3.4. Data Mining Techniques Performance Evaluation

As different data mining techniques can be applied to the problem or dataset, it is important to compare the results of different data mining techniques to apply the best one to the task at hand. To measure the stability of the data mining techniques, the data is divided into training and testing data with 10-fold cross validation. Cross validation reduces the potential for the training to skew the data mining techniques' accuracy through peculiarities in the training data (Bramer 2007). The training dataset allows the data mining techniques to learn from these data. The testing dataset is used to evaluate the performance of the data mining technique in relation to what is learned from the training dataset. To evaluate the performance of the data mining techniques the sensitivity, specificity, and accuracy are calculated.

Sensitivity is the proportion of positive instances that are correctly classified as positive. For example, in the case of a doctor needing to identify if patients are sick or healthy, then the sensitivity is the proportion of sick people that are actually classified as sick. Equation 2.1 shows how the sensitivity is calculated. Specificity is the proportion of negative instances that are correctly classified as negative. Thus, for the doctor, the specificity is the proportion of healthy people that are classified as healthy. Equation 2.2 shows how the specificity is calculated. Accuracy is the proportion of instances that are correctly classified. Again, for the doctor, it is the proportion of the sick people that are identified as sick and the healthy people that are identified as healthy (Han and Kamber 2006, Bramer 2007). Equation 2.3 shows how the accuracy is calculated.

Sensitivity = True Positive / Positive (Equation 2.1)

Specificity = True Negative / Negative (Equation 2.2)

Accuracy = (True Positive + True Negative) / (Positive + Negative) (Equation 2.3)

To measure the stability of every data mining technique, the average and standard deviation of the sensitivity, specificity, and accuracy are calculated. A t-test is

calculated to identify which data mining technique demonstrates better performance. T-test is a well-known statistical method that is used to identify the significance of the performance of different data mining techniques (Fayyad and Keki 1992).

2.4. Data Mining Applications in Healthcare

Data mining researchers have long been concerned with the application of tools to facilitate and improve data analysis on large, complex datasets. The current challenge is to make data mining and knowledge discovery systems applicable to a wider range of domains (Shillabeer and Roddick 2006). Data mining is rapidly growing successfully in a wide range of applications such as analysis of organic compounds, financial forecasting, medical diagnosis and weather forecasting. A supermarket can mine its data to optimize targeting high value customers to increase its profits. Doctors can mine their data to predict the probability that a cancer patient will respond to chemotherapy thus reducing health care costs and improving health care quality (Bramer 2007). The task for data mining researchers is to manipulate data mining technologies to make them applicable to the specific requirements of clients, for example, the healthcare sector (Ashby and A.Smith 2005)

Data mining in healthcare is an emerging field of high importance for providing diagnosis and prognosis and a deeper understanding of medical data. Data mining applications in healthcare include analysis of health care centres for better health policy-making and prevention of hospital errors, early detection and prevention of diseases and preventable hospital deaths, more value for money and cost savings in healthcare delivery, and detection of fraudulent insurance claims (Ruben 2009). Medical data mining has great potential for exploring the hidden patterns in the datasets of the medical domain. Data mining attempts to solve real world health problems in diagnosis and treatment of diseases (Liao and Lee 2002).

The major challenge presented by health and medicine is to develop a technology that can provide trusted hypotheses based on measures that can be relied upon in medical health research and applied in a clinical environment (Ashby and A.Smith 2005). In medical decision making (classification, diagnosing, etc.), there are many situations when decisions must be made effectively and reliably (Podgorelec, Kokol et al. 2002). The discovery of various theoretical implications from training datasets is a difficult process, depending on the researcher’s experience, abilities, and

intuition (Helma, Gottmann et al. 2000). Even the best experts are sometimes overwhelmed by the accumulated data where people's information storing, managing, and computing capabilities are poor in comparison with machines. Data mining can be used to help health professionals in decision making (Helma, Gottmann et al. 2000, Podgorelec, Kokol et al. 2002).

Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes (Porter and Green 2009), stroke (Panzarasa, Quaglini et al. 2010), cancer (Li L 2004), and heart disease (Das, Turkoglu et al. 2009). Motivated by the increasing mortality rates of heart disease, researchers are using several data mining techniques to help healthcare professionals in the diagnosis of heart disease patients.

2.5. Data Mining in Heart Disease

Heart disease professionals store significant amounts of patients' data. This offers the opportunity to analyse these datasets to extract useful knowledge. Researchers are using statistical analysis and data mining techniques to help healthcare professionals to identify patients at risk of heart disease.

Statistical analyses have identified the risk factors associated with heart disease to be age, blood pressure, smoking habit (Heller, Chinn et al. 1984), total cholesterol (Wilson, D'Agostino et al. 1998), diabetes (Simons, Simons et al. 2003), hypertension, family history of heart disease (Salahuddin and Rabbi 2006), obesity, and lack of physical activity (Shahwan-Akl 2010). Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease.

Data mining is an effective tool in analysing data to extract useful knowledge (Helma, Gottmann et al. 2000, Podgorelec, Kokol et al. 2002). Researchers have been using data mining techniques to extract significant patterns and find relationships between different variables in heart disease dataset. Patil and Kumaraswamy (2009) proposed an efficient approach for the extraction of significant patterns from the heart disease warehouses for heart attack prediction. They used maximal frequent item set algorithm with k-means clustering algorithm to extract frequent patterns from the heart disease dataset. The results show that cholesterol, blood pressure, and blood sugar levels are significant factors for heart disease diagnosis (Patil and Kumaraswamy 2009).

2.5.1. Using Data Mining Techniques in Heart Disease Diagnosis

Researchers have been applying different data mining techniques to help health care professionals in the diagnosis of heart disease patients (Helma, Gottmann et al. 2000, Podgorelec, Kokol et al. 2002). The most frequently used data mining techniques focus on classification such as Naïve Bayes, Decision Tree, and K-Nearest Neighbour. Other data mining techniques also focusing on classification are Kernel Density, Neural Network Automatically Defined Groups, Bagging Algorithm, Sequential Minimal Optimization, Direct Kernel Self-Organizing Map, and Support Vector Machine. Table 2.1 shows a comparison of various researchers of different data mining techniques and their accuracy in the diagnosis of heart disease.

Herron (Herron 2004) shows that best results are achieved by Support Vector Machine followed by Naïve Bayes and Decision Tree with accuracies of 83.6%, 83.37%, and 77.56% respectively. Palaniappan and Awang (Palaniappan and Awang 2007) show that best results are achieved by Naïve Bayes, followed by Decision Trees, and Neural Network with accuracies of 95%, 94.93%, and 93.54% respectively. These research results are showing very high accuracies compared to the same data mining techniques applied in other research (Table 2.1). Rajkumar and Reena (Rajkumar and Reena 2010) show that best results are achieved by Naive Bayes, followed by Decision List and K-Nearest Neighbour with accuracies of 52.33%, 52%, and 45.67% respectively. This research results are showing very low accuracies (essentially, chance) compared to the same data mining techniques applied in other research (Table 2.1). Lakshmi, Krishna et al. (Lakshmi, Krishna et al. 2013) show that best results are achieved by Decision Tree, followed by K-Nearest Neighbour, K Means, and Support Vector Machine with accuracies of 84.68 %, 83.95 %, 80.29 %, and 78.10 % respectively. These variable results in data mining technique performance can be explained by the fact that different heart disease datasets have been used.

Researchers have been applying data mining techniques over different heart disease datasets. Each of these datasets has its own input and output attributes. For instance, Yan et al., applied a multi-layer perceptron on a Chinese heart disease dataset. The input attributes involved age, gender, dizziness, weakness, chest pain, blood pressure and heart rate while the output attribute was one of five types of heart disease (Yan, Zheng et al. 2003). Other researchers used the Cleveland dataset whose input

involves age, sex, resting blood pressure, cholesterol level, and exercise included angina while its output is a binary attribute indicating the presence of heart disease or not. The different input variables mean that different indicators are influential in diagnosis. The different output variables constrain the type of analysis that can be made of the data.

Table 2.1: Research Sample of Data Mining Techniques in Heart Disease Diagnosis

Author/Year	Technique	Accuracy
(Hall 2000)	Naïve Bayes	83.24%
	K Nearest Neighbour	82.12%
	Decision Tree	75.32%
(Yan, Zheng et al. 2003)	Multilayer Perceptron	63.6%
(Herron 2004)	Support Vector Machine	83.6%
	J4.8 Decision Tree	77.56%
	Naïve Bayes	83.37%
(Andreeva 2006)	Naïve Bayes	78.56 %
	Decision Tree	75.73 %
	Neural network	82.77 %
	Sequential minimal optimization	84.07 %
	Kernel density	84.44 %
(Polat , Sahan et al. 2007)	Fuzzy-AIRS–k-nearest neighbour	87%
(Palaniappan and Awang 2007)	Naïve Bayes	95%
	Decision Trees	94.93%
	Neural Network	93.54%
(De Beule, Maesa et al. 2007)	Artificial neural network	82%
(Tantimongcolwat, Naenna et al. 2008)	Direct kernel self-organizing map	80.4%
	Multilayer Perceptron	74.5%
(Hara and Ichimura 2008)	Automatically Defined Groups	67.8%

Author/Year	Technique	Accuracy
	Immune Multi-agent Neural Network	82.3%
(Sitar-Taut, Zdrenghea et al. 2009)	Naïve Bayes	62.03%
	Decision Trees	60.40%
(Tu, Shin et al. 2009)	Bagging algorithm	81.41%
(Das, Turkoglu et al. 2009)	Neural network ensembles	89.01%
(Rajkumar and Reena 2010)	Naive Bayes	52.33%
	K nearest neighbour	45.67%
	Decision list	52%
(Srinivas, Rani et al. 2010)	Naïve Bayes	84.14%
	One Dependency Augmented Naïve Bayes	80.46%
(Kangwanariyakul, Nantasenamat et al. 2010)	Back-propagation neural network	78.43%
	Bayesian neural network	78.43%
	Probabilistic neural network	70.59%
	Linear support vector machine	74.51%
	Polynomial support vector machine	70.59%
	Radial basis function support vector machine	60.78%
(Kumari and Godara 2011)	RIPPER	81.08%
	Decision Tree	79.05%
	Artificial Neural Network	80.06%
	Support Vector Machine	84.12%
(Soni, Ansari et al. 2011)	Weighted Associative Classifier	57.75%
	Classification based on Association Rule (CBA)	58.28%
	Classification based on Multiple Class-Association Rules (CMAR)	53.64%
	Classification based on Predictive Association Rules (CPAR)	52.32%

Author/Year	Technique	Accuracy
(Abdullah and Rajalaxmi 2012)	Decision Tree	50.67 %
	Random Forest	63.33 %
(Rajeswari, Vaithiyanathan et al. 2013)	Neural Network	80.53%
	J4.8 Decision Tree	77.89%
	Support Vector Machine	84.16%
	Feature Selection with Neural Network	84.49%
	Feature Selection with Decision Tree	84.16%
	Feature Selection with Support Vector Machine	87.46%
(Lakshmi, Krishna et al. 2013)	Support Vector Machine	78.10 %
	Decision Tree	84.68 %
	K Nearest Neighbour	83.95 %
	K mean	80.29 %
(Pandey, Pandey et al. 2013)	COBWEB	1.98%
	EM	81.51%
	Farthest First	73.59%
	Make Density Based Clusters	81.51%
	Simple K-Means	80.85%

Decision Tree, Naïve Bayes and K-Nearest Neighbour are the three most common data mining techniques used in the diagnosis of heart disease patients. There is a need to identify if one data mining technique can show the best performance on different datasets. Each data mining technique is compared over different heart disease datasets showing different levels of accuracy (see Table 2.2).

Decision Tree is compared over different heart disease datasets showing different levels of accuracy ranging from 50.67 % to 94.93% (see Table 2.2). The Decision Tree (Herron 2004, Kumari and Godara 2011, Rajeswari, Vaithiyanathan et al. 2013) using the Cleveland heart disease dataset shows accuracies of 77.56%, 79.05%, and 77.89% respectively. The slight difference in the Decision Tree performance on the same dataset reported by the three researchers is because of the type of decision tree

used or the discretization method used by each researcher. There is not always enough information about the discretization method used in each research.

Naïve Bayes is compared over different heart disease datasets showing different levels of accuracy ranging from 52.33% to 95% (see Table 2.2). The Naïve Bayes (Hall 2000, Herron 2004, Srinivas, Rani et al. 2010) using the Cleveland heart disease dataset shows accuracies of 83.24%, 83.37%, and 84.14%respectively. The slight difference between the Naïve Bayes performance on the same dataset reported by the three researchers is because of the type of discretization method used by each researcher.

K-Nearest Neighbour is compared over different heart disease datasets showing different levels of accuracy ranging from 45.67% to 83.95 % (see Table 2.2).

Table 2.2: Same Data Mining Technique for Different Heart Disease Datasets

Technique	Author/year	Dataset	Accuracy
Decision Trees	(Herron 2004)	Cleveland Heart disease dataset 303 instances with 15 medical attributes	77.56%
	(Andreeva 2006)	Heart disease dataset	75.73 %
	(Palaniappan and Awang 2007)	Cleveland Heart disease dataset 909 instances with 15 medical attributes	94.93%
	(Sitar-Taut, Zdrenghea et al. 2009)	The DB contained 10 attributes and 303 instances.	60.40%
	(Kumari and Godara 2011)	Cleveland Heart disease dataset 303 instances with 15 medical attributes	79.05%
	(Abdullah and Rajalaxmi 2012)	Coronary Heart Disease Dataset	50.67 %
	(Rajeswari, Vaithiyanathan et al. 2013)	Cleveland Heart disease dataset 303 instances with 15 medical attributes	77.89%
	(Lakshmi, Krishna et al. 2013)	Cleveland Heart disease dataset 2268 instances	84.68 %

Technique	Author/year	Dataset	Accuracy
Naïve Bayes	(Hall 2000)	Cleveland Heart disease dataset 303 instances	83.24%
	(Herron 2004)	Cleveland Heart disease dataset 303 instances	83.37%
	(Andreeva 2006)	Heart disease database	78.563 %
	(Palaniappan and Awang 2007)	Cleveland Heart disease dataset 909 instances with 15 medical attributes	95%
	(Sitar-Taut, Zdrenghea et al. 2009)	The DB contained 10 attributes and 303 instances.	62.03%
	(Rajkumar and Reena 2010)	The Heart disease database contained 3000 instances and 14 attributes.	52.33%
	(Srinivas, Rani et al. 2010)	The database used was obtained from the UCI repository. Supervised discretization	84.14%
K-Nearest Neighbour	(Hall 2000)	Cleveland Heart disease dataset 303 instances	82.12%
	(Rajkumar and Reena 2010)	The Heart disease database contained 3000 instances and 14 attributes.	45.67%
	(Lakshmi, Krishna et al. 2013)	Cleveland Heart disease dataset 2268 instances	83.95 %

The difference in the single data mining technique accuracy over different datasets raises an important question: “Are there a specific data attributes that help data mining techniques enhance performance in the diagnosis of heart disease patients?” This is why the single data mining technique shows different performance on different datasets.

The previous discussion demonstrates that data mining techniques show different results over different datasets. The single data mining technique compared over different heart disease datasets also shows different levels of accuracy. Thus, it is difficult to compare different data mining techniques applied over different datasets. However, a defacto benchmark dataset has emerged in the literature: the Cleveland Heart Disease Dataset (CHDD http://archive.ics.uci.edu/ml/datasets/Heart+Disease). Results of trials on this dataset do allow formal comparison.

2.5.2. Using Single and Hybrid Data Mining Techniques in Cleveland Heart Disease Diagnosis

Section 2.5.1 shows that it is difficult to compare the performance of different data mining techniques. The development and distribution of the CHDD improves the comparison of the performance of different data mining techniques (Hara and Ichimura 2008). Table 2.3 illustrates a sample of data mining techniques used in the diagnosis of heart disease on the CHDD. Palaniappan and Awang investigated applying Decision Tree, Naïve Bayes and K-Nearest Neighbour on the Cleveland heart disease dataset showing very high results (Table 2.2). The Cleveland heart disease dataset used in this research contains 909 rows, however the Cleveland heart disease dataset used by all other researchers contain 303 rows. There is not any explanation given by the researchers how the 909 rows were created from the original 303 rows. This anomalous size of the dataset and the very high accuracy results are unusual and so are discarded from the comparison.

When comparing different single data mining techniques together K-Nearest Neighbour demonstrates the best results followed by Naïve Bayes, Bagging Algorithm and Decision Tree (see Table 2.3). Although Decision Tree shows less accuracy than K-Nearest Neighbour (by approximate 2%), Decision Tree has the ability to explain how a decision is made by its decision rules. K-Nearest Neighbour and Naïve Bayes do not explain how the decision was made as they are based on Euclidian distance and probability.

Hybrid data mining techniques improve the data mining techniques' accuracies in the diagnosis of heart disease patients as shown in Table 2.3. For instance, Polat et al. (Polat , Sahan et al. 2007) use Fuzzy Artificial Immune Recognition System and K-Nearest Neighbour with accuracy of 87% in the detection of heart disease patients (Table 2.3). Das, Turkoglu et al. (Das, Turkoglu et al. 2009) use Neural Network Ensembles showing accuracy of 89.01% (Table 2.3). Rajeswari, Vaithiyanathan et al. (Rajeswari, Vaithiyanathan et al. 2013)use Feature Selection with Neural Network, Decision Tree, and Support Vector Machine data mining techniques showing that this integration could enhance accuracy in the diagnosis of heart disease patients. Ozsen and Gunes applied genetic algorithms with artificial immune systems showing accuracy of 87% (Ozsen and Gunes 2009) (Table 2.3) .

Comparison of single and hybrid data mining techniques with the CHDD shows different accuracies, with the hybrid techniques showing better accuracy than single techniques (see Table 2.3). The best accuracy achieved using single data mining technique is 82.12%% by K-Nearest Neighbour (Hall 2000). However, the best accuracy achieved using hybrid data mining technique is 89.01% by Neural Network Ensembles (Das, Turkoglu et al. 2009). These results suggest that hybridized data mining techniques are more accurate in the diagnosis of heart disease patients.

Table 2.3: A Sample of Data Mining Techniques Used on the Cleveland Heart Disease Dataset

Type	Author/ Year	Technique	Accuracy
Single	(Hall 2000)	K-Nearest Neighbour	82.12%
	(Cheung 2001)	Decision Tree	81.11%
		Naïve Bayes	81.48%
	(Tu, Shin et al. 2009)	J4.8 Decision Tree	78.9%
		Bagging algorithm	81.41%
Hybrid	(Polat , Sahan et al. 2007)	Fuzzy-AIRS – K-Nearest Neighbour	87%
	(Parthiban and Subramanian 2007)	Coactive Neuro-Fuzzy Inference System	Mean Square Error is 0.000842
	(Das, Turkoglu et al. 2009)	Neural Network Ensembles	89.01%
	(Ozsen and Gunes 2009)	Genetic Algorithms with Artificial Immune System	87%
	(Rajeswari, Vaithiyanathan et al. 2013)	Feature Selection with Neural Network	84.49%
		Feature Selection with Decision Tree	84.16%
		Feature Selection with Support Vector Machine	87.46%

2.6. Chapter Summary and Conclusion

This chapter shows that in the last decade heart disease has been the leading cause of death all over the world. Data mining can extract useful knowledge from huge amounts of data in the health sector. Motivated by the high mortality rates of heart disease and

the availability of large amounts of stored data, researchers are applying different data mining techniques to help health care professionals in diagnosing and identifying patients at risk of heart disease.

From the literature, two main facts are evident. The first is that single data mining techniques implemented over different datasets show different levels of accuracies. Thus there are some data characteristics or data attributes that are useful in improving data mining techniques' performance in the diagnosis of heart disease patients. The second element is that recently researchers are suggesting that integrating more than one data mining technique helps in enhancing diagnostic accuracy. There are already 'calculators' used for screening that are based on statistically-derived indicators. Can data mining analysis offer alternative insights for the development of such calculators, particularly if specific indicators (non-invasive attributes) are used? From reviewing the literature of applying different data mining techniques in the diagnosis of heart disease patients, different research questions arise:

1. From a data mining perspective, what are the important attributes for heart disease risk evaluation (diagnosis)?
2. Is different data mining techniques' performance reliable (showing consistent results) across different heart disease datasets' attributes?
3. Can integrating hybrid data mining techniques increase performance across different heart disease datasets?
4. Can a non-invasive attributes prototype for heart disease risk evaluation be developed using data mining analysis?

The research project described in this thesis investigates developing a prototype for heart disease risk evaluation using data mining analysis. It applies different data mining techniques over two different heart disease datasets and on different attribute combinations of the two heart disease datasets. This investigation identifies the important attributes needed by data mining techniques in the diagnosis of heart disease patients. Integrating clustering with data mining techniques is examined to identify which data mining technique combination will provide better accuracy in the diagnosis of heart disease patients on the whole datasets and on different attributes combinations of the two heart disease datasets. Finally, the research develops a heart disease risk evaluation tool to help healthcare professionals in identifying patients at high risk of heart disease.

The next chapter describes the application of different data mining techniques over two heart disease datasets to determine if they provide reliable results. The data mining techniques used include Decision Tree, Naïve Bayes, and K-Nearest Neighbour as common and successful data mining techniques applied to different heart disease datasets. These data mining techniques are applied to the benchmark Cleveland heart disease dataset and to a new Canberra heart disease dataset. The Canberra heart disease dataset is larger than the Cleveland heart disease dataset and contains different attributes than that in the Cleveland dataset. Finally, the same data mining techniques are applied over different attribute combinations of both datasets to identify the significant attributes required for data mining techniques in the diagnosis of heart disease patients and a preferred data mining technique when using limited sets of attributes for this task.

[This Page is Left Blank Intentionally]

Chapter 3

Applying Data Mining Techniques in Heart Disease Diagnosis

3.1. Introduction

The conducted literature review in the previous chapter clearly demonstrates the limitations of the single data mining technique implemented over different datasets: results with different levels of accuracy by the same technique on different datasets. The single data mining technique showed different results across different datasets due to several causes. These causes involve different data distribution of each dataset as well as different data attributes and characteristics for each dataset. The variable results of different data mining techniques across different heart disease datasets is unhelpful as it does not demonstrate a clear guidance on how to improve the performance of different data mining techniques. Different data mining techniques are applied to help healthcare professional in the diagnosis of heart disease. Decision Tree, Naïve Bayes, and K-Nearest Neighbour are three common successful data mining techniques applied to different heart disease datasets for diagnosis. The discussion identifies that some data characteristics or data attributes enhance the performance of data mining techniques in the diagnosis of heart disease patients.

This chapter describes the application of different data mining techniques over two heart disease datasets to determine if they provide reliable results (see Figure 3.1). The data mining techniques used are the Decision Tree, Naïve Bayes, and K-Nearest Neighbour. Applying these data mining techniques to the benchmark Cleveland heart disease dataset establishes base-line accuracy for each technique. Applying the same data mining techniques to the larger Canberra data examines whether they achieve reliable results on the larger dataset. The different data mining techniques results are compared over the two heart disease datasets to identify if each technique provides reliable results. Finally, the same data mining techniques are then compared when applied over all attributes of each dataset, the common attributes, and the Principle component analysis (PCA) attributes of both datasets (see Figure 3.2). This empirical approach identifies some significant attributes required for data mining techniques in

the diagnosis of heart disease patients and a preferred data mining technique when using limited sets of attributes for this task.

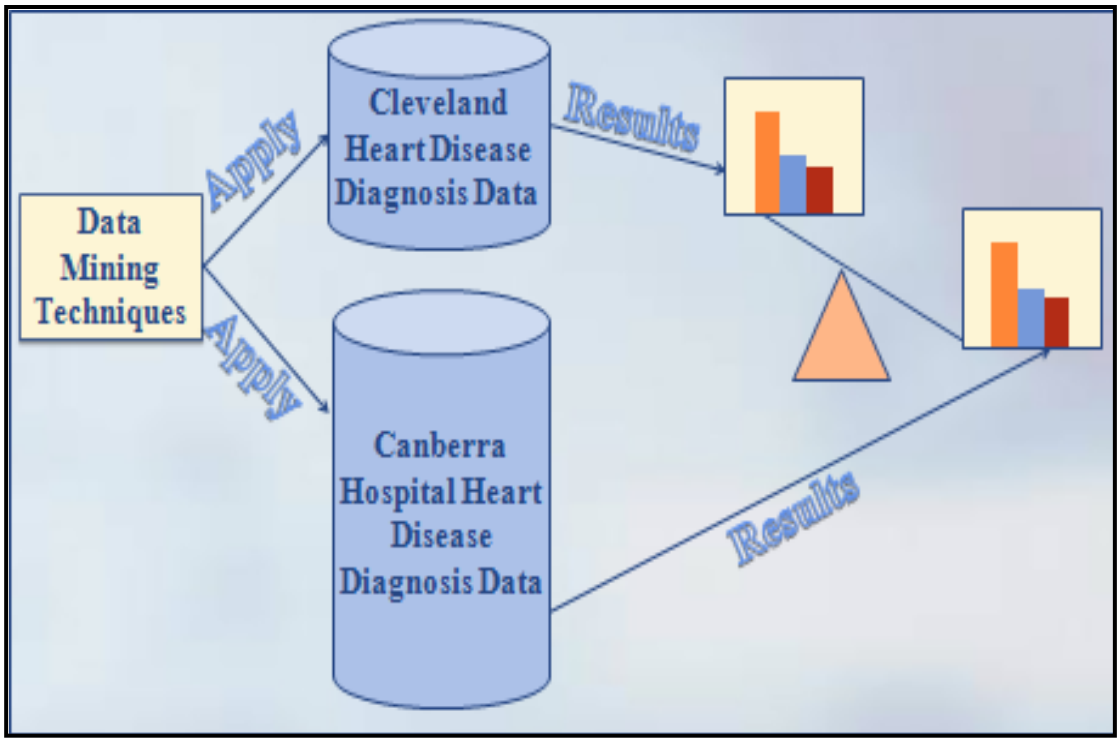

Figure 3.1: Applying Data Mining Techniques on Different Heart Disease Datasets

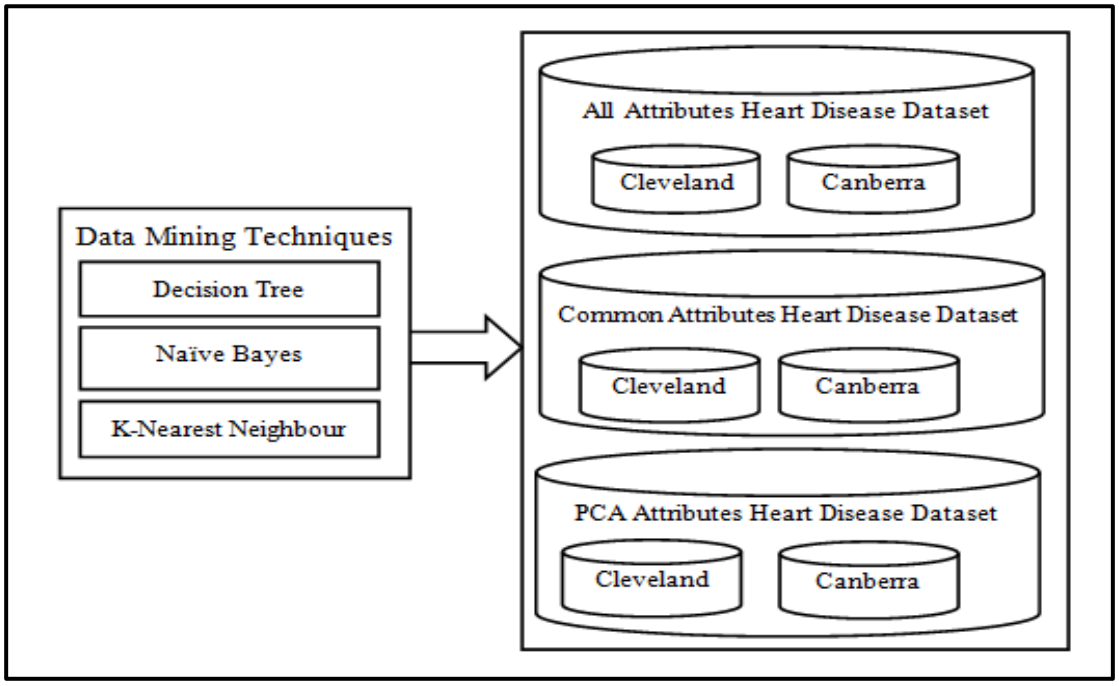

Figure 3.2: Applying Different Data Mining Techniques on Different Heart Disease Datasets Attributes

3.2. Applied Data Mining Techniques

Data attributes of heart disease datasets are divided into categorical (discrete) and continuous data attributes. The discrete data attributes correspond to nominal, binary and ordinal variables. For example, the health status attribute values are high, medium and low risk. Continuous data attributes correspond to integer, interval-scaled and ratio-scaled variables. For example, blood pressure attribute values are continuous values. Some data mining techniques such as Decision Tree and Naïve Bayes cannot deal with continuous attributes (Bramer 2007). Hence continuous attributes need to be converted into discrete ones through a process called discretization (Dougherty, Kohavi et al. 1995).

Discretization methods are classified as supervised or unsupervised (Dougherty, Kohavi et al. 1995). Unsupervised discretization methods do not make use of class membership information during the discretization process (e.g. equal-width interval and equal-frequency discretization methods). Supervised discretization methods use the class labels for carrying out discretization process (e.g. chi-merge and entropy) (Kotsiantis and Kanellopoulos 2006). Dougherty et al. (1995) carried out a comparative study between two unsupervised (equal frequency and equal width) and two supervised (chi-merge and entropy) discretization methods using 16 datasets. They demonstrate that differences between the classification accuracies achieved by different discretization methods are not statistically significant (Dougherty, Kohavi et al. 1995). Different discretization methods involving equal frequency, equal width, chi-merge and entropy were implemented with Decision Tree, Naïve Bayes and K-Nearest Neighbour on the Cleveland heart disease dataset to identify which discretization method would show the better performance. The results show that the equal frequency discretization method is the most reliable (see Appendix B section 1 for more details).

Equal frequency discretization is used as a pre-processing step before applying any of the different data mining techniques to convert the continuous heart disease attributes to discrete attributes. Equal frequency discretization is a popular and successful unsupervised discretization method (Bramer 2007). The equal frequency algorithm determines the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and divides the range into five intervals so that every interval contains the same number of sorted values (Dougherty, Kohavi et al. 1995).

3.2.1. Decision Tree

There are different types of Decision Trees. The difference between them is the mathematical model that is used in selecting the splitting attribute in extracting the Decision Tree rules. The most common and successful type is the gain ratio Decision Tree (Quinlan 1986, Cieslak, Hoens et al. 2012). The Gain Ratio Decision Tree is a relationship between entropy (Information Gain) and splitting information (discussed below).

The entropy (Information Gain) approach selects the splitting attribute that minimizes the value of entropy, thus maximising the Information Gain. To identify the splitting attribute of the Decision Tree, one must calculate the Information Gain for each attribute and then select the attribute that maximizes the Information Gain. The Information Gain for each attribute is calculated using Equation 3.1 (Han and Kamber 2006, Bramer 2007):

$$E = \sum_{i=1}^{k} P_i \log_2 P_i \qquad \text{(Equation 3.1)}$$

Where k is the number of classes of the target attribute

P_i is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring).

The Information Gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values (Han and Kamber 2006). To reduce the effect of the bias resulting from the use of Information Gain, a variant known as Gain Ratio was introduced by the Australian academic Ross Quinlan (Bramer 2007). Gain Ratio adjusts the Information Gain for each attribute to allow for the breadth and uniformity of the attribute values using Equation 3.2.

Gain Ratio = Information Gain / Split Information (Equation 3.2)

Where the split information is a value based on the column sums of the frequency table (Bramer 2007).

After extracting the Decision Tree rules, reduced error pruning is used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules (Esposito,

Malerba et al. 1997). Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

3.2.2. Naïve Bayes

Naïve Bayes classifiers show high accuracy and speed when applied to large datasets (Han and Kamber 2006). The Naïve Bayes technique is easy to implement, not needing any complicated iterative parameter estimation schemes. The Naïve Bayes technique is based on probability theory and seeks to find the most likely possible classification (Wu, Kumar et al. 2007).

Naïve Bayes classifiers are statistical; that is, they can predict class membership on the basis of statistical probabilities (Han and Kamber 2006). Naïve Bayes classification is based on Bayes' theorem. It assumes 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. This assumption is made to simplify the computations involved and, in this sense, is considered "naïve" (Han and Kamber 2006, Bramer 2007).

Naive Bayes classification calculates the prior probability of the target attribute and the conditional probability of the remaining attributes (Han and Kamber 2006, Bramer 2007). For the training data, the prior and conditional probability is calculated. For each testing instance in the testing dataset, the probability is calculated with each of the target attribute values and the target attribute value with the largest probability is then selected. The probability of the testing instance for the target attribute value is calculated using Equation 3.3:

$$P(v = c_i) = P(c_i) \times \sum_{j=1}^{n} P\left(a_j = v_j \middle| \text{class} = c_i\right) \qquad \text{(Equation 3.3)}$$

Where v is the testing instance, ci is the target attribute value, a_j is a data attribute and v_j is its value (Bramer 2007).

3.2.3. K-Nearest Neighbour

K-Nearest Neighbour is one of the most simple and straight forward data mining techniques. It is called a Memory-Based Classification as the training examples need to be in the memory at run-time (Alpaydin 1997).

If a is the first instance denoted by (a_1, a_2, a_3,…a_n) and b is the second instance denoted by (b_1, b_2, b_3, …b_n), the distance between them is calculated using Equation 3.4:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots (a_n - b_n)^2} \qquad \text{(Equation 3.4)}$$

K-Nearest Neighbour usually deals with continuous attributes. However, it can also deal with discrete attributes. When dealing with discrete attributes, if the attribute values for the two instances a_2, b_2 are different, the difference between them is equal to one otherwise it is equal to zero.

A major problem when dealing with the Euclidean distance formula is that the large value frequency swamps the smaller ones. For example, in heart disease records, the cholesterol measure (mg/dl) ranges between 100 and 190 while the age measure (years) ranges between 40 and 80. So, the influence of the cholesterol measure will be greater than the age. To overcome this problem, the continuous attributes are normalized so that they have the same influence on the distance measure between instances. To normalize the value “a1” of the attribute “A” using Equation 3.5:

$$a1 = (a1 - \min) / (\max - \min) \qquad \text{(Equation 3.5)}$$

Where min is the minimum value of the attribute A, and max is the maximum value of the attribute A (Bramer 2007).

Although K-Nearest Neighbour can deal with the continuous and discrete attributes, however all the continues attributes are discretised before applying K-Nearest Neighbour for consistency with the Decision Tree and Naïve Bayes data mining techniques that cannot deal with continuous attributes.

3.3. Cleveland and Canberra Heart Disease Datasets

The benchmark dataset used in this study is the Cleveland Clinic Foundation Heart disease dataset (available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease). The dataset involves 13 data attributes (Table 3.1). The dataset contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiments. The used Cleveland heart disease dataset is shown in Appendix A Table 1.

Table 3.1: Cleveland Heart Disease Data Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-angina pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure in (mm Hg)
Chol	Continuous	Serum cholesterol in (mg/dl)
Fbs	Discrete	Fasting blood sugar > 120 in (mg/dl): 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or defined left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.

Name	Type	Description
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

The comparison dataset used in this study is obtained from the cardiology department of the Canberra Hospital, Canberra, Australia. The dataset involves 13 data attributes (Table 3.2).These data attributes were identified by professional cardiology professor, Leonard Arnolda, as significant attributes for heart disease diagnosis from within the patient records maintained by the cardiology department. The dataset contains 864 rows of which 250 are healthy and 614 are sick. Although the data is taken from cardiology hospital, there are healthy patients that were thought to have heart disease but the medical checks proved that they did not. Data mining techniques need to have proportional balance in the target attribute (Diagnosis) (Han and Kamber 2006). An unbalanced percentage of the target attribute will lead the data mining technique to be more likely to use the high percentage attribute value (Bramer 2007). The benchmark dataset contains 54% records of healthy patients and the remainder of sick patients. So, records were selected from the Canberra dataset to match this proportion for comparison investigations. Random selections from the patient records identified as sick were made to create 460 rows of which 250 are healthy and 210 are sick. Ten different random selections of the main Canberra heart disease dataset (864row) are made to maintain consistency when applying data mining techniques and there is no difference in accuracy found in the different random combinations. The used Canberra heart disease dataset is shown in Appendix A Table 2.

The continuous attributes in both the Cleveland and Canberra heart disease datasets are discretised using the equal frequency discretisation method. The continuous attributes are discretised into five equal frequency discretisation (for more details see Appendix A Table 3).

Table 3.2: Canberra Heart Disease Data Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Postcode	Continuous	Residential Post Code
Height	Continuous	Height in centimetres
Weight	Continuous	Weight in kilogram
Diastole	Continuous	Left Ventricle Diastole in centimetres
Systole	Continuous	Left Ventricle Systole in centimetres
Resting Heart Rate	Continuous	Resting Heart Rate in (bpm)
Peak Heart Rate	Continuous	Peak Heart Rate in (bpm) using treadmill
Resting Blood Pressure High	Continuous	Resting Blood Pressure in (mm Hg)
Resting Blood Pressure Low	Continuous	Resting Blood Pressure in (mm Hg)
Peak Blood Pressure High	Continuous	Peak Blood Pressure in (mm Hg) using treadmill
Peak Blood Pressure Low	Continuous	Peak Blood Pressure in (mm Hg) using treadmill
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

Each of the Cleveland and Canberra heart disease datasets are divided into 10-fold cross-validation training and testing datasets. The sensitivity, specificity, and accuracy for each data mining technique over each of the ten testing datasets of both Cleveland and Canberra heart disease dataset are calculated. Afterwards the mean and

standard deviation of the sensitivity, specificity, and accuracy for each data mining technique is calculated.

3.4. Applying Data Mining Techniques on All Attributes to both Datasets

Decision Tree, Naïve Bayes and K-Nearest Neighbour data mining techniques are applied using all the attributes on the Cleveland heart disease dataset. The different data mining techniques are coded using Visual Studio in a purpose-built analysis program. Different values of K for K-Nearest Neighbour have been used. The mean and standard deviation of sensitivity, specificity, and accuracy in the diagnosis of heart disease using different data mining techniques are shown in Table 3.3. The mean accuracy of the different data mining techniques ranged between 76.5% and 83.5%. Naïve Bayes shows the highest mean accuracy of 83.5% (standard deviation of 5.2%) as shown in Table 3.3.

Table 3.3: Applying Data Mining Techniques on Cleveland Heart Disease Dataset (All Attributes)

Data Mining Technique	**Sensitivity**		**Specificity**		**Accuracy**	
	Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Decision Tree	75.6%	6.1%	81.6%	12.1%	79.1%	5.8%
Naïve Bayes	78%	13.8%	80.8%	12.6%	83.5%	5.2%
KNN K=1	69.2%	16.3%	77.8%	13.5%	76.5%	9.7%
KNN K=9	78.6%	8.9%	84.5%	5.9%	83.4%	2.7%
KNN K=19	76.7%	10.7%	85.1%	7.5%	83.2%	4.1%

Naïve Bayes achieves best results followed by K-Nearest Neighbour and Decision Tree on the Cleveland heart disease data (all attributes). K=9 is showing better accuracy than other values in K-Nearest Neighbour.

The same data mining techniques are applied using all the attributes on the Canberra heart disease dataset. The mean and standard deviation of sensitivity, specificity, and accuracy in the diagnosis of heart disease using different data mining techniques are shown in Table 3.4. The mean accuracy of the different data mining

techniques ranged between 58.8% and 68.9%. Naïve Bayes is showing mean accuracy of 75.4% (standard deviation of 9.1%) as shown in Table 3.4.

Table 3.4: Applying Data Mining Techniques on Canberra Heart Disease Dataset (All Attributes)

Data Mining Technique	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	67.7%	13.6%	64.5%	18%	68.9%	7.8%
Naïve Bayes	65.9%	15.1%	76.8%	23.1%	75.4%	9.1%
KNN K=1	64.2%	19%	57%	17.1%	58.8%	8.6%
KNN K=9	41.9%	20%	85.5%	15.3%	68.8%	8.6%
KNN K=19	45.1%	22.2%	83.5%	15.1%	68.6%	10%

Naïve Bayes achieves the best results followed by Decision Tree and K-Nearest Neighbour on the Canberra heart disease dataset (all attributes). K=9 is showing better accuracy than other values in K-Nearest Neighbour.

Comparison of results from applying different data mining techniques using all the attributes in the Cleveland and Canberra heart disease datasets, shows that different techniques provide reliable results. Some difference in mean accuracy of techniques across the two datasets is evident. The results show the difference in mean accuracy across the two heart disease datasets. Naïve Bayes shows the least difference in mean accuracy of 8.1%, followed by Decision Tree and K-Nearest Neighbour (Table 3.5).

The Cleveland and Canberra heart disease datasets have different data attributes but different techniques still demonstrate reliable performance. (A t-test between different comparative results is presented in the next chapter.) This is evidence that there could be common or corresponding attributes between the two datasets that are influential in diagnosis. It reinforces the strong need for deeper mapping and understanding of the Cleveland and Canberra heart disease dataset attributes.

3.5. Mapping between Cleveland and Canberra Heart Disease Datasets

The Cleveland and Canberra heart disease datasets (see Appendix A) have the same number of attributes (as shown in Table 3.5) but record different data. The datasets have 297 and 460 rows respectively. The percentage of healthy and sick patients in the two datasets is made equivalent as discussed earlier and shown in Table 3.5.

Table 3.5: Cleveland and Canberra Data Analysis

Description	Cleveland	Canberra
No of independent attributes	13	13
No of Rows	297	460
No of Healthy Patients	162 (54%)	250 (54%)
No of Sick Patients	138 (46%)	210 (46%)

Section 3.4 identifies the possibility of common or equivalent attributes. The issue is how the different data attributes are equivalent or corresponding. It is evident that there are three equivalent data attributes (age, sex, and peak heart rate) (Table 3.6). The resting blood pressure in the Cleveland dataset corresponds to the resting blood pressure high and low in the Canberra dataset.

It appeared that there could be some attributes in the Cleveland heart disease dataset that map to other attributes in the Canberra dataset. However, cardiology specialists confirmed four common attributes: age; sex; peak heart rate; and resting blood pressure. From analysis it is evident that there is no mapping between the remaining attributes in both Cleveland and Canberra heart disease datasets. The set of four common attributes raises an important question: “What is the effect of applying different data mining techniques on the four common attributes from the Cleveland and Canberra heart disease datasets?”

Table 3.6: Mapping of Cleveland and Canberra Dataset Attributes

Cleveland Data Attributes	Mapping	Canberra Data Attributes
Age	Same as	Age
Sex		Sex
Peak Heart Rate		Peak Heart Rate
Resting Blood Pressure	Corresponds	Resting Blood Pressure High
		Resting Blood Pressure Low
Resting electrocardiographic	No Mapping	Diastole: Left Ventricle Diastole Pressure
		SYSTOLE: Left Ventricle Systole Pressure
slope of the peak exercise segment		
Exercise induced angina		Resting Heart Rate
Old peak: Depression induced by exercise relative to rest		Peak Blood Pressure High
		Peak Blood Pressure
Chest Pain		Postcode
Cholesterol		Height
Fasting Blood Sugar		Weight
Number of major vessels coloured by fluoroscopy		
Thal		
Diagnosis	Same as	Diagnosis

3.6. Applying Data Mining Techniques on the Common Attributes to both Datasets

As different data mining techniques are showing reliable results across the Cleveland and Canberra datasets, this suggests the possible significant effect of the attributes that are common to both datasets in the diagnosis of heart disease patients. The four common data attributes in each dataset are age, sex, peak heart rate and resting blood

pressure. Section 3.6 describes using Decision Tree, Naïve Bayes and K-Nearest Neighbour data mining techniques over the four common attributes for diagnosis of heart disease patients. The techniques are first applied to the four common attributes of the Cleveland heart disease dataset. The mean and standard deviation of sensitivity, specificity, and accuracy in the diagnosis of heart disease using different data mining techniques are shown in Table 3.7. The mean accuracy ranged between 60.5% and 72.4%. Naïve Bayes shows the highest mean accuracy of 72.4% (standard deviation of 8.2%).

Table 3.7: Applying Data Mining Techniques on Cleveland Heart Disease Dataset (Common Attributes)

Data Mining Technique	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	61.2%	17.6%	76.7%	11.9%	69.6%	7.5%
Naïve Bayes	65%	16.4%	71.2%	12.6%	72.4%	8.2%
KNN K=1	81%	12.2%	41.7%	18.5%	61.2%	10%
KNN K=9	70.1%	12.2%	48.3%	21.1%	60.5%	11.9%
KNN K=19	83%	10.1%	44.2%	19%	63%	10%

Naïve Bayes achieves best results followed by Decision Tree and K-Nearest Neighbour on Cleveland heart disease common data attributes. K=19 shows better accuracy than other values in K-Nearest Neighbour.

The three data mining techniques are then applied to the Canberra heart disease dataset. The mean and standard deviation of sensitivity, specificity, and accuracy of different data mining techniques are shown in Table 3.8. The mean accuracy ranged between 62.2% and 75.1% (Table 3.8). Decision Tree shows the highest mean accuracy of 75.1% (standard deviation of 6.6%) followed by Naïve Bayes and K-Nearest Neighbour. K=1 shows better accuracy than other values in K-Nearest Neighbour.

Comparison of different data mining techniques demonstrates reliable results on Cleveland and Canberra heart disease common data attributes. These results are based on 10-fold cross validation. Table 3.9 shows the difference in mean accuracy of each data mining technique across the two heart disease datasets. Better results are

obtained on the Canberra common data attributes over the Cleveland common data attributes (Table 3.9).

Table 3.8: Applying Data Mining Techniques on Canberra Heart Disease Dataset (Common Attributes)

Data Mining Technique	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	70.4%	14.5%	73.4%	18.8%	75.1%	6.6%
Naïve Bayes	67%	19.1%	72.6%	21.9%	73.3%	6%
KNN K=1	61.5%	17%	71.3%	15%	67.2%	11.7%
KNN K=9	19.3%	10.6%	97%	8%	63.3%	12.3%
KNN K=19	16.6%	10.1%	98.2%	5.4%	62.2%	13.5%

Table 3.9: Comparing Different Data Mining Techniques Accuracy on Cleveland and Canberra Heart Disease Datasets (Common Attributes)

Data Mining Technique	Mean Accuracy		Difference in Mean Accuracy
	Cleveland Common Data Attributes	Canberra Common Data Attributes	
Decision Tree	72.4%	75.1%	2.7%
Naïve Bayes	69.6%	73.3%	3.4%
KNN	63%	67.2%	4.2%

That applying different data mining techniques on the common Cleveland and Canberra heart disease datasets is showing reliable results raises a question of the performance of different data mining techniques across the important attributes in both Cleveland and Canberra heart disease datasets. Principal component analysis (PCA) identifies the significant variables in a large subset of variables (Jolliffe 2002). Jabbar et al., applied PCA as a feature selection method with neural network showing significant result in the diagnosis of heart disease patients (Deekshatulu and Chandra 2013). Giri et

al., used PCA for feature sub-set selection with four different data mining techniques (Support Vector Machine, Gaussian Mixture Model, Probabilistic Neural Network and K-Nearest Neighbor) in Coronary Artery disease detection showing that this integration enhanced data mining techniques accuracy (Giri, Rajendra Acharya et al. 2013). In the next section, the performance of different data mining techniques across the PCA attributes in the Cleveland and Canberra heart disease datasets is discussed.

3.7. Applying Data Mining Techniques on the PCA Attributes to both Datasets

The approach of the principal component analysis (PCA) is essentially to fit a lower-dimensional subspace from a larger one. PCA is a powerful tool for reducing a number of observed variables into a smaller number of artificial variables. PCA identifies the most important variables from all the observed variables by using multivariate data analysis to transform a number of possibly correlated variables into a smaller number of variables (Shlens 2005). PCA is used as a predictor of criterion variables in subsequent analyses (Jolliffe 2002). PCA is one of the most important results from applied linear algebra and one of its most common uses is as the first step in trying to analyse large data sets (Richardson 2009). The weka software is used to extract the PCA attributes for both Cleveland and Canberra heart disease datasets.

The PCA data attributes in the Cleveland heart disease dataset are Thalach, Exang, Oldpeak, Slope, and Thal. Section 3.7 describes applying Decision Tree, Naïve Bayes and K-Nearest Neighbour data mining techniques over the PCA attributes for diagnosis of heart disease patients. The mean and standard deviation of sensitivity, specificity, and accuracy in the diagnosis of the Cleveland heart disease using different data mining techniques are shown in Table 3.10. The mean accuracy ranged between 68.3% and 79.3%.

Naïve Bayes shows the highest mean accuracy of 79.3% (standard deviation of 5.5%) when using PCA attributes as shown in Table 3.11. Naïve Bayes achieves best results followed by K-Nearest Neighbour and Decision Tree on Cleveland heart disease PCA data attributes. K=19 shows better accuracy than other values in K-Nearest Neighbour.

Table 3.10: Applying Data Mining Techniques on Cleveland Heart Disease Dataset (PCA Attributes)

Data Mining Technique	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	62.8%	16.4%	81.7%	12.4%	76.6%	5%
Naïve Bayes	67.2%	28.1%	75.4%	16.5%	79.3%	5.5%
KNN K=1	63.1%	31.5%	53.3%	27.5%	68.3%	8.6%
KNN K=9	65.8%	21.9%	79.8%	9.1%	78.4%	4.1%
KNN K=19	66.3%	22.2%	81.1%	9.2%	79.2%	4.1%

The PCA data attributes in the Canberra heart disease dataset are age, height, weight, peak heart rate, and peak blood pressure high. The three data mining techniques are applied to the Canberra PCA heart disease dataset. The mean and standard deviation of sensitivity, specificity, and accuracy of different data mining techniques are shown in Table 3.11. The mean accuracy ranged between 55.8% and 70.8% (Table 3.11).

Table 3.11: Applying Data Mining Techniques on Canberra Heart Disease Dataset (PCA Attributes)

Data Mining Technique	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	57%	12.9%	75.2%	15.7%	67.4%	10.5%
Naïve Bayes	64.7%	16.1%	71.3%	23%	70.8%	9%
KNN K=1	28.2%	14.3%	93.8%	6.3%	64%	10.5%
KNN K=9	9.5%	7.7%	99.4%	1.8%	58.3%	14.4%
KNN K=19	3%	3.3%	99.2%	1.7%	55.8%	14.5%

Naïve Bayes shows the highest mean accuracy of 70.8% (standard deviation of 9%) followed by Decision Tree and K-Nearest Neighbour as shown in Table 3.12. K=1 shows better accuracy than other values in K-Nearest Neighbour.

Comparison of different data mining techniques demonstrates reliable results on Cleveland and Canberra heart disease PCA data attributes. Table 3.12 shows the difference in mean accuracy of each data mining technique across the two heart disease

datasets. Better results are obtained on the Cleveland PCA data attributes over the Canberra PCA data attributes (Table 3.12).

Table 3.12: Comparing Different Data Mining Techniques Accuracy on Cleveland and Canberra Heart Disease Datasets (PCA Attributes)

Data Mining Technique	Mean Accuracy		Difference in Mean Accuracy
	Cleveland PCA Data Attributes	Canberra PCA Data Attributes	
Decision Tree	76.6%	67.4%	9.2%
Naïve Bayes	79.2%	70.8%	8.4%
KNN	79.3%	64%	15.3%

3.8. Comparing Cleveland and Canberra Different Attributes Combinations

Comparing the different data mining techniques on Cleveland heart disease dataset (including all, common and PCA data attributes), two main issues are discussed here. The first issue is the stability of different data mining techniques across different Cleveland and Canberra heart disease datasets. The stability of data mining technique is its ability to show reliable performance across different datasets as well as across different data attributes. The second issue is the different attributes datasets ability in the diagnosis of heart disease patients.

When investigating the different data mining techniques accuracy on Cleveland and Canberra heart disease dataset (including all, common and PCA data attributes), Naïve Bayes is showing the best mean accuracy on Cleveland (all, common, and PCA) data attributes and Canberra (all and PCA) as shown in Table 3.13. However Decision Tree is showing the best accuracy on the Canberra common data attributes (Table 3.13). K-Nearest Neighbour is showing the least accuracy on Cleveland common data attributes and Canberra (all, common, and PCA) data attributes (Table 3.13).

Table 3.13: Comparing Different Data Mining Techniques Accuracy on Cleveland and Canberra Heart Disease Datasets (All, Common, and PCA) Attributes

Dataset	Data Mining Technique	Mean Accuracy		
		All Data Attributes	Common Data Attributes	PCA Attributes
Cleveland	Decision Tree	79.1%	69.6%	76.6%
	Naïve Bayes	83.5%	72.4%	79.3%
	K-Nearest Neighbour	83.4%	63%	79.3%
Canberra	Decision Tree	68.9%	75.1%	67.4%
	Naïve Bayes	75.4%	73.3%	70.8%
	K-Nearest Neighbour	68.8%	67.2%	64%

When comparing the difference in accuracy between the Decision Tree and Naïve Bayes on the Cleveland and Canberra (all, common and PCA) data attributes, the difference in mean accuracy ranged between 1.8% and 6.5% as shown in Table 3.14. When investigating the t-test significance between the Decision Tree and Naïve Bayes on the Cleveland and Canberra (all, common and PCA) data attributes, the difference for Naïve Bayes or Decision Tree is significance in all cases (Table 3.14). Naïve Bayes is significantly better than Decision Tree on Cleveland (all, common, and PCA) data attributes and Canberra (all and PCA) data attributes while Decision Tree is significantly better than Naïve Bayes on the Cleveland common data attributes.

When comparing the different data mining techniques stability performance on Cleveland heart disease dataset (including all, common and PCA data attributes) the mean accuracy difference ranged between -2.5% and -20.4% (Table 3.15). Comparing all attributes and common data attributes Decision Tree shows the least decrease in accuracy of 9.5%. However, the Naïve Bayes and K-Nearest Neighbour show a decrease of 11.1% and 20.4% respectively (Table 3.15). Comparing all attributes and the PCA attributes Decision Tree shows the least decrease in accuracy of 2.5%. However, both the Naïve Bayes and K-Nearest Neighbour show a decrease of 4.2%

(Table 3.15). Decision Tree is the most stable data mining technique for accuracy followed by Naïve Bayes and K-Nearest Neighbour.

Table 3.14: T-Test Significance between Decision Tree and Naïve Bayes Accuracy on Cleveland and Canberra Datasets (All, Common, and PCA) Attributes

Dataset		**Cleveland**		**Canberra**	
		Decision Tree	**Naïve Bayes**	**Decision Tree**	**Naïve Bayes**
All Data Attributes	Mean Accuracy	79.1%	83.5%	68.9%	75.4%
	Accuracy Difference	4.4%		6.5%	
	T-Test	Yes (t = -9.78, p <= 0.05, degree of freedom=300)		Yes (t = 11.63, p <= 0.05 , degree of freedom = 460)	
Common Data Attributes	Mean Accuracy	69.6%	72.4%	75.1%	73.3%
	Accuracy Difference	2.8%		1.8%	
	T-Test	Yes (t = -4.364, p <= 0.05, degree of freedom=300)		Yes (t = 4.136, p <= 0.05 , degree of freedom = 460)	
PCA Attributes	Mean Accuracy	76.6%	79.3%	67.4%	70.8%
	Accuracy Difference	2.7%		3.4%	
	T-Test	Yes (t = -6.29, p <= 0.05, degree of freedom=300)		Yes (t = -5.27, p <= 0.05 , degree of freedom = 460)	

When comparing the different data mining techniques on Canberra heart disease dataset (including all, common and PCA data attributes) the mean accuracy difference ranged between +6.2% and -5.2% (Table 3.15). Comparing all attributes and common data attributes Decision Tree shows an increase in accuracy of 6.2%. However, the Naïve Bayes and K-Nearest Neighbour show a decrease of 2.1% and 1.6% respectively (Table 3.15). Comparing all attributes and the PCA attributes Decision Tree shows the least decrease in accuracy of 1.5%. The Naïve Bayes and K-Nearest Neighbour show a decrease of 4.6% and 5.2% respectively (Table 3.15). The Decision Tree results are very interesting. It requires just four heart disease attributes for an accuracy of 75.1%. These results imply that for Decision Tree there could be some "noisy" attributes in the Canberra heart disease dataset.

Table 3.15: Comparing Different Data Mining Techniques Mean Accuracy Difference on Cleveland and Canberra Heart Disease Datasets (All, Common, and PCA) Attributes

Dataset	Data Mining Technique	Mean Accuracy			Mean Accuracy Difference	
		All Data Attributes	Common Data Attributes	PCA Attributes	All and Common Data Attributes	All and PCA Data Attributes
Cleveland	Decision Tree	79.1%	69.6%	76.6%	-9.5	-2.5%
	Naïve Bayes	83.5%	72.4%	79.3%	-11.1	-4.2%
	K-Nearest Neighbour	83.4%	63%	79.2%	-20.4	-4.2%
Canberra	Decision Tree	68.9%	75.1%	67.4%	+6.2	-1.5%
	Naïve Bayes	75.4%	73.3%	70.8%	-2.1	-4.6%
	K-Nearest Neighbour	68.8%	67.2%	64%	-1.6	-5.2%

Comparison of different data mining techniques performance over all, common and PCA attributes of the Cleveland and Canberra heart disease datasets shows that the Naïve Bayes achieves the best mean accuracy on each different attribute combination of both Cleveland and Canberra heart disease datasets. However it is also showing that Decision Tree is the most stable data mining technique when compared with Naïve Bayes and K-Nearest Neighbour (Table 3.14).

Although Naïve Bayes is showing the best results on Cleveland (all, common, and PCA) data attributes as well as Canberra (all and PCA) data attributes, Decision Tree is the most stable data mining technique across the two datasets (all, common, and PCA attributes). The difference in mean accuracy between the Decision Tree and Naïve Bayes on the Cleveland and Canberra (all, common and PCA) data attributes ranged between 1.8% and 6.5%. As the Decision Tree technique has unique ability to extract Decision Tree rules from the data and explain how a decision is made and these rules can help healthcare professionals in understanding how can the heart disease dataset

attributes be used in identifying patients at risk of heart disease, the Decision Tree will be used in the further investigation.

When comparing the ability of the different combinations of Cleveland attributes in the diagnosis of heart disease patients, the all attribute combination showed the best results, followed by the PCA attribute combination and then the common attribute combination, as shown in Figure 3.3. When comparing the ability of the different combinations of Canberra attributes in the diagnosis of heart disease patients, the all attribute combination showed the best results, followed by the common attribute combination and then the PCA attribute combination, as shown in Figure 3.4. There is a slight decrease in the performance of different data mining techniques as the number of attributes used is the diagnosis decreases, especially the common attribute combination.

Applying the different data mining techniques to the four data attributes that are common across the Cleveland and Canberra datasets demonstrates reliable results compared to using all of each dataset's attributes. It is very interesting that the common data attributes consist of three non-invasive attributes (age, sex, and resting blood pressure). The fourth attribute, the peak heart rate, is an attribute that requires 'invasive' investigation; in this case, the patient exercises on a tread-mill and the peak heart rate is measured. Although age, sex, and resting blood pressure are known to be significant factors for heart disease, the ability of data mining techniques and especially the Decision Tree to use these attributes in the diagnosis of heart disease patients has not previously been investigated.

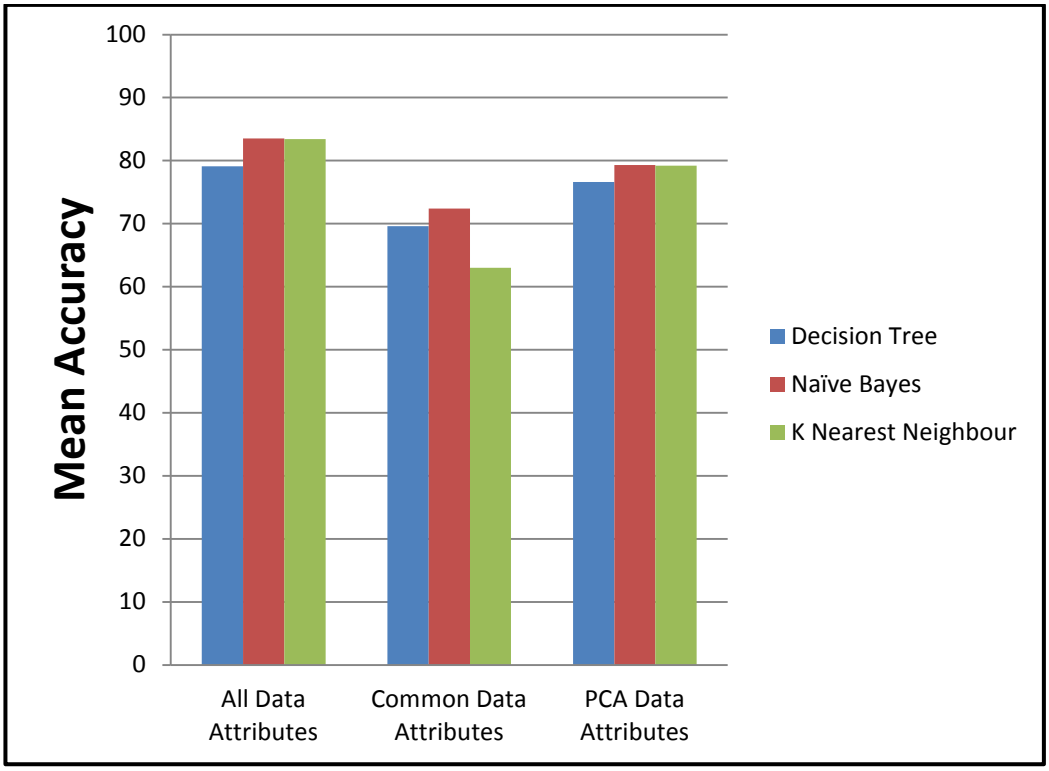

Figure 3.3: Applying Data Mining Techniques on Cleveland Heart Disease Datasets (All, Common, and PCA) Attributes

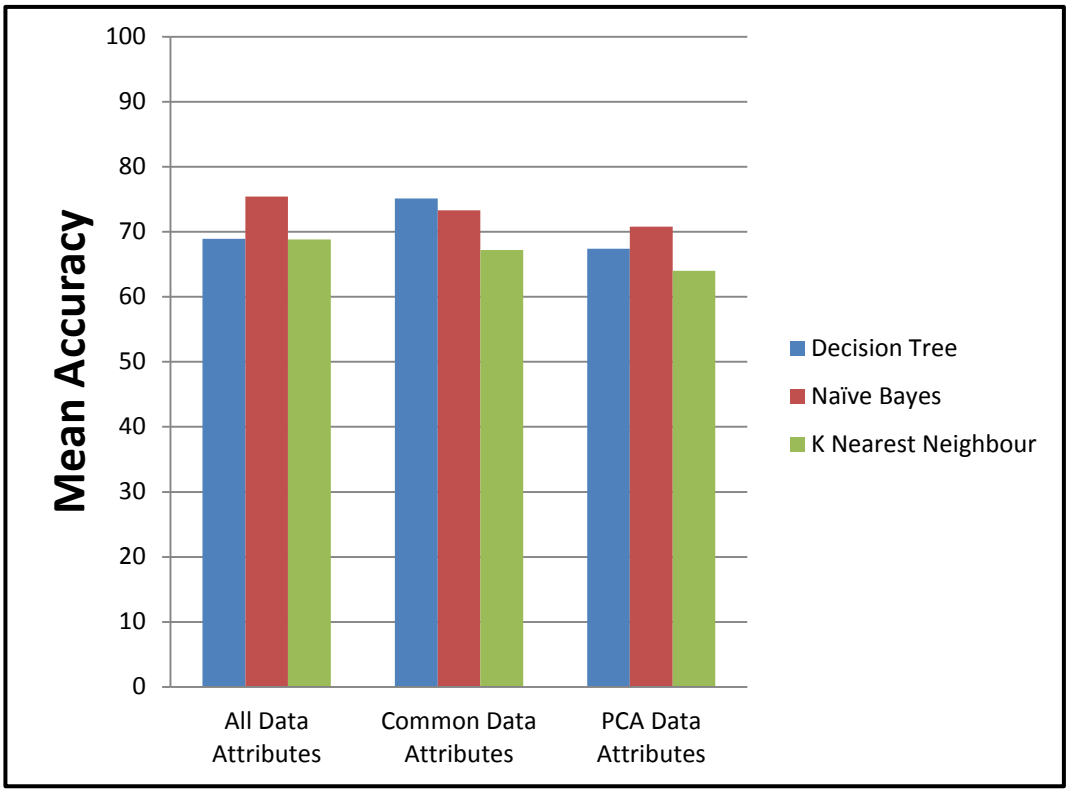

Figure 3.4: Applying Data Mining Techniques on Canberra Heart Disease Datasets (All, Common, and PCA) Attributes

3.9. Chapter Summary and Conclusion

The previous chapter investigated the gaps of applying data mining techniques in the diagnosis of heart disease patients. An important gap is the identification of the significant attributes in the diagnosis of heart disease patients. The single data mining techniques implemented over different datasets show different levels of accuracy. Thus, there are some data characteristics or data attributes that are useful in improving data mining techniques' performance in the diagnosis of heart disease patients.

This chapter presents the result of applying Decision Tree, Naïve Bayes, and K-Nearest Neighbour data mining techniques to all attributes of the Cleveland and Canberra heart disease datasets. These results prove to be reliable. However, the two datasets have different data attributes, so reliability across datasets is interesting. There are four common data attributes (age, sex, resting blood pressure and peak heart rate). Reliable results across datasets imply that these four attributes can be of significant benefit in the diagnosis of heart disease patients.

Applying different data mining techniques to both Cleveland and Canberra heart disease common and PCA data attribute combinations are presented. Comparing the different data mining techniques on Cleveland heart disease dataset (including all, common and PCA data attribute combinations), Decision Tree is the most stable data mining technique for accuracy, followed by Naïve Bayes and K-Nearest Neighbour. Comparing the different data mining techniques on Canberra heart disease dataset (including all, common and PCA data attribute combinations), Decision Tree shows an increase in accuracy of 6.2% when using the four common data attributes as well as showing the least decrease in accuracy of 1.5% when using the PCA attributes. The Decision Tree results are very interesting. The Decision Tree technique requires just four heart disease attributes for an accuracy of 75.1% on the Canberra heart disease dataset.

Decision Tree is the most stable over the four common Cleveland and Canberra dataset attributes, three of which are non-invasive (age, sex, and resting blood pressure). The fourth attribute, the peak heart rate, is an attribute that requires invasive investigation. This raises an important question: "Are there other combinations of non-invasive attributes that can provide (more) reliable performance in the diagnosis of heart

disease patients?". The next chapter investigates applying Decision Tree to identify the non-invasive attribute combination that demonstrates the best performance in the diagnosis of heart disease patients.

[This Page is Left Blank Intentionally]

Chapter 4

Non-Invasive Attributes Significance in Heart Disease Risk Evaluation

4.1. Introduction

There is a need for accurate systematic tools that identify patients at high risk and provide information for early intervention in diseases (Paladugu and Shyu 2010). Heart disease can be detected by several tests, for example, electrocardiogram, stress tests, and cardiac angiogram. However, these tests are expensive and cannot be used as community screening tests (Providence Heart and Vascular Institute 2014). Thus there is an identified need to find cheaper tests that can be applied to community screening to identify heart disease risk.

In the previous chapter, different data mining techniques are tested using all data attributes in the Cleveland and Canberra heart disease datasets. Some show reliable results over both Cleveland and Canberra heart disease datasets but the two datasets have different data attributes. There are four common data attributes (age, sex, resting blood pressure and peak heart rate). Decision Tree demonstrates the best stability with the Cleveland and Canberra heart disease datasets using all attributes and only the common data attributes. Results for the four common attributes, comprising three non-invasive attributes, show the probable significance of non-invasive attributes in the risk evaluation of heart disease patients. As discussed earlier in Chapter 2 (Section 2.2.4) the Framingham test and the Australian Absolute Cardiovascular Risk Calculator use a set of invasive and non-invasive attributes in the risk evaluation of heart disease. The invasive attributes require various data that results from blood tests prior to using the evaluation tool. These scores may be difficult to implement where there are limited resources available. Hence, there is a need to simplify the heart disease risk evaluation tool attributes so that affordable detection strategies can be implemented (Bitton and Gaziano 2010). This need raises a question: “Can combinations of non-invasive attributes provide (more) reliable performance in the diagnosis of patients at risk of heart disease.”

This chapter applies Decision Tree to identify the non-invasive attributes combination that will show the best performance in the diagnosis of heart disease patients. The non-invasive attributes are advantageous because they are low cost attributes. The Decision Tree used on the Cleveland and Canberra heart disease datasets is based on 10 fold cross validation as described earlier in Chapter 3. Hence the mean and standard deviation of sensitivity, specificity, and accuracy is calculated on the 10 samples of the datasets. The importance of non-invasive attributes is explored. The effect of using single non-invasive Cleveland and Canberra datasets attributes is tested (see Figure 4.1). Different combinations of non-invasive attributes from the Cleveland and Canberra datasets are investigated as input to Decision Tree data mining classification for performance evaluation. Equations using different non-invasive attributes are developed and examined to determine if they enhance Decision Tree performance in the diagnosis of heart disease patients (see Figure 4.1).

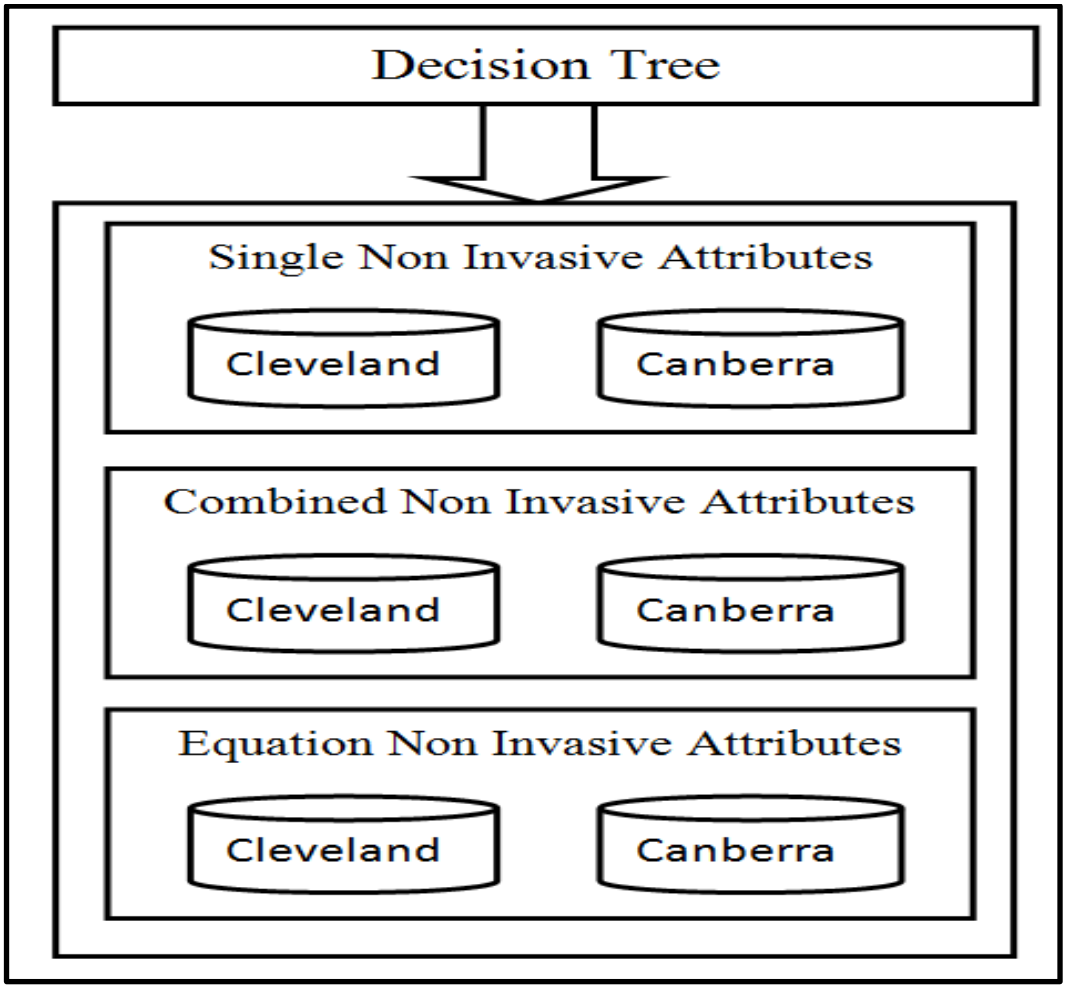

Figure 4.1: Applying Decision Tree to Heart Disease Datasets Non Invasive Attributes

4.2. The Importance of Non-Invasive Attributes

Section 4.1 describes how heart disease can be detected by several tests. However, these tests are very expensive and are not suitable as community screening tests. There is a strong need to find cheaper tests suitable for community screening to identify the risk of heart disease (Providence Heart and Vascular Institute 2014). Accurate systematic tools that identify patients at high risk and provide information to enable early preventative intervention are required (Paladugu and Shyu 2010). Combinations of non-invasive attributes from two datasets (Cleveland and Canberra) provide a basis for examination of their potential.

The non-invasive attributes are those that can be identified easily without complex machines and instruments that need to be done in a hospital; for example: age, sex, height, weight, smoking habits, and resting blood pressure. Although weight needs a scale to be measured and resting blood pressure needs a blood pressure monitor to be measured, these tools can be available at home or in a pharmacy and do not need a hospital to be measured or physical samples to be taken. These attributes can be collected quite cheaply. Age, sex, blood pressure, and smoking habits are major risk factors for developing heart disease (Cupples and D'Agostino 1987). Can these non-invasive attributes be used with data mining techniques to cost effectively identify patients at risk of heart disease?

Researchers and practitioners use invasive attributes such as total and HDL cholesterol and diabetes (usually measured through blood sugar or insulin levels) as well as non-invasive attributes such as age, sex, smoking, and resting blood pressure to identify patients at risk of heart disease. The Framingham Study (2013) and Australian Absolute Cardiovascular Risk Calculator use a combination of invasive and non-invasive attributes in the diagnosis of heart disease patients: age, sex, diabetes, systolic blood pressure, smoking status, BMI, total cholesterol, and HDL cholesterol (Framingham Study. 2013); and age, sex, systolic blood pressure, smoking status, total cholesterol, HDL cholesterol, and diabetes (National Heart Foundation of Australia 2009) respectively.

However, the performance using only non-invasive attributes has not previously been assessed. How accurately can data mining techniques using only non-

invasive attributes classify patients as sick or healthy? Success here would provide a great opportunity for application to community screening tests, thus enabling early intervention in patients at high risk of heart disease and determining suitable treatment regimes for those patients.

4.3. Single Non-Invasive Attributes for Cleveland and Canberra Heart Disease Risk Evaluation

Section 4.2 identified that data mining techniques using non-invasive attributes have not previously been investigated. In this section, the results of applying Decision Tree data mining techniques using only single non-invasive attributes are presented. Equal frequency discretization is used as a pre-processing step before applying the data mining technique to convert the continuous heart disease attributes to discrete attributes as described in section 3.2.

The Cleveland dataset contains three non-invasive attributes: age, sex, and resting blood pressure; and the Canberra heart disease dataset contains five non-invasive attributes: age, sex, resting blood pressure, height, and weight. Table 4.1 shows the mean and standard deviation (St Dev) of sensitivity, specificity, and accuracy of Decision Tree using single non-invasive attributes of the Cleveland dataset where the accuracy ranges between 61.5% and 55.3%. The sex attribute shows best mean accuracy followed by age and resting blood.

Table 4.1: Applying Decision Tree on Single Non-Invasive Cleveland Heart Disease Data Attributes

Cleveland Data Attributes	**Sensitivity**		**Specificity**		**Accuracy**	
	Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Age	55.6%	15.7%	62.8%	20.1%	56%	7.9%
Sex	83%	10.1%	41.7%	18.5%	61.5%	10.1%
Resting Blood Pressure	26.4%	13.3%	79.5%	9.4%	55.3%	12%

Table 4.2 shows the mean and standard deviation of sensitivity, specificity, and accuracy for single non-invasive attributes of the Canberra dataset with accuracy ranging between 71% and 46.1%. The age attribute shows the best mean accuracy followed by sex, resting blood pressure, height, and weight.

Table 4.2: Applying Decision Tree on Single Non-Invasive Canberra Heart Disease Data Attributes

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Age	61.6%	16.5%	75.3%	19.7%	71%	8.2%
Sex	67.4%	14.7%	58.4%	16.3%	66%	8.3%
Resting Blood Pressure	46.6%	11.2%	63.6%	7.3%	55.1%	5.9%
Height	45.7%	18.2%	59.7%	7.9%	50.8%	10.9%
Weight	24.2%	26.8%	71.5%	26.5%	46.1%	9.3%

On both Cleveland and Canberra heart disease datasets, the age and sex single data attributes show the best results among other single non-invasive attributes, with the age attribute attaining 56% and 71% mean accuracy respectively. The sex attribute achieves 61.5% and 66% mean respectively. What is the effect on accuracy of combined non-invasive attributes?

4.4. Different Combinations of Non-Invasive Attributes for Cleveland and Canberra Heart Disease Risk Evaluation

Section 4.4 investigates the performance of combined non-invasive attributes in the diagnosis of Cleveland and Canberra heart disease patients. Table 4.3 shows the performance of different combinations of the non-invasive attributes in diagnosis using the Cleveland heart disease dataset. The combinations of age, sex, and resting blood pressure show best mean accuracy of 65.8% followed by age and sex combination showing mean accuracy of 65.2%.

Table 4.3: Applying Decision Tree on Combined Non-Invasive Cleveland Heart Disease Data Attributes

No of Attributes	Cleveland Data Attributes	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Two	Age, Sex	60.4%	9.1%	67.5%	9.3%	65.2%	7.5%
	Age, RBP	46.6%	12.8%	71.3%	14%	58.3%	9.3%
	Sex, RBP	69%	15.4%	54.2%	13%	60%	8.9%
Three	Age, Sex, RBP	61.2%	10.1%	69.5%	10.1%	65.8%	8.6%

The difference of the mean accuracy between the age and sex combination and age, sex, and resting blood pressure combination is just 0.6%. So, t-test for significance is applied to identify if there is a significant difference between the two combinations. The T-Test shows that there is no significant difference between the two combinations (see Table 4.4). The sensitivity measure is the true positive, meaning sick patients that are identified as sick. The specificity measure is the true negative, meaning healthy patients that are identified as healthy. In this context sensitivity is the most useful in identifying sick patients to ensure appropriate care. Thus the age, sex and resting blood pressure combination demonstrates best results with the highest mean accuracy in the diagnosis of Cleveland heart disease dataset.

Table 4.4: T-Test Significance between Non-Invasive Cleveland Heart Disease Data Combinations

Cleveland Data Attributes	Accuracy			Sensitivity		
	Mean	St Dev	T-Test Significance	Mean	St Dev	T-Test Significance
Age, Sex	65.2%	7.5%	No (t = -0.911, p <= 0.05, degree of freedom = 300)	60.4%	9.1%	No (t = -1.262, p <= 0.05, degree of freedom = 300)
Age, Sex, RBP	65.8%	8.6%		61.2%	10.1%	

The Canberra heart disease dataset contains five non-invasive attributes: age, sex, resting blood pressure (Systolic and Diastolic), height, and weight. Table 4.5 shows

the performance of different combinations of these non-invasive attributes in diagnosis using the Canberra heart disease dataset. The combination of resting blood pressure and height shows best mean accuracy of 79.2%. However, the mean sensitivity of this combination is 35%, a very small value - this combination result is discarded.

The combination of age, sex, and resting blood pressure shows best mean accuracy of 74.8% followed by age and sex with mean accuracy of 74.2%. The difference of the mean accuracy between the age and sex combination and the age, sex, and resting blood pressure combination is just 0.6%. T-Test significance is applied to determine if there is a significant difference between the two combinations. There are no significant differences between the two combinations (see Table 4.6). Applying T-Test between the sensitivity of the two combinations, there is no significant difference (Table 4.6). Although the T-Test significance did not provide clear resolution, age, sex, and resting blood pressure combination is selected because it shows better mean accuracy and mean sensitivity with lower standard deviations. Thus age, sex, and resting blood pressure combination shows better results than other combinations in the diagnosis of Canberra heart disease dataset.

Table 4.5: Applying Decision Tree on Combined Non-Invasive Canberra Heart Disease Data Attributes

No of Attributes	Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Two	Age, Sex	66.3%	14%	73.8%	21.4%	74.2%	7.5%
	Age, RBP	58.8%	15.7%	70.2%	15.3%	65.7%	7.1%
	Age, Height	71%	11.3%	69.9%	19%	72.1%	8.3%
	Age, Weight	67.5%	12.2%	70.1%	19.2%	69.6%	10.3%
	Sex, RBP	54%	9.8%	66.2%	7.7%	61.3%	6.8%
	Sex, Height	66.5%	13.7%	58.4%	16.3%	65.5%	8%
	Sex, Weight	54.3%	9%	72%	13.9%	66.2%	7.7%
	RBP, Height	35%	9.6%	61.3%	11.7%	49.2%	8.9%
	RBP, Weight	38.4%	11.2%	68.4%	14.9%	53.8%	7.4%
	Height, Weight	43.3%	11.5%	63.1%	6.7%	53.3%	9.8%

No of Attributes	Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Three	Age, Sex, RBP	66.7%	13.3%	76.5%	13.7%	74.8%	6.5%
	Age, Sex, Height	26.2%	22.4%	87.3%	12.3%	62.9%	8.3%
	Age, Sex, Weight	42.1%	16%	89%	6.4%	68.2%	12.2%
	Age, RBP, Height	60.7%	10.9%	69.4%	18%	67.1%	8.6%
	Age, RBP, Weight	59.9%	13%	70.2%	17.2%	66.3%	9.7%
	Age, Height, Weight	42.9%	7.8%	80.5%	16.6%	64.2%	7.4%
Four	Age, Sex, RBP, Height	61%	14.1%	73.7%	11.2%	70.4%	7.2%
	Age, Sex, RBP, Weight	60.9%	13%	72.6%	13.4%	69.7%	6.5%
	Age, Sex, Height, Weight	51.9%	9.3%	78.7%	12.8%	68.8%	7.1%
	Age, RBP, Height, Weight	63.7%	10.3%	67.8%	19%	67.3%	9.7%
	Sex, RBP, Height, Weight	53.6%	9.7%	73.8%	8.9%	65.4%	6.6%
Five	Age, Sex, RBP, Height, Weight	59.5%	13.4%	72.9%	11.7%	69%	6.6%

Table 4.6: T-Test Significance between Non-Invasive Canberra Heart Disease Data Combinations

Canberra Data Attributes	Accuracy			Sensitivity		
	Mean	St Dev	T-Test Significance	Mean	St Dev	T-Test Significance
Age, Sex	74.2%	7.5%	No ($t = -1.448$, $p <= 0.05$, degree of freedom = 460)	66.3%	14%	No ($t = -0.444$, $p <= 0.05$, degree of freedom = 460)
Age, Sex, RBP	74.8%	6.5%		66.7%	13.3%	

Including the height and weight attributes with the age, sex, and resting blood pressure non-invasive attributes combination, shows decrease in the mean accuracy to 69% (Table 4.5). Does converting height and weight into body mass index (BMI) or Rohrer's Index (RI) equation enhance the accuracy in the diagnosis of Canberra heart disease patients? The resting blood pressure attribute also includes both the high (Systolic) and low (Diastolic) resting blood pressures, so further analysis is needed to determine if using the difference between the resting blood pressure Systolic and Diastolic (called 'pulse pressure') with the age, sex, and resting blood pressures in combination can enhance accuracy. Hence pulse pressure use as a non-invasive attribute in the risk evaluation of heart disease needs further investigation.

The benchmark Cleveland heart disease dataset does not contain the height and weight attributes, so it is not possible to apply the BMI or Rohrer's Index equation to the Cleveland dataset. This dataset has only high resting blood pressure and not low resting blood pressure, so it will not be possible to apply the 'pulse pressure' equation on the Cleveland dataset. Section 4.5 provides the results of applying the BMI, Rohrer's Index and 'pulse pressure' to the Canberra heart disease dataset.

4.5. Different Equations of Non-Invasive Attributes for Canberra Heart Disease Risk Evaluation

The Canberra dataset non-invasive attributes combinations (age, sex, and resting blood pressure) shows best results in heart disease diagnosis (see Section 4.4). Adding height and weight attributes to this combination decreases the mean accuracy. Section 4.5 investigates including height and weight attributes to calculate BMI and RI equations. The impact of BMI and RI equations in combination with age, sex, and resting blood pressure attribute combinations on mean accuracy in the diagnosis of heart disease patients is discussed. Also investigating if adding the 'pulse pressure' to different non-invasive attribute combinations can enhance mean accuracy in the diagnosis of heart disease patients is presented.

Researchers have compared the RI to the BMI in its ability to predict body fat levels. One study suggested that RI may be a much better choice than BMI at assessing adults overweight (Valdez, Greenlund et al. 1996). However another study found that age specific BMI is better than age specific RI in predicting underweight or overweight

(Mei, Grummer-Strawn et al. 2002). There is a need to identify how BMI and RI can enhance the ability of Decision Tree in the risk evaluation of heart disease.

Equation 4.1 uses the weight and height to calculate the BMI (Sultan, AlObaidy et al. 2009).

$$BMI = Weight\ [kg] / (Height\ [m])^2 \quad \text{(Equation 4.1)}$$

Table 4.7 shows the mean and standard deviation of sensitivity, specificity, and accuracy resulting from adding the BMI equation to different combinations of age, sex, and resting blood pressure non-invasive Canberra heart disease attributes. Integrating BMI with age, sex, and resting blood pressure does not enhance Decision Tree accuracy in the diagnosis of Canberra heart disease patients. The age, sex, resting blood pressure and BMI combination shows mean accuracy and mean sensitivity of 74% and 66.5% respectively. The age, sex, and resting blood pressure combination is showing mean accuracy and mean sensitivity of 74.8% and 66.7 % respectively (see Table 4.7). However, age, sex, resting blood pressure and BMI combination shows more stability in the standard deviation of the accuracy (5%) which is better than the standard deviation of the accuracy of the age, sex, and resting blood pressure combination (6.5%).

Table 4.7: Integrating BMI with Different Non-Invasive Canberra Heart Disease Data Attributes

Canberra Attributes	**Sensitivity**		**Specificity**		**Accuracy**	
	Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Age, Sex, RBP	66.7%	13.3%	76.5%	13.7%	74.8%	6.5%
Age, BMI	61.6%	12.8%	72.4%	16.1%	66.6%	8.6%
Sex, BMI	60.9%	14.2%	69.5%	11.9%	68.8%	6.1%
RBP, BMI	48.8%	10.8%	66.8%	5.2%	59.1%	3.8%
Age, Sex, BMI	24%	17.1%	93.7%	8.1%	62.4%	9.5%
Age, RBP, BMI	57%	14.1%	74.8%	16%	67.4%	9.3%
Sex, RBP, BMI	38.3%	12.4%	75.6%	11.2%	61.2%	6%
Age, Sex, RBP, BMI	66.5%	16.4%	76.1%	10.1%	74%	5%

Rohrer's Index is a measure of leanness of a person calculated as a relationship between mass and height. It was first proposed in 1921 as the "Corpulence Index" by Rohrer and hence also known as Rohrer's Index. It is similar to the body mass index, but the mass is normalized with the third power of body height rather than the second power (Ensminger and Ensminger 1993). Equation 4.2 uses the weight and height to calculate the Rohrer's Index.

$$\text{Rohrer's Index} = \text{Weight [kg]} / (\text{Height [m]})^3 \qquad \text{(Equation 4.2)}$$

Table 4.8 shows the mean and standard deviation of sensitivity, specificity, and accuracy when adding the Rohrer's Index equation to different combinations of age, sex, and resting blood pressure non-invasive Canberra heart disease attributes. Integrating the Rohrer's Index equation with the age, sex, and resting blood pressure attribute combination does not enhance Decision Tree accuracy in the diagnosis of Canberra heart disease patients. However, the mean and standard deviation of the sensitivity and the standard deviation of the accuracy are enhanced. The age, sex, resting blood pressure and Rohrer's Index combination shows mean and standard deviation of the accuracy of 73.8% and 4.9% respectively and mean and standard deviation of the sensitivity of 67.1% and 11.4 % respectively (see Table 4.8).

Table 4.8: Integrating Rohrer's Index with Different Non-Invasive Canberra Heart Disease Data Attributes

Canberra Attributes	**Sensitivity**		**Specificity**		**Accuracy**	
	Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Age, Sex, RBP	66.7%	13.3%	76.5%	13.7%	74.8%	6.5%
Age, Rohrer's Index	66.8%	12.7%	71%	19.1%	70%	6.7%
Sex, Rohrer's Index	59.8%	12.2%	62.4%	15.7%	64.4%	6.5%
RBP, Rohrer's Index	55%	12.1%	70.1%	10.9%	61.9%	6.5%
Age, Sex, Rohrer's Index	33.7%	16.6%	91.8%	4.8%	66.3%	12%
Age, RBP, Rohrer's Index	59.3%	13.3%	74.9%	16.6%	69.1%	9.2%
Sex, RBP, Rohrer's Index	43.4%	10.7%	74%	6.4%	61.2%	8.2%

Canberra Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, Sex, RBP, Rohrer's Index	67.1%	11.4%	75.5%	9.2%	73.8%	4.9%

The resting blood pressure attribute also includes both the high and low resting blood pressures, so further analysis is needed to determine if using the difference between the resting blood pressure high and low (called 'pulse pressure') with the age, sex, and resting blood pressure combination can enhance accuracy. The WHO reports suggests that the pulse pressure can be used as an indicator of heart disease (World Health Organization 2005). Table 4.9 shows the mean and standard deviation of sensitivity, specificity, and accuracy when adding the RBPDiff ('Pulse pressure') Equation (Equation. 4.3) to combinations of age, sex, and resting blood pressure attributes. The RBPDiff equation used in this investigated is:

RBPDiff = RBP High – RBP Low (Equation 4.3)

Integrating RBPDiff with age, sex, and resting blood pressure attributes does not enhance Decision Tree accuracy in the diagnosis of Canberra heart disease patients. The age, sex, resting blood pressure and RBPDiff combination shows mean accuracy and standard deviation of 72.6% and 5.2% respectively, and mean sensitivity and standard deviation of 65.3% and 14.9% respectively (see Table 4.9). Adding the RBPDiff with the age, sex, RBP, and BMI combination and age, sex, RBP, and Rohrer's Index combination does not enhance performance showing mean accuracy of 72.5% and 74.7% respectively (Table 4.9).

Table 4.9: Integrating RBPDiff with Different Non-Invasive Canberra Heart Disease Data Attributes

Canberra Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, Sex, RBP	66.7%	13.3%	76.5%	13.7%	74.8%	6.5%
Age, RBPDiff	65%	15.6%	70.4%	19.2%	68%	9.4%
Sex, RBPDiff	62.2%	13%	60.3%	16.5%	63.1%	8.3%

Canberra Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
RBP, RBPDiff	33.7%	7.7%	67.3%	10.7%	50.7%	6.2%
Age, Sex, RBPDiff	39.3%	20.1%	89.6%	10.3%	68.6%	8.4%
Age, RBP, RBPDiff	56.2%	13%	73.5%	17.4%	67%	7.5%
Sex, RBP, RBPDiff	26.6%	8.9%	83.5%	6.5%	57.7%	11.9%
Age, Sex, RBP, RBPDiff	65.3%	14.9%	73.4%	13.3%	72.6%	5.2%
Age, Sex, RBP, BMI, RBPDiff	63.9%	15.6%	75.1%	12.5%	72.5%	6.5%
Age, Sex, RBP, Rohrer's Index, RBPDiff	67.6%	15.6%	76%	10.1%	74.7%	6.3%

Adding the different BMI, Rohrer's Index, and RBPDiff equations with age, sex, and resting blood pressure attributes combinations, the mean accuracy is not enhanced. However, the standard deviation is decreased demonstrating better stability (see Table 4.10). In this context, sensitivity is more important because patients who are at high risk of heart disease need to be identified and get appropriate care. The age, sex, RBP, and Rohrer's Index attributes combination shows the best mean and standard deviation sensitivity followed by age, sex, and RBP combination (67.1%, 11.4% and 66.7%, 13.3% respectively – see Table 4.10).

Table 4.10: Summarizing Integrating BMI, Rohrer's Index, and RBPDiff with Non-Invasive Canberra Heart Disease Data Attributes

Canberra Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, Sex, RBP	66.7%	13.3%	76.5%	13.7%	74.8%	6.5%
Age, Sex, RBP, BMI	66.5%	16.4%	76.1%	10.1%	74%	5%
Age, Sex, RBP, Rohrer's Index	67.1%	11.4%	75.5%	9.2%	73.8%	4.9%
Age, Sex, RBP, RBPDiff	65.3%	14.9%	73.4%	13.3%	72.6%	5.2%

The T-Test is applied to the accuracy of the two combinations (age, sex, and RBP combination and age, sex, RBP, and Rohrer's Index combination) to identify if there is significant difference and shows significant difference between them (see Table 4.11). However, applying T-Test to the sensitivity of the two combinations shows that there is no significant difference (Table 4.11). In this context, increasing the mean sensitivity and decreasing its standard deviation is more important. The age, sex, RBP, and Rohrer's Index combination shows better results than other non-invasive attributes combinations for the Canberra heart disease dataset.

Table 4.11: T-Test Significance for Adding Rohrer's Index to Non-Invasive Canberra Heart Disease Data Attributes

Canberra Data Attributes	**Accuracy**			**Sensitivity**		
	Mean	**St Dev**	**T-Test Significance**	**Mean**	**St Dev**	**T-Test Significance**
Age, Sex, RBP	74.8%	6.5%	Yes (t = -2.635, p <= 0.05, degree of freedom = 460)	66.7%	13.3%	No (t = 0.490, p <= 0.05, degree of freedom = 460)
Age, Sex, RBP, Rohrer's Index	73.8%	4.9%		67.1%	11.4%	

The Decision Tree rules for the Canberra non-invasive attributes (age, sex, systolic, diastolic, and RI) are extracted to help healthcare professional in understanding how this non-invasive attributes combination are used by Decision Tree in the diagnosis of heart disease patients achieving mean sensitivity of 67.1% (Table 4.11). Figure 4.2 shows a subset of the Decision Tree rules extracted from Canberra non-invasive attributes. All rules are presented in Appendix C. Figure 4.3 shows the attributes' cut-points and the range for each cut-point. The attributes cut-points are used by the Decision Tree method to extract the decision tree rules that applies to Figure 4.2, 4.4, and 4.5.

Decision Tree Rules

If SEX = 0, Age = 0, Systolic= 0, Diastolic = 3 => Then The Target Value Equals 0
If SEX = 0, Age = 0 , Systolic= 0 , Diastolic = 0 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 2 , RI = 3 => Then The Target Value Equals 1
If SEX = 0 , Age = 0 , Systolic= 2 , RI = 2 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 3 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 4 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 1 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 2 => Then The Target Value Equals 1
If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 1 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 0 => Then The Target Value Equals 1
If SEX = 0 , Age = 0 , Systolic= 2 , RI = 1 , Diastolic = 1 => Then The Target Value Equals 0
If SEX = 0 , Age = 0 , Systolic= 2 , RI = 1 , Diastolic = 4 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 2 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 0 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 4 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 3 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 0 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 4 , Systolic= 4 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 4 , Systolic= 0 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 4 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 2 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 0 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 3 => Then The Target Value Equals 1
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 1 => Then The Target Value Equals 0
If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 0 => Then The Target Value Equals 0

Figure 4.2: Sample of Canberra Non-Invasive Decision Tree Rules

The Decision Tree rules extracted from the Canberra non-invasive attributes are used to draw the Decision Tree presented in Figure 4.4 and 4.5. Figure 4.4 represents the Decision Tree for Sex=0 (Male) while Figure 4.5 represents the Decision Tree for Sex=1 (Female). For Figure 4.4 and 4.5 the sex is the root node for a single decision tree. These Decision Trees can help healthcare professional in the diagnosis of heart disease patients using non-invasive attributes. These Decision Trees are very helpful in the diagnosis of heart disease patients. These Decision Trees can evaluate the risk of heart disease using non-invasive attributes (age, sex, systolic, diastolic, and RI)

Age 0: < 65 1: ≥ 65 & < 73 2: ≥ 73 & < 79 3: ≥ 79 & < 86 4: ≥ 86	**Rohrer's Index** 0: < 14 1: ≥ 14 & < 16 2: ≥ 16 & < 17 3: ≥ 17 & < 19 4: ≥ 19
Systolic 0: < 122 1: ≥ 122 & < 134 2: ≥ 134 & < 143 3: ≥ 143 & < 154 4: ≥ 154	**Diastolic** 0: < 66 1: ≥ 66 & < 74 2: ≥ 74 & < 80 3: ≥ 80 & < 90 4: ≥ 90

Figure 4.3: Attribute cut-point and range in Heart Disease Decision Tree Rules

For example, if the sex = male, age = 60, systolic =140, diastolic = 85, RI = 18 then to evaluate the degree of risk of heart disease for this patient, first the non-invasive attributes need to be scaled using Figure 4.3. Then sex = 0, age = 0, sys = 2, dys = 3, and RI = 3. As the sex = 0 then Figure 4.4 will be used. When applying the scaled values of the non-invasive attributes to the Decision Tree in Figure 4.4, then the degree of risk will be yes.

Another example, if the sex = female, age = 80, systolic = 150, diastolic = 85, RI = 13 then to evaluate the degree of risk of heart disease for this patient, first the non-invasive attributes need to be scaled using Figure 4.3. Then sex = 1, age = 3, sys =3, dys = 3, and RI = 0. As the sex = 1 then Figure 4.5 will be used. When applying the scaled values of the non-invasive attributes to the Decision Tree in Figure 4.5, then the degree of risk will be no.

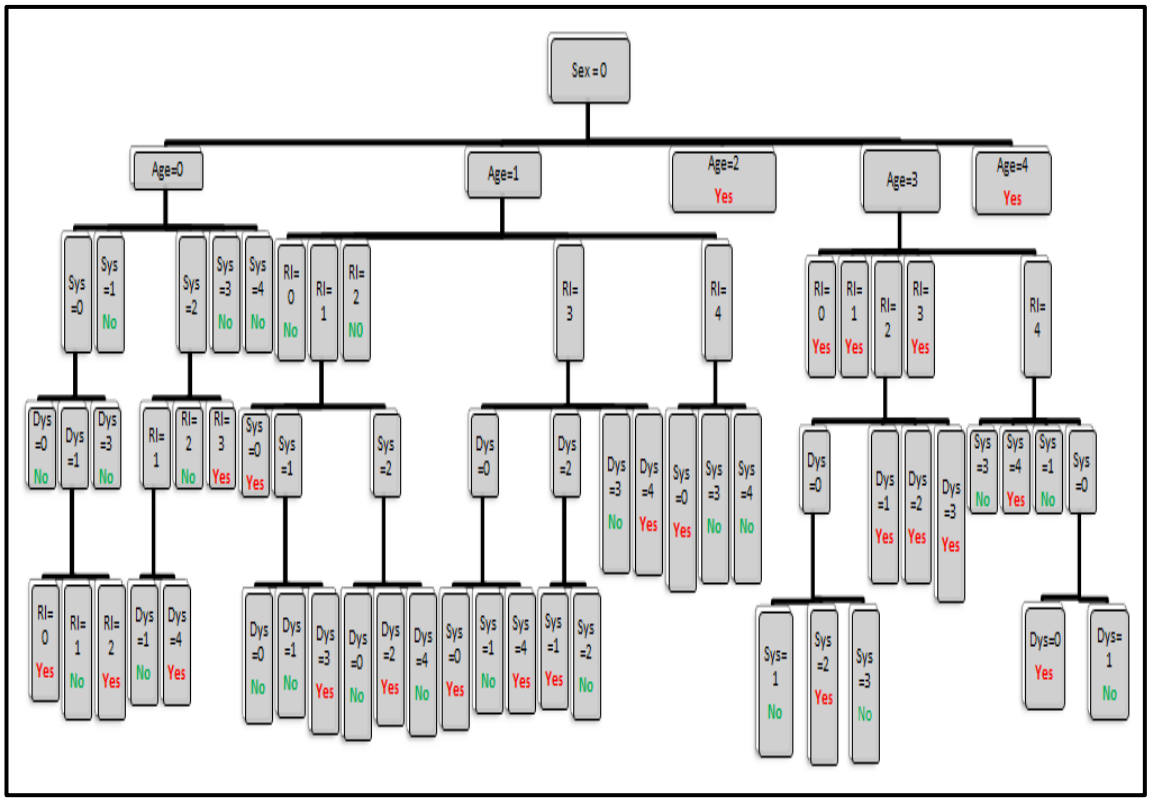

Figure 4.4: Male Decision Tree Using Non-Invasive Canberra Attributes

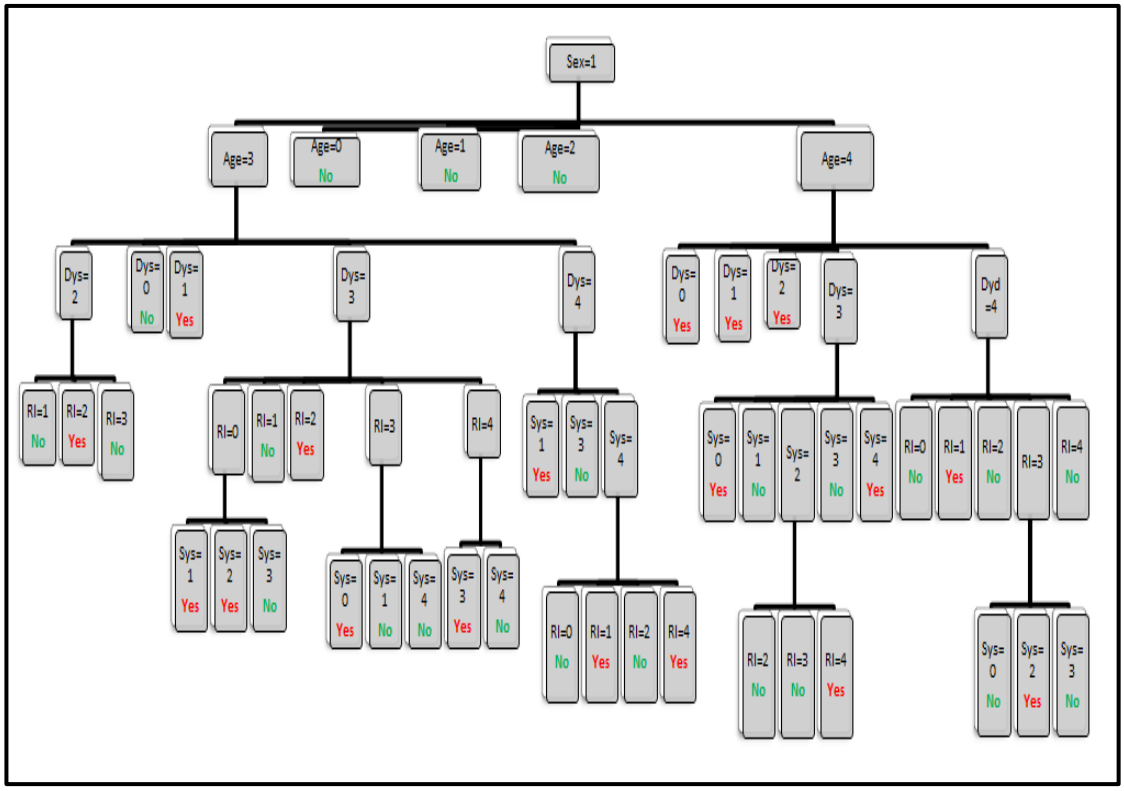

Figure 4.5: Female Decision Tree Using Non-Invasive Canberra Attributes

4.6. Chapter Summary and Conclusion

This chapter identifies cost-effective non-invasive attributes that can be used in community screening tests to identify patients at risk of heart disease. It discusses the investigation of the Decision Tree data mining technique to identify if non-invasive attributes show reliable performance in the diagnosis of heart disease patients. It also investigates those combinations of attributes that show the best accuracy in the diagnosis of heart disease patients.

The effect of using single non-invasive Cleveland and Canberra dataset attributes in the diagnosis of heart disease patients is investigated. On both the Cleveland and Canberra heart disease dataset, the age and sex single data attributes show the best results (Table 4.1 and 4.2). Different combinations of non-invasive attributes are investigated. The age, sex, and resting blood pressure combination shows best results (Table 4.4 and 4.6). Equations (see Equation 4.1, 4.2 and 4.3) derived from different non-invasive attributes (e.g. BMI, Rohrer's Index and RBPDiff) are applied to identify if Decision Tree performance is enhanced. The results show that the best combination is age, sex, resting blood pressure and Rohrer's Index equation with mean accuracy and standard deviation of 73.8% and 4.9% respectively (Table 4.11).

As discussed in Chapter 2 (section 2.5.2), researchers are suggesting that integrating more than one data mining technique improves accuracy in the diagnosis of heart disease patients. K-Means clustering is one of the most popular and well-known clustering techniques (Wu, Kumar et al. 2007). Its simplicity and reliable behaviour make it popular in many applications (Bramer 2007). However, initial centroid selection is a critical issue in K-Means clustering and strongly affects results (Poomagal and Hamsapriya 2011). The next chapter explores a hybrid model that integrates K-Means clustering using different initial centroid selection methods with Decision Tree in the diagnosis of heart disease patients. The hybrid model is applied to all attributes and non-invasive attribute combinations of the Cleveland and Canberra heart disease datasets to identify if this integration can enhance Decision Tree performance in the risk evaluation of heart disease patients.

Chapter 5

Integrating Clustering With Decision Tree in Heart Disease Diagnosis

5.1. Introduction

The previous chapter discusses Decision Tree performance against different single, combined, and calculated non-invasive attribute combinations on the Cleveland and Canberra datasets in the diagnosis of heart disease patients. That work shows that the age attribute is the second-most important attribute in the Cleveland dataset and the most important attribute in the Canberra dataset. The age, sex, and resting blood pressure combination shows the best results among other combinations in the diagnosis of heart disease on both the Cleveland and Canberra datasets. With different equations, the age, sex, resting blood pressure and Rohrer's Index equation (Equation 4.2) combination shows the best results for the Canberra dataset.

Recent research suggests that integrating more than one data mining technique can enhance the performance in the diagnosis of heart disease patients. As reviewed in chapter 2 section 2.5.2 that applying hybrid data mining techniques enhances their performance in the diagnosis of heart disease patients. For instance, Das et al. (2009) use Neural Network ensembles for the Cleveland heart disease dataset showing that this integration enhances Neural Network accuracy (Das, Turkoglu et al. 2009). Also Rajeswari et al. (2013) use feature selection with Neural Network, Decision Tree, and Support Vector Machine data mining techniques showing that this integration enhances different data mining techniques' accuracy (Rajeswari, Vaithiyanathan et al. 2013).

K-Means clustering is one of the most popular and well-known clustering techniques. Initial centroid selection is a critical factor that strongly affects K-Means clustering performance. Selecting initial centroids in an intelligent way helps to optimize the performance of the K-Means clustering algorithm. This chapter investigates integrating K-Means clustering (with different initial centroid selection methods) with Decision Tree in the diagnosis of heart disease patients. The hybrid model is applied to Cleveland and Canberra datasets using all attributes, and non-invasive, attribute combinations (see Figure 5.1). The Decision Tree used on the

Cleveland and Canberra heart disease datasets is based on 10-fold cross-validation as described earlier in Chapter 3. Hence the mean and standard deviation of sensitivity, specificity, and accuracy is calculated on the 10 samples of the datasets. This investigation determines if integrating K-Means clustering with Decision Tree enhances its performance in the diagnosis of heart disease patients.

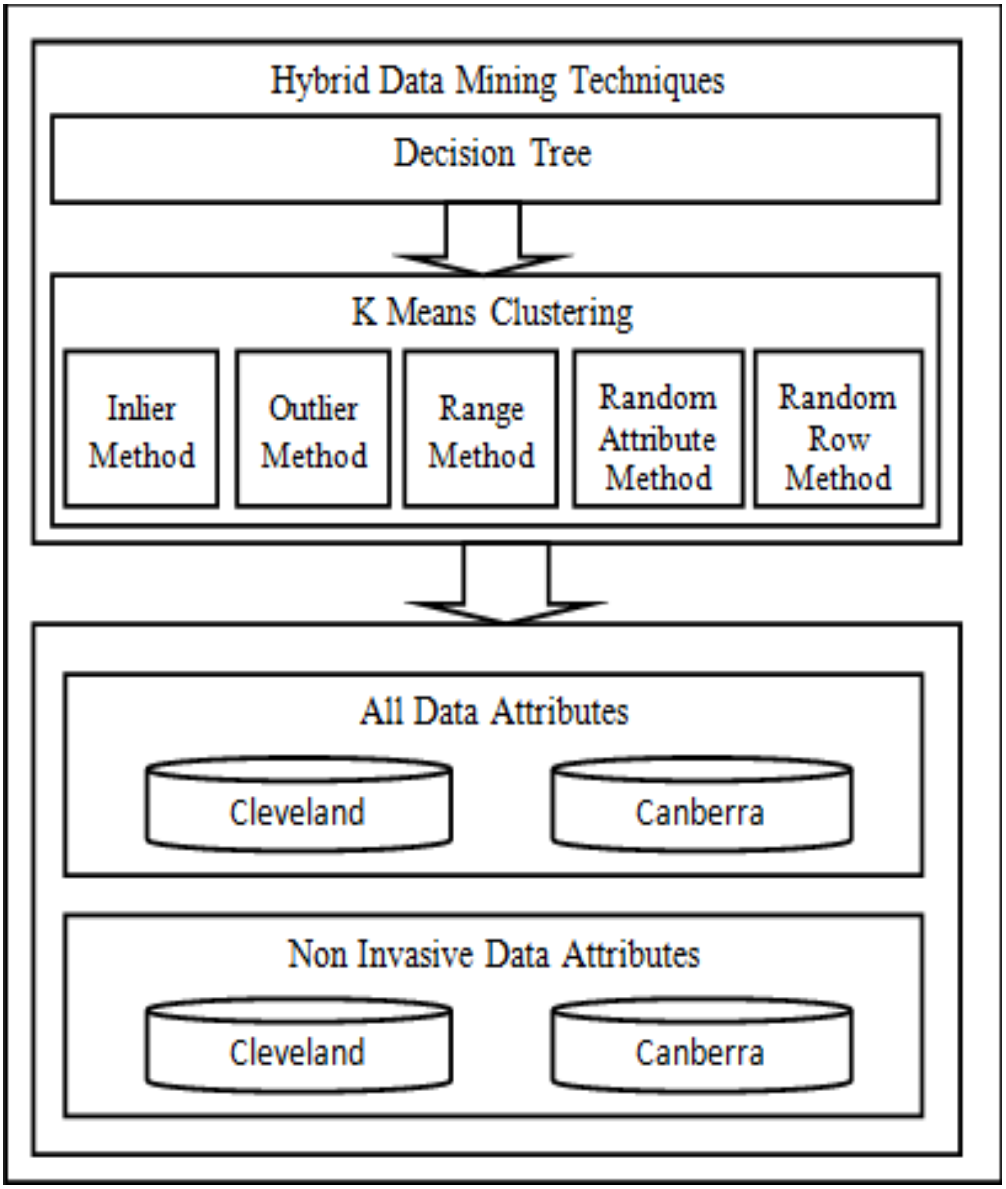

Figure 5.1: Applying K-Means Clustering Decision Tree to Heart Disease Datasets

5.2. Understanding K-Means Clustering with Different Initial Centroid Selection Methods

K-Means clustering is simple and demonstrates good behaviour in many applications (Bramer 2007, Wu, Kumar et al. 2007). Initial centroid selection is a critical factor that strongly affects K-Means clustering performance. Selecting the initial centroids in an intelligent way helps optimize the performance of the K-Means clustering algorithm (Pavan, Rao et al. 2011, Poomagal and Hamsapriya 2011, Santhi, Sai Leela et al. 2011, Tajunisha and Saravanan 2011). Researchers have been investigating enhancing K-Means clustering performance through intelligent initial centroid selection methods with various techniques. Poomagal and Hamsapriya (2011) propose optimizing K-Means clustering by selecting the initial centroids in an intelligent mathematical model showing that this method produces high quality results (Poomagal and Hamsapriya 2011). Immaculate Mary and Kasmir Raja (2009) propose improving K-Means clustering through the integration of ant colony optimization. This model creates the initial centroids based on the mode value of the data then applies the K-Means algorithm. The ant colony optimization algorithm is then applied to refine the clusters' quality (Immaculate Mary and Kasmir Raja 2009). Pavan, Rao et al. (2011) proposes a single-pass seed-selection initial centroid selection method that produces a single, optimal solution that is outlier insensitive. The algorithm is an extension to K-Means in that it selects initial seeds with specific probabilities and demonstrates effectiveness in the clustering results (Pavan, Rao et al. 2011).

K-Means clustering performance is dependent on the dataset (Alsabti, Ranka et al. 1997). However, no previous research on enhancing initial centroid selection methods for K-Means clustering is evident for the diagnosis of heart disease patients. Consequently, this chapter explores integrating K-Means clustering with different conventional initial centroid selection methods with Decision Tree. Integrating Naïve Bayes and K-Nearest Neighbour data mining techniques with K-Means clustering and initial centroid selection methods are also presented (see Appendix B).

The steps of applying k-Means clustering are described in Figure 5.2. The generation of initial centroids is based on actual sample data points using Inlier Method, Outlier Method, Range Method, Random Attribute Method, and Random Row Method

(Khan and Mohamudally August, 2010). The difference between these initial centroid selection methods is discussed in the following sections.

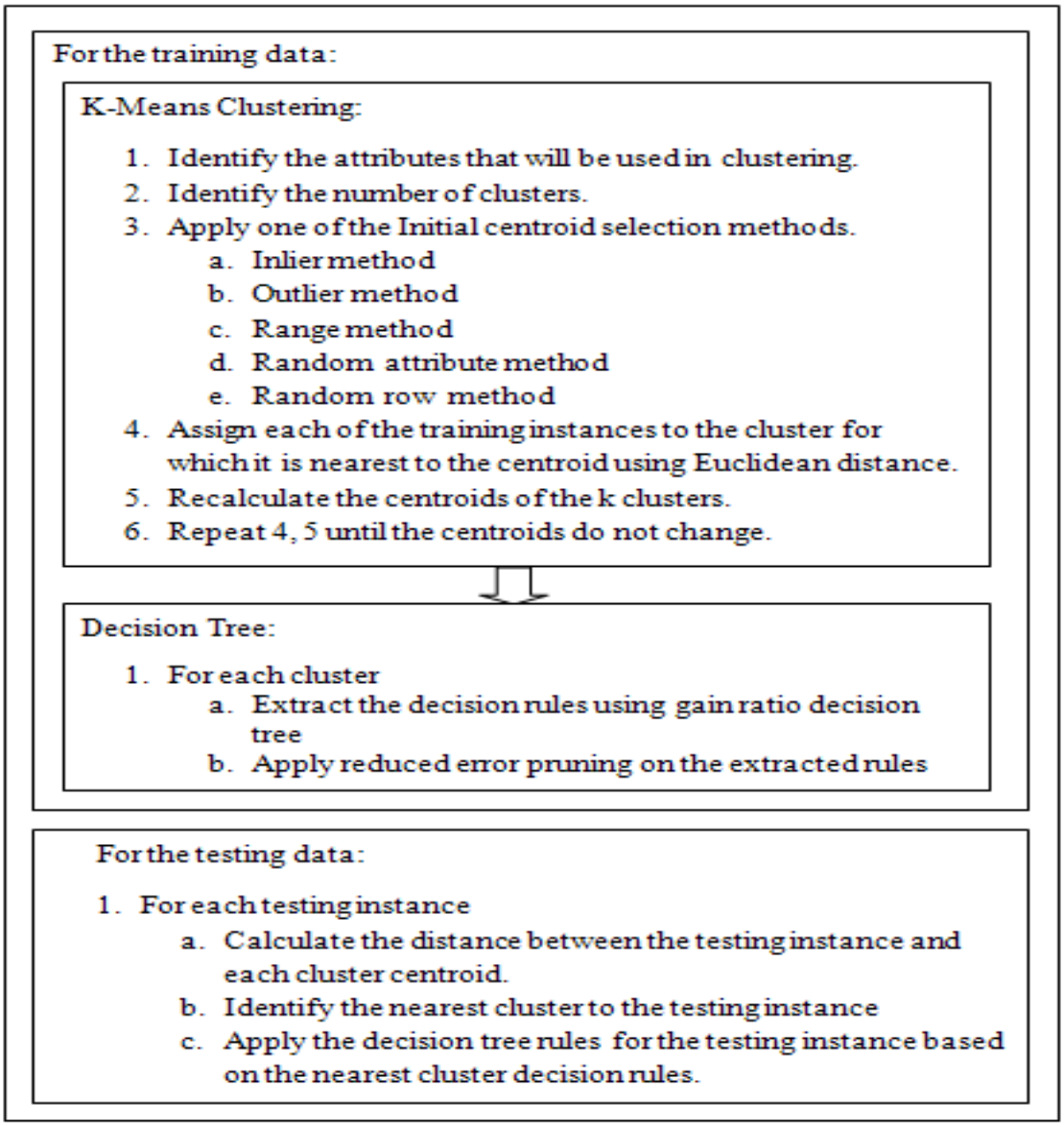

Figure 5.2: K-Means Clustering Decision Tree Methodology

5.2.1. Inlier Method Initial Centroid Selection

In generating the initial K centroids using the Inlier method, equation 5.1 and 5.2 are used:

$$ci = Min\ (X) - i \text{ where } 0 \leq i \leq k \qquad \text{(Equation 5.1)}$$

$$cj = Min\ (Y) - j \text{ where } 0 \leq j \leq k \qquad \text{(Equation 5.2)}$$

Where the initial centroid is C (c_i, c_j) and min (X) and min (Y) are the

minimum value of attribute X, and attribute Y, respectively and k represents the number of clusters.

5.2.2. Outlier Method Initial Centroid Selection

In generating the initial K centroids using the Outlier method, equation 5.3 and 5.4 are used:

$$c_i = \text{Max}(X) - i \text{ where } 0 \leq i \leq k \quad \text{(Equation 5.3)}$$

$$c_j = \text{Max}(Y) - j \text{ where } 0 \leq j \leq k \quad \text{(Equation 5.4)}$$

Where the initial centroid is C (c_i, c_j) and max (X) and max (Y) are the maximum value of attribute X, and attribute Y, respectively and k represents the number of clusters.

5.2.3. Range Method Initial Centroid Selection

In generating the initial K centroids using the Range method, equation 5.5 and 5.6 are used:

$$c_i = ((\text{Max}(X) - \text{Min}(X)) / K) * n \text{ where } 0 \leq i \leq k \quad \text{(Equation 5.5)}$$

$$c_j = ((\text{Max}(Y) - \text{Min}(Y)) / K) * n \text{ where } 0 \leq j \leq k \quad \text{(Equation 5.6)}$$

The initial centroid is C (ci, cj). Max (X) and min (X) are maximum and minimum values of attribute X, and max (Y) and min (Y) are maximum and minimum values of attribute Y. k represents the number of clusters and n varies from 1 to k.

5.2.4. Random Attribute Method Initial Centroid Selection

In generating the initial K centroids using the Random Attribute method, equation 5.7 and 5.8 are used:

$$c_i = \text{random}(X) \text{ where } 1 \leq i \leq k \quad \text{(Equation 5.7)}$$

$$c_j = \text{random}(Y) \text{ where } 1 \leq j \leq k \quad \text{(Equation 5.8)}$$

The initial centroid is C (c_i, c_j). The values of i and j vary from 1 to k.

5.2.5. Random Row Method Initial Centroid Selection

In generating the initial K centroids using the Random Row method, equation 5.9, 5.10 and 5.11 are used:

$$I = \text{random}(V) \text{ where } 1 \leq V \leq N \quad \text{(Equation 5.9)}$$

$c_i = X(I)$ (Equation 5.10)

$c_j = Y(I)$ (Equation 5.11)

The initial centroid is C (c_i, c_j). N is the number of instances in the training dataset. X (I) and Y(I) are the values of the attributes X and Y, respectively for the instance I.

For the Random Attribute and Random Row methods, ten runs are executed and the average and best for each method are calculated and used as the results.

5.3. Integrating Clustering With Decision Tree on the Cleveland and Canberra Heart Disease Datasets (All Attributes)

Several researchers have identified that age is a critical risk factor associated with heart disease (Heller, Chinn et al. 1984, Salahuddin and Rabbi 2006, Shahwan-Akl 2010). When considered singly, the age attribute is the first or second most explanatory attribute in the diagnosis of Cleveland and Canberra heart disease datasets (see chapter 4). Thus the age attribute is applied as a clustering attribute for heart disease patients. Between two and five clusters are used for K-Means clustering in this investigation, with different initial centroid selection methods. When using all the heart disease data attributes, the mean accuracy (the percentage of patients that are correctly diagnosed) acts as the most important performance measure between different initial centroid selection methods while maintaining reliable mean sensitivity.

Table 5.1 presents the results of integrating Decision Tree, K-Means clustering with different initial centroid selection methods, and different numbers of clusters on all data attributes of the Cleveland heart disease dataset. The best results for the Decision Tree K-Means clustering is achieved by the two clusters Inlier method showing a mean accuracy of 81.2% (standard deviation of 6.2%) (Table 5.1). A 2.1% increase in mean accuracy is achieved compared to the single Decision Tree of 79.1% (see Table 5.1). Similarly, a 0.3% increase in mean sensitivity is achieved compared to the single Decision Tree of 75.6% (Table 5.1). Although the two clusters Inlier K-Means clustering is not showing the best sensitivity, when using all the data attributes the mean accuracy is more critical than the mean sensitivity. The best mean sensitivity is achieved by the three clusters Range K-Means clustering showing mean sensitivity of 76.2%

(Table 5.1). The number of test samples in each cluster of the Cleveland dataset is relatively small (as the testing data is 30 records out of 300) especially in the five clusters and can range from four to nine samples in each cluster.

Table 5.1: Integrating Decision Tree with K-Means Clustering On Cleveland Dataset (All Attributes)

Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Decision Tree		75.6%	6.1%	81.6%	12.1%	79.1%	5.8%
Number of Clusters = 2	Inlier	75.9%	7.2%	85.1%	11.4%	81.2%	6.2%
	Outlier	76%	9.6%	80.3%	13.4%	78.7%	6.7%
	Range	76%	9.6%	80.3%	13.4%	78.7%	6.7%
	Random Row	76.1%	8.9%	82.8%	9.6%	80.1%	4.7%
	Random Attribute	74.3%	8%	84%	12.2%	80.1%	7%
Number of Clusters = 3	Inlier	75.8%	8.5%	83.4%	14.6%	80.8%	8.6%
	Outlier	75.8%	8.5%	83.4%	14.6%	80.8%	8.6%
	Range	76.2%	10.8%	76%	13.2%	78.9%	8.5%
	Random Row	69.8%	16.5%	77%	13.1%	76.3%	9.3%
	Random Attribute	70.5%	10.8%	81.7%	12.4%	79.4%	8.8%
Number of Clusters = 4	Inlier	71.5%	11.8%	78.1%	12.3%	77.5%	8.6%
	Outlier	73.2%	11.2%	84.1%	11.8%	79.7%	8.2%
	Range	71.9%	12%	78.1%	12.3%	77.8%	8.6%
	Random Row	73.2%	9%	79.1%	10.5%	78.4%	8.3%
	Random Attribute	68.1%	15.9%	80.8%	11.3%	77%	9.1%
Number of Clusters = 5	Inlier	71.5%	11.8%	78.1%	12.3%	77.5%	8.6%
	Outlier	73.2%	11.2%	84.1%	11.8%	79.7%	8.2%
	Range	71.9%	12%	78.1%	12.3%	77.8%	8.6%
	Random Row	69.3%	13.9%	79.5%	9.6%	77.1%	7.8%

	Random Attribute	69.3%	13.9%	79.5%	9.6%	77.1%	7.8%

A T-Test is applied to identify if the increase in mean accuracy and sensitivity is significant. When applying a T-Test for significance for accuracy between the two clusters Inlier K-Means clustering Decision Tree and the single Decision Tree, there is a significant difference between the accuracy of the two methods (see Table 5.2). However, when applying a T-Test for significance for the sensitivity in the same comparison, there is no significant difference between the two methods.

Table 5.2: T-Test Significance for K-Means clustering to Cleveland Heart Disease Dataset (All Attributes)

Canberra Data Attributes	**Accuracy**			**Sensitivity**		
	Mean	**St Dev**	**T-Test Significance**	**Mean**	**St Dev**	**T-Test Significance**
Decision Tree	79.1%	5.8%	Yes ($t = 4.284$, $p \leq 0.05$, degree of freedom=300)	75.6%	6.1%	No ($t = 0.551$, $p \leq 0.05$, degree of freedom=300)
Two Clusters Inlier K-Means Clustering Decision Tree	81.2%	6.2%		75.9%	7.2%	

Figure 5.3 shows the mean accuracy of integrating Decision Tree with K-Means clustering with different initial centroid selection methods and with different numbers of clusters in the diagnosis records in the Cleveland heart disease dataset when all data attributes are used. There is no specific trend in diagnosis accuracy when increasing the number of clusters integrated with the Decision Tree (Figure 5.3).

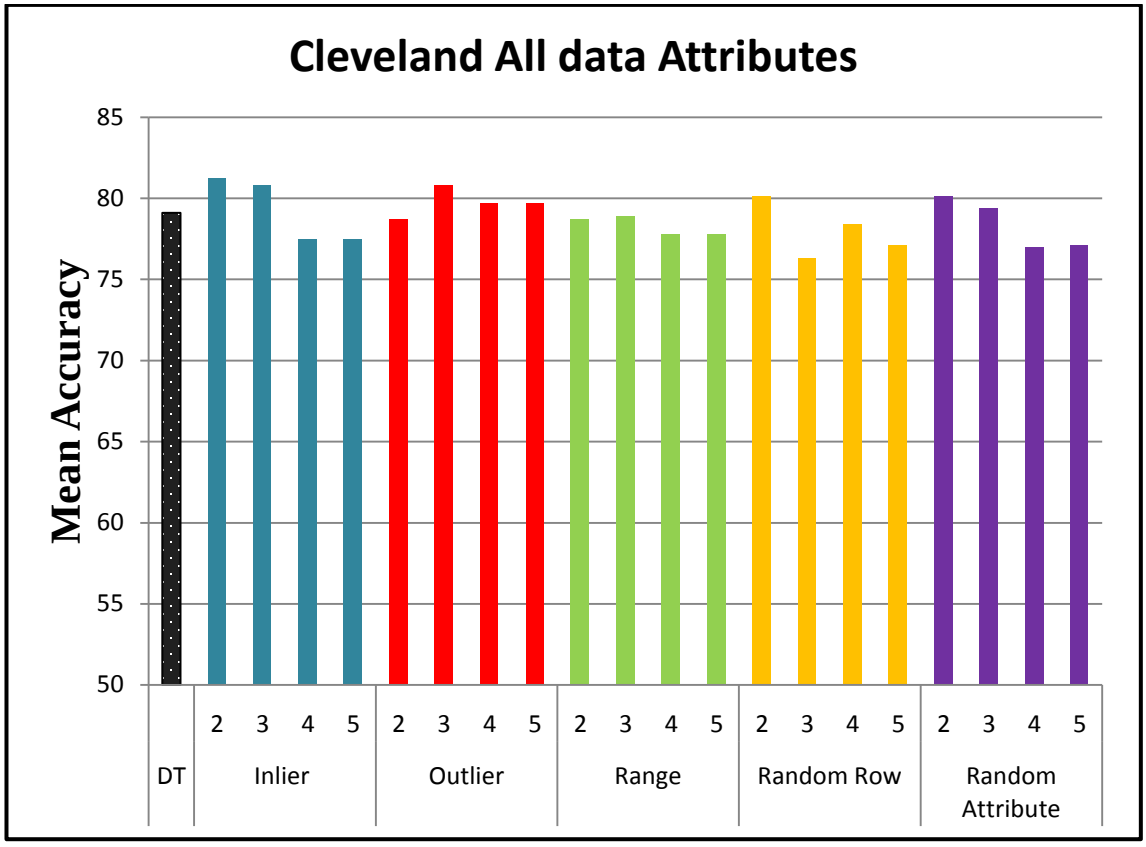

Figure 5.3: Applying K-Means Clustering Decision Tree to Cleveland Heart Disease Dataset (All Attributes)

Table 5.3 presents the results of integrating Decision Tree with K-Means clustering with different initial centroid selection methods and with different numbers of clusters on the Canberra heart disease dataset using all data attributes. The best results are achieved by the three clusters Inlier, Outlier, and Random Row K-Means clustering showing a mean accuracy of 71.5% (standard deviation of 6.7%) (see Table 5.3). That result represents a 2.6% increase in mean accuracy when compared to the single Decision Tree of 68.9% (Table 5.3). It is also shows a 4.6% decrease in mean sensitivity compared to the single Decision Tree of 67.7% (Table 5.3). Although the three cluster Inlier K-Means clustering is not showing the best sensitivity, when using all the data attributes the mean accuracy is more critical than the mean sensitivity. The best mean sensitivity is achieved by the three clusters Random Attribute K-Means clustering showing mean sensitivity of 67.2% (Table 5.3). The number of test samples in each cluster of the Canberra data is relatively small (as the testing data is 46 records out of 460) especially in the five clusters and can range from ten to twenty samples in each cluster.

Table 5.3: Integrating Decision Tree with K-Means Clustering On Canberra Dataset (All Attributes)

Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Decision Tree		67.7%	13.6%	64.5%	18%	68.9%	7.8%
Number of Clusters = 2	Inlier	61.8%	11.5%	73.8%	11.6%	70.1%	7.5%
	Outlier	60.3%	13.9%	73%	11.6%	69.6%	7%
	Range	60.3%	13.9%	73%	11.6%	69.6%	7%
	Random Row	64.5%	13%	71.5%	11.1%	70%	7%
	Random Attribute	63.2%	12.9%	72%	12.4%	70%	7.3%
Number of Clusters = 3	Inlier	63.1%	13.1%	72.9%	14.6%	71.5%	6.7%
	Outlier	63.1%	13.1%	72.9%	14.6%	71.5%	6.7%
	Range	59.7%	13.3%	71.4%	10.6%	68.4%	5.6%
	Random Row	63.1%	13.1%	72.9%	14.6%	71.5%	6.7%
	Random Attribute	67.2%	14%	69.5%	15.8%	70.9%	8.8%
Number of Clusters = 4	Inlier	61.1%	12.5%	71.5%	14.2%	69.5%	5.8%
	Outlier	65.1%	16.5%	67.5%	20.1%	69.6%	8.5%
	Range	65.3%	12.6%	67.7%	13.6%	69.5%	6.3%
	Random Row	64%	14.2%	66.2%	14.7%	68.3%	6.9%
	Random Attribute	64.4%	13.5%	67.3%	17.1%	67.6%	7.5%
Number of Clusters = 5	Inlier	61.1%	12.5%	71.5%	14.2%	69.5%	5.8%
	Outlier	65.1%	16.5%	67.5%	20.1%	69.6%	8.5%
	Range	65.3%	12.6%	67.7%	13.6%	69.5%	6.3%
	Random Row	64.3%	14.1%	65.6%	17.6%	67.9%	7.6%
	Random Attribute	59.7%	12.5%	67%	16.8%	66.6%	7.1%

A T-Test is applied to identify if the increase in mean accuracy is significant. When applying a T-Test for significance for accuracy between the three-cluster Inlier, Outlier, and Random Row K-Means clustering Decision Tree and the single Decision Tree, there is a significant difference between the two methods (see Table 5.4). When applying a T-Test for significance for sensitivity between the three-cluster Inlier, Outlier, and Random Row K-Means clustering Decision Tree and the single Decision Tree, there is a significant difference between the two methods (see Table 5.4). When using all the heart disease data attributes, the mean accuracy (the percentage of patients that are correctly diagnosed) acts as the most important performance measure between different initial centroid selection methods while maintaining reliable mean sensitivity.

Table 5.4: T-Test Significance for K-Means clustering to Canberra Heart Disease Dataset (All Attributes)

Canberra Data Attributes	**Accuracy**			**Sensitivity**		
	Mean	**St Dev**	**T-Test Significance**	**Mean**	**St Dev**	**T-Test Significance**
Decision Tree	68.9%	7.8%	Yes ($t = 5.423$, $p \leq 0.05$, degree of freedom = 460)	67.7%	13.6%	Yes ($t = 5.225$, $p \leq 0.05$, degree of freedom = 460)
Three Clusters Inlier Outlier, and Random Row K-Means Clustering	71.5%	6.7%		63.1%	13.1%	

Figure 5.4 shows the mean accuracy of integrating Decision Tree with K-Means clustering with different initial centroid selection methods and with different numbers of clusters in the diagnosis of Canberra heart disease dataset using all data attributes. There is no specific trend in diagnosis accuracy when increasing the number of clusters integrated with the Decision Tree (see Figure 5.4).

Integrating K-Means clustering with different initial centroid selection methods and decision tree shows improved mean accuracy over Cleveland and Canberra heart disease dataset when using all data attributes. However, there is a need to identify if this

integration will also enhance the mean accuracy over Cleveland and Canberra when only using the non-invasive attribute combinations.

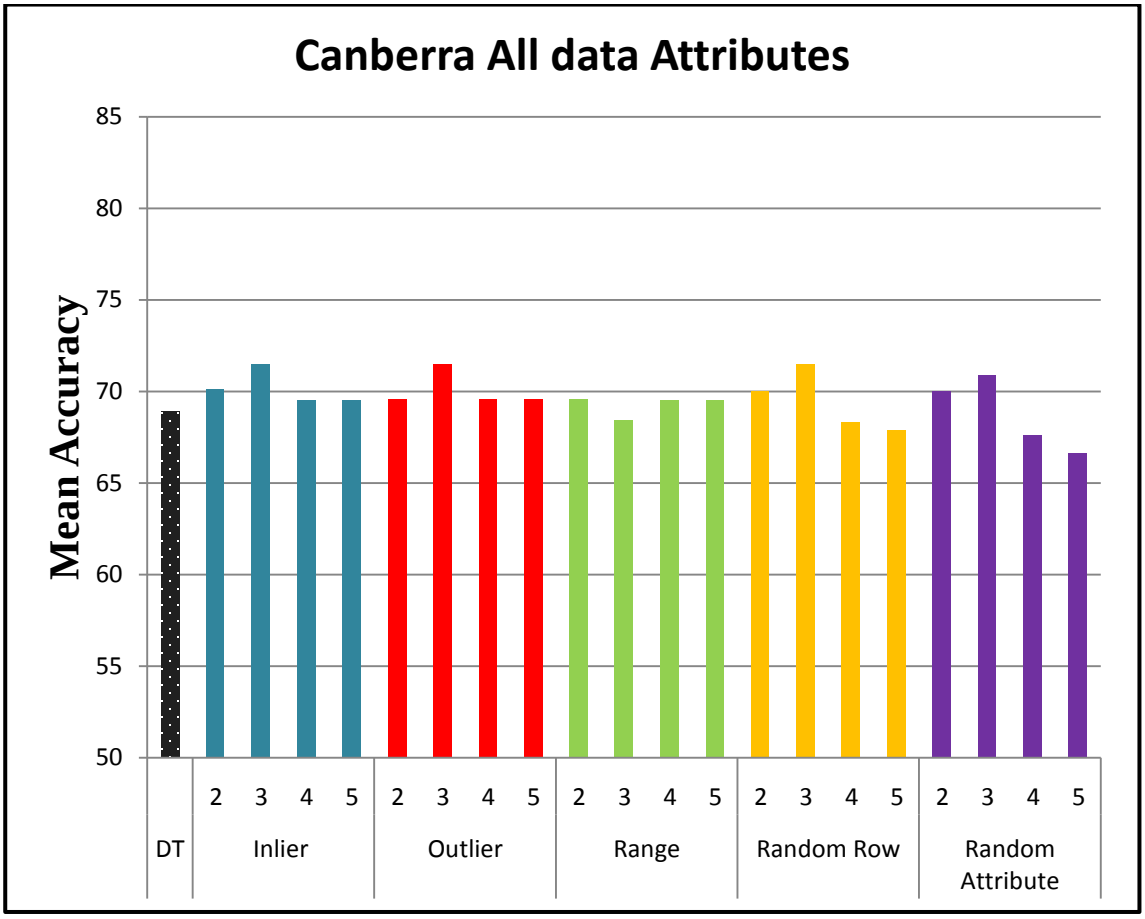

Figure 5.4: Applying K-Means Clustering Decision Tree to Canberra Heart Disease Dataset (All Attributes)

5.4. Integrating Clustering with Decision Tree on the Cleveland and Canberra Heart Disease Datasets (Non-Invasive Attributes)

Section 5.4 investigates integrating Decision Tree, K-Means clustering with different initial centroid selection methods, and different numbers of clusters on the Cleveland and Canberra heart disease datasets using only non-invasive data attribute combinations. The age attribute is used as a clustering attribute and the number of clusters ranges between two and five for each initial centroid selection method. When using the non-invasive heart disease data attributes, the mean sensitivity (the percentage sick patients that are diagnosed as sick) acts as the most important performance measure between different initial centroid selection methods while maintaining reliable mean accuracy.

Table 5.5 presents the results of integrating Decision Tree with K-Means clustering with different initial centroid selection methods and different numbers of clusters on the non-invasive heart disease data attributes on the Cleveland dataset: Age, sex, and resting blood pressure (see Chapter 4). Integrating Decision Tree with K-Means clustering did not enhance accuracy in the diagnosis of heart disease patients using the Cleveland non-invasive heart disease data attributes. The best results are achieved by the two-cluster Outlier initial centroid selection method with a mean accuracy of 64.5% (standard deviation of 12.5%) as shown in Table 5.5. There is a 1.3% decrease in mean accuracy compared to the single Decision Tree of 65.8% (Table 5.5). Worse, there is a 24.5% decrease in mean sensitivity when compared to the single Gain Ratio Decision Tree of 61.2% (Table 5.5).

Although maintaining reliable mean accuracy is important, showing a slightly degradation in the mean accuracy while enhancing the mean sensitivity is considered an enhancement. In the risk evaluation of heart disease patients, the mean sensitivity is more important than the mean accuracy as the mean sensitivity is the proportion of sick patients that are identified as sick. The mean sensitivity represents the patients that really need to be addressed and need more attention in helping them to overcome the heart disease. So maintaining a higher mean sensitivity using integrated K-Means clustering and Decision tree is considered an enhancement in the ability of the data mining technique in the diagnosis of heart disease patients.

Figure 5.5 shows the mean accuracy of integrating Decision Tree with K-Means clustering with different initial centroid selection methods and different numbers of clusters with the Cleveland non-invasive heart disease data attributes dataset. There is no specific trend evident in increasing the number of clusters and the increase/decrease of accuracy (see Figure 5.5).

Table 5.5: Integrating Decision Tree with K-Means Clustering On Cleveland Dataset (Non-Invasive Attributes)

Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Decision Tree		61.2%	10.1%	69.5%	10.1%	65.8%	8.6%
Number of Clusters = 2	Inlier	36.4%	15.7%	84.7%	7.8%	63.9%	11.5%
	Outlier	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
	Range	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
	Random Row	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
	Random Attribute	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
Number of Clusters = 3	Inlier	14.4%	11.1%	93.9%	4.8%	57.2%	16%
	Outlier	14.4%	11.1%	93.9%	4.8%	57.2%	16%
	Range	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Row	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Attribute	38.4%	17.3%	84.7%	7.8%	64.2%	12%
Number of Clusters = 4	Inlier	36.4%	15.7%	84.7%	7.8%	63.9%	11.5%
	Outlier	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Range	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Row	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Attribute	38.4%	17.3%	84.7%	7.8%	64.2%	12%
Number of Clusters = 5	Inlier	36.4%	15.7%	84.7%	7.8%	63.9%	11.5%
	Outlier	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Range	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Row	38.4%	17.3%	84.7%	7.8%	64.2%	12%
	Random Attribute	38.4%	17.3%	84.7%	7.8%	64.2%	12%

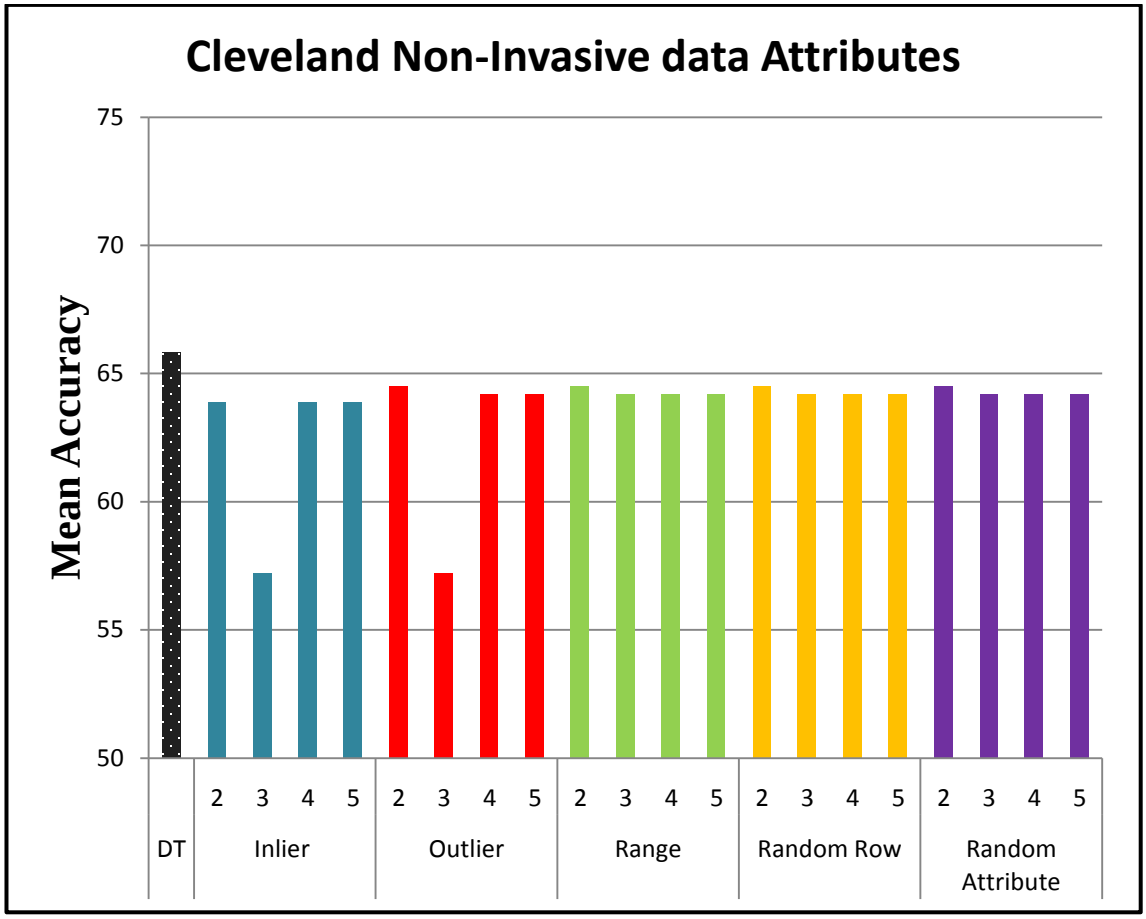

Figure 5.5: Applying K-Means Clustering Decision Tree to Cleveland Heart Disease Dataset (Non-Invasive Attributes)

Table 5.6 presents the results of integrating Decision Tree with K-Means clustering, different initial centroid selection methods, and different numbers of clusters on the non-invasive heart disease attributes of the Canberra dataset. Age, sex, resting blood pressure, and Rohrer's Index equation (Equation 4.2) show the best accuracy (see Chapter 4).

Table 5.6: Integrating Decision Tree with K-Means Clustering On Canberra Dataset (Non- Invasive Attributes)

Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Decision Tree		67.1%	11.4%	75.5%	9.2%	73.8%	4.9%
Number of Clusters = 2	Inlier	67.8%	13.6%	75%	10%	74.4%	4%
	Outlier	69.7%	9.9%	71.9%	11.3%	72.1%	5.1%
	Range	67.8%	11.2%	74.2%	8.1%	73.1%	4.6%
	Random Row	69.5%	10.3%	72.2%	10.7%	72.3%	5.1%
	Random Attribute	67.6%	13%	71.2%	10.6%	71%	4.2%
Number of Clusters = 3	Inlier	68.7%	10.3%	74.4%	11.4%	73.2%	5.2%
	Outlier	68.7%	10.3%	74.4%	11.4%	73.2%	5.2%
	Range	67.8%	11.2%	74.2%	8.1%	73.1%	4.6%
	Random Row	66.6%	9.4%	75.8%	8.6%	73.1%	4.7%
	Random Attribute	68.1%	10.6%	74.2%	10.9%	73%	4.3%
Number of Clusters = 4	Inlier	67.2%	13%	75.9%	8.4%	74.1%	3.8%
	Outlier	69.7%	9.9%	73.1%	10.5%	72.6%	5.2%
	Range	67.9%	12.8%	74.5%	8%	73.4%	3.9%
	Random Row	68.4%	11.1%	76.2%	7.8%	74.4%	5.2%
	Random Attribute	67.7%	13.1%	72.2%	10.9%	72.2%	5%
Number of Clusters = 5	Inlier	67.2%	13%	75.9%	8.4%	74.1%	3.8%
	Outlier	69.7%	9.9%	73.1%	10.5%	72.6%	5.2%
	Range	67.9%	12.8%	74.5%	8%	73.4%	3.9%
	Random Row	67.8%	11.2%	73.9%	8.1%	73%	4.7%
	Random Attribute	67.8%	11.2%	73.9%	8.1%	73%	4.7%

Best accuracy is achieved by the two clusters Inlier initial centroid selection method showing mean accuracy of 74.4% (standard deviation of 4%) as shown in Table 5.5. A 0.6% increase in mean accuracy is evident when comparing it with the single Decision Tree of 73.8% (Table 5.6). A mean sensitivity of 67.8% (standard deviation of 13.6%) is a decrease of 0.7% on the single decision tree (Table 5.6). The mean sensitivity is more critical than the mean accuracy that is, the sick patients need to be identified as sick. The two, four, and five clusters Outlier K-Means clustering decision tree shows an increase in the mean sensitivity 69.7% (standard deviation 9.9) over the single decision tree (see Table 5.6). The mean accuracy (72.1% or 72.6%) of this integration decreased. However, the mean sensitivity is more important in this context. The increase in mean sensitivity between the single decision tree and the two, four, and five clusters Outlier K-Means clustering decision tree is 2.6%.

A T-Test for significance is applied to the two, four, and five clusters Outlier K-Means clustering decision tree and the single decision tree to determine if this increase is significant. There is significant difference between the sensitivity of the two methods (see Table 5.7). The four and five clusters K-Means clustering decision tree have the same mean accuracy of 72.6% while the two clusters K-Means clustering decision tree shows mean accuracy of 72.1% (Table 5.6). The T-Test for significance is then applied to the four, and five clusters Outlier K-Means clustering decision tree and two clusters Outlier K-Means clustering decision tree to determine if the difference in mean accuracy is significant. There is no significant different between the accuracy of the two methods (see Table 5.7).

Figure 5.6 shows the means accuracy of integrating Decision Tree, K-Means clustering, different initial centroid selection methods, and different numbers of clusters using the non-invasive heart disease data attributes on the Canberra dataset. There is no specific trend evident in increasing the number of clusters with integrating Decision Tree and the increase/decrease of the accuracy (see Figure 5.6).

Figure 5.7 shows the means sensitivity of integrating Decision Tree, K-Means clustering, different initial centroid selection methods, and different numbers of clusters using the non-invasive heart disease data attributes on the Canberra dataset. Although there is no specific trend evident in increasing the number of clusters with the integrated Decision Tree and the increase/decrease of the accuracy, however this integration could enhance Decision Tree mean sensitivity (see Figure 5.7).

Table 5.7: T-Test Significance for K-Means clustering to Canberra Heart Disease Dataset (Non-Invasive Attributes)

Canberra Non-Invasive	Sensitivity			Accuracy		
	Mean	St Dev	T-Test Significance	Mean	St Dev	T-Test Significance
Decision Tree	67.1%	11.4%	Yes ($t = 3.693$, $p \leq 0.05$, degree of freedom = 460)	73.8%	4.9%	-
Two Clusters Outlier	69.7%	9.9%		72.1%	5.1%	No ($t = 1.472$, $p \leq 0.05$, degree of freedom = 460)
Four Clusters Outlier				72.6%	5.2%	
Five Clusters Outlier						

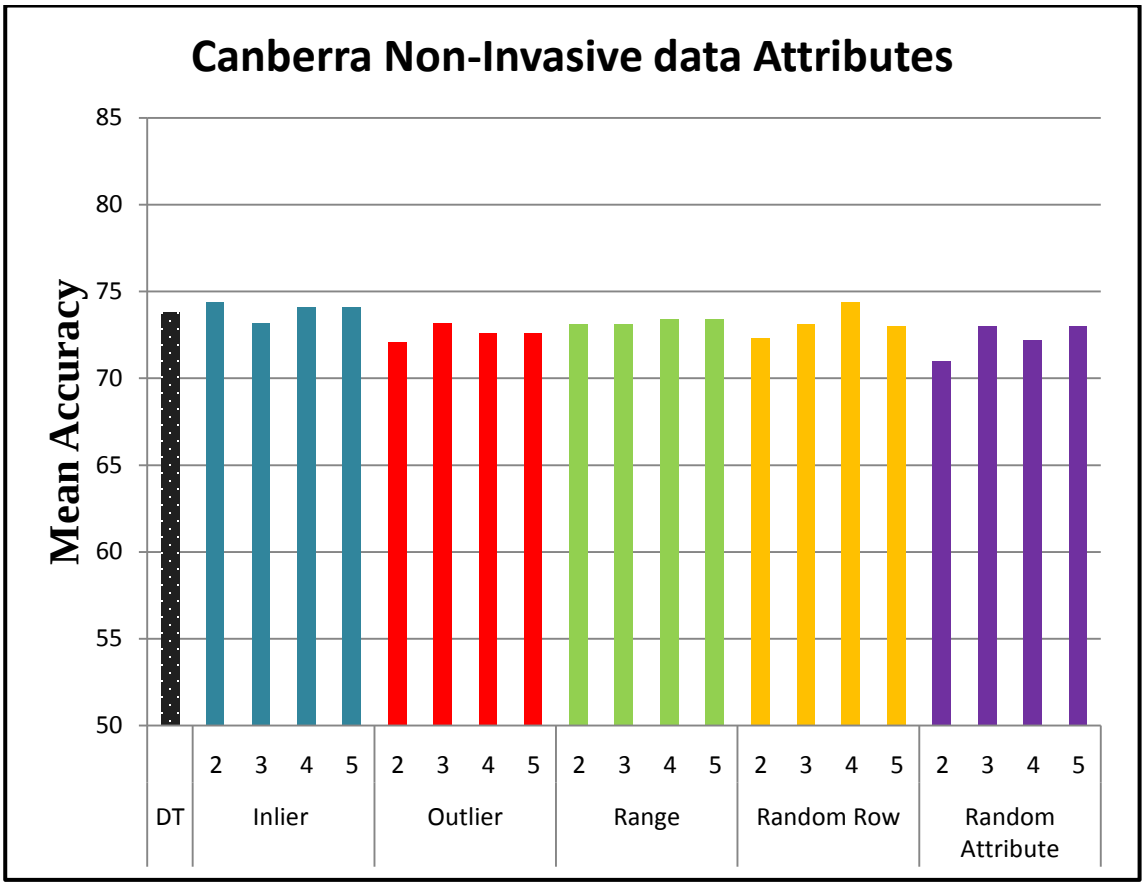

Figure 5.6: Mean Accuracy of Applying K-Means Clustering Decision Tree to Canberra Heart Disease Dataset (Non-Invasive Attributes)

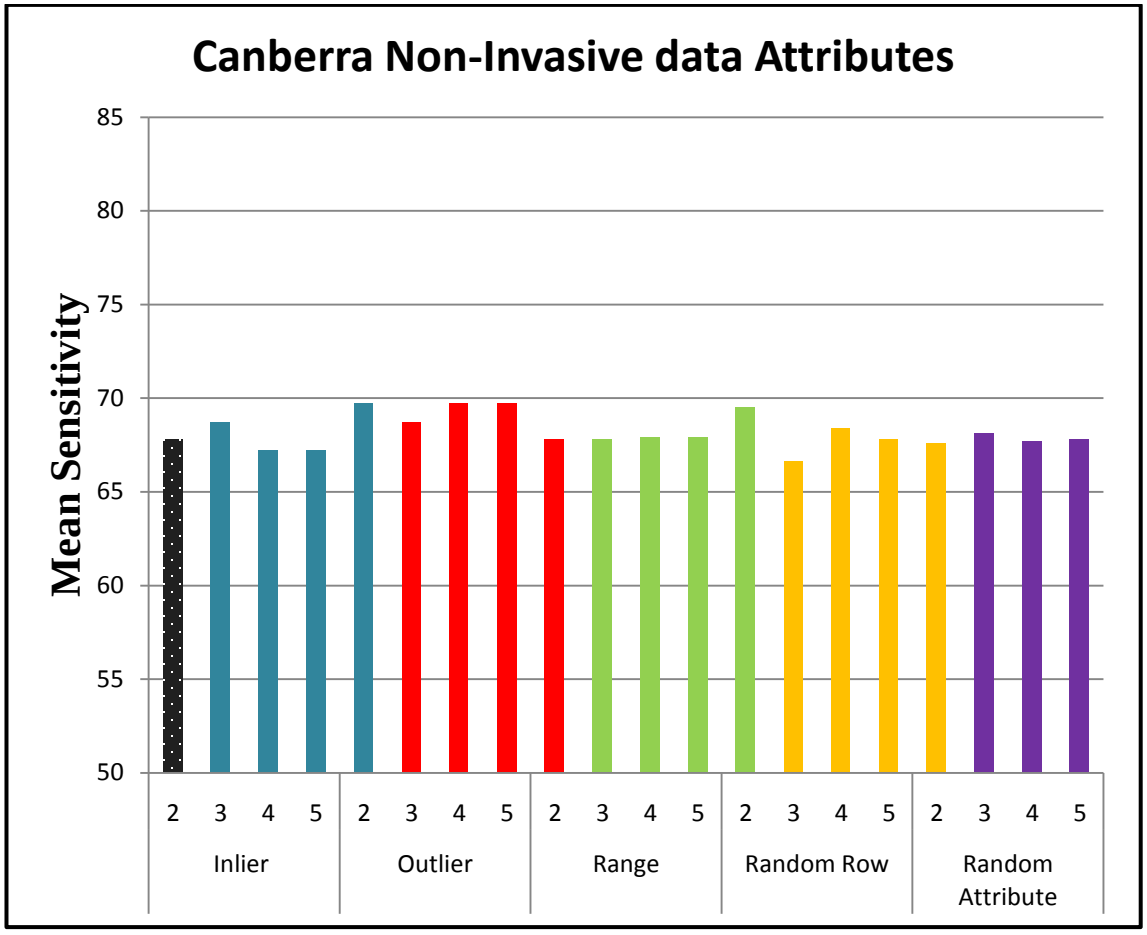

Figure 5.7: Mean Sensitivity of Applying K-Means Clustering Decision Tree to Canberra Heart Disease Dataset (Non-Invasive Attributes)

Integrating K-Means clustering initial centroid selection methods with Decision Tree shows differing results across the Cleveland and Canberra datasets when using all attributes and non-invasive data attribute combinations. These results are discussed in the Section 5.5.

5.5. Comparing Integrating Clustering With Decision Tree on the Cleveland and Canberra Datasets Different Attributes Combinations Results

The best results achieved across the Cleveland and Canberra datasets are presented in Table 5.8. The Inlier initial centroid selection method demonstrates best results for both datasets all attributes while the outlier initial centroid selection method demonstrates best results for both datasets non-invasive attributes (Table 5.8).

For the all attributes Cleveland and Canberra heart disease attributes, integrating K-Means clustering with Decision Tree could enhance decision tree accuracy showing mean accuracy of 81.2% and 71.5 % respectively. The Inlier initial centroid selection method shows best accuracy among other initial centroid selection methods (Table 5.8). For the non-invasive Cleveland heart disease attributes, integrating K-Means clustering with decision tree did not enhance decision tree sensitivity. The decrease in the sensitivity of the hybrid model on the non-invasive Cleveland dataset may be due to the few numbers of attributes as well as the small number of rows available for decision tree after clustering. Another reason for the low sensitivity could be the possibility of overfitting in the data. The overfitting of data is when a data mining technique searches for the best parameters for one particular model using a limited set of data, it may model not only the general patterns in the data but it may also model the noise specific to that data, causing poor performance of the model on test data (Fayyad, Piatetsky-Shapiro et al. 1996). The Outlier initial centroid selection method shows best accuracy among other initial centroid selection methods for the Cleveland dataset using non-invasive attribute combinations. However, the Outlier method did not enhance decision tree mean accuracy neither did the Outlier method improve the mean sensitivity (Table 5.8).

For the non-invasive Canberra heart disease attribute combinations, integrating K-Means clustering with Decision tree could enhance decision tree performance. The Outlier initial centroid selection method shows best mean sensitivity results among other initial centroid selection methods for Canberra non-invasive attributes. However, the Outlier method did not enhance decision tree accuracy (Table 5.8). In the context of non-invasive attributes, mean sensitivity is more critical than the mean accuracy for the purposes of informing a community screening test.

Although integrating K-Means clustering for Canberra non-invasive heart disease attributes enhances Decision Tree mean sensitivity, the same integration did not enhance the mean sensitivity for the Cleveland non-invasive heart disease attributes. This result could be because the small number of instances in the Cleveland dataset (decreased further in each cluster when applying K-Means clustering) does not allow Decision Tree enough data to create discriminatory decision tree rules.

Table 5.8: Integrating Decision Tree with K-Means Clustering on Cleveland and Canberra Datasets (All and Non-Invasive) Attributes

Dataset	Data Mining Technique	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Cleveland All Data	Decision Tree	75.6%	6.1%	81.6%	12.1%	79.1%	5.8%
	Two Clusters Inlier K-Means Decision Tree	75.9%	7.2%	85.1%	11.4%	81.2%	6.2%
Canberra All Data	Decision Tree	67.7%	13.6%	64.5%	18%	68.9%	7.8%
	Three Clusters Inlier K-Means Decision Tree	63.1%	13.1%	72.9%	14.6%	71.5%	6.7%
Cleveland Non-Invasive Data	Decision Tree	61.2%	10.1%	69.5%	10.1%	65.8%	8.6%
	Two Clusters Outlier K-Means Decision Tree	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
Canberra Non-Invasive Data	Decision Tree	67.1%	11.4%	75.5%	9.2%	73.8%	4.9%
	Two Clusters Outlier K-Means Decision Tree	69.7%	9.9%	71.9%	11.3%	72.1%	5.1%

From this investigation, integrating the two cluster Outlier K-Means clustering could enhance decision tree mean sensitivity in diagnosis for the Canberra heart disease dataset when using non-invasive data attributes. Furthermore, this hybrid model can readily be used for community screening test to help healthcare professionals in the diagnosis of heart disease patients.

5.6. Chapter Summary and Conclusion

Recent research suggests that integrating more than one data mining technique can enhance the performance in the diagnosis of heart disease patients. K-Means clustering is one of the most popular and well-known clustering techniques. However, evidence suggests that initial centroid selection is a critical factor that strongly affects K-Means

clustering performance. This chapter investigates integrating K-Means clustering with different initial centroid selection methods with Decision Tree in the diagnosis of heart disease patients across the Cleveland and Canberra heart disease datasets using all attributes and only the non-invasive data attributes.

Integrating K-Means clustering with Decision Tree enhances decision tree accuracy in the diagnosis of the Cleveland and Canberra datasets showing 81.2% and 71.5% respectively. The Inlier initial centroid selection method shows best results for Cleveland and Canberra datasets using all data attributes. Integrating K-Means clustering with different initial centroid selection methods and Decision Tree on non-invasive data attributes in the Cleveland and Canberra datasets did not enhance decision tree accuracy. Moreover, the mean sensitivity degraded to 38.4% on the Cleveland dataset. However, integrating K-Means clustering with different initial centroid selection methods and decision tree over the non-invasive Canberra dataset attributes shows an increase of the mean sensitivity achieving 69.7% and mean accuracy of 72.1%. These results are very interesting. It shows that the non-invasive Canberra heart disease attributes involving age, sex, resting blood pressure and Rohrer's Index equation (Equation 4.2) can be readily used to create a community screening test for the evaluation of heart disease risk.

The next chapter describes the development and implementation of a heart disease risk evaluation tool using non-invasive Canberra heart disease data attributes to help healthcare professionals in the diagnosis of heart disease patients. This evaluation tool can act as a community-level screening test to identify the risk of heart disease as being high or low. Chapter 6 will extract the decision tree rules to identify how a decision is made using the non-invasive data attributes.

Chapter 6

Heart Disease Risk Evaluation Tool

6.1. Introduction

Although heart disease is the leading cause of death all over the world causing 7.25 million deaths representing 12.8% of all the deaths (World Health Organization 2013c), it has also been identified as among the most preventable and controllable diseases (Centers for Disease Control and Prevention 2013). At least 80% of heart disease could be prevented by healthy diet, regular physical activity, and avoidance of tobacco products (World Health Organization 2013c). The World Health Organization (2010) reported that early detection and treatment are aimed to reduce progression to severe and costly illness and complications of heart disease.

The relative success of chronic disease treatments are dependent on the early detection of those diseases (Paladugu and Shyu 2010). Although heart disease can be detected by several tests such as chest X-rays, coronary angiograms, electrocardiograms, and exercise stress tests (National Center for Chronic Disease Prevention and Health Promotion 2013), those tests are very costly and require sophisticated equipment and a visit to a medical facility for heart disease detection. There is a vital need for accurate and systematic tools that provide information for early detection of heart disease to identify those patients at high risk (Paladugu and Shyu 2010).

There is a need to find less costly tests and accurate systematic tools that can be used for community screening to identify patients at high risk of heart disease and provide information to enable early intervention (Paladugu and Shyu 2010). Community screening tests play an especially important role in the early detection of heart disease (Kotnik 2010). These community screening tests can be applied in pharmacies or public health clinics where non-medical healthcare professionals could screen the community at large for potential heart disease sufferers and refer those potential sufferers to more thorough screening tests. Recent research focuses on discovering new specific, sensitive and cheap community screening tests (Kotnik 2010).

The level of accepted accuracy for screening tests varied from one tool to another showing 70% or 80% accuracy. A screening tool for mild cognitive impairment

is developed showing mean accuracy of 76% for Alzheimer disease (Nasreddine, Phillips et al. 2005). Another research used dobutamine stress as a screening tool in detecting coronary artery showing accuracy that ranged between 84% and 89% (Marcovitz and Armstrong 1992).

It is widely accepted that age, sex, blood pressure, smoking, cholesterol and diabetes are the major risk factors for developing heart diseases (Cupples and D'Agostino 1987). The Framingham Heart Disease Risk Calculator (Framingham Study. 2013) and Australian Absolute Cardiovascular Risk Calculator (National Heart Foundation of Australia 2009) are two famous heart disease screening tests that use major risk factors for identifying degree of risk of heart disease. They use different invasive and non-invasive data attributes in the risk evaluation of heart disease. The use of invasive attributes such as cholesterol and diabetes require blood tests investigation which slows the use of these screening tests. In Chapter 4, Decision tree performance using different non-invasive Canberra data attributes in the diagnosis of heart disease patients is investigated. The results demonstrate that the combination of age, sex, resting blood pressure (Systolic and Diastolic) and Rohrer's Index (RI) provides the best results. In the previous chapter, integrating K-Means clustering with different initial centroid selection methods and Decision tree on the Canberra heart disease non-invasive data attributes is investigated. The results demonstrate that integrating the two clusters Outlier K-Means clustering methods and Decision tree gives an increase of mean sensitivity and mean accuracy, achieving 69.7% and 72.1% respectively. These results seem sufficiently high that the two clusters Outlier K-Means clustering Decision tree methods could be used to create a screening test for the evaluation of heart disease risk for use at a community level. The results of applying non-invasive attributes with these results are satisfactory at this exploratory stage.

This chapter develops the idea of a community-level screening test based on the data mining results reported here. First, it discusses expert systems including their components, advantages and limitations and their application in heart disease diagnosis. Second, an expert system risk evaluation tool is investigated, using the two clusters Outlier K-Means clustering Decision tree method on the Canberra heart disease non-invasive data attributes. Third, the Decision tree rules are extracted to create a decision

chart as a community screening test to help healthcare professionals diagnose high or low risk heart disease patients.

6.2. Expert System Overview

Experience owned by human experts is critically important in solving real time problems. This experience not only takes very long time to be learnt but also involves a large amount of knowledge. Expertise enables people to find solutions much faster and at lower costs. Researchers are trying to find tools that help in collecting, capturing and storing this expertise in a computer system that can help people in solving relevant problems (Giarratano and Riley 2004).

Expert systems are one of the most successful applications of artificial intelligence to real–world problems found in medicine, finance, weather, education and health. Artificial intelligence is the branch of computer science that develops machines and software with intelligence. It includes the study and design of intelligent agents that apply reasoning, knowledge, planning, learning and perception to solve problems (Giarratano and Riley 2004). Using expert systems in clinical laboratories is revolutionizing the practice of medicine providing computerized imaging techniques, assisting health professionals using advances in artificial intelligence technology, automating clinical laboratories and developing hospital information systems (Bronzino 1992). Expert systems are able to analyse complex medical data and can be applied in almost every field of medicine. The potential to exploit meaningful relationships within a dataset can be used in diagnosis, treatment and prediction in many clinical situations (Patel, Shortliffe et al. 2009).

Expert systems simulate the chain of reasoning of an expert in a specific problem domain (Wooldridge 2009). Feigenbaum and McCorduck et al. (Feigenbaum, McCorduck et al. 1988).defined expert systems as intelligent computer programs that use experts' knowledge to solve problems difficult enough to require significant expertise They are knowledge-based systems that employ knowledge of a specific application domain then use reasoning procedures to solve problems that require human competence or expertise. The significance of expert systems stems primarily from the specific knowledge about a narrow domain stored in the expert system's knowledge base (Ruan, D'hondt et al. 2006).

An expert system has three main interacting components (see Figure 6.1): the knowledge base; the inference engine; and the user interface (Giarratano and Riley 2004, Ruan, D'hondt et al. 2006). The knowledge base is the knowledge obtained from databases, books, and experts. It stores the facts and rules about a particular problem domain and makes them available to the inference engine in a form that it can use. The inference engine is the program that locates the appropriate knowledge in the knowledge base, and infers new knowledge by applying logical processing and problem-solving strategies. The inference engine is responsible for drawing conclusions from the knowledge base and presenting them to the user. The user interface is the interface that the user uses to interact with the system. It is the means of communication between the users and the expert systems problem-solving processes. The expert system is not very useful unless it has an effective interface. It accepts the queries (users' input) that the inference engine translates into working instructions for the rest of the system.. Careful attention is given to the interface design to make the expert system appear 'friendly' to the user (Giarratano and Riley 2004, Ruan, D'hondt et al. 2006).

The potential benefits of expert systems have encouraged researchers to implement them in several disease diagnosis applications. Different data mining techniques can be used to build expert systems to help healthcare professionals in the diagnosis of heart disease patients. Palaniappan and Awang (Palaniappan and Awang 2007) use Decision tree, Naïve Bayes and Neural Networks to build an expert system on the Cleveland heart disease dataset. Polat, Sahan et al., (Polat , Sahan et al. 2007) investigate applying an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism in the diagnosis of heart disease patients again using the Cleveland heart disease dataset suggesting that the results be used to build an expert system. However, these researches use a set of invasive and no-invasive attributes that limit their application as a community screening test.

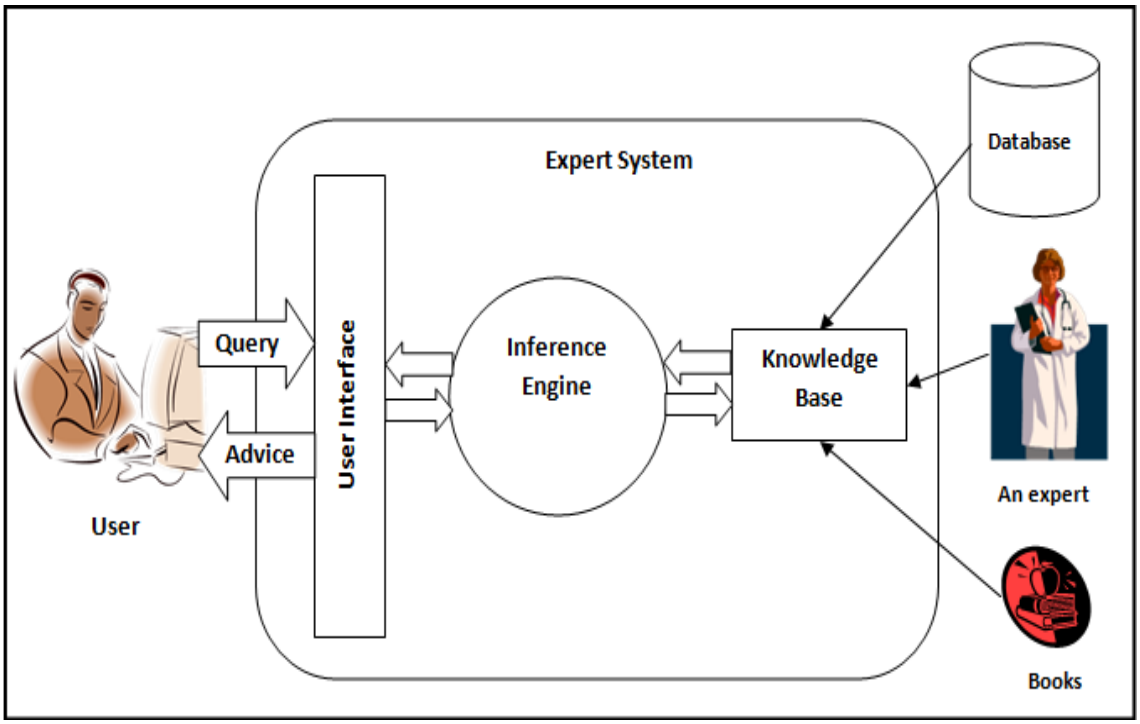

Figure 6.1: Expert System Components

6.3. Heart Disease Expert System Risk Evaluation Tool

In this section the development of an expert system for heart disease risk evaluation that uses non-invasive attributes is discussed. It assists healthcare professionals in identifying patients at risk of heart disease but also achieves this objective at low cost. The use of non-invasive data attributes also allows an expert system to identify the degree of risk of heart disease patients (see Chapter 4). The Framingham test and Australian Absolute Cardiovascular Risk Calculator (see Chapter 2 section 2.2.4) are well known heart disease screening tests that use combinations of invasive and non-invasive data attributes. This approach has been identified as limiting for a community-screening test as it needs pre-blood tests investigation. The proposed heart disease expert system risk evaluation tool is innovative in that it identifies the degree of risk of heart disease patients using ONLY non-invasive data attributes, thus supporting its application as a community screening test. For simplicity, we have called this system HD – ESRET (Heart Disease Expert System Risk Evaluation Tool).

Figure 6.2 shows the three main components of HD - ESRET: the knowledge base; inference engine; and the interface. The knowledge base applies the two-cluster Outlier Decision tree on the Canberra non-invasive heart disease data attributes to extract the expert system rules. The non-invasive attributes are the age, sex, systolic,

diastolic, and Rohrer's Index. The inference engine uses the extracted rules and the users' input to draw conclusions from the knowledge base and presents them to the user via the user interface. The user interface allows for "communication" screens where the user enters input data and the expert system returns the degree of heart disease risk as calculated by the inference engine.

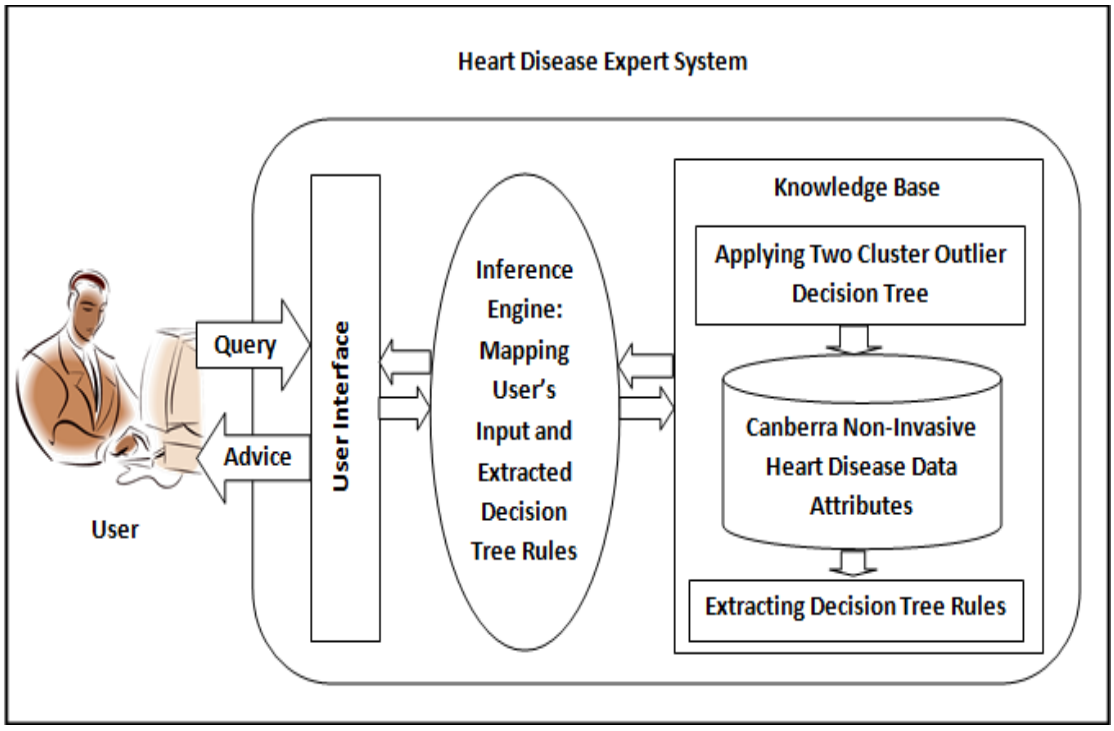

Figure 6.2: Heart Disease Expert System Evaluation Tool Components

The HD – ESRET has been implemented in two different forms: a proof-of-concept computer program and a proof-of-concept diagnostic chart which are presented in the following two sections respectively. The implementation of the HD – ESRET in the two different forms demonstrates how broad and useful the HD – ESRET application can be in identifying patients at risk of heart disease using non-invasive attributes.

6.4. HD - ESRET Implementation

The construction plan for HD - ESRET consists of two main phases (Figure 6.3). The first phase applies the Decision tree to the non-invasive attributes of the Canberra dataset. This phase include loading the Canberra non-invasive attributes (age, sex, systolic, diastolic, and Rohrer's Index), applying the two cluster Outlier K-Means clustering Decision tree to the loaded Canberra non-invasive data attributes and then the

diagnostic rules are extracted and stored. In the second phase the user enters his/her data: age, sex, height, weight, Systolic, and Diastolic. The height and weight values are used to calculate the Rohrer's Index using Equation 4.2 (see Chapter 4section 4.5). These attributes are used by the stored diagnostic rules to calculate the user's degree of risk of heart disease which is displayed to the user.

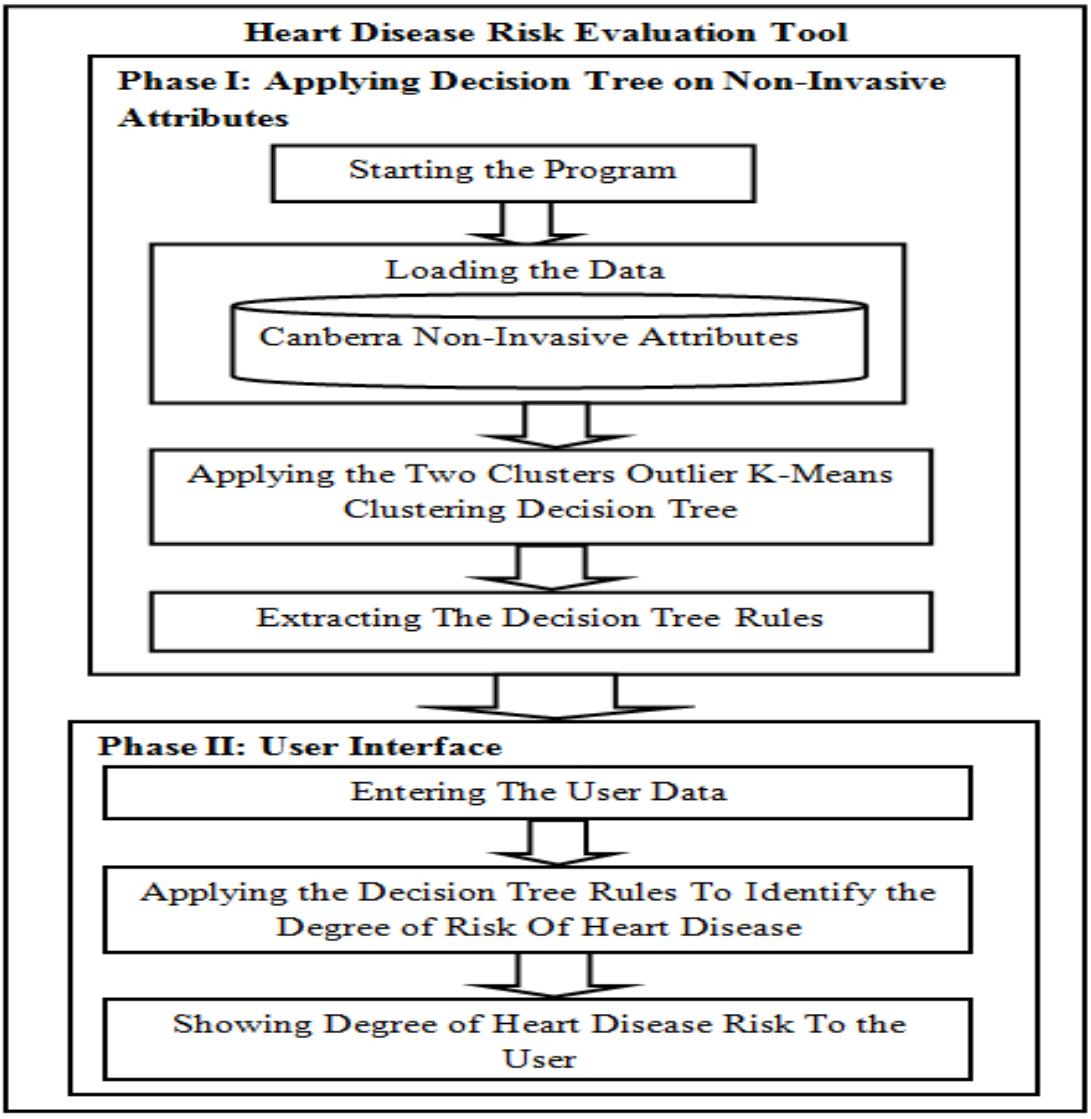

Figure 6.3: The Heart Disease Risk Evaluation Tool Design

An example of the implementation of the heart disease risk evaluation tool is described. The HD – ESRET is implemented using C# Microsoft Visual Studio 2008 for Windows.

Figure 6.4 shows the opening or start screen of HD – ESRET that loads the Canberra non-invasive data attributes to the expert system, applies the two clusters

Outlier K-Means clustering Decision tree, and extracts the diagnostic rules to be used in the next form.

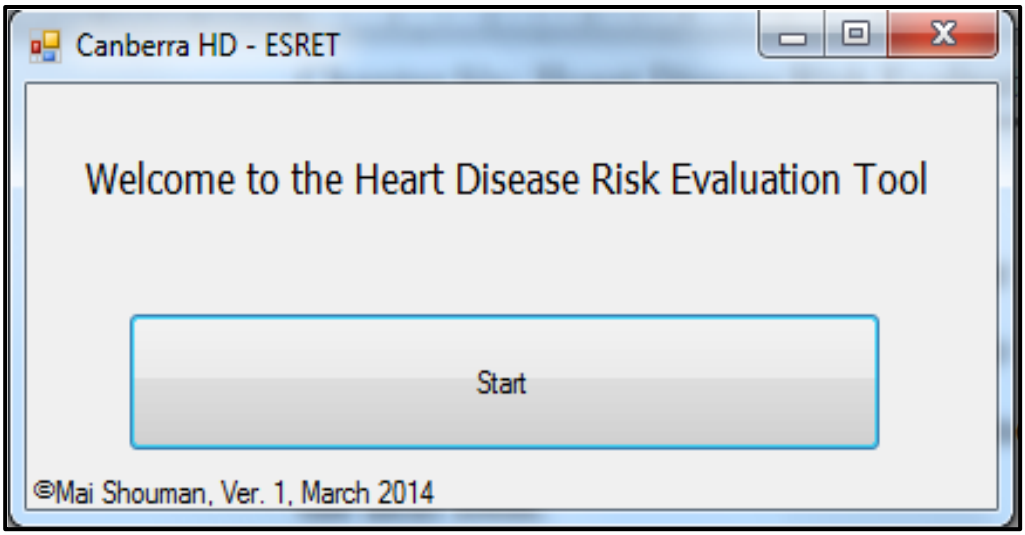

Figure 6.4: Starting the Heart Disease Risk Evaluation Tool

Figure 6.5 shows the user interface form where the user enters his/her data the degree of heart disease risk is calculated and displayed.

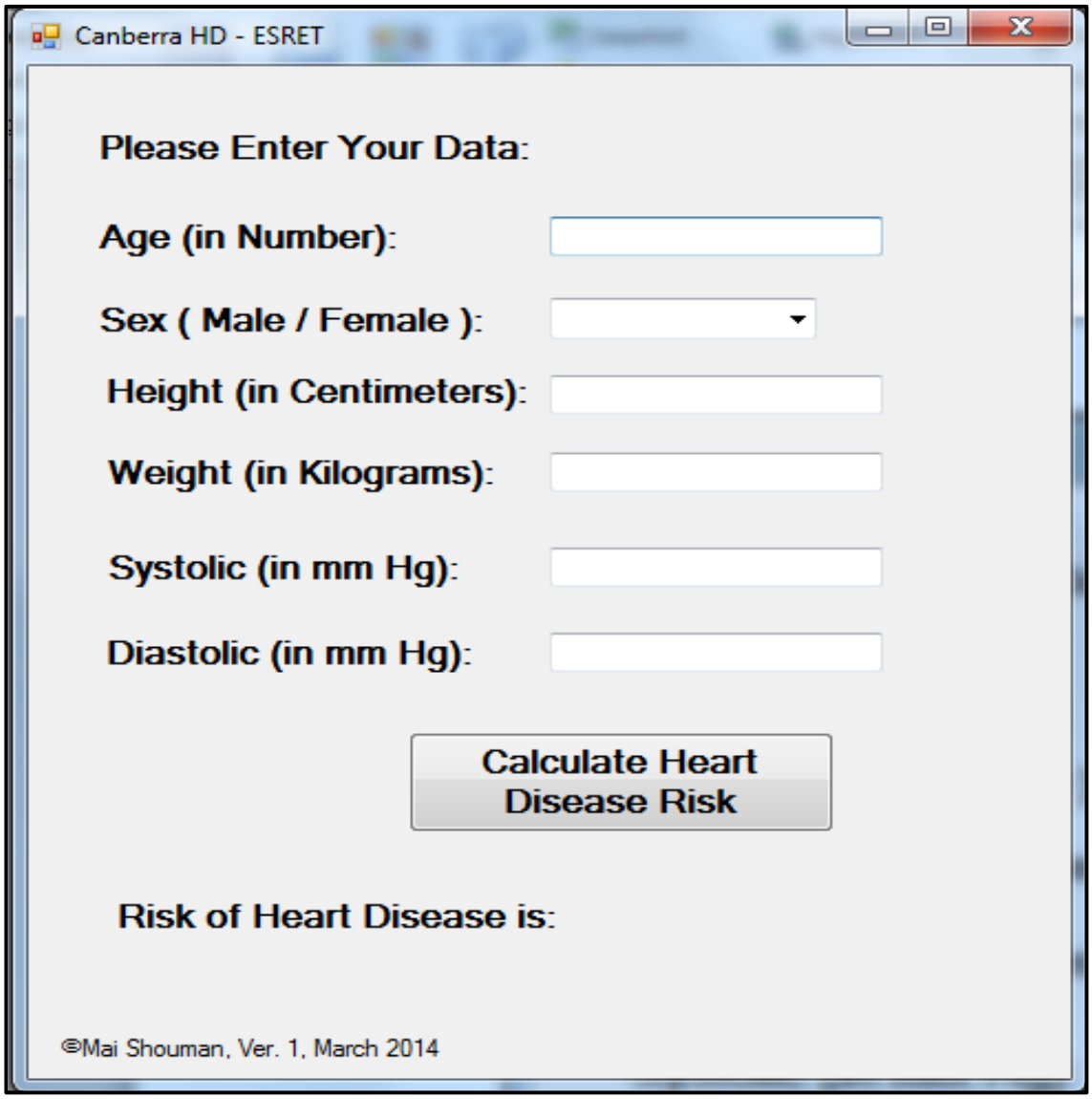

Figure 6.5: The Heart Disease Risk Evaluation Tool Interface

Figure 6.6 is a demonstration of the user data entry screen that allows HD – ESRET to calculate the degree of heart disease risk (in this case high for the entered data).

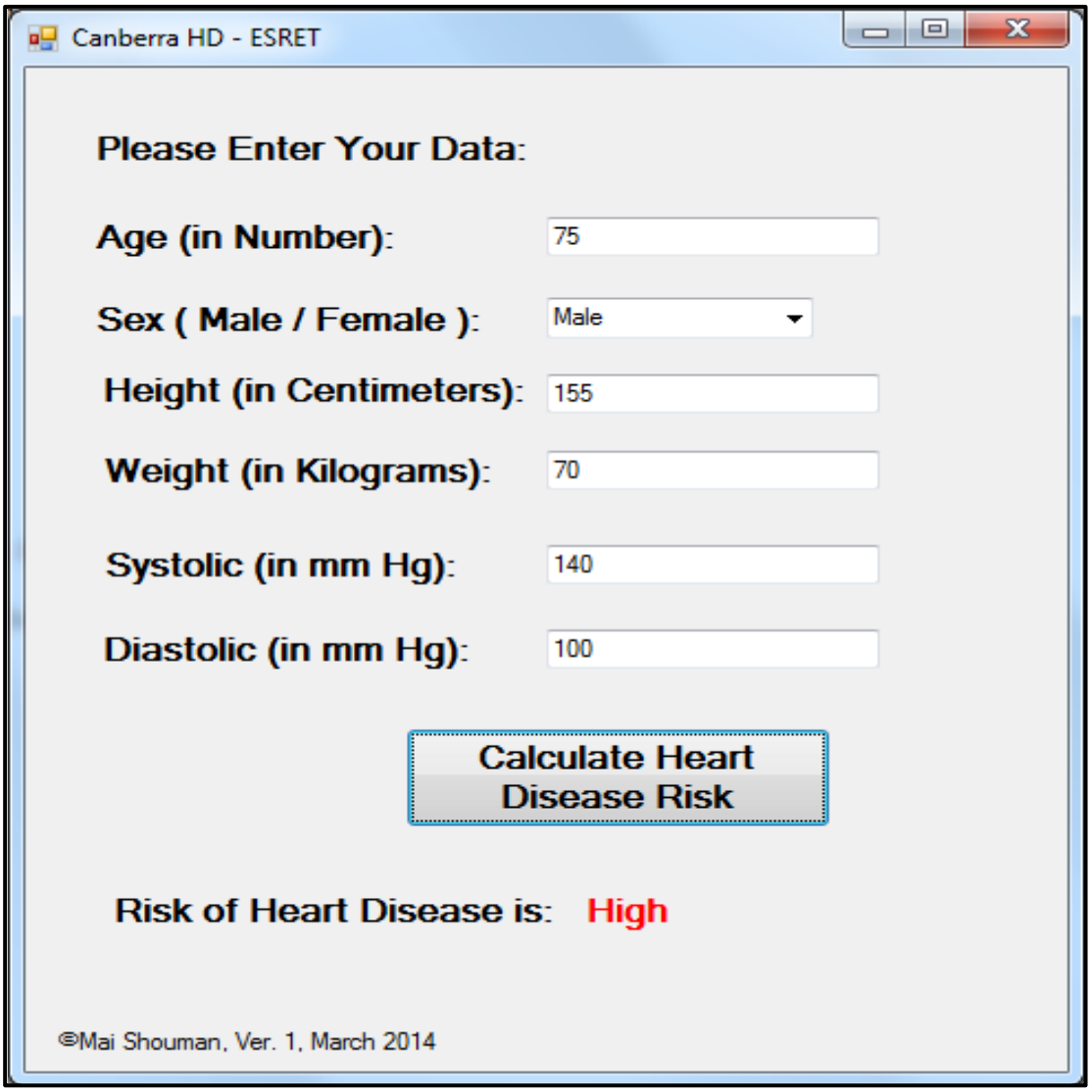

Figure 6.6: High Risk Heart Disease Risk Evaluation Example

Figure 6.7 is a second example of HD – ESRET, in this case a low risk of heart disease based on the entered data.

These examples demonstrate that HD - ESRET can act as a community-level screening test. The simplicity of the user interface allows healthcare professionals to identify patients at high risk of heart disease using very low cost non-invasive attributes. The HD – ESRET could be implemented on mobile applications as well as desktop applications too.

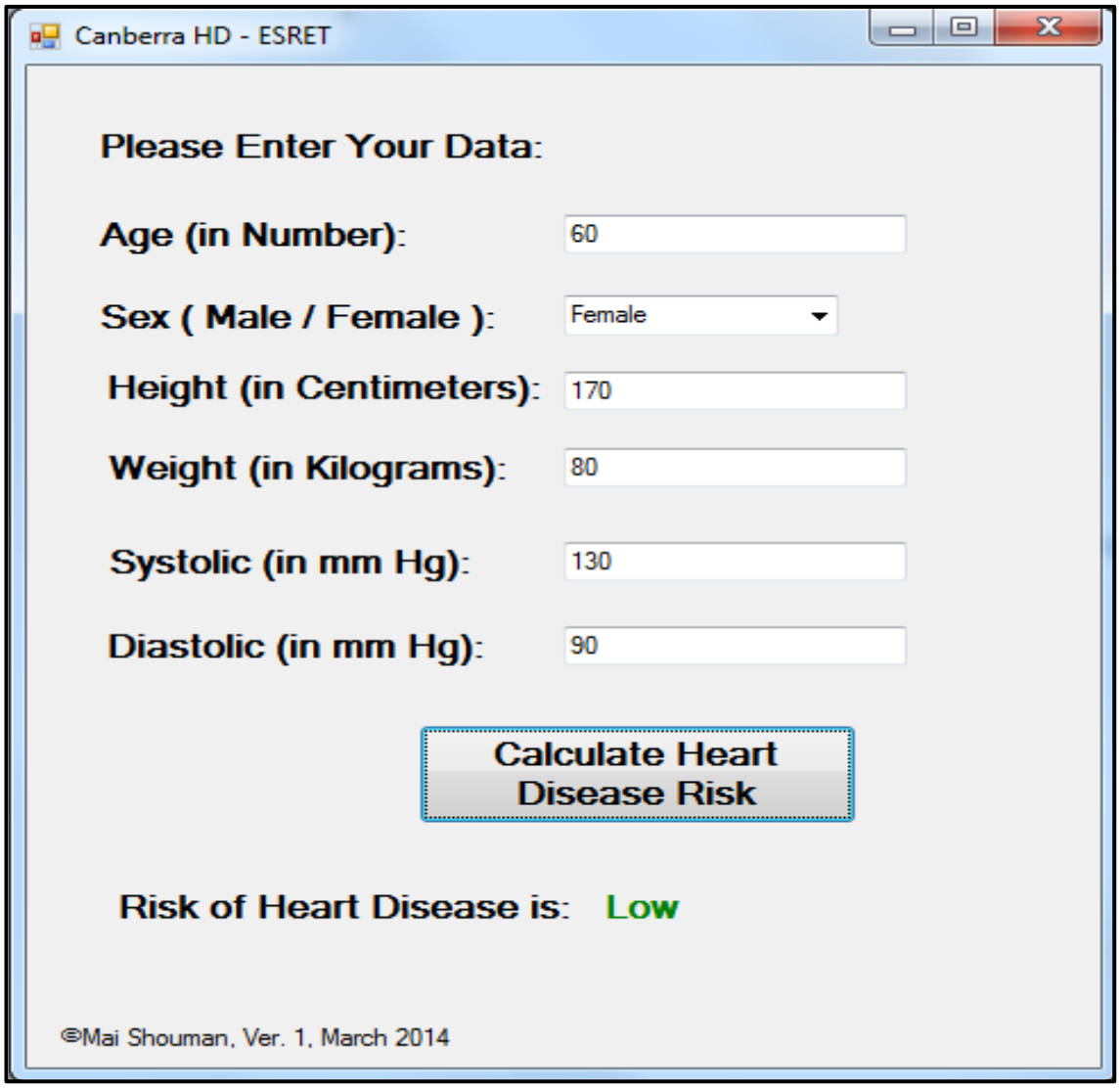

Figure 6.7: Low Risk Heart Disease Risk Evaluation Example

6.5. HD – ESRET DT Decision Rules, Tree and Chart

The results from Chapter 4 and 5 results demonstrate that the two clusters Outlier K-Means clustering Decision tree method provides reliable performance using the age, sex, resting blood pressure (Systolic and Diastolic) and Rohrer's Index non-invasive Canberra attributes (mean sensitivity and mean accuracy of 69.7% and 72.1% respectively), and can be used to create a community-level screening test. To build the diagnostic chart and Decision Tree rules are extracted.

The Canberra non-invasive heart disease data attributes combination contains four continuous attributes age, resting blood pressure (Systolic and Diastolic), and Rohrer's Index and one discrete attribute (sex). A Decision tree is unable to deal with continuous attributes (see Chapter 3, section 3.2). Equal-frequency discretisation is

applied to convert attributes from continuous to discrete. Figure 6.8 shows the attributes' cut-points and the range for each cut-point. The attributes cut-points are used by the two clusters Outlier K-Means clustering Decision Tree method to create the Decision tree rules.

Age 0: < 65 1: ≥ 65 & < 73 2: ≥73 & < 79 3: ≥79 & < 86 4: ≥86	**Rohrer's Index** 0: < 14 1: ≥ 14 & < 16 2: ≥ 16 & < 17 3: ≥ 17 & < 19 4: ≥19
Systolic 0: < 122 1: ≥ 122 & < 134 2: ≥ 134 & <143 3: ≥ 143 & <154 4: ≥ 154	**Diastolic** 0: < 66 1: ≥ 66 & < 74 2: ≥ 74 & < 80 3: ≥ 80 & < 90 4: ≥90

Figure 6.8: Attribute cut-point and range in Heart Disease Risk Evaluation Rules

Figure 6.9 and Figure 6.10 are samples of the Decision tree rules for the first and second cluster. All rules are presented in Appendix C. Figure 6.9 is for the first cluster where age is less than seventy-five; and Figure 6.10 is for the second cluster where age is greater than or equal to seventy-five. These examples of extracted rules explain how the non-invasive attributes are used to identify the degree of risk of heart disease patients. In the sex attribute, zero corresponds to male, and one corresponds to female. The values for age, Rohrer's Index, Systolic and Diastolic in Figure 6.9 and 6.10 are described in Figure 6.8.

Although Figure 6.9 and 6.10 are clustered based on the age attribute, the age attribute is also still used in the Decision Tree rules for further analysis. For instance, in Figure 6.9, the cluster is for age less than 75. Discretized Age values equal to zero, one, and two are used in the Decision Tree rules representing people who are less than 65, between 65 and 73, and between 73 and 79 years old respectively (Figure 6.8). In Figure 6.10, the cluster is for age greater than or equal to 75. Discretized Age values equal to three and four are used in the Decision Tree rules representing people aged between 79 and 86, and greater than 86 years old respectively (Figure 6.8).

The First Cluster the Age < 75

SEX = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 1, Diastolic = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 1, Diastolic = 4 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 0, Diastolic = 3 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 0 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 3 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 3 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 4 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 4 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 3 => Then the Target Value Equals 0

SEX = 0, Age = 2 => Then the Target Value Equals 1

Figure 6.9: Sample of the First Cluster Heart Disease Risk Evaluation Rules

The Second Cluster Age >=75

SEX = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 0, Age = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 0, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 1 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 1 => Then the Target Value Equals 0

SEX = 1, Diastolic = 2, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 1, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 1, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 0 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 4, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 4, Age = 4 => Then the Target Value Equals 0

Figure 6.10: Sample of the Second Cluster Heart Disease Risk Evaluation Rules

Figure 6.11 and 6.12 shows the diagnostic rules for the first and second clusters for the HD – ESRET. The diagnostic rules in Figures 6.11 and 6.12 are drawn from the Decision Tree rules in Figures 6.9 and 6.10 respectively. The value of zero in the Sex attribute corresponds to male, and one corresponds to female. The values for age, Rohrer's Index, Systolic and Diastolic in Figure 6.11 and 6.12 are described in Figure 6.8. The diagnostic rules of Figures 6.11 and 6.12 can be used to identify the degree of risk of heart disease patients using the non-invasive attributes. The No and Yes in Figure 6.11 and 6.12 correspond to no risk and at risk in the risk evaluation of heart disease respectively.

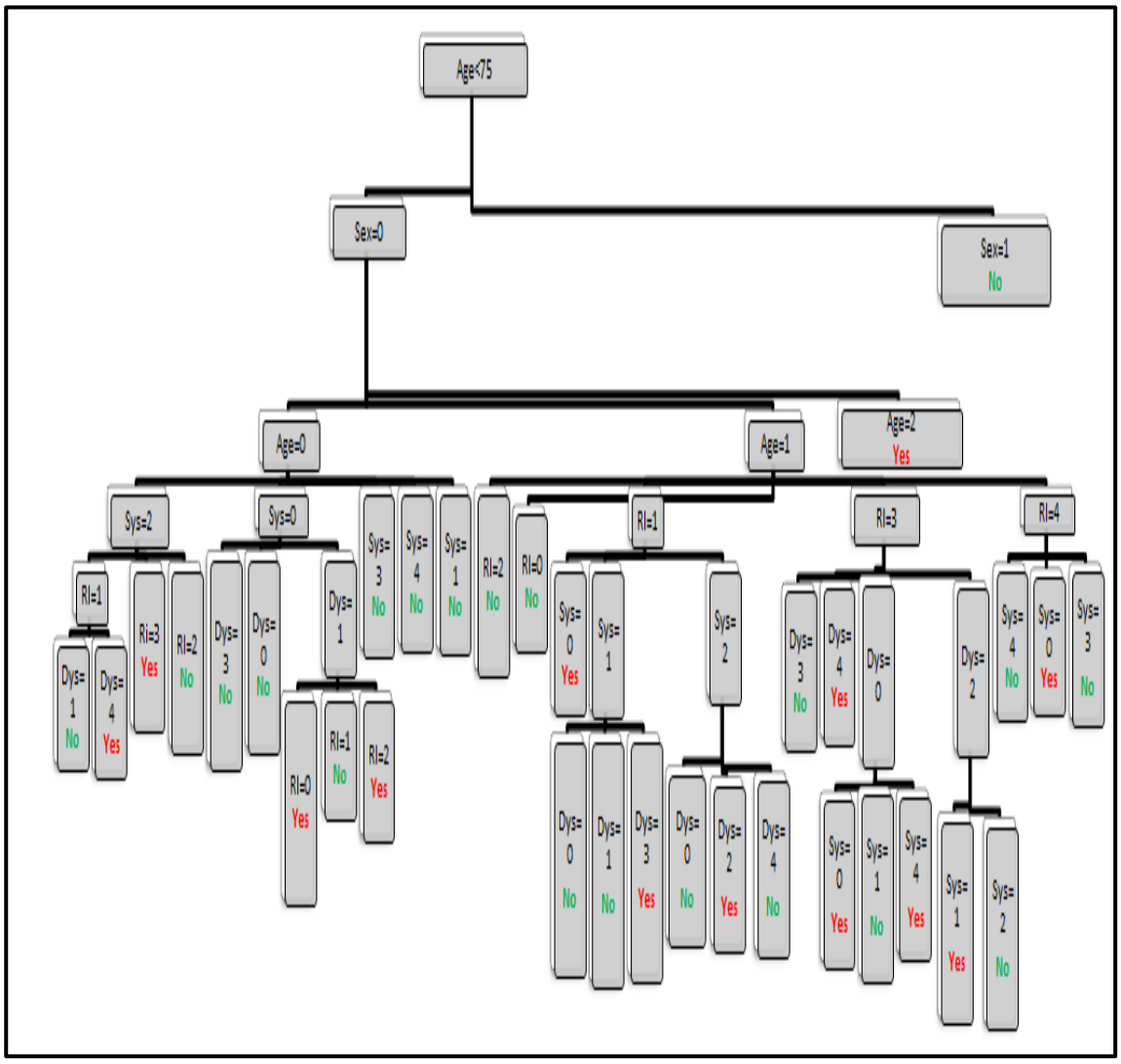

Figure 6.11: The First Cluster Heart Disease Risk Evaluation Decision Tree

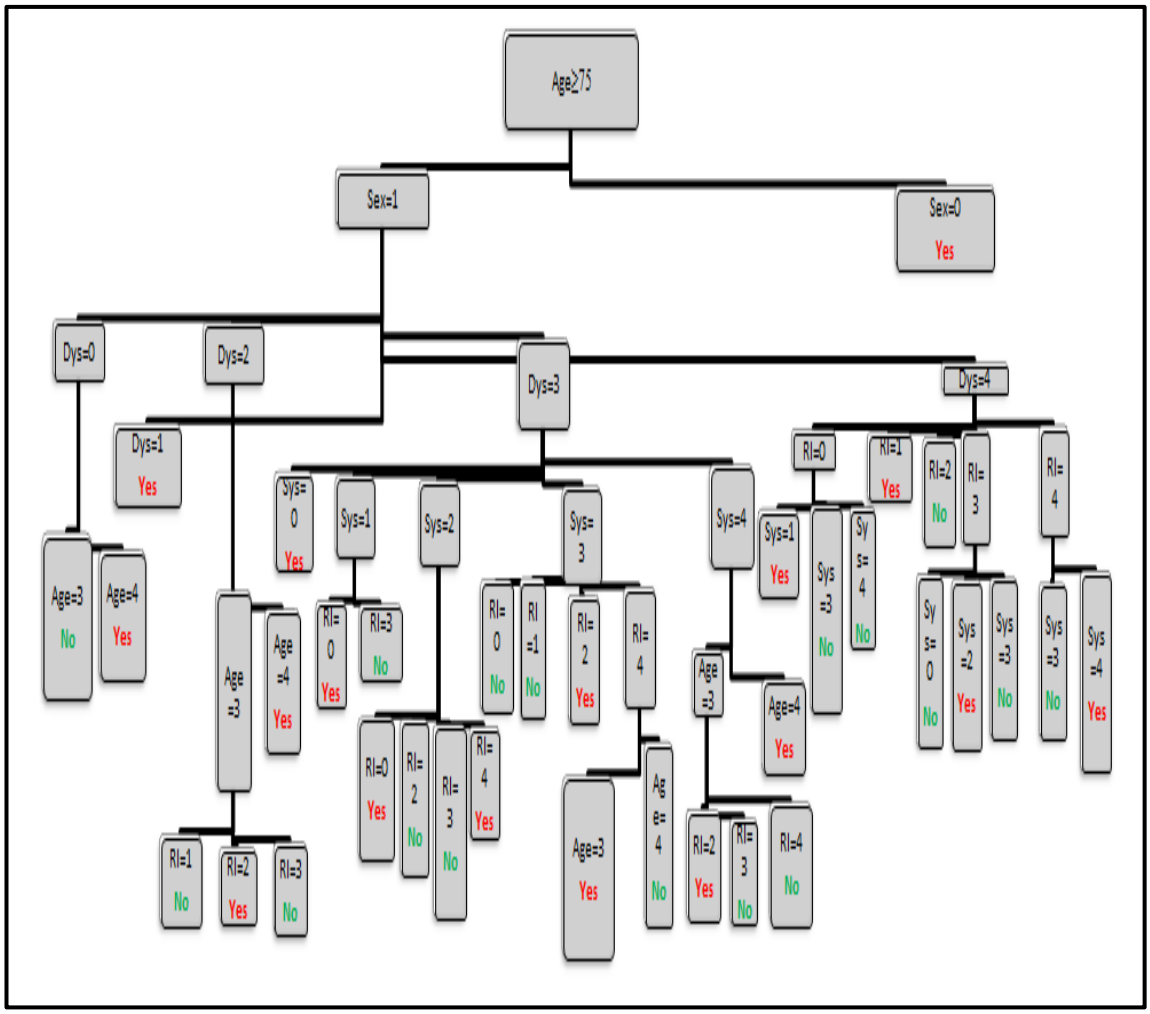

Figure 6.12: The Second Cluster Heart Disease Risk Evaluation Decision Tree

The two clusters Outlier K-Means clustering Decision Tree rules for the non-invasive heart disease data attributes of the Canberra dataset are used to create the diagnostic chart that identifies patients at high risk of heart disease. Figure 6.13 shows the definition of symbols for the non-invasive heart disease evaluation chart. It shows the range for each of Systolic (ranging from “a” to “e”) and Diastolic (ranging from “f” to “j”) attributes. It also shows the colours of the cells used in Figure 6.14 and Figure 6.15 showing green for low risk and red for high risk.

Figure 6.14 and Figure 6.15 shows the first and second clusters for the heart disease risk evaluation chart respectively. Figure 6.14 shows the heart disease risk evaluation for patients whose age is less than seventy-five. It shows that females whose age is less than seventy-five are at low risk of heart disease. However, for male patients whose age is less than seventy-five the attributes demonstrate varying degrees of risk of having heart disease. Figure 6.15 shows the heart disease risk evaluation for patients

whose age is greater than seventy-five. It shows that males of greater than seventy-five are at high risk of heart disease. However, the females whose age is greater than seventy-five the attributes demonstrate varying degrees of risk of having heart disease.

Systolic:	**Diastolic:**
a: < 122	**f:** < 66
b: ≥ 122 & < 134	**g:** ≥ 66 & < 74
c: ≥ 134 & <143	**h:** ≥ 74 & < 80
d: ≥ 143 & <154	**i:** ≥ 80 & < 90
e: ≥ 154	**j:** ≥90

Colours:
Green: Low Risk
Red: High Risk

Figure 6.13: Needed Symbols in Non-Invasive Heart Disease Evaluation Chart

For example, assume patient data is:

Age = 74, Sex = Male, Rohrer's Index = 15, Systolic = 130, and Diastolic = 75

To identify the degree of heart disease risk for the above patient, we need to identify to which cluster this patient belongs. Age is less than 75, so patient belongs to the first cluster (Figure 6.14). Sex is male so the patient belongs to the Tables in Figure 6.12. The Systolic measure is 130, so the patient belongs to the "d" category (Figure 6.13). The Diastolic measure is 75, so the patient belongs to the "h" category (Figure 6.13). Figure 6.15 identifies the degree of heart disease risk (given sex= Male, Rohrer's Index = 15, Systolic = c, and Diastolic = h) as high.

A second example confirms the process:

Age = 80, Sex = Female, Rohrer's Index = 18, Systolic = 145, and Diastolic = 95

Age is greater than 75 so this patient belongs to the second cluster (Figure 6.15). Sex is female, so patient belongs to the Tables in Figure 6.15. The Systolic measure is 145, so the patient belongs to the "d" category (Figure 6.13). The Diastolic measure is 95 so patient belongs to the "j" category (Figure 6.13). Figure 6.13 identifies

the degree of heart disease risk (given sex = Female, Rohrer's Index = 18, Systolic = d, and Diastolic = j) is low.

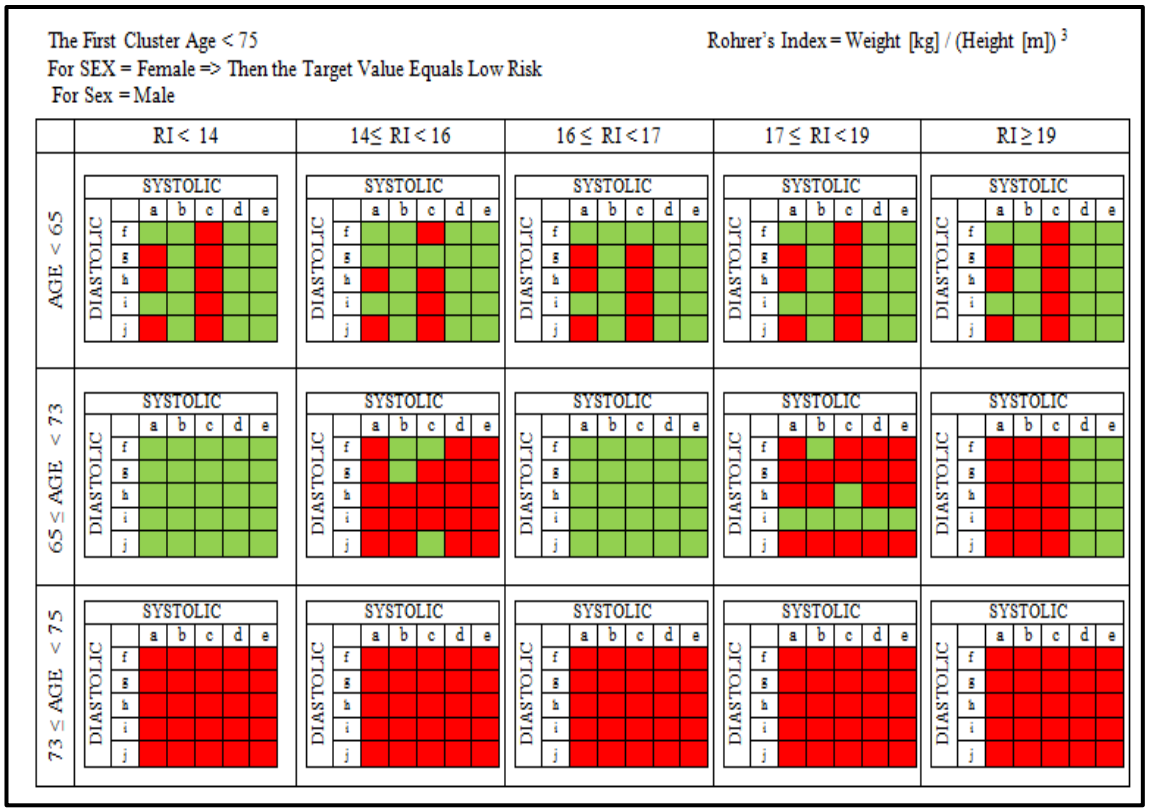

Figure 6.14: First Cluster Heart Disease Risk Evaluation Chart

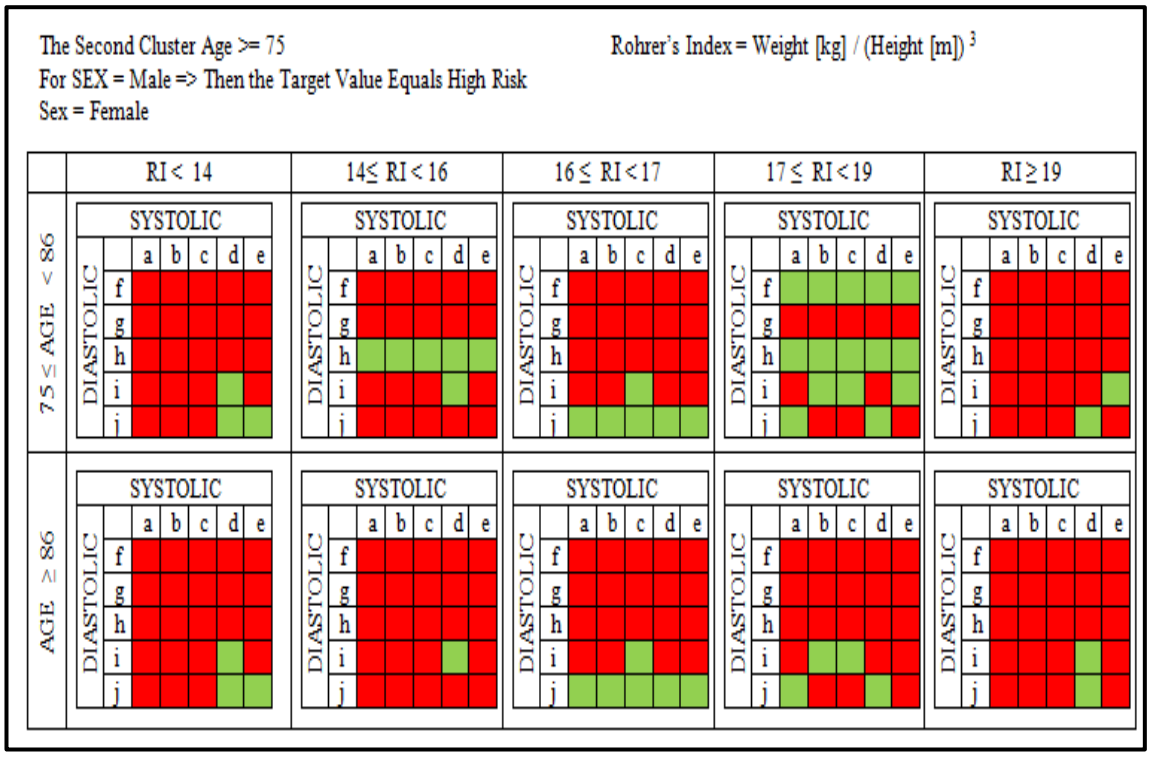

Figure 6.15: Second Cluster Heart Disease Risk Evaluation Chart

Thus, the extracted decision tree rules can be used in medical environments to help healthcare professionals establish a patient's risk of heart disease using non-invasive attributes with a low-cost, reliable and effective community-level screening prototype evaluation tool.

6.6. Chapter Summary and Conclusion

Researchers have been using expert systems in the diagnosis of several diseases such as breast cancer, lungs and heart disease. Although heart disease can be detected by several tests such as electrocardiogram, stress tests, and cardiac angiogram, these tests are expensive and cannot be used as community-screening tests. So there is a need to find less expensive tests that can be used for community-level screening to identify patients at high risk of heart disease. There is a need for accurate systematic tools that identify patients at high risk and provide information for early intervention in heart diseases.

This chapter discusses the construction of a heart disease expert system risk evaluation tool (HD – ESRET) using the two clusters Outlier K-Means clustering Decision Tree method on the Canberra heart disease dataset non-invasive data attributes. HD - ESRET is novel in that it is able to identify the degree of risk of heart disease patients using a novel low-cost non-invasive attributes combination. Decision tree rules are extracted to create a diagnostic chart that identifies if a patient is at high or low risk. HD - ESRET acts as a community-level screening test that identifies the degree of risk of heart disease in patients as being high or low, thus providing healthcare professionals with inexpensive, reliable screening of potential heart disease patients.

Importantly, the HD – ESRET implementation results as well as the extracted diagnostic rules and charts are indicative but not definitive as they are based on only the Canberra heart disease dataset. This limits the geographic and socio-economic range of the patients for whom records have been created. The Canberra heart disease dataset contains 460 rows as discussed earlier (see Chapter 3, section 3.5), which is still a small number of patients' rows. However, the ability to use non-invasive attributes in the risk evaluation of heart disease patients showing mean sensitivity and mean accuracy of 69.7% and 72.1% respectively is an important contribution of this thesis.

The next chapter concludes the main findings of this thesis. It presents the research objectives, summarizes the research conclusions followed by discussion of the research limitations and future directions. Finally, the main contributions of this thesis are presented.

[This Page is Left Blank Intentionally]

Chapter 7

Conclusions and Future Work

This chapter briefly presents the research objectives, summarizes the research conclusions, as well as discussing the limitations, describing the possible future research directions followed by presenting the main contributions of this thesis.

7.1. Introduction

Heart disease is the leading cause of death in the world over the past decade in different continents and countries regardless of their income (World Health Organization 2011b). The World Health Organization (WHO) reported that heart disease is the leading cause of death all over the world, causing 7.25 million deaths, representing 12.8% of all deaths (World Health Organization 2013c). Although heart diseases are among the most common chronic diseases causing a high rate of death all over the world, they have also been identified as among the most preventable and controllable diseases (Centers for Disease Control and Prevention 2013). Early detection of heart disease patients can help in recovering patients' health and decreasing the mortality rate from heart disease (Centers for Disease Control and Prevention 2013). The relative success of chronic disease treatments are dependent on the earliness of detection of those diseases (Paladugu and Shyu 2010). There is a vital need for accurate systematic tools that identify patients at high risk and provide information for early detection of heart disease (Paladugu and Shyu 2010).

Although heart disease can be detected by several tests such as electrocardiogram, stress tests, and cardiac angiogram, these tests are expensive and cannot be used as community-level screening tests. The Framingham Heart Disease Risk Evaluation tool and the Australian Absolute Cardiovascular Risk Calculator are two famous heart disease risk evaluation tools that help to identify patients at risk of heart disease. Both of the two heart disease risk evaluation tests use a set of invasive and non-invasive attributes, such as age, sex, systolic blood pressure, total cholesterol, diabetes and smoking status, to identify if a patient is at high, moderate or low risk of heart disease (National Heart Foundation of Australia 2009, Framingham Study. 2013). However, the tools need prior blood test investigations to identify the cholesterol and

diabetes levels. Using invasive attributes slows down and adds expense to the tests for the risk evaluation of heart disease. There is a need to find lower cost tests that can be used for community-level screening to identify the risk of heart disease.

7.2. Research Objective

Motivated by the increasing mortality rates of heart disease all over the world and the fact that early detection helps in recovering patients' health and decreasing the mortality rate from heart disease, the main objective of this thesis is helping healthcare professionals in the early detection and risk evaluation of heart disease patients. To achieve this objective, this research poses the question:

Can data mining assist healthcare professionals in the early detection of heart disease in a community setting?

Although researchers have been applying different data mining techniques to help healthcare professionals in the diagnosis of heart disease patients, there is not a clear view of different data mining techniques' performance across different data attribute combinations in the diagnosis of heart disease patients. The main objective of this thesis is, in answering the above question,:

To provide healthcare professionals with a community-level screening tool for the early detection and risk evaluation of heart disease patients

To achieve this, the research is focussed on the following key questions:

1. Can significant attributes in the diagnosis of heart disease patients be identified?
    - Chapter 3 demonstrates that significant attributes are identifiable.
2. Can applying data mining techniques on non-invasive attributes be usefully applied to the diagnosis of heart disease patients?
    - Chapter 4 demonstrates that data mining techniques can be usefully applied on non-invasive attributes.
3. Can hybrid data mining techniques be usefully applied to enhance performance on non-invasive data attributes in the diagnosis of heart disease patients?
    - Chapter 5 demonstrates that hybrid data mining techniques can be usefully applied to enhance performance.

4. Can a reliable heart disease expert system risk evaluation tool, using non-invasive heart disease data attributes, be constructed?
    - Chapter 6 demonstrates that a reliable expert system tool, using non-invasive data attributes can be constructed.

The results from Chapters 3 – 6 that support and confirm answers to the key questions posed in this thesis also demonstrate support for the principal question that "Data Mining can Assist Healthcare Professionals in the Early Detection of Heart Disease in a Community Screening Environment".

7.3. Research Conclusions

Motivated by the increasing heart disease mortality rates all over the world, researchers are using data mining techniques to help healthcare professionals in the diagnosis of heart disease patients. Although heart disease can be detected by several tests such as electrocardiogram, stress tests, and cardiac angiogram, these tests are expensive and cannot easily be used as community-level screening tests. There is a need to find lower cost tests that can be used for community-level screening to identify the risk of heart disease. The main objective of this thesis is to help healthcare professionals in the risk evaluation of heart disease patients using low cost data attributes. The research here has investigated whether data mining techniques can be applied with reliable accuracy to only non-invasive attributes of heart disease patient records to create diagnostic rules to inform a community-level heart disease risk evaluation tool. The details of the contributions of this work are discussed in the next few sub-sections.

7.3.1. Significant Attributes in Heart Disease Risk Evaluation

Chapter 3 identifies the significant attributes needed by data mining techniques in the diagnosis of heart disease patients. It investigates applying three common and successful data mining techniques (Decision Tree, Naïve Bayes, and K-nearest Nearest Neighbour) to two different heart disease datasets (the Cleveland Heart Disease Dataset and a new Canberra Heart Disease dataset). Applying different data mining techniques to both the Cleveland and Canberra heart disease datasets shows reliable results. Although different data mining techniques show reliable results over both heart disease datasets using all attributes of each dataset, the two datasets have different data attribute sets. When mapping the Cleveland and Canberra heart disease data set attributes, it

appears that there are four common data attributes: age, sex, resting blood pressure and peak heart rate. The results attained imply that these four attributes can be of significant effect in the diagnosis of heart disease patients.

Decision Tree, Naïve Bayes, and K-nearest Nearest Neighbour data mining techniques are then applied using the common data attributes of the Cleveland and Canberra datasets to identify if the common attributes have significant effect the reliability of the techniques in the diagnosis of heart disease patients. The results show that the techniques remain reliable. The results show that the Decision Tree is the most stable data mining technique using the common attributes of the Cleveland and Canberra heart disease dataset. The accuracy attained using only the four common attributes, of which three are non-invasive attributes, raises an important question about the effect of different non-invasive attributes' performance in the diagnosis of heart disease patients using data mining techniques. Table 7.1 summarises the performance of the different data mining techniques applied to different Cleveland and Canberra dataset attributes.

Table 7.1: Applying Different Data Mining Techniques on Cleveland and Canberra Datasets using Different Attribute Combinations

Dataset	**Data Mining Technique**	**Mean Accuracy**					
		All Data Attributes		**Common Data Attributes**		**PCA Attributes**	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Cleveland	Decision Tree	79.1%	5.8%	69.6%	7.5%	76.6%	5%
	Naïve Bayes	83.5%	5.2%	72.4%	8.2%	79.3%	5.5%
	K-Nearest Neighbour	83.4%	2.7%	63%	10%	79.2%	4.1%
Canberra	Decision Tree	68.9%	7.8%	75.1%	6.6%	67.4%	10.5%
	Naïve Bayes	75.4%	9.1%	73.3%	6%	70.8%	9%
	K-Nearest Neighbour	68.8%	8.6%	67.2%	11.7%	64%	10.5%

7.3.2. Non-Invasive Attributes' Significance in Risk Evaluation of Heart Disease

Chapter 4 identifies the significance of different non-invasive attributes in the diagnosis of heart disease patients. The main importance of non-invasive attributes is that they are low cost attributes and can be used in community-level screening tests to identify patients at risk of heart disease. Chapter 4 investigates if there is a combination of non-invasive data attributes that can provide reliable data mining performance in the diagnosis of heart disease patients. Chapter 4 investigates applying the Gain Ratio Decision Tree technique to different single, combined, and calculated non-invasive data attribute combinations to identify which combination will show the best performance in the diagnosis of heart disease patients.

Table 7.2 summarizes the results of applying the Decision Tree technique on different non-invasive heart disease attribute combinations. When investigating the effect of different single non-invasive Cleveland and Canberra attributes in the diagnosis of heart disease patients, the age and sex attributes show the best results compared to other single non-invasive attributes. When using different combinations of non-invasive attributes of the Cleveland and Canberra datasets, the age, sex, and resting blood pressure combination shows the best results compared to other combinations. Finally, when calculated non-invasive attributes (BMI, Rohrer's Index and RBPDiff) are combined with the other non-invasive attributes, the results shows that the best combination in the diagnosis of heart disease in the Canberra dataset is the age, sex, resting blood pressure and Rohrer's Index combination, with mean accuracy of 73.8% (standard deviation 4.9%).

Table 7.2: Data Mining Diagnosis Results using Non-Invasive Attribute Combinations on the Cleveland and Canberra Heart Disease Datasets

Dataset	Data Attributes	Accuracy		Sensitivity	
		Mean	St Dev	Mean	St Dev
Cleveland	Age, Sex	65.2%	7.5%	60.4%	9.1%
	Age, Sex, RBP	65.8%	8.6%	61.2%	10.1%
Canberra	Age, Sex, RBP	74.8%	6.5%	66.7%	13.3%
	Age, Sex, RBP, Rohrer's Index	73.8%	4.9%	67.1%	11.4%

7.3.3. Integrating Clustering with Decision Tree in Heart Disease Risk Evaluation

Recently researchers are suggesting that integrating more than one data mining technique can enhance data mining performance in the diagnosis of heart disease. Chapter 5 investigates integrating K-Means clustering with different initial centroid selection methods with Decision Tree in the diagnosis of heart disease patients across the Cleveland and Canberra heart disease datasets. Although K-Means clustering is one of the most popular and well-known clustering techniques, initial centroid selection is a critical factor that strongly affects performance. Different clustering attributes were tested and the age attribute as the clustering column shows the best results.

Table 7.3 summarizes the results of integrating Decision Tree with K-Means clustering on the Cleveland and Canberra datasets using all attributes and non-invasive data attribute combinations. The Inlier initial centroid selection method shows the best results for both the Cleveland and Canberra heart disease datasets when using all attributes. Integrating K-Means clustering with Decision Tree could enhance diagnostic accuracy in both Cleveland and Canberra datasets, showing a mean accuracy of 81.2% and 71.5% respectively.

Chapter 5 goes on to investigate integrating K-Means clustering with different initial centroid selection methods and Decision Tree using non-invasive attributes of the Cleveland and Canberra heart disease datasets. Integrating K-Means clustering with Decision Tree could not enhance diagnostic accuracy or sensitivity when using non-invasive attributes of the Cleveland heart disease dataset. Moreover, the mean sensitivity degraded to 38.4%. However, integrating K-Means clustering and Decision Tree using the non-invasive attributes of the Canberra heart disease dataset shows an increase of the mean sensitivity, achieving 69.7%, with a mean accuracy of 72.1%. These results are very interesting as it shows that the non-invasive Canberra heart disease attributes (involving age, sex, resting blood pressure and Rohrer's Index) could be used to create a community-level screening test for the evaluation of heart disease risk.

Table 7.3: Integrating Decision Tree with K-Means Clustering on Cleveland and Canberra all and Non Invasive Data Attributes

Dataset	**Data Mining Technique**	**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Cleveland All Data	Decision Tree	75.6%	6.1%	81.6%	12.1%	79.1%	5.8%
	Two Clusters Inlier K-Means Decision Tree	75.9%	7.2%	85.1%	11.4%	81.2%	6.2%
Canberra All Data	Decision Tree	67.7%	13.6%	64.5%	18%	68.9%	7.8%
	Three Clusters Inlier K-Means Decision Tree	63.1%	13.1%	72.9%	14.6%	71.5%	6.7%

Dataset	Data Mining Technique	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Cleveland Non Invasive Data	Decision Tree	61.2%	10.1%	69.5%	10.1%	65.8%	8.6%
	Two Clusters Outlier K-Means Decision Tree	38.4%	17.3%	85.1%	8.1%	64.5%	12.5%
Canberra Non Invasive Data	Decision Tree	67.1%	11.4%	75.5%	9.2%	73.8%	4.9%
	Two Clusters Outlier K-Means Decision Tree	69.7%	9.9%	71.9%	11.3%	72.1%	5.1%

7.3.4. Building A Heart Disease Expert System Risk Evaluation Tool

Although heart disease can be detected by several tests such as electrocardiogram, stress tests, and cardiac angiogram, these tests are expensive and cannot be used as community-level screening tests. Community-level screening tests play an important role in the early detection of heart diseases. So, there is a strong need to find low-cost tests that can be used for community-level screening to identify patients at high risk of heart disease. There is a need for accurate systematic tools that identify patients at high risk and provide information for early intervention of heart diseases.

Chapter 6 investigates building a heart disease expert system risk evaluation tool based on the two clusters Outlier K-Means clustering Decision Tree method applied to the non-invasive data attributes of the Canberra heart disease dataset. The main contribution in this heart disease expert system risk evaluation tool is its ability to identify the degree of risk of heart disease patients using only low-cost non-invasive attributes. This expert system risk evaluation tool can act as a community-level screening test that can identify the risk of heart disease as being high or low. The decision tree rules are extracted to create a diagnostic chart (Figures 6.14 and 6.15) to

identify if a patient is at high or low risk of heart disease using only non-invasive data attributes.

7.4. Research Limitations

This research has adopted a thorough and systematic approach to investigating how data mining techniques could inform healthcare professionals in the risk evaluation of heart disease patients using low cost data attributes. There are some limitations to the research, including:

- Using only three common data mining techniques: Decision Tree, Naïve Bayes, and K-Nearest Neighbour.
- Using the Decision Tree for its ability to produce rules that can explain to healthcare professionals how the risk evaluation is decided, when the Naïve Bayes technique showed slightly better performance in the risk evaluation.
- Using non-invasive attributes such as age, sex, resting blood pressure, height, and weight only while not using other non-invasive attributes such as smoking, family history of heart disease, physical activity, and socio-economic levels.
- Testing the data mining techniques on the Cleveland and Canberra datasets, which have relatively small numbers of rows (297 and 460 respectively).
- Using the Canberra heart disease dataset to build the HD – ESRET as it is based on a specific city population dataset.
- No usability testing of the proof-of-concept evaluation tools with community-level screening healthcare providers (e.g. pharmacists).

7.5. Future Work

To overcome the limitations listed above, further investigation in the following directions is needed:

- Investigating the performance of other data mining techniques such as Neural Network, Support Vector Machine, and Genetic Algorithm using all available attributes and only non-invasive attributes in the diagnosis of heart disease patients against the same datasets to develop comparative performance results.

- Investigating applying other hybrid data mining techniques on all attributes and only non-invasive data attributes in the diagnosis of heart disease patients against the same datasets to develop comparative performance results.
  - Studying the significance of adding other non-invasive attributes (such as family history of heart disease, physical activity, smoking and socio-economic level) on the performance of different data mining techniques in the risk evaluation of heart disease patients.
  - Identifying the significance of controlled non-invasive attributes such as weight and smoking on different age and sex groups in the risk evaluation of heart disease.
  - Affirming the results found here by using larger datasets and from a variety of places.
  - Investigating applying other initial centroid selection methods for K-Means clustering in the diagnosis of heart disease patients.
- Testing the realistic usability and acceptability of the HD – ESRET among pharmacist and other community healthcare providers.

Finally, applying data mining techniques in identifying suitable treatments for heart disease patients is an area that has not been tested before and warrants further investigation.

7.6. And, Finally…

The main goal of this thesis is using data mining analysis to build a heart disease expert system risk evaluation tool (HD - ESRET) using non-invasive attributes to help healthcare professionals in the risk evaluation of heart disease patients using low cost data attributes. This goal involves contributions to both practice and research including:

- Identifying Decision Tree as one of the most stable data mining techniques across different attributes combination of heart disease datasets.
- Identifying that age, sex, resting blood pressure and peak heart rate are significant attributes among different heart disease attributes f needed by data mining techniques in the risk evaluation of heart disease.
- Identifying non-invasive attributes to be significant in the risk evaluation of heart disease.

- Identifying age, sex, resting blood pressure and Rohrer's Index to be significant non-invasive attributes in the risk evaluation of heart disease.
- Identifying that integrating K-Means clustering with different initial centroids selection methods can enhance Decision Tree performance in the risk evaluation of heart disease patients.
- Developing the HD – ESRET using non-invasive heart disease data attributes to help healthcare professionals in the risk evaluation of heart disease.

[This Page is Left Blank Intentionally]

References:

Abdullah, A. S., & Rajalaxmi, R. R. (2012). A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (ICON3C 2012), ICON3C(3), 22-25.

Alpaydin, E. (1997). Voting over multiple condensed nearest neighbors. Artificial Intelligence Review, 11(1-5), 115-132.

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

American Heart Association (2011). American Heart Association Live and Learn. Accessed 15 February 2011, from http://www.americanheart.org/presenter.jhtml?identifier=4478

American Heart Association (2013). What is Cardiovascular Disease (Heart Disease)? Accessed 25 October 2013, from http://www.heart.org/HEARTORG/Caregiver/Resources/WhatisCardiovascularDisease/What-is-Cardiovascular-Disease_UCM_301852_Article.jsp

Andreeva, P. (2006). Data modelling and specific rule generation via data mining techniques. Paper presented at the International Conference on Computer Systems and Technologies.

Ashby, D., & Smith, A. (2005). The Best Medicine? Plus Magazine - Living Mathematics.

Australian Bureau of Statistics. (2013). Accessed 12 March 2014, from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3303.0Media%20Release12011?opendocument&tabname=Summary&prodno=3303.0&issue=2011&num=&view=

Avci, E. (2012). A New Expert System for Diagnosis of Lung Cancer: GDA—LS_SVM. Journal of Medical Systems, 36(3), 2005-2009.

Bitton, A., & Gaziano, T. (2010). The Framingham Heart Study's impact on global risk assessment. Progress in cardiovascular diseases, 53(1), 68-78.

Bramer, M. (2007). Principles of data mining: Springer.

Brindle, P. M., McConnachie, A., Upton, M. N., Hart, C. L., Smith, G. D., & Watt, G. C. (2005). The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. British Journal of General Practice, 55(520), 838-845.

Bronzino, J. D., & Austin-LaFrance, R. J. (1992). Management of medical technology: a primer for clinical engineers: Butterworth-Heinemann.

Canlas Jr, R. D. (2009). Data Mining in Healthcare: Current Applications and Issues. [MS thesis in Information Technology].

Center for Disease Control and Prevention. (2014). Heart Disease and Family History. Accessed 13 January 2014, from http://www.cdc.gov/genomics/resources/diseases/heart.htm

Centers for Disease Control and Prevention. (2013). Chronic Disease Prevention and Health Promotion. Accessed 27 September 2013, from http://www.cdc.gov/nccdphp/

Cheung, N. (2001). Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering. B. Sc. Thesis, University of Queenland.

Cieslak, D., Hoens, T. R., Chawla, N., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. Data Mining and Knowledge Discovery, 24(1), 136-158.

Cupples, L., & D'Agostino, R. (1987). Some risk factors related to the annual incidence of cardiovascular disease and death in pooled repeated biennial measurements. US Department of Health and Human Services.

D'Agostino, R., Vasan, R. S., Pencina, M., Wolf, P., Cobain, M., Massaro, J., & Kanne, l. W. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. Journal of the American Heart Association, 117(6), 743-753.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 36 (2009), 7675–7680.

De Beule, M., Maes, E., De Winter, O., Vanlaere, W., & Van Impe, R. (2007). Artificial neural networks and risk stratification: A promising combination. Mathematical and computer modelling, 46(1), 88-94.

Deekshatulu, B., & Chandra, P. (2013). Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. Global Journal of Computer Science and Technology, 13(3).

Department of Health & Aging, A. G. (2012). Seniors and Aged Care Australia websites have been replaced. Accessed 16 August 2012, from http://www.agedcareaustralia.gov.au/internet/agedcare/publishing.nsf/Content/Prevention+and+awareness-2

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann, 194–202.

Ensminger, M. E., & Ensminger, A. H. (1993). Foods & nutrition encyclopedia (Vol. 1): CRC press.

ESCAP. (2010). Accessed 9 December 2010, from http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 19(5), 476-491.

European Public Health Alliance. (2013). Cardiovascular health takes centre stage in Brussels. Accessed 12 March 2014, from http://www.epha.org/a/5899

Fayyad, U. (1997). Data mining and knowledge discovery in databases: implications for scientific databases, Ninth International Conference on Scientific and Statistical Database Management.

Fayyad, U. M., & Keki, B. I. (1992). On the handing of Continuous-Valued Attributes in Decision Tree Generation. Machine Learning, 8, 87-102.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. Paper presented at the KDD.

Feigenbaum, E., McCorduck, P., & Nii, H. S. (1988). The Rise of the Expert Company: Vintage Books.

Framingham Heart Study. (2013). About the Framingham Heart Study. Accessed 5 October 2013, from http://www.framinghamheartstudy.org/about/history.html

Giarratano, J. C., & Riley, G. D. (2004). Expert Systems: Principles and Programming, Fourth Edition

Giri, D., Rajendra Acharya, U., Martis, R. J., Vinitha Sree, S., Lim, T.-C., Ahamed VI, T., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-Based Systems, 37, 274-282.

Gupta, R. (2010). "Data Mining for Discovery of Clinical and Genomic Disease Markers."

Hall, M. A. (2000). "Feature Selection for Discrete and Numeric Class Machine Learning." Seventeenth International conference on Machine Learning.

Han, J., & Kamber, M. (2006). Data Mining Concepts and Techniques: Morgan Kaufmann Publishers.

Hara, A., & Ichimura, T. (2008). Data Mining by Soft Computing Methods for The Coronary Heart Disease Database. Paper presented at the Fourth International Workshop on Computational Intelligence & Applications, IEEE.

Heller, R. F., S. Chinn, H. D. Tunstall Pedoe and G. Rose (1984). "How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project." British Medical Journal, May 12;288(6428):1409–1411.

Helma, C., Gottmann, E., & Kramer, S. (2000). Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research, 9(4), 329-358.

Herron, P. (2004). Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms, INLS 110, Data Mining,Spring

Hubert, H. B., Feinleib, M., McNamara, P. M., & Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation, 67(5), 968-977.

Jolliffe, I. (2005). Principal component analysis: Wiley Online Library.

Kangwanariyakul, Y., Naenna, T., Nantasenamat, C., & Tantimongcolwat, T. (2010). Data mining of magnetocardiograms for prediction of ischemic heart disease.

Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2), 3465-3469.

Khan, D. M., & Mohamudally, N. (2010). A Multiagent System (MAS) for the Generation of Initial Centroids for kmeans Clustering Data Mining Algorithm based on Actual Sample Datapoints. Paper presented at the 2nd International Conference In Software Engineering and Data Mining (SEDM), IEEE.

Kotnik, T. (2010). Prevention Programs. Challenges in Family Medicine.

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, 32(1), 47-58.

Krose, B., & Van der Smagt, P. (1996). An introduction to neural network: The University of Amesterdam.

Kumari, M., & Godara, S. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction

Lakshmi, K., Krishna, M. V., & Kumar, S. P. (2013). Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.

Lee, S.-C. L., M. Embrechts, I-N. (2000). Data mining techniques applied to medical information. Informatics for Health and Social Care, 25(2), 81-102.

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., . . . Clark, R. A. (2004). Data mining techniques for cancer detection using serum proteomic profiling. Artificial intelligence in medicine, 32(2), 71-83.

Liao, S.-C., & Lee, I.-N. (2002). Appropriate medical data categorization for data mining classification techniques. Informatics for Health and Social Care, 27(1), 59-67.

Marcovitz, P. A., & Armstrong, W. F. (1992). Accuracy of dobutamine stress echocardiography in detecting coronary artery disease. The American Journal of Cardiology, 69(16), 1269-1273.

MARY, C., & Raja, S. K. (2009). Refinement of clusters from k-means with ant colony optimization. Journal of Theoretical & Applied Information Technology, 6(4).

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030.

Matkovsky, I., & Nauta, K. (1998). Overview of data mining techniques. Presented at the Federal Database Colloquium and Exposition, San Diego, CA.

Mei, Z., Grummer-Strawn, L. M., Pietrobelli, A., Goulding, A., Goran, M. I., & Dietz, W. H. (2002). Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. The American journal of clinical nutrition, 75(6), 978-985.

Moore, T., Jesse, C., & Kittler, R. (2001). An Overview and Evaluation of Decision Tree Methodology. ASA Quality and Productivity Conference.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., . . . Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. Journal of the American Geriatrics Society, 53, 695–699.

National Center for Chronic Disease Prevention and Health Promotion. (2013). Know the facts about heart disease. Accessed 9 October 2013, from http://www.cdc.gov/heartdisease/docs/consumered_heartdisease.pdf

National Heart Foundation of Australia. (2009). Guidelines for the assessment of Absolute cardiovascular disease risk.

Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology, 25(8), 690-695.

Özşen, S., & Güneş, S. (2009). Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. Expert Systems with Applications, 36(1), 386-392.

Paladugu, S. (2010). Temporal mining framework for risk reduction and early detection of chronic diseases. University of Missouri--Columbia.

Panzarasa, S., Quaglini, S., Sacchi, L., Cavallini, A., Micieli, G., & Stefanelli, M. (2010). Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.

Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3), 157-160.

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., & Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. Artificial intelligence in medicine, 46(1), 5-17.

Patil, S. B., & Kumaraswamy, Y. (2009). Extraction of significant patterns from heart disease warehouses for heart attack prediction. IJCSNS, 9(2), 228-235.

Pavan, K. K., Rao, A. A., Rao, A., & Sridhar, G. (2011). Robust seed selection algorithm for k-means type algorithms. International Journal of Computer Science & Information Technology, 3(5), 147-163.

Piction, P. (2000). Neural Networks.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. Journal of Medical Systems, 26(5), 445-463.

Polat, K., Şahan, S., & Güneş, S. (2007). Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Systems with Applications, 32(2), 625-631.

Poomagal, S., & Hamsapriya, T. (2011). Optimized k-means clustering with intelligent initial centroid selection for web search using URL and tag contents. Paper presented at the Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway.

Porter, T., & Green, B. (2009). Identifying Diabetic Patients: A Data Mining Approach.

Providence Heart and Vascular Institute. (2014). Congenital heart disease is the No. 1 cause of sudden death in young people., Accessed 16 February 2014, from http://oregon.providence.org/patients/programs/providence-heart-and-vascular-institute/Pages/playsmart/health-professionals.html

Purnami, S., Zain, J., & Embong, A. (2010). A New Expert System for Diabetes Disease Diagnosis Using Modified Spline Smooth Support Vector Machine. In D. Taniar, O. Gervasi, B. Murgante, E. Pardede & B. Apduhan (Eds.), Computational Science and Its Applications – ICCSA 2010 (Vol. 6019, pp. 83-92): Springer Berlin Heidelberg.

Quinlan, J. R. (1986). Decision trees and multi-valued attributes. Hayes & D. Michie (Eds.), Machine intelligence. Oxford University Press.

Rajeswari, K., Vaithiyanathan, V., & Pede, S. V. (2013). Feature Selection for Classification in Medical Data Mining. International Journal of Emerging Treands and Technology in Computer Science, 492-497.

Rajkumar, M., & Reena, G. S. (2010). Diagonsis of Heaer Disease using Datamining Algorithm. Global Journal of Computer Science and Technology, 10(10), 38-43.

Richardson, M. (2009). "Principal Component Analysis."

Ruan, D., D'hondt, P., F. Fantoni, P., De Cock, M., Nachtegael, M., & E.Kerre, E. (2006). Applied Artificial Intelligence: World Scientific Publishing Co. Pte. Ltd.

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng., 8, 537-565.

Din, S., Rabbi, F., Qadir, F., & Khattak, M. (2007). Statistical analysis of risk factors for cardiovascular disease in Malakand division. Pakistan Journal of Statistics and Operation Research, 3(2), 117-124.

Sandhya, J., & Shenoy, P. D. (2010). Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering and Technology, 2(4).

Santhi, M., Sai Leela, V., Anitha, P., & Nagamalleswari, D. (2011). Enhancing K-Means Clustering Algorithm. International Journal of Computer Science & Technology, IJCST, 2(4), 73-77.

Scales, R., & Embrechts, M. (2002). Computational intelligence techniques for medical diagnostics. Paper presented at the Graduate Research Conference Proceedings of Walter Lincoln Hawkins.

Shahwan-Akl, L. (2010). Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 1(1), 1-7.

Shillabeer, A., & Roddick, J. F. (2006). Towards role based hypothesis evaluation for health data mining. electronic Journal of Health Informatics, 1(1), 1-9.

Shlens, J. (2005). A Tutorial on Principal Component Analysis. Accessed 10 February 2014, from http://www.brainmapping.org/NITP/PNA/Readings/pca.pdf

Simons, L. A., Simons, J., Friedlander, Y., McCallum, J., & Palaniappan, L. (2003). Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia, 178(3), 113-116.

Sitar-Taut, V., Zdrenghea, D., Pop, D., & Sitar-Taut, D. (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 5(3), 29-32.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. International Journal on Computer Science & Engineering, 3(6), 2385-2392.

Srinivas, K., Rani, B. K., & Govrdha, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science & Engineering, 2(2), 250-255.

Statistics South Africa. (2008). Mortality and causes of death in South Africa, 2006: Findings from death notification. 7 February 2011, from http://www.statssa.gov.za/publications/P03093/P030932006.pdf

Sultan, K. M., AlObaidy, M. W., & Hussein, A. I. (2009). The Prevalence of Weight Loss Assessed by Body Mass Index in Patients with Stable Chronic Obstructive Pulmonary Disease. The Iraqi postgraduate medical journal.

Tajunisha, N., & Saravanan, V. (2011). A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets. International Journal of Advanced Science & Technology, 27, 85-94.

Tang, J. T. C. (2008). The Role of Pharmacists in Asia and Africa: A Comparative Study to the UK and Sweden.

Tantimongcolwat, T., Naenna, T., Isarankura-Na-Ayudhya, C., Embrechts, M. J., & Prachayasittikul, V. (2008). Identification of ischemic heart disease via machine learning analysis on magnetocardiograms. Computers in biology and medicine, 38(7), 817-825.

The Expert Panel. (1994). National Cholesterol Education Program Second Report.The expert panel on detection, evaluation, and treatment of high blood cholesterol inadults (Adult Treatment Panel II). Circulation, 89:1333–1445.

Thuraisingham, B. (2000). A primer for understanding and applying data mining. IT Professional, 2(1), 28-31.

Tu, M. C., Shin, D., & Shin, D. (2009). Effective Diagnosis of Heart Disease through Bagging Approach. Paper presented at the 2nd International Conference on Biomedical Engineering and Informatics.

Tunstall-Pedoe, H., Kuulasmaa, K., Mahonen, M., Tolonen, H., Ruokokoski, E., & Amouyel, P. (1999). Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. Lancet, 353 :1547 – 1557.

U.S department of health and human services. (2005). High Blood Cholesterol What you need to know. Accessed 15 January 2014, from http://www.nhlbi.nih.gov/health/public/heart/chol/wyntk.pdf

Valdez, R., Greenlund, K., Wattigney, W., Bao, W., & Berenson, G. (1996). Use of weight-for-height indices in children to predict adult overweight: the Bogalusa Heart Study. International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity, 20(8), 715-721.

Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. Data Science Journal, 5(19), 119-124.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837-1847.

Wooldridge, M. (2009). An Introduction to Multiagent Systems: John Wiley & Sons Ltd.

World Bank Disease Control Priorities Project. (2013). Health Priority Setting in the Southern Cone:Action Needed on Lifestyle Risk Factors. Accessed 7 November 2013, from http://www.dcp2.org/file/80/

World Health Organization. (2005). Clinical guidelines for the management of hypertension: WHO regional office for the eastern mediterranean.

Organization, W. H. (2010). Global status report on noncommunicable diseases 2010. Accessed 24 March 2013, from http://www.who.int/nmh/publications/ncd_report_full_en.pdf.

World Health Organization. (2011a). The top ten causes of death. Accessed 24 March 2013, from http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

World Health Organization. (2011b). Burden: mortality, morbidity and risk factors. Accessed 15 December 2012, from http://www.who.int/nmh/publications/ncd_report_chapter1.pdf

World Health Organization. (2013a). The impact of chronic disease in Spain. Accessed 13 January 2014, from http://www.who.int/chp/chronic_disease_report/spain.pdf

World Health Organization. (2013b). The impact of chronic disease in Egypt. Accessed 23 January 2013, from http://www.who.int/chp/chronic_disease_report/media/impact/egypt.pdf

World Health Organization. (2013c). Deaths from coronary heart disease. Accessed 2 September 2013, from http://www.who.int/cardiovascular_diseases/en/cvd_atlas_14_deathHD.pdf

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 algorithms in data mining analysis. Knowledge and Information Systems, 14(1), 1-37.

Yan, H., Zheng, J., Jiang, Y., Peng, C., & Li, Q. (2003). Development of a decision support system for heart disease diagnosis using multilayer perceptron. Paper presented at the Proceedings of the 2003 International Symposium on Circuits and Systems.

Zahan, S., Bogdan, R., & Capalneanu, R. (2000). Fuzzy expert system for cardiovascular disease diagnosis-tests and performance evaluation. Paper presented at the Proceedings of the 5th Seminar on Neural Network Applications in Electrical Engineering.

Zaiane, O. R. (1999). Principles of knowledge discovery in databases. Dept. of Computing Science, University of Alberta.

Appendix A

1. Cleveland Heart Disease Data Set Data

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
60	0	3	120	178	1	0	96	0	0	1	0	3	0
54	1	4	110	206	0	2	108	1	0	2	1	3	1
53	1	4	142	226	0	2	111	1	0	1	0	7	0
66	1	2	160	246	0	0	120	1	0	2	3	6	1
59	1	1	160	273	0	2	125	0	0	1	0	3	1
71	0	3	110	265	1	2	130	0	0	1	1	3	0
41	1	2	135	203	0	0	132	0	0	2	0	6	0
45	0	2	112	160	0	0	138	0	0	2	0	3	0
62	1	2	128	208	1	2	140	0	0	1	0	3	0
70	1	2	156	245	0	2	143	0	0	1	0	3	0
59	0	4	174	249	0	0	143	1	0	2	0	3	1
45	1	4	142	309	0	2	147	1	0	2	3	7	1
42	1	3	130	180	0	0	150	0	0	1	0	3	0
48	1	4	130	256	1	2	150	1	0	1	2	7	1
47	1	3	108	243	0	0	152	0	0	1	0	3	1
46	0	4	138	243	0	2	152	1	0	2	0	3	0
58	0	2	136	319	1	2	152	0	0	1	2	3	1
41	1	2	110	235	0	0	153	0	0	1	0	3	0
51	1	3	94	227	0	0	154	1	0	1	1	7	0
64	0	4	180	325	0	0	154	1	0	1	0	3	0
55	1	2	130	262	0	0	155	0	0	1	0	3	0
46	1	2	101	197	1	0	156	0	0	1	0	7	0
47	1	3	138	257	0	2	156	0	0	1	0	3	0
35	1	4	126	282	0	2	156	1	0	1	0	7	1
54	1	2	108	309	0	0	156	0	0	1	0	7	0
50	0	4	110	254	0	2	159	0	0	1	0	3	0
54	0	2	132	288	1	2	159	1	0	1	1	3	0
57	0	4	128	303	0	2	159	0	0	1	1	3	0
52	1	4	112	230	0	0	160	0	0	1	1	3	1
53	0	4	138	234	0	2	160	0	0	1	0	3	0
60	0	3	102	318	0	0	160	0	0	1	1	3	0
43	1	4	110	211	0	0	161	0	0	1	0	7	0
52	1	4	128	255	0	0	161	1	0	1	1	7	1
60	0	4	158	305	0	2	161	0	0	1	0	3	1
59	1	4	140	177	0	0	162	1	0	1	1	7	1
49	0	2	134	271	0	0	162	0	0	2	0	3	0

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
42	1	2	120	295	0	0	162	0	0	1	0	3	0
77	1	4	125	304	0	2	162	1	0	1	3	3	1
50	1	3	129	196	0	0	163	0	0	1	0	3	0
54	0	3	160	201	0	0	163	0	0	1	1	3	0
62	0	4	124	209	0	0	163	0	0	1	0	3	0
49	0	4	130	269	0	0	163	0	0	1	0	3	0
41	0	2	126	306	0	0	163	0	0	1	0	3	0
59	1	2	140	221	0	0	164	1	0	1	0	3	0
58	1	3	140	211	1	2	165	0	0	1	0	3	0
54	0	3	108	267	0	2	167	0	0	1	0	3	0
57	1	4	132	207	0	0	168	1	0	1	0	7	0
44	1	3	120	226	0	0	169	0	0	1	0	3	0
37	0	3	120	215	0	0	170	0	0	1	0	3	0
44	1	2	120	220	0	0	170	0	0	1	0	3	0
54	0	3	135	304	1	0	170	0	0	1	0	3	0
45	1	2	128	308	0	2	170	0	0	1	0	3	0
58	1	4	125	300	0	2	171	0	0	1	2	7	1
46	0	2	105	204	0	0	172	0	0	1	0	3	0
63	0	3	135	252	0	2	172	0	0	1	0	3	0
41	0	3	112	268	0	2	172	1	0	1	0	3	0
67	0	3	152	277	0	0	172	0	0	1	1	3	0
42	0	3	120	209	0	0	173	0	0	2	0	3	0
53	1	3	130	246	1	2	173	0	0	1	3	3	0
34	1	1	118	182	0	2	174	0	0	1	0	3	0
48	1	3	124	255	1	0	175	0	0	1	2	3	0
44	1	4	110	197	0	2	177	0	0	1	1	3	1
63	0	2	140	195	0	0	179	0	0	1	2	3	0
39	0	3	94	199	0	0	179	0	0	1	0	3	0
41	1	3	112	250	0	0	179	0	0	1	0	3	0
44	1	3	140	235	0	2	180	0	0	1	0	3	0
40	1	4	152	223	0	0	181	0	0	1	0	7	1
59	1	4	138	271	0	2	182	0	0	1	0	3	0
39	1	3	140	321	0	2	182	0	0	1	0	3	0
52	1	2	128	205	1	0	184	0	0	1	0	3	0
45	1	4	115	260	0	2	185	0	0	1	0	3	0
48	1	4	122	222	0	2	186	0	0	1	0	3	0
51	1	4	140	261	0	2	186	1	0	1	0	3	0
52	1	1	118	186	0	2	190	0	0	2	0	6	0
54	1	2	192	283	0	2	195	0	0	1	1	7	1
29	1	2	130	204	0	2	202	0	0	1	0	3	0

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
69	1	1	160	234	1	2	131	0	0.1	2	1	3	0
66	1	4	112	212	0	2	132	1	0.1	1	1	3	1
47	1	4	112	204	0	0	143	0	0.1	1	0	3	0
52	1	4	108	233	1	0	147	0	0.1	1	3	7	0
58	1	4	100	234	0	0	156	0	0.1	1	1	7	1
52	0	3	136	196	0	2	169	0	0.1	2	0	3	0
64	1	4	128	263	0	0	105	1	0.2	2	1	7	0
74	0	2	120	269	0	2	121	1	0.2	1	1	3	0
64	0	3	140	313	0	0	133	0	0.2	1	0	7	0
45	0	4	138	236	0	2	152	1	0.2	2	0	3	0
59	1	1	170	288	0	2	159	0	0.2	2	0	7	1
67	1	4	125	254	1	0	163	0	0.2	2	2	7	1
43	0	3	122	213	0	0	165	0	0.2	2	0	3	0
52	1	2	120	325	0	0	172	0	0.2	1	0	3	0
57	1	3	150	126	1	0	173	0	0.2	1	1	7	0
48	1	2	130	245	0	2	180	0	0.2	2	0	3	0
57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
67	0	4	106	223	0	0	142	0	0.3	1	2	3	0
44	0	3	118	242	0	0	149	0	0.3	2	1	3	0
53	0	4	130	264	0	2	143	0	0.4	2	0	3	0
54	1	3	120	258	0	2	147	0	0.4	2	0	7	0
57	1	3	128	229	0	2	150	0	0.4	2	1	7	1
66	1	4	120	302	0	2	151	0	0.4	2	0	3	0
51	0	3	130	256	0	2	149	0	0.5	1	0	3	0
48	1	4	124	274	0	2	166	0	0.5	2	0	7	1
42	0	4	102	265	0	2	122	0	0.6	2	0	3	0
58	1	3	105	240	0	2	154	1	0.6	2	0	7	0
51	0	3	120	295	0	2	157	0	0.6	1	0	3	0
65	1	4	110	248	0	2	158	0	0.6	1	2	6	1
44	0	3	108	141	0	0	175	0	0.6	2	0	3	0
45	0	2	130	234	0	2	175	0	0.6	2	0	3	0
34	0	2	118	210	0	0	192	0	0.7	1	0	3	0
58	1	4	150	270	0	2	111	1	0.8	1	0	7	1
49	1	3	118	149	0	2	126	0	0.8	1	3	3	1
46	1	4	120	249	0	2	144	0	0.8	1	0	7	1
55	1	4	160	289	0	2	145	1	0.8	2	1	7	1
65	0	3	155	269	0	0	148	0	0.8	1	0	3	0
65	0	3	160	360	0	2	151	0	0.8	1	0	3	0
52	1	2	134	201	0	0	158	0	0.8	1	1	3	0

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
42	1	1	148	244	0	2	178	0	0.8	1	2	3	0
42	1	3	120	240	1	0	194	0	0.8	3	0	7	0
67	1	4	100	299	0	2	125	1	0.9	2	2	3	1
50	1	4	144	200	0	2	126	1	0.9	2	0	7	1
60	0	1	150	240	0	0	171	0	0.9	1	0	3	0
67	1	4	120	237	0	0	71	0	1	2	0	3	1
47	1	4	110	275	0	2	118	1	1	2	1	3	1
58	0	4	100	248	0	2	122	0	1	2	0	3	0
57	1	4	165	289	1	2	124	0	1	2	3	7	1
61	0	4	145	307	0	2	146	1	1	2	0	7	1
68	1	3	118	277	0	0	151	0	1	1	1	7	0
66	0	4	178	228	1	0	165	1	1	2	2	7	1
52	1	4	125	212	0	0	168	0	1	1	2	7	1
76	0	3	140	197	0	1	116	0	1.1	2	0	3	0
50	0	2	120	244	0	0	162	0	1.1	1	0	3	0
57	1	4	152	274	0	0	88	1	1.2	2	1	7	1
62	0	3	130	263	0	0	97	0	1.2	2	1	7	1
39	1	4	118	219	0	0	140	0	1.2	2	0	7	1
59	1	4	110	239	0	2	142	1	1.2	2	1	7	1
51	0	4	130	305	0	0	142	1	1.2	2	0	7	1
51	1	3	100	222	0	0	143	1	1.2	2	0	3	0
56	1	4	125	249	1	2	144	1	1.2	2	1	3	1
62	0	4	140	394	0	2	157	0	1.2	2	0	3	0
60	1	4	140	293	0	2	170	0	1.2	2	2	7	1
52	1	1	152	298	1	0	178	0	1.2	2	0	7	0
43	1	4	115	303	0	0	181	0	1.2	2	0	3	0
62	1	2	120	281	0	2	103	0	1.4	2	1	7	1
54	1	4	120	188	0	0	113	0	1.4	2	1	7	1
62	0	4	150	244	0	0	154	1	1.4	2	0	3	1
46	0	3	142	177	0	2	160	1	1.4	3	0	3	0
55	0	2	135	250	0	2	161	0	1.4	2	0	3	0
65	1	1	138	282	1	2	174	0	1.4	2	1	3	1
35	0	4	138	183	0	0	182	0	1.4	1	0	3	0
68	0	3	120	211	0	2	115	0	1.5	2	0	3	0
57	1	4	110	201	0	0	126	1	1.5	2	0	6	0
51	0	3	140	308	0	2	142	0	1.5	1	1	3	0
56	1	4	130	283	1	2	103	1	1.6	3	0	7	1
71	0	4	112	149	0	0	125	0	1.6	2	0	3	0
35	1	4	120	198	0	0	130	1	1.6	2	0	7	1

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
68	1	3	180	274	1	2	150	1	1.6	2	0	7	1
54	0	3	110	214	0	0	158	0	1.6	2	0	3	0
67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
54	1	3	150	232	0	2	165	0	1.6	1	0	7	0
51	1	4	140	299	0	0	173	1	1.6	1	0	7	1
62	1	4	120	267	0	0	99	1	1.8	2	2	7	1
46	1	4	140	311	0	0	120	1	1.8	2	2	7	1
42	1	4	136	315	0	0	125	1	1.8	2	0	6	1
64	1	3	125	309	0	0	131	1	1.8	2	0	7	1
63	1	4	130	330	1	2	132	1	1.8	1	3	7	1
62	1	3	130	231	0	0	146	0	1.8	2	3	7	0
63	0	4	108	269	0	0	169	1	1.8	2	2	3	1
62	0	4	138	294	1	0	106	0	1.9	2	3	3	1
61	1	4	140	207	0	2	138	1	1.9	1	1	7	1
56	0	4	134	409	0	2	150	1	1.9	2	2	7	1
56	1	1	120	193	0	2	162	0	1.9	2	0	7	0
43	1	3	130	315	0	0	162	0	1.9	1	1	3	0
53	1	4	123	282	0	0	95	1	2	2	2	7	1
58	1	4	146	218	0	0	105	0	2	2	1	7	1
64	0	4	130	303	0	0	122	0	2	2	2	3	0
64	1	4	145	212	0	2	132	0	2	2	2	6	1
49	1	3	120	188	0	0	139	0	2	2	3	7	1
69	1	3	140	254	0	2	146	0	2	2	3	7	1
41	1	3	130	214	0	2	168	0	2	2	0	3	0
56	1	4	132	184	0	2	105	1	2.1	2	1	6	1
64	1	4	120	246	0	2	96	1	2.2	3	1	3	1
58	1	4	128	216	0	2	131	1	2.2	2	3	7	1
59	1	3	126	218	1	0	134	0	2.2	2	1	6	1
66	1	4	160	228	0	2	138	0	2.3	1	0	6	0
70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
51	1	3	125	245	1	2	166	0	2.4	2	0	3	0
70	1	4	145	174	0	0	125	1	2.6	3	0	7	1
61	1	1	134	234	0	0	145	0	2.6	2	2	3	1
60	0	4	150	258	0	2	157	0	2.6	2	2	7	1
54	1	4	110	239	0	0	126	1	2.8	2	1	7	1
65	1	4	135	254	0	2	127	0	2.8	2	1	7	1
60	1	4	125	258	0	2	141	1	2.8	2	1	7	1
60	1	4	145	282	0	2	142	1	2.8	2	2	7	1
70	1	3	160	269	0	0	112	1	2.9	2	1	7	1

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
58	1	4	128	259	0	2	130	1	3	2	2	7	1
43	0	4	132	341	1	2	136	1	3	2	0	7	1
45	1	4	104	208	0	2	148	1	3	2	0	3	0
60	1	3	140	185	0	2	155	0	3	2	0	3	1
54	1	4	122	286	0	2	116	1	3.2	2	2	3	1
55	0	4	180	327	0	1	117	1	3.4	2	0	3	1
59	1	4	170	326	0	2	140	1	3.4	3	0	7	1
61	1	4	138	166	0	2	125	1	3.6	2	1	3	1
61	1	4	120	260	0	0	140	1	3.6	2	1	7	1
46	1	3	150	231	0	0	147	0	3.6	2	0	3	1
38	1	1	120	231	0	0	182	1	3.8	2	0	7	1
56	0	4	200	288	1	2	133	1	4	3	2	7	1
63	0	4	150	407	0	2	154	0	4	2	3	7	1
51	1	4	140	298	0	0	122	1	4.2	2	3	7	1
59	1	1	178	270	0	2	145	0	4.2	3	0	7	0
55	1	4	140	217	0	0	111	1	5.6	3	0	7	1
62	0	4	160	164	0	2	145	0	6.2	3	3	7	1
59	1	1	134	204	0	0	162	0	0.8	1	2	3	1
64	1	1	170	227	0	2	155	0	0.6	2	0	7	0
66	0	3	146	278	0	2	152	0	0	2	1	3	0
39	0	3	138	220	0	0	152	0	0	2	0	3	0
57	1	2	154	232	0	2	164	0	0	1	1	3	1
58	0	4	130	197	0	0	131	0	0.6	2	0	3	0
57	1	4	110	335	0	0	143	1	3	2	1	7	1
47	1	3	130	253	0	0	179	0	0	1	0	3	0
55	0	4	128	205	0	1	130	1	2	2	1	7	1
35	1	2	122	192	0	0	174	0	0	1	0	3	0
61	1	4	148	203	0	0	161	0	0	1	1	7	1
58	1	4	114	318	0	1	140	0	4.4	3	3	6	1
58	0	4	170	225	1	2	146	1	2.8	2	2	6	1
56	1	2	130	221	0	2	163	0	0	1	0	7	0
56	1	2	120	240	0	0	169	0	0	3	0	3	0
67	1	3	152	212	0	2	150	0	0.8	2	0	7	1
55	0	2	132	342	0	0	166	0	1.2	1	0	3	0
44	1	4	120	169	0	0	144	1	2.8	3	0	6	1
63	1	4	140	187	0	2	144	1	4	1	2	7	1
63	0	4	124	197	0	0	136	1	0	2	0	3	1
41	1	2	120	157	0	0	182	0	0	1	0	3	0
59	1	4	164	176	1	2	90	0	1	2	2	6	1

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
57	0	4	140	241	0	0	123	1	0.2	2	0	7	1
45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
68	1	4	144	193	1	0	141	0	3.4	2	2	7	1
57	1	4	130	131	0	0	115	1	1.2	2	1	7	1
57	0	2	130	236	0	2	174	0	0	2	1	3	1
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	1
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	1
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
58	0	1	150	283	1	2	162	0	1	1	0	3	0
58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
58	1	3	132	224	0	2	173	0	3.2	1	2	7	1
60	1	4	130	206	0	2	132	1	2.4	2	2	7	1
50	0	3	120	219	0	0	158	0	1.6	2	0	3	0
58	0	3	120	340	0	0	172	0	0	1	0	3	0
66	0	1	150	226	0	0	114	0	2.6	3	0	3	0
43	1	4	150	247	0	0	171	0	1.5	1	0	3	0
40	1	4	110	167	0	2	114	1	2	2	0	7	1
69	0	1	140	239	0	0	151	0	1.8	1	2	3	0
60	1	4	117	230	1	0	160	1	1.4	1	2	7	1
64	1	3	140	335	0	0	158	0	0	1	0	3	1
59	1	4	135	234	0	0	161	0	0.5	2	0	7	0

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	State
44	1	3	130	233	0	0	179	1	0.4	1	0	3	0
42	1	4	140	226	0	0	178	0	0	1	0	3	0
43	1	4	120	177	0	2	120	1	2.5	2	0	7	1
57	1	4	150	276	0	2	112	1	0.6	2	1	6	1
55	1	4	132	353	0	0	132	1	1.2	2	1	7	1
61	1	3	150	243	1	0	137	1	1	2	0	3	0
65	0	4	150	225	0	2	114	0	1	2	3	7	1
40	1	1	140	199	0	0	178	1	1.4	1	0	7	0
71	0	2	160	302	0	0	162	0	0.4	1	2	3	0
59	1	3	150	212	1	0	157	0	1.6	1	0	3	0
61	0	4	130	330	0	2	169	0	0	1	0	3	1
58	1	3	112	230	0	2	165	0	2.5	2	1	7	1
51	1	3	110	175	0	0	123	0	0.6	1	0	3	0
50	1	4	150	243	0	2	128	0	2.6	2	0	7	1
65	0	3	140	417	1	2	157	0	0.8	1	1	3	0
53	1	3	130	197	1	2	152	0	1.2	3	0	3	0
41	0	2	105	198	0	0	168	0	0	1	1	3	0
65	1	4	120	177	0	0	140	0	0.4	1	0	7	0
44	1	4	112	290	0	2	153	0	0	1	1	3	1
44	1	2	130	219	0	2	188	0	0	1	0	3	0
60	1	4	130	253	0	0	144	1	1.4	1	1	7	1
54	1	4	124	266	0	2	109	1	2.2	2	1	7	1
50	1	3	140	233	0	0	163	0	0.6	2	1	7	1
41	1	4	110	172	0	2	158	0	0	1	0	7	1
54	1	3	125	273	0	2	152	0	0.5	3	1	3	0
51	1	1	125	213	0	2	125	1	1.4	1	1	3	0
53	1	3	130	197	1	2	152	0	1.2	3	0	3	0
71	0	2	160	302	0	0	162	0	0.4	1	2	3	0
57	1	4	150	276	0	2	112	1	0.6	2	1	6	1

2. Canberra Heart Disease Data Set Data

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
91	2603	0	175	78	0	0	80	0	140	90	195	80	1
80	2370	0	184	80	0	0	84	80	150	95	200	104	0
68	2587	1	160	97.9	0	0	64	99	130	80	190	100	1
85	2607	0	172	89	0	0	46	106	162	72	181	82	1
91	2611	0	173	87	0	0	52	108	190	70	194	70	1
89	2601	0	175	95	0	0	67	109	140	80	170	80	1
87	2603	0	165	87	0	0	65	112	129	77	178	93	1
84	2720	0	180	77	0	0	70	115	100	60	140	90	0
65	2550	1	152.4	83.6	4.5	0	64	116	152	79	183	104	0
81	2615	0	164	60	0	0	57	117	130	85	160	100	1
76	2550	0	169	54	0	0	60	118	122	76	178	127	1
80	2586	0	163	72	0	0	59	118	136	75	164	84	1
87	2611	1	154	46	0	0	60	119	169	105	244	103	0
91	2620	1	152	65	3.5	0	75	120	120	80	170	93	1
79	2602	0	173	74	5.9	0	55	122	142	61	164	77	1
79	2903	0	170	87	0	0	79	122	118	82	150	90	1
79	2903	0	170	89.6	0	0	53	125	120	70	189	100	1
63	2617	0	181	105	0	0	76	127	120	80	120	80	0
82	2615	0	168	95	0	0	85	127	198	93	214	105	1
78	2900	0	182	79	0	0	83	127	109	64	147	80	1
84	2607	0	165.5	85.7	0	0	65	128	156	68	239	102	1
78	2620	0	185	77	0	0	70	129	158	85	223	100	0
87	2611	0	171	74	0	0	91	129	161	44	263	117	1
83	2550	1	154.5	66	0	0	76	133	108	74	202	44	1
83	2901	0	169	77.8	0	0	63	135	168	89	227	90	1
82	2903	0	175	85.1	0	0	54	135	135	46	227	180	1
71	2902	0	174	75	0	0	74	137	138	70	180	73	1
73	2590	0	172	75	0	0	68	138	130	70	160	75	1
92	2611	0	169	72	1.5	0	66	140	140	90	180	75	1
97	2627	1	160	0	0	0	75	140	152	73	198	80	1
84	2903	0	173	71.5	0	0	75	141	114	69	155	96	1
91	2615	1	161	83	0	0	99	142	150	90	250	100	0
85	2602	0	175	68	49	0	93	144	146	98	192	106	1
87	2604	1	160	55	0	0	84	144	179	68	194	101	1
91	2603	1	162	73	0	0	85	145	130	80	150	70	0
75	2607	0	180	92	0	0	60	146	139	86	232	100	1
87	2904	0	172	77	4.6	0	64	148	140	80	200	90	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
92	2600	0	172	98	0	0	82	149	163	95	215	117	1
86	2606	1	161	59	0	0	81	149	134	66	185	72	1
80	2537	0	158	57	0	0	77	150	140	70	220	100	1
69	2611	0	168	82	0	0	84	150	164	110	227	122	1
81	2615	0	188	94	0	0	70	150	182	90	215	75	1
87	2617	1	158	58	0	0	67	151	160	69	206	90	0
92	2601	0	187	86.8	0	0	73	152	105	63	163	89	1
68	2611	1	157	77	0	0	66	153	130	70	130	70	0
76	2615	1	150	55	0	0	93	153	160	73	199	87	0
76	2536	0	175	126	4.3	0	75	154	164	96	234	99	1
59	2585	0	172	90.9	5.7	0	71	156	140	75	230	100	1
61	2902	1	145	85	0	0	100	159	152	71	176	81	0
77	2904	1	152	59	0	0	78	160	153	70	228	82	0
84	2602	1	167	79	0	0	79	161	150	78	191	93	0
70	2611	1	149	90	3.8	0	81	164	103	70	240	70	0
69	2904	1	167	68	0	0	86	164	135	68	209	84	0
87	2606	0	171	72	0	0	70	165	142	85	170	80	0
83	2586	0	170	88	0	0	79	168	148	68	161	80	1
68	2902	0	0	0	0	0	71	170	151	90	235	102	0
52	2604	0	174	91	0	0	74	171	150	73	212	91	0
71	2611	0	180	86.4	0	0	74	172	139	94	224	107	0
69	2619	1	152	68	0	0	79	172	120	61	141	48	0
79	2611	0	182	76.5	0	0	69	172	125	80	200	85	1
78	2600	0	182	76	0	0	73	176	134	80	190	80	0
62	2905	1	174	83	0	0	87	178	148	91	183	82	0
64	2904	1	161	97	4.9	0	91	180	106	68	169	64	0
70	2602	1	165	57	0	0	121	181	112	64	149	60	0
43	2903	0	175	90	0	0	82	181	141	61	236	84	0
68	2903	0	188	109	0	0	95	181	152	79	207	87	0
69	2582	1	155	94	0	0	77	182	143	90	257	102	0
64	2584	1	163	73	0	0	107	182	120	80	140	80	1
68	2590	0	174	78	0	0	80	183	135	75	176	93	1
75	2906	1	147	57	0	0	113	184	155	90	180	95	0
63	2611	1	154	92.4	0	0	102	185	98	61	121	74	0
43	2913	1	156	74	0	0	93	213	108	85	169	73	0
49	2906	1	157	94	3.6	1.5	114	169	129	62	164	41	0
62	2904	0	152	59	3.4	1.7	94	157	171	66	193	84	0
87	2607	1	166	69	3.5	1.7	79	160	165	72	240	76	1
70	2607	1	165	56	3.8	1.7	79	181	166	75	201	120	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
62	2607	1	155.6	73	3.7	1.9	82	169	140	80	160	90	0
74	2602	1	162	66.9	4	1.9	77	172	130	69	195	87	1
65	2606	1	178	80	4.2	2	83	100	132	66	166	81	0
64	2905	1	164	60	3.9	2	71	150	118	82	160	100	0
85	2603	1	163	67	3.4	2	89	161	170	90	200	110	1
83	2587	1	158	64	4.1	2	84	163	152	91	198	92	0
75	2620	0	176	80	3.4	2	80	179	166	91	222	115	1
79	2607	1	162	55	4.2	2	71	183	145	87	292	123	0
72	2615	1	154	59	3.2	2.1	67	109	125	64	157	82	1
81	2607	0	178	78	3.1	2.1	72	114	162	81	235	99	1
65	2550	1	157	99	4.2	2.1	63	120	152	95	208	115	0
87	2536	1	157	74.4	4.3	2.1	78	168	149	89	215	102	0
68	2615	0	167	82	4	2.2	54	98	92	53	122	72	1
73	2606	1	143	67	3.9	2.2	79	138	166	94	197	111	0
68	2600	1	167	58	4.8	2.2	83	139	136	87	206	86	0
85	2623	1	156	55	3.6	2.2	67	151	144	69	188	82	1
69	2631	1	153	71	4.7	2.2	58	153	130	77	159	96	1
66	2904	0	175	64	3.9	2.2	66	168	123	59	206	48	0
79	2905	1	155	76	4.3	2.2	89	170	168	87	223	105	0
52	2615	1	173	92.5	3.6	2.2	80	185	145	45	170	75	0
83	2603	1	162	68	4.3	2.3	71	117	145	76	205	90	1
74	2611	1	147	64	3.6	2.3	66	152	152	84	214	124	0
75	2613	0	178	77	4.2	2.3	90	152	103	61	200	95	1
89	2605	1	165	75	3.8	2.3	81	154	138	78	203	78	1
73	2902	1	165	71	4.5	2.3	60	161	155	76	264	94	0
87	2604	1	160	70	4.6	2.3	78	167	157	83	220	103	1
81	2537	1	157	52	4	2.3	80	171	140	80	195	90	1
73	2904	1	165	67	4.8	2.3	97	175	154	68	215	77	0
63	2615	1	160	86	5	2.3	90	182	143	82	184	82	0
63	2904	1	161	89	3.4	2.3	65	188	123	67	153	75	0
52	2905	0	175	104.5	4.3	2.3	89	197	156	70	227	86	0
88	2601	1	162	61	4.1	2.3	57	226	159	73	194	90	0
88	2602	1	162	61	4.1	2.3	57	226	159	73	194	90	0
85	2611	0	173	67	4.4	2.4	53	73	110	70	150	70	1
94	2604	0	170	88	4.1	2.4	60	80	130	90	130	90	1
84	7170	1	155	68	4.4	2.4	58	98	121	55	185	68	0
59	2607	1	158	49	4	2.4	74	115	134	77	183	81	0
60	2905	1	164	67	4.2	2.4	96	116	146	78	173	98	0
86	2580	1	163	67.5	4.2	2.4	94	144	167	99	199	90	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
86	2614	1	163	67.5	4.2	2.4	94	144	167	99	199	90	0
85	2607	0	165	68	4.8	2.4	85	146	159	75	206	86	0
90	2621	1	158	84	4.4	2.4	75	150	165	84	230	102	1
89	2607	1	160	64	4.4	2.4	67	152	138	80	150	70	0
62	2902	0	178	89.04	5	2.4	73	179	145	95	210	100	0
81	2900	1	162	67	3.9	2.4	75	180	200	100	200	100	0
95	2606	1	168	59	4.1	2.5	62	118	150	90	185	90	0
95	2903	0	173	69	4.3	2.5	63	122	100	65	125	65	1
61	2904	1	165	79	4.5	2.5	69	134	150	100	170	100	0
78	2913	1	152	59	4.5	2.5	68	140	158	80	180	90	0
76	2620	1	174	80	4.3	2.5	87	152	130	76	164	60	0
74	2607	1	155	69	4.3	2.5	64	159	118	80	170	60	0
75	2900	1	167	66	4.2	2.5	73	161	143	67	203	79	0
93	2605	1	165	68	4.8	2.5	76	162	150	80	180	100	0
88	2905	0	164	89	3.2	2.5	0	165	160	98	220	98	0
73	2611	1	155	69	3.9	2.5	94	169	124	74	238	82	0
74	2603	1	165	70	3.9	2.5	74	170	110	75	180	90	0
74	2601	0	163	65	3.4	2.5	77	172	121	79	202	87	1
55	2906	0	177	105	4.5	2.5	61	173	128	79	200	82	0
50	3888	0	178	94	4.3	2.5	102	180	122	73	181	55	0
76	2611	0	182	83	3.8	2.5	79	183	189	86	200	82	0
57	2617	0	170	65	4.8	2.5	80	189	139	75	308	75	0
74	2607	1	163	80	4.5	2.6	57	108	111	65	181	58	0
91	2611	1	150	54	3.9	2.6	69	110	130	76	170	70	1
94	2630	0	165	77	5	2.6	63	110	140	75	160	80	1
77	2602	1	155	82	3.8	2.6	76	130	118	58	162	63	0
91	2904	1	160	76	4.7	2.6	68	130	140	80	210	80	0
78	2632	0	173	108	5.2	2.6	65	150	165	100	240	90	1
76	2606	1	0	0	4.8	2.6	94	155	125	72	159	73	0
79	2537	0	171	96	5.3	2.6	62	155	110	60	200	70	1
72	2906	1	163	78	3.8	2.6	55	166	113	75	156	89	0
69	2551	0	176	79.6	4.6	2.6	85	167	136	54	198	83	0
68	2606	1	160	75	4.3	2.6	88	173	142	74	223	88	0
52	2615	0	176	100	5.2	2.6	90	179	145	89	224	89	0
71	2632	0	177	92	5.1	2.6	91	193	142	92	160	70	0
90	2612	1	165	65	3.4	2.7	53	93	160	85	160	88	1
92	2619	0	178	62	4.3	2.7	82	107	140	80	160	90	1
71	2606	1	165	65	4.7	2.7	76	120	159	66	205	53	0
96	2600	0	160	73	4.6	2.7	67	125	138	80	170	80	1

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
64	2913	1	157	64	4.8	2.7	64	126	150	80	195	80	0
75	2550	1	156	60.2	4.6	2.7	56	128	143	86	191	93	0
81	2900	0	168	89	5.9	2.7	70	140	150	95	160	90	1
83	2602	1	142	62	5.1	2.7	81	141	149	68	201	89	0
79	2602	1	160	64	4.8	2.7	62	144	170	80	190	80	1
65	2905	1	167	77	4.6	2.7	74	149	106	61	175	79	0
86	2615	1	158	67	4.5	2.7	96	151	75	116	67	124	0
72	2905	1	165	80	4.2	2.7	83	151	146	93	213	163	0
74	2551	1	151	69	4.6	2.7	87	153	125	80	166	88	0
84	2600	1	172	75	4.1	2.7	99	153	141	71	189	89	0
56	2602	1	170	114	4.2	2.7	76	157	130	92	192	86	0
74	2620	1	157	86	4.7	2.7	61	160	144	78	229	96	0
69	2904	1	156	61	4.4	2.7	88	165	127	60	174	77	0
69	2904	1	156	61	4.4	2.7	88	165	127	60	174	77	0
80	2582	1	162	80	4.7	2.7	82	168	154	78	227	88	0
68	2903	1	157	62	4.2	2.7	78	173	158	97	200	108	0
47	2601	0	178	101.4	4.8	2.7	69	178	127	54	183	69	0
68	2606	1	157.4	86.6	4.4	2.7	75	182	125	90	190	90	0
77	2614	1	160	65	4.6	2.7	64	183	134	66	179	63	0
82	2583	0	168.07	80	4.5	2.7	74	242	140	44	208	114	1
66	2902	0	178	85	4.2	2.8	53	97	120	80	140	80	1
79	2611	0	175	98	3.5	2.8	64	114	115	90	170	70	1
83	2602	1	0	0	4.5	2.8	79	131	130	84	170	90	1
75	2904	0	181	96	4.4	2.8	55	141	150	105	210	110	0
82	2605	1	160	79	4.9	2.8	61	141	152	73	219	73	1
72	2611	1	160	87	4.8	2.8	75	142	126	57	246	107	0
81	2620	0	152	72	4.5	2.8	70	143	150	80	180	80	0
92	2600	0	173	76	4.4	2.8	80	150	150	90	210	95	1
91	2614	0	170	65	4	2.8	80	155	115	70	170	75	1
65	2620	0	172	98	4.3	2.8	70	157	165	95	200	80	0
63	2607	0	172	76	4.3	2.8	85	159	140	72	164	75	0
90	2611	1	169	78	4.3	2.8	71	159	126	70	210	110	1
62	2606	1	156	48	4.7	2.8	59	161	119	69	154	78	1
81	2904	0	180	74	4.5	2.8	69	162	126	88	202	114	0
74	2905	0	173	81	4.7	2.8	82	162	128	67	198	76	0
78	2537	1	168	61	3.7	2.8	103	165	112	86	185	98	0
95	2537	0	169	77	4.7	2.8	90	165	120	80	210	90	1
72	2905	1	163	62	4.3	2.8	76	166	111	65	167	80	0
70	2903	1	162	87	4.8	2.8	74	179	131	55	252	72	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
65	2904	0	183	160	5.6	2.8	96	180	145	88	171	88	0
57	2902	0	186	86	4.9	2.8	85	186	156	50	214	103	1
45	2601	0	178	98	4.7	2.8	96	199	136	73	191	79	1
34	2906	0	171	94	5	2.8	60	202	130	60	200	90	0
67	2902	1	155	66	5	2.9	178	65	130	70	190	80	0
74	2607	1	166	92.5	5.2	2.9	63	103	120	60	160	80	1
83	2607	0	170	73	4.4	2.9	61	108	151	42	173	57	1
59	2550	0	158	66	5.6	2.9	67	123	118	69	156	93	1
88	2606	1	160	65	4	2.9	79	123	110	70	170	80	1
86	2617	0	168	83	5.2	2.9	73	126	160	95	240	90	1
63	2902	0	179	86	5.7	2.9	57	132	132	86	256	73	1
75	2905	1	161.5	58.1	4.6	2.9	66	133	160	70	160	80	0
80	2605	1	168	55	4.3	2.9	77	138	163	91	220	42	0
60	2663	1	165	69	4.8	2.9	70	138	101	53	168	60	1
81	2903	0	175	85.6	5.3	2.9	68	140	148	60	191	57	0
59	2905	0	180	105	4.7	2.9	59	140	162	84	189	90	0
79	2611	1	163	84	4.9	2.9	70	140	150	80	203	120	1
75	2611	1	165	91	4.6	2.9	97	146	152	76	212	90	0
63	2611	1	165	126	4.3	2.9	76	151	123	67	173	78	0
68	2605	1	162	74	4.8	2.9	73	153	98	64	158	76	0
78	2615	0	175	93	4.8	2.9	64	153	138	80	212	110	1
81	2606	0	179	78.4	4.4	2.9	65	154	125	85	170	70	1
61	2616	1	167	87	4.4	2.9	98	157	160	90	180	100	1
76	2580	0	174	81	4.3	2.9	78	160	161	88	210	88	1
71	2620	1	174	91	4.4	2.9	71	161	130	90	185	95	0
55	2606	0	166	84	4.6	2.9	74	163	114	65	269	94	0
67	2601	0	170	80	5.1	2.9	66	164	136	90	184	85	0
67	2602	1	158	59	4.3	2.9	85	166	154	74	197	77	0
76	2607	1	160	90.8	4.9	2.9	85	166	138	89	210	84	0
70	2606	1	154	70	4.3	2.9	94	166	162	83	198	100	1
76	2601	1	170	107	4.7	2.9	83	168	120	68	168	44	0
76	2611	0	180	80	4.4	2.9	78	170	124	67	204	82	1
56	2615	1	155	81	4.4	2.9	74	183	111	96	180	96	0
67	2611	1	161	78	5.1	2.9	79	188	120	90	160	90	0
87	2903	1	162.5	71.3	5	3	53	94	87	162	81	171	0
98	2902	1	162	78	4.9	3	66	96	140	90	210	100	1
82	2904	1	155	63	4.3	3	77	120	158	85	200	82	0
85	2546	0	169	78.7	4.6	3	75	122	155	85	215	105	0
88	2536	1	0	0	5.1	3	83	122	120	70	130	65	1

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
79	2903	0	181	85	4.6	3	78	125	136	62	216	88	1
75	2617	1	171	96.6	4.7	3	68	135	111	79	170	91	0
92	2603	0	170	76	4.8	3	75	136	133	62	222	82	1
77	2902	1	160	84	4.9	3	70	144	147	78	215	75	0
84	2549	1	164	79	4.6	3	60	150	130	80	165	80	0
85	2600	1	162	84.8	5.5	3	56	155	144	90	180	100	0
84	2905	0	180	87.6	5.5	3	78	158	130	80	180	90	0
70	2603	1	153	73	4.8	3	68	163	160	90	230	100	0
65	2906	1	161	78	4.3	3	76	163	140	90	200	100	0
78	2603	0	175	86	4.9	3	89	175	130	80	190	80	1
85	2615	1	162	72	4.6	3	91	182	159	74	213	94	0
61	2536	0	178	73	4.4	3	79	183	134	76	247	97	1
63	2904	1	162	65	4.7	3	100	186	147	86	236	92	0
62	2537	1	160.05	80.07	5.1	3	71	197	153	86	246	194	0
57	2581	1	175	58	4.5	3	101	197	138	85	232	90	0
83	2536	1	160	62	5.1	3.1	62	92	160	90	150	90	1
74	2620	1	148	103	5.5	3.1	64	100	128	67	168	60	0
81	2904	0	173	108.7	5.5	3.1	69	102	110	70	160	90	0
86	2605	1	160	70	5.3	3.1	73	117	142	69	207	70	1
73	2602	0	170	75	4.5	3.1	64	121	140	70	150	90	1
82	2615	1	149	64	4.5	3.1	67	127	140	64	161	70	0
84	2614	1	160	59	4.5	3.1	76	128	142	74	169	76	0
75	2607	0	167	73	5.1	3.1	89	137	142	56	194	107	1
94	2904	1	155	85	4.6	3.1	62	142	170	100	232	110	0
90	2620	1	145	51	5.2	3.1	73	144	126	69	161	71	1
71	2617	0	168	94	5.4	3.1	59	147	149	82	242	118	1
78	2604	0	167	85	5.2	3.1	72	159	134	81	191	81	0
63	2620	1	160	52	4.1	3.1	87	160	137	80	172	109	0
76	2605	1	150	46.8	4.6	3.1	58	161	130	8	175	85	1
72	2902	1	165	75	5.2	3.1	82	167	160	90	200	90	0
78	2603	1	169	80	5.2	3.1	93	169	117	62	169	180	0
55	2913	0	160	67	4.7	3.1	59	172	126	88	173	93	0
74	2902	0	167	78	5	3.1	73	175	170	94	196	88	0
55	2904	1	164	88	4.8	3.1	110	183	137	42	155	101	0
62	2903	0	178	79	5.3	3.1	93	201	154	90	198	111	0
97	2605	1	155	51	4.6	3.2	70	124	140	70	190	80	1
85	2611	0	175	110	5	3.2	78	141	150	95	190	90	0
91	2606	0	178	83	5.3	3.2	79	146	157	81	163	86	1
70	2611	0	173	106	5.3	3.2	76	149	150	100	240	100	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
70	2611	0	173	106	5.3	3.2	76	149	150	100	240	100	0
72	2614	0	195	94	5.7	3.2	79	150	163	75	187	92	0
83	2611	1	163	81	4.6	3.2	92	152	134	62	198	104	1
77	2603	0	177	102	5.1	3.2	60	160	129	78	199	83	0
63	2620	1	155	61	4.7	3.2	72	161	142	80	196	57	0
63	2611	0	180	86	5.4	3.2	79	169	154	83	182	75	0
83	2604	0	173	84	5.2	3.2	63	174	150	80	156	80	1
81	2550	0	174	78.7	5.1	3.3	65	65	126	84	150	80	1
90	2611	1	151	100.6	5.2	3.3	62	109	140	80	160	90	1
70	2620	0	178	81	4.8	3.3	58	118	130	80	160	80	1
73	2606	0	177	80	4.9	3.3	88	120	120	85	160	80	1
90	2620	1	170	61	4.9	3.3	58	121	153	61	181	68	1
88	2582	1	165	62	5.5	3.3	61	122	159	63	315	81	1
89	2903	1	165	75	4.8	3.3	61	125	125	75	198	88	0
83	2615	1	151	59	0	3.3	89	125	120	85	170	70	1
74	2905	1	167	101	5.2	3.3	67	132	104	55	189	82	0
68	2913	0	182	78	5	3.3	84	144	126	80	168	79	0
92	2587	0	172	76	5.6	3.3	69	148	155	75	200	70	1
72	2905	0	168	67	5.2	3.3	93	153	126	70	163	70	0
78	2604	0	162	67	5.4	3.3	74	158	147	66	208	62	0
69	2620	1	167	84	5	3.3	116	159	135	94	203	87	0
73	2607	0	170	79	4.5	3.3	70	159	140	87	223	103	1
64	2601	0	175	78	4.3	3.3	79	160	120	58	161	81	0
77	2904	1	158	82	4.9	3.3	666	160	163	72	218	95	0
60	2905	1	160	110	5.3	3.3	80	162	133	70	184	94	0
70	2611	0	178	80	5.4	3.3	78	164	120	70	170	90	1
69	2902	1	150	96	5.1	3.3	88	166	119	53	177	48	0
66	2602	1	175	99	5.5	3.3	70	167	70	167	170	85	0
64	2606	1	157	109	4.8	3.3	113	168	120	65	153	78	0
77	2605	0	172	83.5	5.2	3.3	109	169	139	64	197	68	0
67	2486	0	177	121	4.6	3.3	88	171	157	43	266	80	0
67	2612	1	151	74	5.4	3.3	80	175	115	73	144	40	0
64	2611	0	170	89	5.2	3.3	69	180	140	110	220	100	1
71	2611	0	0	0	5	3.3	69	181	140	83	216	110	0
66	200	0	180	93.9	5.3	3.3	71	248	114	74	267	86	0
86	2550	1	155	62	4.4	3.4	48	109	120	80	175	85	1
55	2628	1	160	71	5.4	3.4	75	149	140	95	190	120	0
78	2607	0	173	83	5.5	3.4	59	152	163	92	213	100	0
72	2607	0	183	78	5.2	3.4	74	159	105	105	200	105	1

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
89	2611	1	159	53.9	4.5	3.4	130	161	140	60	170	60	1
89	2606	1	160	61	4.8	3.4	137	162	150	90	160	100	1
66	2604	0	185	86	5	3.4	80	167	104	71	282	83	0
66	2546	1	170	79	5	3.4	65	171	162	75	268	76	0
63	2617	0	170	79	5.4	3.4	85	174	144	84	217	76	1
75	2607	0	173	76	5.7	3.4	75	175	144	71	199	107	1
59	2607	1	168	88	5.1	3.4	76	200	120	80	160	80	0
68	2607	0	165	77	5.7	3.5	80	104	156	60	157	48	1
78	2902	1	152	60	4.9	3.5	80	107	156	88	209	62	0
92	2622	0	167	77	4.7	3.5	58	112	110	66	160	70	1
86	2601	0	171	57	5	3.5	79	120	150	45	169	71	1
67	2605	0	172	75	5.2	3.5	60	139	110	90	170	90	1
88	2602	0	184	95	4.2	3.5	69	142	143	92	231	127	1
73	2607	1	164	97	5.3	3.5	92	150	138	76	150	95	0
81	2600	0	182	102	5.2	3.5	66	154	145	77	236	107	1
66	2324	0	184	78	4.7	3.5	66	161	145	100	190	80	0
76	2546	0	173	88	4.9	3.5	87	180	153	79	324	297	1
62	2550	1	162	70	5.3	3.5	81	188	125	80	140	90	0
89	2601	0	158	64	5.3	3.6	70	120	135	71	179	83	1
88	2620	1	158	66.08	5.6	3.6	92	143	190	86	210	70	1
65	2611	1	469	79	4.8	3.6	80	144	152	87	217	102	0
72	2614	0	170	70	5.5	3.6	59	144	124	72	183	92	0
61	2902	1	166.5	109	5.1	3.6	83	149	154	72	176	98	0
69	2904	1	170	69	5	3.6	73	150	171	76	238	64	0
64	2606	1	162.5	81	5.3	3.6	80	157	104	66	186	85	0
71	2586	1	155	89.03	5.4	3.6	92	163	152	95	224	95	0
63	2600	0	180	100	5.4	3.6	76	169	153	97	190	110	0
58	2904	0	185	122	5.5	3.6	71	169	120	80	160	90	0
66	2611	1	167	80	5.4	3.6	82	171	142	100	257	94	0
64	2904	0	177	102	4.7	3.6	89	180	167	57	270	109	0
72	2905	0	170.1	72.4	5.7	3.6	98	180	137	66	268	70	0
90	2903	0	162	54	4.6	3.7	82	121	130	75	150	80	1
89	2602	0	168	72.5	5	3.7	55	123	127	90	215	105	1
92	2537	0	170	73	5.2	3.7	80	129	130	80	180	80	1
64	2903	0	160	73	5.1	3.7	91	139	130	70	150	80	0
94	2615	0	157	68	4.4	3.7	71	144	150	70	170	70	1
93	2618	0	173	82	4.7	3.7	57	145	137	81	216	104	1
75	2607	1	158	84	5.2	3.7	76	156	166	66	200	92	0
81	2620	1	163	75	5.2	3.7	72	159	141	76	183	90	0

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
54	2611	0	183	93	5.5	3.7	51	166	120	73	194	84	0
57	2902	0	177	80	5.4	3.7	67	168	129	77	182	95	1
82	2611	0	181	77	5.6	3.7	75	169	157	90	197	108	0
86	2602	0	173	86	5.4	3.7	66	183	140	85	202	116	1
83	2615	0	168	80	4.7	3.7	110	203	111	85	132	88	1
87	2603	1	152	68	5.3	3.7	97	214	165	89	222	95	1
87	2614	0	182	81	6.6	3.8	44	98	134	95	221	106	1
68	2615	0	187	115	5.8	3.8	74	105	125	78	170	85	1
97	2550	1	159	74.7	5.4	3.8	64	112	165	74	216	78	1
75	2620	0	175	93	5.8	3.8	65	115	112	80	160	90	1
88	2615	0	188	93.5	5.3	3.8	45	128	145	75	160	80	1
62	2902	0	185	119	5.7	3.8	69	154	131	72	302	86	0
93	2614	1	160	112	4.6	3.8	106	156	182	91	226	69	1
70	2620	1	170.5	101	5.8	3.8	105	158	150	93	196	75	1
81	2546	0	164	67.6	5.8	3.8	57	159	163	83	228	102	1
80	2587	1	161	66	5.1	3.8	71	160	152	82	202	82	1
69	2605	0	171	73	6.2	3.8	71	161	125	65	175	90	0
59	2607	1	161	70	5.3	3.8	77	164	115	66	151	76	0
44	2905	1	170	73	5.7	3.8	93	169	126	54	150	56	0
76	2810	0	172	75	5.6	3.8	65	170	122	82	160	90	1
61	2902	0	165	76	5.5	3.8	85	175	130	95	140	100	0
80	2604	1	163	63	4.7	3.8	87	188	146	85	218	102	0
58	2905	0	187	91	5.6	3.8	74	192	141	77	199	87	0
56	2620	0	204	103	5.8	3.8	79	200	120	80	160	80	0
82	2580	1	168	64	5	3.9	77	0	150	80	190	100	0
87	2614	0	178	75	5.3	3.9	69	129	140	65	187	87	1
77	2913	0	176	109	5.5	3.9	72	132	166	99	228	54	0
73	2903	1	160	106	5.2	3.9	110	138	160	104	160	90	0
66	2607	0	182	86	5.2	3.9	57	142	137	68	195	97	1
77	2605	0	186	95	6.2	3.9	90	144	132	54	195	102	1
73	2620	0	182	80	5.5	3.9	105	152	120	78	190	93	1
69	2606	1	165	83	5	3.9	53	154	114	65	189	82	0
65	2902	0	174	88.2	5.2	3.9	80	155	130	70	160	80	0
75	2546	0	172.5	96.8	5.5	3.9	96	165	166	87	233	102	1
36	2606	0	193	103	5.6	3.9	76	174	149	68	252	72	0
63	2611	0	189	92	5.5	3.9	72	176	131	78	170	72	0
85	2602	0	167	67	5.8	4	79	0	134	70	175	70	1
91	2607	0	182	70	4.8	4	58	80	159	55	175	68	1
81	2605	0	170	79	5.6	4	83	113	130	75	150	90	1

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
91	2600	0	180	97	5.8	4	55	117	165	78	223	88	1
78	2605	0	182	75	5.2	4	75	129	105	65	140	70	1
93	2548	1	168	69	5.2	4	93	137	150	80	170	80	1
50	2545	0	172	60	6	4	87	138	171	82	219	91	0
71	2603	0	0	0	6.1	4	75	155	124	74	298	159	1
68	2611	0	180	105	6.1	4	73	160	140	79	194	94	0
84	2607	0	183	91	6.2	4	82	169	136	89	186	100	1
62	2604	0	179	92	5.1	4	57	173	128	44	196	113	1
80	2536	1	172	66.09	5.5	4	111	174	140	70	210	90	1
49	2621	0	182	101	5.8	4	80	181	153	95	191	92	0
68	2580	0	166	81	5.4	4.1	45	127	130	60	170	90	0
47	2905	0	181	91	5.9	4.1	91	135	126	58	169	56	0
71	2611	0	168	104	5.1	4.1	87	139	120	90	160	80	1
50	2315	0	170	94	5.6	4.1	123	173	153	62	202	101	0
64	2615	0	190	116	6.5	4.1	71	176	140	90	200	80	1
88	2536	0	175	81	5.7	4.2	50	112	100	50	160	70	1
75	2605	0	177.8	66.5	5.8	4.2	61	135	167	77	199	111	1
83	2902	0	175	84	6.5	4.2	63	146	125	57	221	168	0
81	2607	0	181	83	5.5	4.2	77	164	126	65	199	98	1
80	2614	0	177	78	5.5	4.2	92	171	138	106	302	168	1
78	2903	0	164	75	5.5	4.2	71	176	137	95	210	115	1
72	2602	0	173	96.4	5.3	4.2	105	207	132	85	174	103	0
75	2620	1	164	70	5.6	4.3	79	107	94	62	170	59	0
75	2620	1	164	70	5.6	4.3	79	107	94	62	170	59	0
64	2605	1	176	84	6.2	4.3	59	121	117	63	154	75	0
78	2602	0	177	90	6.7	4.3	74	144	140	90	160	90	1
75	2604	0	175	92.5	6.2	4.3	92	145	139	79	185	74	1
85	2600	0	181	92	5.5	4.3	67	155	128	79	216	83	1
78	2600	0	180	78	6.4	4.3	70	160	139	63	207	93	0
58	2902	1	165	76.5	6	4.3	65	186	137	72	157	86	0
89	2794	1	173	95	6.5	4.4	76	106	150	100	190	90	0
85	2550	1	166	69	5.3	4.4	88	136	112	72	165	72	1
75	2606	0	188	83	6	4.4	77	155	119	73	174	83	1
96	2605	0	170	64	5.5	4.4	78	161	120	65	145	60	1
60	2902	0	167	72	5.7	4.4	87	172	139	90	199	91	1
99	2810	0	182	85	6.2	4.5	70	85	120	60	160	86	1
92	2536	1	160	59	5.7	4.5	79	124	140	90	190	110	1
92	2605	0	167	64	6	4.5	54	126	120	60	120	60	1
81	2546	0	177	76	6	4.5	73	151	130	90	170	90	1

Age	Post code	Sex	Height	Weight	Diastole	Systole	Resting heart rate	Peak heart rate	RBP high	RBP low	PBP high	PBP low	State
93	2605	0	167	82.1	6.1	4.5	66	155	140	80	199	80	1
74	2630	0	179	90.2	5.4	4.5	83	159	127	67	208	89	0
75	2621	0	168.8	81.8	6.1	4.6	53	142	130	70	200	85	1
88	2607	0	170	81	5.8	4.7	59	89	130	82	170	85	1
78	2582	0	170	81.05	5.4	4.7	90	132	145	80	190	80	1
57	2903	0	167	90	6.1	4.7	101	148	154	75	180	92	0
60	2903	0	178	104	7.1	4.7	73	166	150	81	177	98	0
92	2604	0	185	84.1	3.3	4.7	81	182	140	80	220	121	0
80	2550	0	175	102	6.9	4.8	71	116	120	70	160	80	1
79	2615	0	187	81	5.8	4.8	66	124	135	82	177	107	0
87	2537	0	166	74.02	6.1	4.8	87	136	150	90	200	100	1
81	2630	0	172	78	6.1	4.8	103	163	158	81	198	88	0
65	2621	0	182	78	5.8	4.8	74	169	115	74	199	83	1
67	2607	0	177	97	5.7	4.8	148	198	140	90	200	90	1
77	2456	0	185	93	6.4	5	74	108	150	98	160	95	1
75	2580	0	167	104	6.3	5	97	136	149	88	173	110	1
87	2902	1	161	84	7	5.1	56	122	138	82	160	90	1
87	2537	0	173	77	6.5	5.2	87	143	168	55	177	89	1
63	2619	0	170	60	6.7	5.3	57	142	100	70	120	70	1
82	2537	1	163	60.8	6.6	5.3	68	160	130	95	170	100	1
70	2620	1	165	62	6.8	5.4	61	124	129	65	202	83	0
66	2607	1	167	68	6.4	5.5	86	171	146	92	192	70	0
69	2615	0	170	83	5.9	5.5	76	191	120	80	200	90	0
64	2902	0	177	107	7	5.6	82	141	149	95	196	110	1
81	2606	0	178	91	6.4	5.6	69	163	154	67	212	110	1
81	2603	0	163	83	6.9	5.8	68	162	133	54	149	63	0
90	2546	0	164	58	6.7	6.1	55	84	150	80	180	80	1
84	2582	0	161	76	6.8	6.1	48	146	150	85	214	107	1
74	2602	0	184	105	7.7	6.1	75	160	145	95	165	85	1
70	2903	0	171	83	6.6	6.2	64	170	152	76	199	116	0
79	2537	1	164	85	7.5	6.4	100	150	160	90	200	90	1
88	2903	1	80.8	161	6.8	6.8	71	140	153	81	262	97	0
83	2611	0	190	81	6.9	6.8	74	164	135	67	151	91	1

3. Cleveland and Canberra Continuous Data Attributes Discretization

Dataset	Attribute	Intervals
Cleveland	Age	**0:** < 46 **1:** ≥ 46 < 53 **2:** ≥ 53 < 58 **3:** ≥ 58 < 63 **4:** ≥ 63
	Trestbps	**0:** < 120 **1:** ≥ 120 < 126 **2:** ≥ 126 < 134 **3:** ≥ 134 < 145 **4:** ≥ 145
	Chol	**0:** < 205 **1:** ≥ 205 < 233 **2:** ≥ 233 < 257 **3:** ≥ 257 < 288 **4:** ≥ 288
	Thalach	**0:** < 130 **1:** ≥ 130 < 147 **2:** ≥ 147 < 159 **3:** ≥ 159 < 170 **4:** ≥ 170
	Old peak ST	**0:** < 0 **1:** ≥ 0 < 0.3 **2:** ≥ 0.3 < 1.2 **3:** ≥ 1.2 < 1.9 **4:** ≥ 1.9
Canberra	Age	**0:** < 65 **1:** ≥ 65 < 72 **2:** ≥ 72 < 79 **3:** ≥ 79 <86 **4:** ≥ 86
	Postcode	**0:** < 2602 **1:** ≥ 2602 < 2607 **2:** ≥ 2607 < 2615 **3:** ≥ 2615 <2902 **4:** ≥ 2902
	Height	**0:** < 158 **1:** ≥ 158 < 164 **2:** ≥ 164 < 170 **3:** ≥ 170 < 177 **4:** ≥ 177

Dataset	Attribute	Intervals
Canberra	Weight	**0:** <66 **1:** ≥ 66 < 75 **2:** ≥ 75 < 81 **3:** ≥ 81 < 92 **4:** ≥ 92
	Diastole	**0:** <3.8 **1:** ≥ 3.8 < 4.5 **2:** ≥ 4.5 < 4.9 **3:** ≥ 4.9 < 5.3 **4:** ≥ 5.3
	Systole	**0:** < 2.1 **1:** ≥ 2.1 < 2.7 **2:** ≥ 2.7 < 3.1 **3:** ≥ 3.1 < 3.6 **4:** ≥ 3.6
	Resting Heart Rate	**0:** < 64 **1:** ≥ 64 < 71 **2:** ≥ 71 < 79 **3:** ≥ 79 <87 **4:** ≥ 87
	Peak Heart Rate	**0:** < 126 **1:** ≥ 126 < 147 **2:** ≥ 147 < 160 **3:** ≥ 160 < 172 **4:** ≥ 172
	Resting Blood Pressure High	**0:** < 123 **1:** ≥ 123 < 134 **2:** ≥ 134 < 145 **3:** ≥ 145 < 156 **4:** ≥ 156
	Resting Blood Pressure Low	**0:** < 67 **1:** ≥ 67 < 74 **2:** ≥ 74 < 80 **3:** ≥ 80 < 90 **4:** ≥ 90
	Peak Blood Pressure High	**0:** < 163 **1:** ≥ 163 < 181 **2:** ≥ 181 < 199 **3:** ≥ 199 < 217 **4:** ≥ 217

Dataset	Attribute	Intervals
Canberra	Peak Blood Pressure Low	**0:** < 75 **1:** ≥ 75 < 82 **2:** ≥ 82 < 90 **3:** ≥ 90 < 100 **4:** ≥ 100

Appendix B

1. Different Discretization Methods on Cleveland All Data Attributes

Data Mining Technique	Discretization Method	Sensitivity		Specificity		Accuracy	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Decision Tree	Equal Frequency	75.6	6.1	81.6	12.1	79.1	5.8
	Equal Width	70.5	8.7	80.7	11.6	76.3	4.9
	Chi-Merge	68.6	13	83.9	12.8	77.3	4.5
	Entropy	71.7	7.5	82.4	8.9	76.3	6.1
Naïve Bayes	Equal Frequency	78	13.8	80.8	12.6	83.5	5.2
	Equal Width	75.3	17.2	81.7	13	83.3	4.2
	Chi-Merge	71.9	17.5	84	12.5	82.6	5.9
	Entropy	73.8	15.5	79.6	11.6	81	4.2
KNN K = 1	Equal Frequency	69.2	16.3	77.8	13.5	76.5	9.7
	Equal Width	69.2	16.3	77.8	13.5	76.5	9.7
	Chi-Merge	69.2	16.3	77.8	13.5	76.5	9.7
	Entropy	69.2	16.3	77.8	13.5	76.5	9.7
KNN K = 9	Equal Frequency	78.6	8.9	84.5	5.9	83.4	2.7
	Equal Width	78.6	8.9	84.5	5.9	83.4	2.7
	Chi-Merge	78.6	8.9	84.5	5.9	83.4	2.7
	Entropy	78.6	8.9	84.5	5.9	83.4	2.7
KNN K= 19	Equal Frequency	76.7	10.7	85.1	7.5	83.2	4.1
	Equal Width	76.7	10.7	85.1	7.5	83.2	4.1
	Chi-Merge	76.7	10.7	85.1	7.5	83.2	4.1
	Entropy	76.7	10.7	85.1	7.5	83.2	4.1

2. Cleveland Single, Combined Non-Invasive Attributes

2.1. Naïve bayes

Cleveland Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	52	18.8	66.8	21.7	56.3	7.9
Sex	83	10.1	41.7	18.5	61.5	10.1
Resting Blood Pressure	26.4	13.3	78.2	12.3	54.2	12.1
Age, Sex	59	13	71.4	10.9	64.8	7.5
Age, RBP	51	15.8	66.8	13.5	58.6	5.5
Sex, RBP	60.6	23	54.8	17.5	55.1	11.4
Age, Sex, RBP	62.1	10.3	72.4	9.1	67.9	7.2

2.2. K = 1 Nearest Neighbour

Cleveland Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	11.2	8	91.4	5.9	54.5	16.5
Sex	75	26.9	47.5	25.5	62.6	9.9
Resting Blood Pressure	1.7	5.1	100	0	54.7	20.7
Age, Sex	20.3	27.5	90.4	21.7	54.1	17.2
Age, RBP	0	0	100	0	54.3	20.1
Sex, RBP	9.7	24	94.2	17.4	53.6	20.3
Age, Sex, RBP	17.5	19.8	82.6	18	49.4	16.5

2.3. K = 9 Nearest Neighbour

Cleveland Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	0	0	100	0	54.3	20.1
Sex	0	0	100	0	54.3	20.1
Resting Blood Pressure	0	0	100	0	54.3	20.1
Age, Sex	28.5	33.5	75.9	29.9	53.3	16.8
Age, RBP	0	0	100	0	54.7	20.7
Sex, RBP	0	0	100	0	54.3	20.1
Age, Sex, RBP	7	15.5	95.1	8.6	52.5	18.9

2.4. K = 19 Nearest Neighbour

Cleveland Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	0	0	100	0	54.3	20.1
Sex	0	0	100	0	54.3	20.1
Resting Blood Pressure	0	0	100	0	54.3	20.1
Age, Sex	2	6	100	0	54.6	20.6
Age, RBP	0	0	100	0	54.7	20.7
Sex, RBP	52.9	43.7	65.2	33.5	56.8	13.7
Age, Sex, RBP	26.3	22	81.8	18.1	55.5	10.9

3. Canberra Single, Combined, equation Non-Invasive Attributes

3.1. Naïve bayes

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	61.6	16.5	75.3	19.7	71	8.2
Sex	67.4	14.7	58.4	16.3	66	8.3
Resting Blood Pressure	29.4	7.8	71.6	8	51	8.8
Height	53.9	16.3	52.4	10.6	52.3	10.3
Weight	26	25.7	69.7	25.1	46.5	9.1
Age, Sex	68.3	14	73.8	21.4	74.2	7.5
Age, RBP	60.2	13.9	73.8	16.9	68.9	6.8
Age, Height	63.5	11.6	72.2	18.8	69.4	9.7
Age, Weight	63.2	13.5	71.4	19.8	68.4	8
Sex, RBP	62.2	11.9	62.5	15	65	7.6
Sex, Height	64.8	12.8	59.2	15.9	64.9	7.3
Sex, Weight	66	13.9	59.3	16.4	65.2	7.8
RBP, Height	41.1	9	63	9.9	53.9	9.6
RBP, Weight	36.5	11.3	71.4	12.1	53.7	10.9
Height, Weight	44.4	11.3	61.4	10.8	53	10
Age, Sex, RBP	65.5	12.8	74.3	20	72.8	6.6
Age, Sex, Height	61.2	15.1	76.5	16	72.6	7.7
Age, Sex, Weight	65.3	13.2	74.9	19.7	72.9	5.8
Age, RBP, Height	60.5	12.1	72.8	19.8	69.1	8.5
Age, RBP, Weight	60.4	12.6	71.8	18.9	67.7	6
Age, Height, Weight	63.5	12.1	71.6	20.3	68.9	9.9
Age, Sex, RBP, Height	63.2	9.8	75.4	13.7	71.9	5.6
Age, Sex, RBP, Weight	64.7	13	74.9	17.3	72.3	6.1
Age, Sex, Height, Weight	63.3	9.1	76.8	11.4	72.3	6.4

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, RBP, Height, Weight	61.5	12	71.8	21.1	68.4	9.2
Sex, RBP, Height, Weight	63.1	10.4	61.4	14.6	64.9	5.9
Age, Sex, RBP, Height, Weight	62.4	11.8	75.6	13.2	71.9	6.7
Age, Sex, RBP	65.5	12.8	74.3	20	72.8	6.6
Age, BMI	63.4	14.5	71.7	18.5	68.7	6.9
Sex, BMI	60.9	14.2	69.5	11.9	68.8	6.1
RBP, BMI	56	12.5	58.2	10.8	54.3	3.7
Age, Sex, BMI	70.5	12.8	74.6	18.1	74.5	6.4
Age, RBP, BMI	62.5	10.3	71.9	16.3	68.1	5.9
Sex, RBP, BMI	56.7	11.2	69	11	66	4.6
Age, Sex, RBP, BMI	66.7	11.7	72.7	16.2	71.8	6
Age, Sex, RBP	65.5	12.8	74.3	20	72.8	6.6
Age, Rohrer’s index	64.5	12	73	19.7	70.3	7.7
Sex, Rohrer’s index	63	15	63.7	15.5	67.1	6.5
RBP, Rohrer’s index	50.5	7.8	63.3	8.1	56.5	6.5
Age, Sex, Rohrer’s index	70.4	10.6	73.9	17.6	74.2	5.7
Age, RBP, Rohrer’s index	62.4	10.5	74.2	17.6	69.8	5.7
Sex, RBP, Rohrer’s index	59.6	14.6	65.1	14.1	66	5.7
Age, Sex, RBP, Rohrer’s index	69.1	12.3	73.1	16.4	73.3	6.1
Age, Sex, RBP	65.5	12.8	74.3	20	72.8	6.6
Age, RBPDiff	62.3	15.3	72.9	18.5	69.6	6.4
Sex, RBPDiff	64.2	13.8	59.6	15.4	64.5	9
RBP, RBPDiff	35.6	8.6	69.8	13	52	7.8
Age, Sex, RBPDiff	65.2	12.1	75.8	19.8	73.6	7.3
Age, RBP, RBPDiff	59.3	14.2	71.4	19.5	67.3	8.4
Sex, RBP, RBPDiff	56.4	10.8	63.1	14.8	62.2	9

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, Sex, RBP, RBPDiff	64.9	12	76.5	14.7	72.7	6.6
Age, Sex, RBP, BMI, RBPDiff	68.7	12.3	73.6	16	72.6	6.2
Age, Sex, RBP, Rohrer's index,	63.9	12.4	73.9	16.6	73.9	4.9

3.2. K = 1 Nearest Neighbour

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	8.2	7.4	97.9	2.7	57	14
Sex	98.2	5.4	5.5	16.5	45	15.1
Resting Blood Pressure	53.8	21.1	48.9	20.7	52.1	10.4
Height	63.1	11.1	44.9	13.9	52.6	6.8
Weight	50.6	10.8	44.9	12	46.3	8
Age, Sex	9.4	7.8	99.1	1.9	58.1	15.6
Age, RBP	42.8	21.3	82.2	16.8	62.2	10.4
Age, Height	8.5	7	97.3	4.1	57	16.2
Age, Weight	7.5	8.3	96.8	3.7	56.7	15.6
Sex, RBP	71.2	15.4	54.8	16.7	65	9.3
Sex, Height	67.6	9.2	60.9	8.6	64	6.6
Sex, Weight	51.2	11.4	56.1	8.6	54.3	8.8
RBP, Height	55.5	20.6	47.8	19.8	52.6	9.2
RBP, Weight	47.6	23.4	51.1	18.9	50.7	6.3
Height, Weight	47.6	10.1	46.2	9.5	46.7	4
Age, Sex, RBP	54.8	10.8	77.2	15.7	66.5	11.1
Age, Sex, Height	13.8	5.6	97.3	4.1	59.8	14
Age, Sex, Weight	13.2	10.1	98.5	2.3	59.8	14.9
Age, RBP, Height	45.3	20	81.9	16.8	63.1	10.4

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, RBP, Weight	44	18.6	81.4	17.1	62.1	9.8
Age, Height, Weight	8.1	6.3	97.6	2.7	57.3	13.4
Age, Sex, RBP, Height	54.4	10.7	77.2	15.7	66.2	10.8
Age, Sex, RBP, Weight	55.4	14.6	78.8	15.1	66.5	10.5
Age, Sex, Height, Weight	12.4	8.4	98.5	2.3	59.6	14.4
Age, RBP, Height, Weight	45.8	18.6	80.5	17.7	62.6	10.5
Sex, RBP, Height, Weight	64.9	18.2	59.2	16.5	64.3	9.1
Age, Sex, RBP, Height, Weight	55.7	15.5	79.1	15.3	66.7	10.6
Age, Sex, RBP	54.8	10.8	77.2	15.7	66.5	11.1
Age, BMI	11.1	11.5	97.4	3.6	58.1	14.5
Sex, BMI	54.6	12.2	65.6	13.9	59.5	8.4
RBP, BMI	51.5	22.3	48.6	20.4	50.8	10
Age, Sex, BMI	10.9	10.6	98.5	2	58.4	15.2
Age, RBP, BMI	47.8	20.8	81	17.3	63.9	10.8
Sex, RBP, BMI	70.9	15.3	54.5	16.3	64.6	9
Age, Sex, RBP, BMI	56.4	11.6	77.2	15.3	67.2	11
Age, Sex, RBP	54.8	10.8	77.2	15.7	66.5	11.1
Age, Rohrer’s index	10.4	8.6	97.4	3	58	12.8
Sex, Rohrer’s index	61.6	14.8	51.2	11.9	54.5	6.1
RBP, Rohrer’s index	53	22.1	49.2	21.3	52	10.2
Age, Sex, Rohrer’s index	10	7.7	98.5	2	58.3	13.9
Age, RBP, Rohrer’s index	44.2	21.2	81.6	17.2	62.4	10.2
Sex, RBP, Rohrer’s index	71.2	15.4	54.8	16.7	65	9.3
Age, Sex, RBP, Rohrer’s index	55.7	11.3	77.2	15.7	67	11
Age, Sex, RBP	54.8	10.8	77.2	15.7	66.5	11.1
Age, RBPDiff	10.4	9.4	90	4.2	54.7	9.4

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Sex, RBPDiff	51	12.1	57.6	11.8	55	10.9
RBP, RBPDiff	66.8	13.8	31.6	11.4	47.1	6.7
Age, Sex, RBPDiff	12.1	8.5	94.2	5.1	58	10
Age, RBP, RBPDiff	47.7	16.6	75.3	13.6	61.4	8.4
Sex, RBP, RBPDiff	73.2	12.7	39.4	12.7	56.5	7
Age, Sex, RBP, RBPDiff	50.9	15	78.5	11.8	64.2	7.2
Age, Sex, RBP, BMI, RBPDiff	51.9	15.3	78.5	11.8	64.6	7
Age, Sex, RBP, Rohrer's index,	51.6	15.4	78.5	11.8	64.4	7.2

3.3. K = 9 Nearest Neighbour

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	0.8	2.4	100	0	55.1	15.5
Sex	90	30	10	30	39.9	12.4
Resting Blood Pressure	14.7	19.4	85.3	15.7	54.9	11.4
Height	51.1	12.2	59.4	17.7	54.5	7.4
Weight	35.6	14.3	56.1	10.8	45.6	10.6
Age, Sex	0.8	1.6	100	0	55.3	14.9
Age, RBP	0	0	100	0	54.9	15.3
Age, Height	0	0	100	0	54.9	15.3
Age, Weight	0	0	100	0	54.9	15.3
Sex, RBP	67.4	14.7	58.4	16.3	66	8.3
Sex, Height	61.4	12.1	64.8	17.3	62.8	8.4
Sex, Weight	62	11.7	64.6	11	65.8	6.6
RBP, Height	15.3	20.3	85.1	14.9	55.2	13.1
RBP, Weight	26.2	11.6	76.9	12.8	54.8	7.7

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Height, Weight	51.1	9.3	58.5	8.6	55.2	6.5
Age, Sex, RBP	14.1	9.7	98.2	5.4	60.9	13.8
Age, Sex, Height	0	0	100	0	54.9	15.3
Age, Sex, Weight	3.6	3.6	99.7	0.9	55.9	15.6
Age, RBP, Height	0	0	100	0	54.9	15.3
Age, RBP, Weight	0	0	100	0	54.9	15.3
Age, Height, Weight	0	0	100	0	54.9	15.3
Age, Sex, RBP, Height	14.6	10.7	98.2	5.4	61.2	13.9
Age, Sex, RBP, Weight	20.8	10.5	97.2	5.5	63.6	12.8
Age, Sex, Height, Weight	4	3.4	99.7	0.9	56.1	15.3
Age, RBP, Height, Weight	1.4	2.8	99	1.5	54.7	15.2
Sex, RBP, Height, Weight	65.1	14.8	59.2	15.6	65.4	7.1
Age, Sex, RBP, Height, Weight	21.2	10.3	97.2	5.5	63.9	12.4
Age, Sex, RBP	14.1	9.7	98.2	5.4	60.9	13.8
Age, BMI	0.4	1.2	100	0	55.1	15.2
Sex, BMI	54.4	9.1	68.3	10.8	63.5	6.9
RBP, BMI	15.2	19.6	85.5	15.3	55.1	12.1
Age, Sex, BMI	6	3.9	100	0	57.4	14.4
Age, RBP, BMI	0	0	100	0	54.9	15.3
Sex, RBP, BMI	67.4	14.7	58.4	16.3	66	8.3
Age, Sex, RBP, BMI	14.1	9.7	98.2	5.4	60.9	13.8
Age, Sex, RBP	14.1	9.7	98.2	5.4	60.9	13.8
Age, Rohrer’s index	0.8	2.4	99.7	0.9	54.9	15.3
Sex, Rohrer’s index	52.3	13.5	65.5	9.4	61.5	3.5
RBP, Rohrer’s index	14.7	20.1	84.5	15.2	54.4	11.3
Age, Sex, Rohrer’s index	0.7	2.1	100	0	55.1	15.4

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, RBP, Rohrer's index	0	0	100	0	54.9	15.3
Sex, RBP, Rohrer's index	67.4	14.7	58.4	16.3	66	8.3
Age, Sex, RBP, Rohrer's index	14.1	9.7	98.2	5.4	60.9	13.8
Age, Sex, RBP	14.1	9.7	98.2	5.4	60.9	13.8
Age, RBPDiff	0	0	100	0	54.9	15.3
Sex, RBPDiff	50.8	16	59.1	15.3	57.9	10.8
RBP, RBPDiff	22.8	14	77.6	10	54.3	10.3
Age, Sex, RBPDiff	0	0	100	0	54.9	15.3
Age, RBP, RBPDiff	0	0	100	0	54.9	15.3
Sex, RBP, RBPDiff	62.5	13.3	61.3	13.8	64.7	8.4
Age, Sex, RBP, RBPDiff	10.8	8.8	99.1	2.7	59.5	14.7
Age, Sex, RBP, BMI, RBPDiff	10.8	8.8	99.1	2.7	59.5	14.7
Age, Sex, RBP, Rohrer's index,	10.8	8.8	99.1	2.7	59.5	14.7

3.4. K = 19 Nearest Neighbour

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age	0	0	100	0	54.9	15.3
Sex	59.2	24.1	62.9	20.5	60.9	14.5
Resting Blood Pressure	22.4	31.7	82.2	24.2	51.6	13.2
Height	51	12.4	62.6	16.1	58.1	7.3
Weight	35.8	17.3	56.3	9.8	44.8	9.4
Age, Sex	0	0	100	0	54.9	15.3
Age, RBP	0	0	100	0	54.9	15.3
Age, Height	0	0	100	0	54.9	15.3
Age, Weight	0	0	100	0	54.9	15.3

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Sex, RBP	67.1	14.5	58.4	16.3	65.8	8.1
Sex, Height	62.3	13.3	67.7	14.3	66.9	8.4
Sex, Weight	61.5	11.7	66	10.1	66.1	5.8
RBP, Height	22.3	30.7	83.2	20.7	52.6	14.2
RBP, Weight	30.6	21.9	68.8	19.4	48.6	9.7
Height, Weight	50.5	9.9	58.6	7.3	53.9	6.2
Age, Sex, RBP	10.4	8.4	100	0	59.4	14.6
Age, Sex, Height	0	0	100	0	54.9	15.3
Age, Sex, Weight	0.4	1.2	100	0	55.1	15.2
Age, RBP, Height	0	0	100	0	54.9	15.3
Age, RBP, Weight	0	0	100	0	54.9	15.3
Age, Height, Weight	0	0	100	0	54.9	15.3
Age, Sex, RBP, Height	10.7	8.3	100	0	59.6	14.2
Age, Sex, RBP, Weight	13.6	8.5	99.4	1.8	60.8	13.5
Age, Sex, Height, Weight	0.7	2.1	100	0	55.1	15.4
Age, RBP, Height, Weight	0	0	100	0	54.9	15.3
Sex, RBP, Height, Weight	67	15	58.8	16.3	66	8
Age, Sex, RBP, Height, Weight	13.6	8.5	99.4	1.8	60.8	13.5
Age, Sex, RBP	10.4	8.4	100	0	59.4	14.6
Age, BMI	0	0	100	0	54.9	15.3
Sex, BMI	57.4	12.2	67.9	13.2	65.2	8
RBP, BMI	22.9	32.6	81.2	24.5	51.2	14.4
Age, Sex, BMI	0	0	100	0	54.9	15.3
Age, RBP, BMI	0	0	100	0	54.9	15.3
Sex, RBP, BMI	67.1	14.5	58.4	16.3	65.8	8.1
Age, Sex, RBP, BMI	10.4	8.4	100	0	59.4	14.6
Age, Sex, RBP	10.4	8.4	100	0	59.4	14.6

Canberra Data Attributes	Sensitivity		Specificity		Accuracy	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
Age, Rohrer's index	0	0	100	0	54.9	15.3
Sex, Rohrer's index	55.8	13.1	66.3	12.3	64.5	5.9
RBP, Rohrer's index	22.3	31.6	81.7	23.9	51.4	13.5
Age, Sex, Rohrer's index	0	0	100	0	54.9	15.3
Age, RBP, Rohrer's index	0	0	100	0	54.9	15.3
Sex, RBP, Rohrer's index	67.1	14.5	58.4	16.3	65.8	8.1
Age, Sex, RBP, Rohrer's index	10.4	8.4	100	0	59.4	15.6
Age, Sex, RBP	10.4	8.4	100	0	59.4	14.6
Age, RBPDiff	0	0	100	0	54.9	15.3
Sex, RBPDiff	50.9	15.8	62.4	17.3	59.7	7.8
RBP, RBPDiff	13.2	15.5	88.6	14.6	52.9	13.3
Age, Sex, RBPDiff	0.6	1.8	100	0	55.1	15.3
Age, RBP, RBPDiff	0	0	100	0	54.9	15.3
Sex, RBP, RBPDiff	64.1	13.7	60.7	17.2	65.5	8.9
Age, Sex, RBP, RBPDiff	7.6	8	99.1	2.7	58.3	14.7
Age, Sex, RBP, BMI, RBPDiff	7.6	8	99.1	2.7	58.3	14.7
Age, Sex, RBP, Rohrer's index,	7.6	8	99.1	2.7	58.3	14.7

4. Cleveland All Data Attributes Results

4.1. Integrating K-Means clustering with Naïve Bayes

Cleveland All Data Attributes							
Integrating Clustering with Naïve Bayes							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		78	13.8	80.8	12.6	83.5	5.2
Number of Clusters = 2	Inlier	73.2	22	83.3	12.2	83.1	4.7
	Outlier	76.4	15.2	82.3	13.1	83.4	4.5
	Range	76.4	15.2	82.3	13.1	83.4	4.5
	Random Row	76.5	14.6	82	12.7	83.1	3.4
	Random Attribute	76.4	15.2	83.8	12.5	84	4.4
Number of Clusters = 3	Inlier	75.9	14.7	83.7	12.6	83.6	4.4
	Outlier	75.9	14.7	83.7	12.6	83.6	4.4
	Range	78	13.1	81.3	14.6	84.1	4.3
	Random Row	76.6	14.4	85	13.2	84.8	4.7
	Random Attribute	74.3	17.6	82.2	14.1	82.3	6
Number of Clusters = 4	Inlier	76.9	13.5	81.7	13.1	83.5	3.1
	Outlier	70.1	17.4	83.1	12.9	80.7	6.8
	Range	76.9	13.5	81.7	13.1	83.5	3.1
	Random Row	71.4	19.6	83	13.1	81.6	6.3
	Random Attribute	70.9	20.5	83.5	12.3	82	5.7
Number of Clusters = 5	Inlier	76.9	13.5	81.7	13.1	83.5	3.1
	Outlier	70.1	17.4	83.1	12.9	80.7	6.8
	Range	76.9	13.5	81.7	13.1	83.5	3.1
	Random Row	71.1	14.8	81.1	13.1	80.4	4.7
	Random Attribute	71.1	14.8	81.1	13.1	80.4	4.7

4.2. Integrating K-Means clustering with K = 1 Nearest Neighbour

Cleveland All Data Attributes							
Integrating Clustering with K =1 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		69.2	16.3	77.8	13.5	76.5	9.7
Number of Clusters = 2	Inlier	70.8	10.6	78	13.7	77.9	7.1
	Outlier	71.4	10.8	79.9	9.2	78	7.2
	Range	70.8	10.6	78	13.7	77.9	7.1
	Random Row	71.5	11.3	80.6	9.8	78.7	7.6
	Random Attribute	69.6	10.5	79.6	9.6	77.3	7.2
Number of Clusters = 3	Inlier	84.2	13.3	65.4	14.7	76.2	6.7
	Outlier	74.9	18.1	66	17.3	75.7	6.8
	Range	80.9	13.1	66.7	14.8	76.6	6.3
	Random Row	75.5	18.2	67.2	14.7	75.2	7.7
	Random Attribute	77.3	18.9	63	23.4	75.7	6.5
Number of Clusters = 4	Inlier	78.6	13.1	72	11.1	76.5	7.9
	Outlier	78.1	10.4	61.8	25	72.2	7.6
	Range	80.9	13.1	66.7	14.8	76.6	6.3
	Random Row	74.9	16.7	74	7.6	75.5	7.1
	Random Attribute	75.6	12	69.4	16.5	74.5	5.6
Number of Clusters = 5	Inlier	78	13.1	76.3	7.6	78.3	7.7
	Outlier	73.9	15.9	69.1	17.5	73.9	4.1
	Range	80.3	14.6	70.7	11.6	75.8	8.2
	Random Row	70.9	18.7	66.2	12.4	71.2	11.6
	Random Attribute	69.2	13.5	69.9	10.1	72.4	7.4

4.3. Integrating K-Means clustering with K = 9 Nearest Neighbour

Cleveland All Data Attributes							
Integrating Clustering with K = 9 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		78.6	8.9	84.5	5.9	83.4	2.7
Number of Clusters = 2	Inlier	76.9	10.6	85.3	11	82.1	7.4
	Outlier	79.6	13.6	81.8	7.6	82.7	6
	Range	76.9	10.6	85.3	11	82.1	7.4
	Random Row	78.9	14.2	83.1	9	83.2	6.1
	Random Attribute	77.1	11.4	80.8	7.2	80.6	6.4
Number of Clusters = 3	Inlier	83	14.6	74	12.7	79.7	6.2
	Outlier	69.7	16	71.6	17.2	75.8	6.2
	Range	81.3	13.4	73.1	13	78.7	7
	Random Row	80.4	15.4	74	13	78	7
	Random Attribute	78.2	14.8	72.7	12.7	77.3	7.7
Number of Clusters = 4	Inlier	55.8	21.1	83.2	9.6	72.6	8.7
	Outlier	81.1	15.1	68.9	16.8	77	6.4
	Range	81.3	13.4	73.1	13	78.7	7
	Random Row	74.3	12.1	80.4	9.3	78.7	5
	Random Attribute	74.4	13	79.2	20.1	81.1	5.1
Number of Clusters = 5	Inlier	53.8	19.8	82.4	11.5	71.4	8.6
	Outlier	78.8	14.3	70.3	10.4	75.7	4.8
	Range	60.8	24.8	82.8	9.9	73.3	8.9
	Random Row	72.8	20.3	75.8	12.6	74.1	5
	Random Attribute	69.1	18.2	81.8	12.1	74.4	3.7

4.4. Integrating K-Means clustering with K = 19 Nearest Neighbour

Cleveland All Data Attributes							
Integrating Clustering with K = 19 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		76.7	10.7	85.1	7.5	83.2	4.1
Number of Clusters = 2	Inlier	77.8	13	88.8	5.9	85.7	5.4
	Outlier	75.9	17.7	84.2	9.8	83.6	5.6
	Range	77.8	13	88.8	5.9	85.7	5.4
	Random Row	77.9	12.9	86	10.3	84.7	4.9
	Random Attribute	79.8	11.7	86.6	6.4	85.7	5.4
Number of Clusters = 3	Inlier	63.8	15.7	78.1	12.9	75.3	6.4
	Outlier	61.3	13.6	76.1	11.9	72.8	5.8
	Range	67.2	11.9	79	12.9	76.7	5.7
	Random Row	61.9	14	77.4	12.4	74.2	5.2
	Random Attribute	63.4	14.2	79.2	12.5	75.2	6.3
Number of Clusters = 4	Inlier	50.5	15	81.1	8.5	70	7.8
	Outlier	67.4	13.2	71.6	11.2	72.1	6.2
	Range	67.2	11.9	79	12.9	76.7	5.7
	Random Row	57.3	16.2	82.9	11.1	72.8	8.2
	Random Attribute	57.5	17	78.9	14.2	74.1	5.9
Number of Clusters = 5	Inlier	50.5	15.9	81.8	8.7	70.4	7.5
	Outlier	66.5	14.9	71.4	11.2	70.7	5.2
	Range	50.5	15	80.7	8.7	69.7	7.9
	Random Row	56	16.1	80	16.2	70.1	7.4
	Random Attribute	52.7	21.4	75.6	13.5	67.8	6.4

5. Cleveland Non-Invasive Data Attributes Results

5.1. Integrating K-Means clustering with Naïve Bayes

Cleveland Non-Invasive Data Attributes							
Integrating Clustering with Naïve Bayes							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		62.1	10.3	72.4	9.1	67.9	7.2
Number of Clusters = 2	Inlier	60.2	11.6	67.7	10.3	64.9	9
	Outlier	57.9	11.7	70.8	11	65.6	9.6
	Range	57.9	11.7	70.8	11	65.6	9.6
	Random Row	58.7	13.5	67.1	10.6	63.5	9.1
	Random Attribute	59.7	11.6	67.7	9.5	64.1	8.7
Number of Clusters = 3	Inlier	52.5	15.3	69.6	7.9	61.4	9.3
	Outlier	52.5	15.3	69.6	7.9	61.4	9.3
	Range	64.7	10.9	66	11.6	66.4	9.2
	Random Row	59.5	10.6	65.8	10.4	63.5	9.8
	Random Attribute	56.3	13.7	66.4	10.2	61.5	8.6
Number of Clusters = 4	Inlier	58.9	13.8	65.8	7.5	62.6	8.2
	Outlier	53.6	12.9	66.8	8.3	60.1	8.7
	Range	58.9	13.8	65.8	7.5	62.6	8.2
	Random Row	59.3	13.1	65.8	8.2	62.5	8.8
	Random Attribute	58.2	14.4	63.6	7.1	60.4	7.7
Number of Clusters = 5	Inlier	58.9	13.8	65.8	7.5	62.6	8.2
	Outlier	53.6	12.9	66.8	8.3	60.1	8.7
	Range	58.9	13.8	65.8	7.5	62.6	8.2
	Random Row	59.9	12.1	62.9	7.1	61.2	7.7
	Random Attribute	59.9	12.1	62.9	7.1	61.2	7.7

5.2. Integrating K-Means clustering with K = 1 Nearest Neighbour

Cleveland Non-Invasive Data Attributes							
Integrating Clustering with K = 1 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		17.5	19.8	82.6	18	49.4	16.5
Number of Clusters = 2	Inlier	12.5	20.1	90.8	17.8	51.5	17.7
	Outlier	17.4	14.7	79.2	18.7	47	15.7
	Range	12.5	20.1	90.8	17.8	51.5	17.7
	Random Row	13.1	20.2	86.9	18	49.5	16.6
	Random Attribute	13	20.5	88.1	17.3	49.1	15.6
Number of Clusters = 3	Inlier	65.8	19.1	65.5	16.3	64.7	11
	Outlier	36.8	20	74.5	11.5	55.9	7.6
	Range	62.5	15.4	67.7	11.7	65.8	8.9
	Random Row	45.3	19.2	72.8	11.9	59.6	9.7
	Random Attribute	52.4	22.8	66.4	12.8	61.7	9.1
Number of Clusters = 4	Inlier	39.2	22.8	80.9	16.5	61.6	11.3
	Outlier	56.2	24.9	66.5	22.3	58	9.6
	Range	62.5	15.4	67.7	11.7	65.8	8.9
	Random Row	40.8	21.8	75.6	17.2	59.4	11.7
	Random Attribute	42.2	25.8	72.9	20.7	55.4	8.2
Number of Clusters = 5	Inlier	36	19.7	84.1	9.3	61.9	11.4
	Outlier	51.4	15.9	60.2	17.9	54.1	8.6
	Range	47.5	27.8	74.8	22	58.4	11.3
	Random Row	52.1	12	54	22.2	55.2	12
	Random Attribute	43.8	11.2	60.7	12.2	50.8	10.5

5.3. Integrating K-Means clustering with K = 9 Nearest Neighbour

Cleveland Non-Invasive Data Attributes							
Integrating Clustering with K = 9 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		7	15.5	95.1	8.6	52.5	18.9
Number of Clusters = 2	Inlier	44.8	30	65.7	20.4	51	16.8
	Outlier	35.1	14.2	67.1	15.5	50.7	14.5
	Range	44.8	30	65.7	20.4	51	16.8
	Random Row	37	23.9	67.7	16.1	50	17.5
	Random Attribute	36.1	16.1	65.9	16.9	50.3	14.5
Number of Clusters = 3	Inlier	63.2	23.1	52.3	16.6	57.5	10.8
	Outlier	35.5	11.3	65	14.9	52.9	10.5
	Range	63.2	23.1	53.1	15.2	58.1	9.6
	Random Row	49.1	19.6	56.9	19.9	54.7	10.7
	Random Attribute	53.2	23.1	57.5	12.8	54.2	9.9
Number of Clusters = 4	Inlier	38.5	17.7	70.6	10.6	58.4	6.9
	Outlier	61.2	20.8	57.1	18.8	57.2	12.6
	Range	63.2	23.1	53.1	15.2	58.1	9.6
	Random Row	53.6	22.9	61.4	23.8	58.3	12.9
	Random Attribute	62.5	21.3	53.6	22.2	55.1	14.9
Number of Clusters = 5	Inlier	35.6	14.7	72.2	13.4	58	7.7
	Outlier	58.9	20.1	47.6	16.8	52.4	15.6
	Range	45.1	20.5	65.8	16.9	56	9.1
	Random Row	62.4	21.9	56	23.5	55.6	12.8
	Random Attribute	56	26.8	54.5	22	53.7	13.7

5.4. Integrating K-Means clustering with K = 19 Nearest Neighbour

Cleveland Non-Invasive Data Attributes							
Integrating Clustering with K = 19 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		26.3	22	81.8	18.1	55.5	10.9
Number of Clusters = 2	Inlier	0	0	100	0	54.3	20.1
	Outlier	41.4	22.6	82.3	11.6	59.9	19.8
	Range	0	0	100	0	54.3	20.1
	Random Row	25.7	26.4	91.7	9.3	57.6	21
	Random Attribute	20.7	26.8	93.9	7.8	54.9	20.9
Number of Clusters = 3	Inlier	38.9	12.9	74.1	7.9	60.7	6.8
	Outlier	29.2	11.7	76.9	8.1	57.2	8.7
	Range	37.2	16.4	76.3	11.1	62.1	8.7
	Random Row	32.5	13.2	75.1	8.7	58.2	8
	Random Attribute	36.9	15.4	74.6	8.1	60.3	7
Number of Clusters = 4	Inlier	35	13.9	75	8.3	59	7.2
	Outlier	43.9	10.1	71.3	8.6	59.4	8.8
	Range	37.2	16.4	76.3	11.1	62.1	8.7
	Random Row	39.9	26.1	76	5.3	60.1	7.3
	Random Attribute	46.8	23.6	70.3	14.4	60.3	8.6
Number of Clusters = 5	Inlier	36.2	15.6	72.6	13.5	58.6	7.5
	Outlier	45.5	17.8	73.7	13.4	59.3	16.1
	Range	35	13.9	75	8.3	59	7.2
	Random Row	58.5	21.7	65.6	10.6	63.9	6.8
	Random Attribute	59.4	13.5	65.5	10.4	62.8	8.5

6. Canberra All Data Attributes Results

6.1. Integrating K-Means clustering with Naïve Bayes

Canberra All Data Attributes							
Integrating Clustering with Naïve Bayes							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		65.9	15.1	76.8	23.1	75.4	9.1
Number of Clusters = 2	Inlier	58.2	15.4	77.5	19.3	72.3	7.6
	Outlier	63.9	20.1	73.1	20.2	71.4	9.5
	Range	63.9	20.1	73.1	20.2	71.4	9.5
	Random Row	61.2	16	78.1	16.2	72.9	8.2
	Random Attribute	58.2	18.5	77.3	17	71.2	8.9
Number of Clusters = 3	Inlier	60.6	16.5	77.4	14.9	71.9	7.8
	Outlier	60.6	16.5	77.4	14.9	71.9	7.8
	Range	63.9	18.3	73.5	18.9	71.4	8.2
	Random Row	65	17.4	75.7	18.5	72.7	9.1
	Random Attribute	61.7	16.9	73.7	16.4	70.9	7.3
Number of Clusters = 4	Inlier	61.6	15.2	77.4	13.9	72.3	6.8
	Outlier	62.4	17.4	74.4	14.4	70.5	7.4
	Range	61.6	15.2	77.4	13.9	72.3	6.8
	Random Row	63.1	17	76.6	15.3	71.9	7.8
	Random Attribute	60.3	16.4	76.7	14.4	71.1	6.7
Number of Clusters = 5	Inlier	61.6	15.2	77.4	13.9	72.3	6.8
	Outlier	62.4	17.4	74.4	14.4	70.5	7.4
	Range	61.6	15.2	77.4	13.9	72.3	6.8
	Random Row	63.3	15.5	74.5	13.4	70.8	6.5
	Random Attribute	63.3	15.5	74.5	13.4	70.8	6.5

6.2. Integrating K-Means clustering with K = 1 Nearest Neighbour

Canberra All Data Attributes							
Integrating Clustering with K = 1 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		64.2	19	57	17.1	58.8	8.6
Number of Clusters = 2	Inlier	52.8	21.6	61.8	13.8	61.3	9.3
	Outlier	63	17.5	55.7	12.9	61.4	9.6
	Range	52.8	21.6	61.8	13.8	61.3	9.3
	Random Row	60.9	18.6	57.7	14.1	61.9	9.9
	Random Attribute	59.3	22.2	58	12.4	62	9.2
Number of Clusters = 3	Inlier	50.8	27.2	52.6	25.3	50.4	12.5
	Outlier	54.4	23.4	61.5	15.5	59.8	11.2
	Range	50.8	27.2	52.6	25.3	50.4	12.5
	Random Row	49.8	27.7	65.1	15.8	54.1	13.7
	Random Attribute	48.1	18.5	56.8	18.4	56.3	6.6
Number of Clusters = 4	Inlier	56.3	20.8	67.4	19.2	62.1	10.4
	Outlier	62.1	14.8	48.3	9.8	55.3	6.7
	Range	54.4	20.6	69.9	18.4	62.5	10.3
	Random Row	74.7	16.1	44	22.9	59	14.4
	Random Attribute	53.6	25.5	46.2	22.7	53.3	11.3
Number of Clusters = 5	Inlier	72.4	27.1	51.1	22.8	64.8	9.5
	Outlier	69.8	20.6	43.9	24.8	57.5	10.3
	Range	56.3	20.8	67.4	19.2	62.1	10.4
	Random Row	78.3	19.2	34.6	22	57	14.6
	Random Attribute	72.8	19.3	37.6	28.6	56.8	12.3

6.3. Integrating K-Means clustering with K = 9 Nearest Neighbour

Canberra All Data Attributes							
Integrating Clustering with K = 9 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		41.9	20	85.5	15.3	68.8	8.6
Number of Clusters = 2	Inlier	58.4	16.4	76.1	15.7	71.5	9.2
	Outlier	55.9	19.8	78	14.9	71.7	8.9
	Range	58.4	16.4	76.1	15.7	71.5	9.2
	Random Row	58.4	16.4	77.2	14.6	72	9
	Random Attribute	58.4	16.4	77.2	14.6	72	9
Number of Clusters = 3	Inlier	58.1	20.2	75.2	13.9	68.6	9.1
	Outlier	50.7	12.3	78.1	11	65.4	8.9
	Range	58.1	20.2	75.2	13.9	68.6	9.1
	Random Row	55.7	16.7	75.2	8.8	67.9	8.9
	Random Attribute	48.3	19.4	75.5	12	64.1	5.1
Number of Clusters = 4	Inlier	56.1	16.9	79.8	14.7	70	6.1
	Outlier	40.8	16.4	83.1	14.3	65.8	10.7
	Range	55.2	17.4	79.4	14.7	69.3	6.6
	Random Row	43.8	20.3	78.6	21.1	63.5	13.6
	Random Attribute	54.8	9.2	74.6	18.6	67.7	10.6
Number of Clusters = 5	Inlier	57.3	19.9	79.3	14.5	71.9	6.8
	Outlier	60.5	25.2	70.7	19.3	63.9	11.9
	Range	56.1	16.9	79.8	14.7	70	6.1
	Random Row	57.5	27	73.8	23.1	66.6	11.6
	Random Attribute	53.2	20.7	78	18.7	68	11

6.4. Integrating K-Means clustering with K = 19 Nearest Neighbour

Canberra All Data Attributes							
Integrating Clustering with K = 19 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		45.1	22.2	83.5	15.1	68.6	10
Number of Clusters = 2	Inlier	56.5	16.3	79.1	14.8	72.4	8
	Outlier	61.1	14.4	76.7	15.7	72	9.5
	Range	56.5	16.3	79.1	14.8	72.4	8
	Random Row	58.5	15.3	77.6	14	71.6	7.5
	Random Attribute	60.7	14.7	75.9	15.3	71.4	9.7
Number of Clusters = 3	Inlier	58.4	14.5	75.2	19.6	69.7	8.2
	Outlier	50.7	14.2	79.6	16.5	67.7	9.7
	Range	58.4	14.5	75.2	19.6	69.7	8.2
	Random Row	53.9	14.3	82.6	13.6	69.5	8.3
	Random Attribute	55.6	17.3	78.3	18.8	70.1	7.4
Number of Clusters = 4	Inlier	50.7	16.9	85.9	10.7	69.9	6
	Outlier	35.5	13.5	87.9	13.4	65.5	11.8
	Range	51.7	16.1	83.8	9.7	69.2	6.4
	Random Row	49.7	22.2	86.2	9.2	70.1	10
	Random Attribute	44.4	20	82.6	12.7	65.3	12.4
Number of Clusters = 5	Inlier	56.6	20.6	80.4	18.6	71	9.7
	Outlier	50.5	25.5	82.2	13.3	67.3	13.4
	Range	50.7	16.9	85.9	10.7	69.9	6
	Random Row	65.1	19.1	76.5	17.7	74.1	6.8
	Random Attribute	64.1	17.5	73.5	21.7	71.8	7.6

7. Canberra Non-Invasive Data Attributes Results

7.1. Integrating K-Means clustering with Naïve Bayes

Canberra Non-Invasive Data Attributes							
Integrating Clustering with Naïve Bayes							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		69.1	12.3	73.1	16.4	73.3	6.1
Number of Clusters = 2	Inlier	68.3	11.6	73.7	15.7	73.4	4.6
	Outlier	69.5	12.2	70.1	15	72.7	4.2
	Range	69.5	12.2	70.1	15	72.7	4.2
	Random Row	69.7	11.9	73.4	15.8	74.3	4.4
	Random Attribute	69.7	11.9	71.8	16.1	73.4	4.2
Number of Clusters = 3	Inlier	68.8	11.1	73.6	16.7	73.7	5
	Outlier	68.8	11.1	73.6	16.7	73.7	5
	Range	64.1	11.2	70.1	14.6	70.2	4.4
	Random Row	66.3	12.1	72.1	16.1	72.2	4.4
	Random Attribute	66.5	9.7	71.4	15.3	71.5	5.2
Number of Clusters = 4	Inlier	64.6	9.4	72.9	15.9	71.5	4.9
	Outlier	71.8	11.3	72	16.3	73.8	4.7
	Range	64.6	9.4	72.9	15.9	71.5	4.9
	Random Row	67.3	9.6	74.1	16.3	73.3	4.9
	Random Attribute	68.7	10.4	73.6	16	73.4	4.5
Number of Clusters = 5	Inlier	64.6	9.4	72.9	15.9	71.5	4.9
	Outlier	71.8	11.3	72	16.3	73.8	4.7
	Range	64.6	9.4	72.9	15.9	71.5	4.9
	Random Row	67.5	9.7	71.2	15.5	71.6	4.9
	Random Attribute	67.5	9.7	71.2	15.5	71.6	4.9

7.2. Integrating K-Means clustering with K = 1 Nearest Neighbour

Canberra Non-Invasive Data Attributes							
Integrating Clustering with K = 1 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		55.7	11.3	77.2	15.7	67	11
Number of Clusters = 2	Inlier	48	22.3	82.6	15.5	70.8	9.2
	Outlier	51.3	15.1	80.6	17.6	69.8	11.4
	Range	48	22.3	82.6	15.5	70.8	9.2
	Random Row	47.7	21.9	81.5	18.3	69.7	11.3
	Random Attribute	52.1	15.2	82.8	14.3	71.6	9
Number of Clusters = 3	Inlier	56.4	14.3	59.8	19.2	60.5	10
	Outlier	52.7	23.4	79.1	14.7	69.8	7.4
	Range	56.4	14.3	59.8	19.2	60.5	10
	Random Row	44.9	23	80.8	11.9	65.9	10
	Random Attribute	52.9	20.6	73.8	20.5	66.4	10.4
Number of Clusters = 4	Inlier	30.2	30.3	83.5	13.6	58.8	14.6
	Outlier	62.5	22.4	71.1	19.5	69.2	9.3
	Range	30.7	30.5	84.3	12.9	59.4	14.6
	Random Row	72.4	15.4	72.2	5.8	71.7	10.3
	Random Attribute	54.2	21	73.1	16.2	65	12.8
Number of Clusters = 5	Inlier	51.6	22.2	66.9	22.5	63.5	9.8
	Outlier	63.3	18.3	64.9	23.5	66.1	12.5
	Range	30.2	30.3	83.5	13.6	58.8	14.6
	Random Row	62	32.2	57.3	31	62.5	13.3
	Random Attribute	54.7	22.6	71.5	13.7	65.2	12.4

7.3. Integrating K-Means clustering with K = 9 Nearest Neighbour

Canberra Non-Invasive Data Attributes							
Integrating Clustering with K = 9 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		14.1	9.7	98.2	5.4	60.9	13.8
Number of Clusters = 2	Inlier	51.3	16.2	84.9	11.5	72.7	7.3
	Outlier	45.9	19.7	86.6	10.6	71.7	7.8
	Range	51.3	16.2	84.9	11.5	72.7	7.3
	Random Row	47.8	18.4	85.8	10.4	71.9	6.9
	Random Attribute	50	15.5	85.8	10.8	72.4	7.8
Number of Clusters = 3	Inlier	49.6	18.7	83.1	11.8	70.4	8.1
	Outlier	43.8	18.8	83	17.5	66.3	7.8
	Range	49.6	18.7	83.1	11.8	70.4	8.1
	Random Row	50.8	15.1	75.2	17.4	65	9.7
	Random Attribute	44	19.4	85.6	14.2	67.6	7.3
Number of Clusters = 4	Inlier	43.7	14.7	85.4	10.7	67.3	8.3
	Outlier	38.2	15.8	85.7	15.5	65.8	9.6
	Range	44.2	14.4	83.7	9.5	66.6	8.5
	Random Row	40	16.4	81	18.7	64.7	8.5
	Random Attribute	45.1	21.8	81.7	13.1	65.1	10
Number of Clusters = 5	Inlier	44.3	23.6	83.2	19.2	69.3	9.4
	Outlier	61.7	20.3	70.8	16.1	65.2	9.5
	Range	43.7	14.7	85.4	10.7	67.3	8.3
	Random Row	52.6	26.5	79.1	21.6	68	8.6
	Random Attribute	54	14.8	73.3	18.9	64.6	8.2

7.4. Integrating K-Means clustering with K = 19 Nearest Neighbour

Canberra Non-Invasive Data Attributes							
Integrating Clustering with K = 19 Nearest Neighbour							
Data Mining Technique		**Sensitivity**		**Specificity**		**Accuracy**	
		Mean	**St Dev**	**Mean**	**St Dev**	**Mean**	**St Dev**
Without Clustering		10.4	8.4	100	0	59.4	14.6
Number of Clusters = 2	Inlier	51.3	16.2	84.9	11.5	72.7	7.3
	Outlier	45.9	19.7	86.6	10.6	71.7	7.8
	Range	51.3	16.2	84.9	11.5	72.7	7.3
	Random Row	48.7	19	85.7	11.4	72.3	7.6
	Random Attribute	48.1	19.2	85.2	11.4	71.7	7.9
Number of Clusters = 3	Inlier	41.2	16.2	86.5	10.7	68.9	7.6
	Outlier	41.5	18.9	88.2	12.4	66.2	11.3
	Range	41.2	16.2	86.5	10.7	68.9	7.6
	Random Row	49.5	14.4	85.2	13.1	69	8.9
	Random Attribute	42.1	18.4	83.9	18.4	67.6	7
Number of Clusters = 4	Inlier	52.8	16.9	81.9	16.6	69.7	7.9
	Outlier	34.9	13.4	89.3	13.3	65.9	11.6
	Range	52.8	16.9	80.7	15.6	69	8.2
	Random Row	36.7	19.8	85.6	13.6	63.9	11
	Random Attribute	40.4	14.2	86	14.4	66.1	10.9
Number of Clusters = 5	Inlier	46	14.7	84.4	16.9	67.4	9.5
	Outlier	43.2	22.6	86.2	11	65.7	12.9
	Range	52.8	16.9	81.9	16.6	69.7	7.9
	Random Row	46.4	20.6	82.4	12.5	65.6	12.7
	Random Attribute	47.5	14.8	80	14.1	65	11.5

[This Page is Left Blank Intentionally]

Appendix C

1. Canberra Non-Invasive Attributes Rules (Age, Sex, Systolic, Diastolic, and Rohrer Index (RI)) Decision Tree Rules

If SEX = 0, Age = 0, Systolic= 0, Diastolic = 3 => Then The Target Value Equals 0

If SEX = 0, Age = 0 , Systolic= 0 , Diastolic = 0 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 2 , RI = 3 => Then The Target Value Equals 1

If SEX = 0 , Age = 0 , Systolic= 2 , RI = 2 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 4 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 2 => Then The Target Value Equals 1

If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 0 , Diastolic = 1 , RI = 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 0 , Systolic= 2 , RI = 1 , Diastolic = 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 0 , Systolic= 2 , RI = 1 , Diastolic = 4 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 2 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 4 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 3 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 0 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 4 , Systolic= 4 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 4 , Systolic= 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 4 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 2 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 2 , Diastolic = 0 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 3 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 1 , Systolic= 1 , Diastolic = 0 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 0 , Systolic= 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 0 , Systolic= 4 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 0 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 2 , Systolic= 1 => Then The Target Value Equals 1

If SEX = 0 , Age = 1 , RI = 3 , Diastolic = 2 , Systolic= 2 => Then The Target Value Equals 0

If SEX = 0 , Age = 1 , RI = 4 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 0 , Age = 2 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 3 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 2 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 1 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 1 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 4 , Systolic= 4 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 4 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 0 , Age = 3 , RI = 4 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 3 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 0 , Systolic= 2 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 0 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 0 , Age = 3 , RI = 2 , Diastolic = 0 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 3 , RI = 4 , Systolic= 0 , Diastolic = 0 => Then The Target Value Equals 1

If SEX = 0 , Age = 3 , RI = 4 , Systolic= 0 , Diastolic = 1 => Then The Target Value Equals 0

If SEX = 0 , Age = 4 => Then The Target Value Equals 1

If SEX = 1 , Age = 1 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 0 => Then The Target Value Equals 1

If SEX = 1 , Age = 2 => Then The Target Value Equals 0

If SEX = 1 , Age = 0 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 0 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 2 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 0 => Then The Target Value Equals 1

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 4 => Then The Target Value Equals 1

If SEX = 1 , Age = 4 , Diastolic = 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 4 , Diastolic = 2 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 2 , RI = 2 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 2 , RI = 1 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 2 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 1 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 0 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 4 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 3 , Systolic= 0 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 3 , Systolic= 2 => Then The Target Value Equals 1

If SEX = 1 , Age = 4 , Diastolic = 4 , RI = 3 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 2 , RI = 2 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 2 , RI = 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 4 , Diastolic = 3 , Systolic= 2 , RI = 4 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 2 , RI = 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 4 , RI = 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 4 , RI = 2 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 4 , RI = 0 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 4 , Systolic= 4 , RI = 4 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 0 , Systolic= 3 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 0 , Systolic= 2 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 0 , Systolic= 1 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 4 , Systolic= 4 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 4 , Systolic= 3 => Then The Target Value Equals 1

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 3 , Systolic= 4 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 3 , Systolic= 1 => Then The Target Value Equals 0

If SEX = 1 , Age = 3 , Diastolic = 3 , RI = 3 , Systolic= 0 => Then The Target Value Equals 1

2. 2 Clusters K means Clustering Outlier Method Decision Tree Rules

2.1. The First Cluster the Age < 75

SEX = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 1, Diastolic = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 1, Diastolic = 4 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 0, Diastolic = 3 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 0 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 3 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 2, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 3 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 4 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 1 => Then the Target Value Equals 0

SEX = 0, Age = 0, Systolic= 0, Diastolic = 1, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 4 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 3 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 0 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 4, Systolic= 4 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 4, Systolic= 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 2, Diastolic = 4 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 2, Diastolic = 2 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 2, Diastolic = 0 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 1, Diastolic = 3 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 1, Diastolic = 1 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 1, Systolic= 1, Diastolic = 0 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 0, Systolic= 0 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 0, Systolic= 4 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 0, Systolic= 1 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 2, Systolic= 1 => Then the Target Value Equals 1

SEX = 0, Age = 1, Rohrer's Index = 3, Diastolic = 2, Systolic= 2 => Then the Target Value Equals 0

SEX = 0, Age = 1, Rohrer's Index = 4, Systolic= 3 => Then the Target Value Equals 0

SEX = 0, Age = 2 => Then the Target Value Equals 1

2.2. The Second Cluster Age >=75

SEX = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 0, Age = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 0, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 1 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 1 => Then the Target Value Equals 0

SEX = 1, Diastolic = 2, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 2, Age = 3, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 1, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 1, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 0 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 0 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 2, Rohrer's Index = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 4, Age = 4 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 4, Age = 3 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 3, Rohrer's Index = 1 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 4, Age = 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 4, Age = 3, Rohrer's Index = 4 => Then the Target Value Equals 0

SEX = 1, Diastolic = 3, Systolic= 4, Age = 3, Rohrer's Index = 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 3, Systolic= 4, Age = 3, Rohrer's Index = 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 1 => Then the Target Value Equals 1

SEX = 1, Diastolic = 4, Rohrer's Index = 2 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 0, Systolic= 4 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 0, Systolic= 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 0, Systolic= 1 => Then the Target Value Equals 1

SEX = 1, Diastolic = 4, Rohrer's Index = 3, Systolic= 0 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 3, Systolic= 2 => Then the Target Value Equals 1

SEX = 1, Diastolic = 4, Rohrer's Index = 3, Systolic= 3 => Then the Target Value Equals 0

SEX = 1, Diastolic = 4, Rohrer's Index = 4, Systolic= 4 => Then the Target Value Equals 1

SEX = 1, Diastolic = 4, Rohrer's Index = 4, Systolic= 3 => Then the Target Value Equals 0