Heart Disease Risk Evaluation Model Using Data Mining Techniques

Submitted in partial fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

by

Syed Immamul Ansarullah

Enrolment Number (A160920)

Under the Supervision of

Dr. Pradeep Kumar

Associate Professor & Head,

Department of CS & IT

September-2019

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

MAULANA AZAD NATIONAL URDU UNIVERSITY

(A Central University)

Gachibowli, Hyderabad, Telangana INDIA

मौलाना आज़ाद नेशनल उर्दू यूनिवर्सिटी مولانا آزاد نیشنل اُردو یونیورسٹی
MAULANA AZAD NATIONAL URDU UNIVERSITY
(A Central University established by an Act of Parliament in 1998)

Department of Computer Science and Information Technology

CERTIFICATE

It is certified that research work presented in the thesis entitled **"Heart Disease Risk Evaluation Model Using Data Mining Techniques"** in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** in **Computer Science** has been carried out under my guidance and supervision. He has fulfilled all the requirements for submission of the thesis, which to the best of my knowledge has reached the requisite standard. This thesis presented by him, to the best of my knowledge and belief, did not form the basis for the award of any other degree earlier.

Dr. Pradeep Kumar
Supervisor
Associate Professor and Head
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Gachibowli, Hyderabad, INDIA

DECLARATION

I, **Syed Immamul Ansarullah**, solemnly declare that the thesis entitled "**Heart Disease Risk Evaluation Model Using Data Mining Techniques"** is my original work. The study has been conducted under the guidance of **Dr. Pradeep Kumar** with the Department of Computer Science and Information Technology with **Maulana Azad National Urdu University** (A Central University), Gachibowli, Hyderabad, India. It is further declared that to the best of my knowledge and belief, it has not been submitted earlier for the award of any other degree, by anyone.

Dated: ___/___/2019

Syed Immamul Ansarullah
Research Scholar
Department of Computer Science & Information Technology
School of Technology
Maulana Azad National Urdu University
Gachibowli, Hyderabad, INDIA

ACKNOWLEDGEMENTS

To begin with, I would like to express my humble gratitude and prostrate before the ALMIGHTY ALLAH for HIS unlimited, unending BLESSINGS showered on me and bestowing me the courage, patience, and wisdom to correctly decipher right path from the wrong path in my life. This piece of work would not have been possible, but for HIS GUIDANCE and BLESSINGS.

I want to acknowledge the untiring and prompt help of my supervisor and Head, Department of Computer Science and Information Technology, Dr. Pradeep Kumar, who allowed me complete freedom to define and explore my directions in research. While this proved difficult and somewhat mystifying, to begin with, yet I highly value and appreciate the wisdom of his way. He was always ready to provide critical but constructive comments.

I want to express my gratitude to Prof. Abdul Wahid, Dean, School of Technology, Dr. Mudasir Manzoor Kirmani, Dr. Syed Mohsin Saif and Ziema Mushtaq, for their untiring support, constructive criticism and valuable suggestions as and when needed during this research. I also want to express my heartfelt gratitude to Mr. Mehboob Basha, a professional from software industry, for his technical guidance and valuable feedback during this research.

I am indebted and highly thankful to all faculty members and other supporting staff of the Department of CS & IT, School of Technology, MANUU for their consistent support. I also want to express my appreciation to all the research scholars of the Department of CS&IT and other individuals for providing their moral support during my study.

I express my sincere thanks to my beloved parents and other family members who provided me their unlimited affection, love, and support during the entire period of my study. Their consistent support has let me to success in my life and facilitated me in realizing the goal of my research study.

SYED IMMAMUL ANSARULLAH

ABSTRACT

Heart disease is emerging as the single most critical cause of death worldwide and is one of the costliest chronic conditions. Despite tremendous improvements, heart diseases continue to impose a major burden on patients and the healthcare systems. Regardless of damaging complications, heart disease is the most preventable and controllable disease; therefore, it is important to predict and diagnose it ahead of time. Even though the latest diagnostic and restorative advances have now become the standard of care but, these modalities are invasive, require physical samples, and are relatively costly which restrains their use in rural areas and at public-level screening evaluations.

After reviewing the literature and technical reports, it is found that disability and mortality rates by heart disease are rising. Stimulated by the increasing mortality rate incidents, a heart disease risk evaluation model is developed to help physicians in the early prediction and diagnosis. The significant non-invasive risk attributes like (Age, Systolic BP, Diastolic BP, BMI, Hereditary Factor, Smoking, Alcohol, and Physical Inactivity) are identified by the help of medical domain experts, and their reliability in the prediction of heart disease is investigated through different feature selection techniques. The enhancements of applying specific investigated techniques like Decision Tree, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Naive Bayes to the risk factors are tested. Further, to predict heart disease more accurately with minimum misdiagnosis rate, the hyperparameter optimization is performed.

The heart disease risk model is developed using the Jupyter notebook web application, and its performance is tested not only through medical domain measures like sensitivity, specificity, precision, misclassification rate, accuracy, AUROC, and cross-validation but also through the model performance measures like computational complexity and comprehensibility. Experimental results show that the random forest heart disease risk

evaluation model outperforms other heart disease risk models on the default parameter settings with the sensitivity of 85%, the specificity of 83%, accuracy of 84%, precision of 85%, miss classification rate of 15.1%, and AUROC score of 85%. To increase the efficiency and minimize the misdiagnosis rate the heart disease risk models are optimized. The hyperparameter optimized results show that the random forest model supersedes other optimized heart disease risk models with the increase in sensitivity to 87%, specificity to 84%, accuracy to 87%, precision to 86%, AUROC score to 86% and decrease in miss diagnosis rate to 13%. The heart disease risk feature combination subset [systolic BP, Diastolic BP, Age, BMI, and Heredity] showed the highest scores using random forest with sensitivity of 72%, specificity of 78% and accuracy of 78.9%. The simulation results of the developed heart disease risk evaluation model show that it outperforms the existing risk evaluation models with admirable predictive accuracy and prove its usefulness in the initial prediction and examination of heart disease.

The developed non-invasive heart disease risk evaluation model would help medical practitioners and would give patients a warning about the probable presence of heart disease even before he/she visits a hospital or goes for costly medical checkups. The model is usable where people do not have the advantage of integrated primary healthcare technologies for early prediction. The heart disease rules generated by the model are evaluated and validated by various medical domain experts. The extracted heart disease diagnostic rules are suggestive but not definitive as they are based on the specific ethnicity (Kashmir). Although this research develops a low-cost heart disease risk evaluation model using data mining techniques on novel non-invasive risk attributes combination; however, additional research is needed to understand newly identified discoveries about the disease.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABS	Australian Bureau of Statistics
AHA	American Heart Association
AIRS	Artificial Immune Recognition System
AUROC	Area Under the Receiver Operating Characteristics
BHF	British Heart Foundation
BMI	Body Mass Index
CAD	Coronary Artery Disease
CANFIS	Coactive Neuro Fuzzy Inference System
CHD	Coronary Heart Disease
CHF	Coronary Heart Failure
CRISP_DM	Cross-Industry Standard Process for Data Mining
CT	Computerized Tomography
CVD	Cardiovascular Disease
DALY	Disability Adjusted Life Years
DMP	Default Model Parameter
DMX	Data Mining Extension
DSS	Decision Support System
ECG	Electrocardiogram
EDA	Exploratory Data Analysis
EF	Ejection Fraction
ESCAP	Economic and Social Commission of Asia and the Pacific
ETC	Extra Tree Classifier
GBC	Gradient Boosting Classifier
GNI	Gross National Income

HBP	High Blood Pressure
HDREM	Heart Disease Risk Evaluation Model
HPT	Hyper-Parameter Tuning
HVD	Heart Valve Disease
IHDPS	Intelligent Heart Disease Prediction System
KDD	Knowledge Discovery from Data
KNN	K Nearest Neighbor
LAD	Left Anterior Descending
LCX	Left Circumflex
LMICs	Low-and-Middle-Income-Countries
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
OOB	Out-of-Bag
PCA	Principal Component Analysis
RCF	Right Coronary Artery
RFE	Recursive Feature Elimination
SAS	Statistical Analysis Software
SEMMA	Sample Explore Modify Model Assess
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
XGB	EXtreme Gradient Boosting
YLL	Years of Life Lost

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

List of Publications

1. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** "Heart Disease Prediction and Diagnosis by Finding the Correlation and Significance Among Different Risk Attributes Through Machine Learning Techniques", Journal of Advanced Research in Dynamical and Control Systems (JARDCS) , Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 1943023X)

2. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** " Performance and evaluation of heart disease risk assessment model using machine learning classification techniques", Journal of Advanced Research in Dynamical and Control Systems (JARDCS) , Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 1943023X)

3. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** " A Systematic Literature Review On Cardiovascular Disorder Identification Using Knowledge Mining and Machine Learning Methods", International Journal of Recent Technology and Engineering (IJRTE), Revised Manuscript Received on December 22, 2018,Vol. 10, 02-Special Issue, 2018 (**Scopus Indexed**) (ISSN Number: 2277-3878)

4. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** "Heart Disease Prognosis and Identification Using Different Machine Learning Techniques", at 6th International Conference on Recent Challenges in Engineering and Technology held on 24th to 25th Nov, 2018 at Nagpur, Maharashtra, INDIA.

5. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** "Heart Disease Risk Assessment Model Development Using Machine Learning Techniques", at 6th International Conference on Recent Challenges in Engineering and Technology held on 24th to 25th Nov, 2018 at Nagpur, Maharashtra, INDIA.

6. **Syed Immamul Ansarullah, Dr. Pradeep Kumar,** "Heart Disease Prediction using ensemble Classification Technique" at National conference on emerging trends and issues in information technology & communication, ETIITC-18, MANUU Hyderabad, INDIA.

7. **Syed Immamul Ansarullah, Pradeep Kumar, Abdul Wahid, Mudasir M Kirmani,** "Heart Disease Prediction System using Data Mining Techniques: A study" International Research Journal of Engineering and Technology (IRJET) (ISSN: 2395 -0056), Volume: 03 Issue: 08 | Aug-2016.

CHAPTER 1

Introduction

1.1 Background and Motivation

In recent years, the foremost reason for mortality and disability has shifted from infectious diseases to chronic diseases such as cancer, diabetes, and heart diseases. This move from infectious diseases to chronic diseases is called the 'epidemiologic transition.' At any given time, various countries of the world or even different regions within a country are at the epidemiologic transition [1]. The global health data demonstrate that heart diseases are the major source of death, with 17.3 million fatalities each year [2]. The death ratio is expected to rise to 23.6 million by the year 2030 [3]. Global health societies report that heart and respiratory diseases are the major sources of death in different countries [4] [5] [6].

Regardless of being among the most widespread chronic condition leading to a large percentage of disability and mortality across the globe, it is recognized as among the most avoidable and controllable diseases [7]. Initial identification of cardiac disorder victims can benefit from recuperating patients' health and diminishing the death ratio [8]. The World Health Organization (WHO) reported that initial prediction and treatment of heart disease decreases advancement to critical conditions and complexities [9]. Hence there is a critical demand for accurate systematic tools that can recognize patients at severe risk and provide knowledge for initial prediction [10] [11]. Different researchers use data mining techniques in health care industries to support health care professionals in diagnosing heart disease at its initial stages.

1.2 Heart Disease Overview

There have been various attempts to define disease but articulating a satisfactory definition of disease is surprisingly difficult. The disease can loosely be defined "as a condition of the body

or some part or organ of the body in which its functions are disrupted or deranged" [12]. Heart disease is a significant death reason in all low, middle, and high-income countries and is the foremost source of death for both men and women [13] [14]. Heart disease is an umbrella term for any disorder that affects the heart. Diseases under the heart disease umbrella include Coronary Artery Disease (CAD), arrhythmias, congenital heart defects, among others. Heart disease is often used interchangeably with cardiovascular disease (CVDs) [15] [16]. It occurs when coronary arteries become narrowed by a build-up of plaque. The plaque (atheroma) is an accumulation of cholesterol, fat, and other substances that result in the reduction of the blood supply to the heart over time [17] [18]. The pain perceived from such narrowing of blood vessels can lead to stroke, angina, or heart attack [19] [20]. The symptoms associated with heart disease vary from one type of heart disease to another, but generally include the common symptoms like discomfort in the arms, elbows, left shoulder, jaw or back, anxiety or distress in the center of the chest [21].

This chapter presents an outline of heart disease and its mortality rates. The global burden of heart disease, its recognition and diagnosis, and its risk evaluation are also explained. Afterward, in this chapter, the data mining overview, data mining as a step in knowledge discovery in data, data mining applications in health care, statement of the problem, research objectives and finally thesis outline are discussed.

1.2.1 Heart Disease Mortality Rates

Heart Disease is listed as an underlying reason for death in the general population of both the developing and advanced countries. WHO reported that heart disease death rates are uniformly disseminated among men and women with 3.8 and 3.4 million deaths [2]. Different health organizations report heart diseases are the foremost reason for mortality in different countries and continents. The British Heart Foundation (BHF) reported that heart disease causes 26% of all deaths in the UK [22]. The Australian Bureau of Statistics (ABS) presented

that heart and circulatory diseases are the primary reason for mortality in Australia, resulting in 33.7% of total deaths [23] [24] [25] [26]. The Economic and Social Commission of Asia and the Pacific (ESCAP 2010) report that 1/5th of Asian countries are afflicted with non-communicable diseases like cancer, heart diseases, and chronic respiratory diseases [27].

Statistical health reports of different regions show varying mortality rates by heart disease. The East *Asia and Pacific* region accounts for 35.2% of all deaths in the region half of those deaths resulted from ischemic heart disease [28] and in the Middle East and North Africa regions 47% of all deaths are due to heart disease [11]. Similarly, in *South Asia*, heart disease is the foremost reason for deaths accountable for 10.6% of total reported fatalities [29] and *Sub Saharan Africa in Western Africa* reported that 13% of all deaths were due to CVD [30] [31]. Similarly, *Eastern Europe and Central Asia*, *Latin America and the Caribbean* and developed countries like *Asia-Pacific, Australasia, Western Europe, and North America* regions showed that heart disease dominates among circulatory diseases [32] [33]. The global health statistical reports show that heart diseases are the contributing source of death in all countries regardless of their income.

1.2.2 Global Burden of Heart Disease

Heart disease is a significant global problem that has substantial consequences at many stages: individual mortality and disability, family suffering, and stunning economic expenses. The burdens of heart disease are diverse, which are explained as follows:

The British Heart Foundation (BHF) estimates that the cost of heart disease in the UK is 9 billion pounds per year. This economic cost includes costs associated with premature death and disability caused by heart diseases [24]. The annual expenditure of stroke and heart disorder in the United States is estimated at 312.6 billion dollars, and by 2035; the cost will skyrocket to 1.1 trillion dollars [10]. In China, annual expenses of heart diseases are more than $40 billion or approximately 4% of Gross National Income (GNI) [34]. South Africa spends,

2% to 3% of GNI on the cure of heart disease, which is approximately equivalent to a quarter of South African primary care costs [11]. Globally, the health care expenditures of heart disease for the year 2001 were estimated at $370 billion U.S dollars; which represents roughly 10% of total global health care expenses for that year [32] [35]. Similarly, in the Eastern European region, the hypertension costs were estimated at nearly 25% of total medical care costs [13] [14].

The American Heart Association (AHA) built up a methodology to predict future expenditures of medical management for High Blood Pressure, Coronary Artery Disease (CAD), Stroke, Angina and other forms of CVDs [35]. The proposed methodology showed that by the year 2030, 40.8% of the U.S. population is predicted to have some form of heart disease. Between the years 2013 and 2030, total healthcare costs of heart disease are forecasted to escalate from 320 to 818 U.S billion dollars. These reports specify that heart disease prevalence and expenditures are forecasted to rise considerably. These reports also show that heart disease epidemic has a considerable effect on the world and is one of the dominant health and development challenges in terms of both the human suffering they induce and the loss they impose on the socioeconomic foundation of countries [36] [37]. After recognizing the social, economic, and public health effects of heart disease, it is important to be diagnosed at early stages and delay in taking action will result in worsening of the situation [38].

1.2.3 Heart Disease Recognition and Diagnosis

Heart Disease is identified as among the most preventable and controllable disease. At least 80% of heart disease could be avoided by taking a healthy diet, doing regular physical activity, and restraining smoking [39] [40]. The general population dying from heart disease has some common key risk factors that are influenced by lifestyle [41]. The fundamental aim of heart disease detection and prevention is to maintain a strategic distance from the disease and to intrude on the improvement of the disease. Heart disease prevention activities are

performed at three different levels, like primary prevention, secondary prevention, and tertiary prevention [42]. Primary prevention is concerned with healthy people and how to reduce the risk factors that could result in a disease's occurrence [35]. Secondary prevention is concerned with the risk factors and early disease detection, increasing the probability of successful medical treatments. Tertiary prevention is concerned with the medical treatment of the disease and controlling the risk factors [17] [43].

Primary and secondary preventions are two important factors in controlling heart disease. Primary prevention plays a crucial role in controlling the effects of heart disease. Untimely diagnosis of heart disorder victims can help in recuperating patients' wellbeing and diminishing the death rate [17].

American Heart Association (AHA) reports that 11.4 million heart disease deaths in between the age group of 30-69 years and 15.9 million heart disease mortalities between individuals 70 years and older could be prevented in the year 2025 if objectives like cessation of tobacco and alcohol, decreasing salt intake, controlling obesity and lowering blood pressure are met [10]. The WHO reported that prior identification and treatment of heart disease are determined to decrease progression to critical and expensive illnesses and problems. So to achieve this objective of the initial prediction and treatment of the cardiac disorder, there is a fundamental requirement for a correct and systematic tool that classifies those victims who are at an elevated risk of heart condition [39].

Secondary level preventions can be identified by several heart disease tests such as ECG (Electrocardiogram), Coronary Angiography and Exercise Stress tests, etc. However, these physical examinations are expensive and need complicated types of equipment and a visit to a healthcare facility for risk detection. Unfortunately, it is supposed that 82% of the forthcoming rise in heart disorder fatality will take place in Lower-and-Middle-Income-Countries (LMICs) [12]. The economic circumstances of LMICs tend to limit the availability of the sophisticated equipment and medical facilities needed to meet the demand for heart

disease diagnosis. Because of insufficient resources, identifying low-cost prevention approaches is a primary preference. Using Community level screening tests to identify persons at higher risk is a well-incorporated secondary prevention procedure and would be proved cost-effective in LMICs [44].

1.2.4 Heart Disease Risk Evaluation

Even though a large percentage of heart diseases are controllable, but they continue to advance because preventive methods are insufficient. As the magnitude of heart disease continues to accelerate globally, the need for health maintenance and screening tests in pharmacies is required to help to enhance patients' health care. Public-level screening tests can help in untimely prognosis and examination of heart disease [28]. Although the newer diagnostic technologies are utilized for the initial prognosis of heart condition, however, these tests are costly and can't be utilized as community-level screening tests. Hence there is a need to find less expensive tests to be conducted as a community-level screening test [36].

If we are to reduce the rising burden of heart disease, it is critical to recognize its underlying risk factors that drag the world to the unfavourable situation [45]. It is widely accepted that the risk factors like Age, harmful intake of alcohol, unhealthy diet, smoking, and physical inactivity are the significant risk features of heart disease [46] [47] and continuing exposure to these risk features results in raised Hypertension [48], diabetes [49], Dyslipidaemia [50], Obesity [51] and Stroke [52].

Many risk prediction tools are widely available like The Reynolds Risk Score [53], Framingham Heart Disease Risk Evaluation Tool [54] [55], the Australian Absolute Cardiovascular Risk Calculator [56], etc, however, all tools result in risk predictions that are less appropriate and also require prior blood sample examinations, an invasive and relatively costly process, which reduces their usability in other than medical settings. Hence, there is a need to simplify these tests and use only non-invasive risk attributes for the early prognosis of

heart disease victims. The use of entirely non-invasive features in heart disease risk calculation has not been examined before. Furthermore, if non-invasive features demonstrate substantial achievement in the risk assessment of cardiac disorder cases, then this analysis would be of immense benefit to the initial prognosis of heart disease.

1.3 Data Mining Overview

Several researchers agreed that data mining is a multidisciplinary area and can be defined from different perspectives. According to [57], "Data mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." Researchers like [58] added that "data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases." Similarly, [59] describes data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database." Researchers [60] defined data mining "as an attempt to discover hidden patterns where these patterns are difficult to detect with traditional statistical methods." These definitions allow us to conclude that data mining is the extraction of useful knowledge from a tremendous quantity of raw data to identify deeply hidden interesting and valid patterns, relationships, and knowledge.

1.3.1 Data Mining as a Step in Knowledge Discovery from Data

Researchers define data mining in various perspectives some researchers' describe data mining as a synonym for Knowledge Discovery from Data (KDD), however other researchers characterize data mining as simply a fundamental step in KDD. Below given figure1.1 shows the knowledge discovery in databases as an iterative series of following steps [59]. Data cleansing and integration is the first step in KDD that is applied to remove the noise and

correct inconsistencies in data, while data integration combines data from heterogeneous sources into a coherent data store.

The second step is data selection and transformation in which the suitable data is extracted from the data store for the analysis purpose and then consolidated and transformed into forms appropriate for mining purposes by performing the summarization or aggregation techniques.

The third step is data mining, where insightful methods are applied to mine data patterns. The fourth step is pattern evaluation that identifies the most appealing patterns signifying knowledge based on interestingness measures.

The final step in the KDD process is knowledge representation, where visualization and knowledge representation methods are utilized to provide mined knowledge to users [59]. Data mining is predicted to be the most innovative advances of the next decade because it becomes more widespread every day in a large range of applications [61].

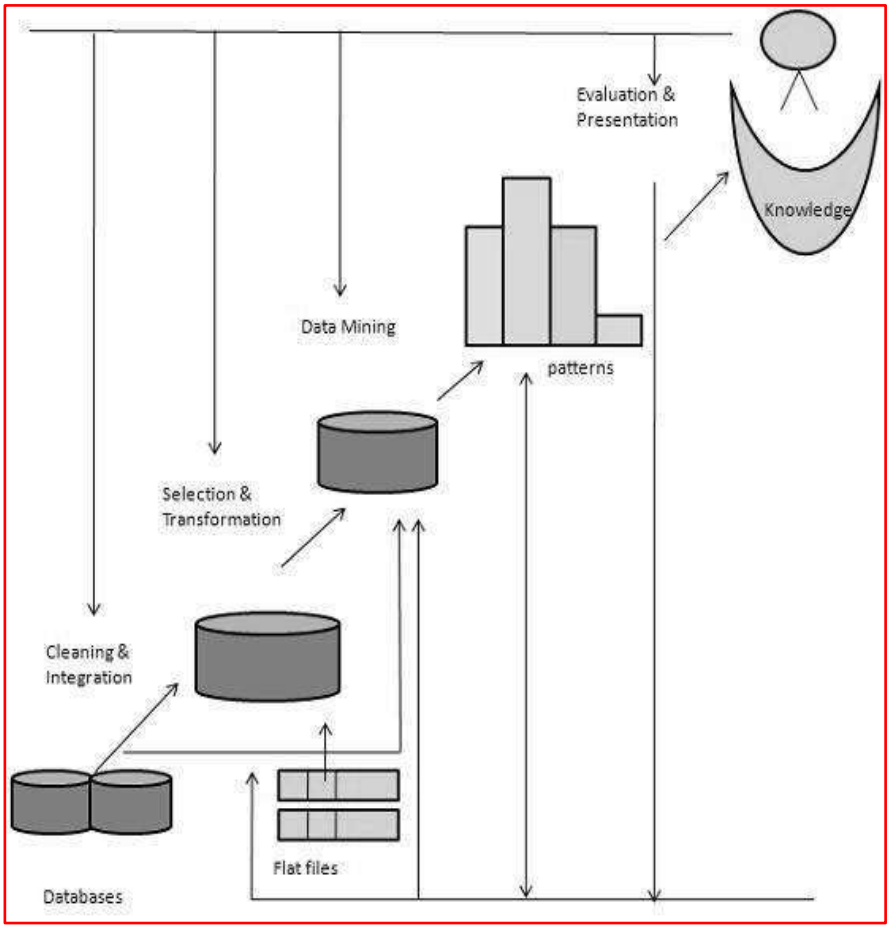

Figure 1.1 The Process of Knowledge Discovery in Data [Han and Kamber]

1.3.2 Applications of Data Mining in Healthcare

Because of the extremely application-driven subject, data mining has seen extraordinary achievements in various disciplines. Data mining is growing successfully in the extensive scope of applications such as web mining, business intelligence, load forecasting, diagnosis, marketing and sales, oil refinement and screening images [62] [63] [64]. Data mining in health care is a rising area of high significance for providing prediction and diagnosis [65] [66] [67]. Data mining applications in health care incorporate the investigation of health care institutes for improved health strategy-formulation and avoidance of hospital failures, initial prevention, and detection of diseases and unnecessary hospital deaths, more value for money and cost savings in health care delivery, and recognition of fraudulent insurance claims [68].

Healthcare data mining has the extraordinary capability for investigating the concealed patterns in the datasets of the medical domain. The most important challenge presented by health and medicine is to build up a technology that can present trusted hypotheses based on measures that can be relied upon in medical health research and applied in a clinical environment [69]. The discovery of various theoretical implications from training datasets is a difficult process; even the best experts are overwhelmed by the accumulated data. Hence data mining is used to benefit health professionals in decision making [70].

Researchers are using data mining techniques in diagnosing numerous diseases such as diabetes [71], Stroke [72], Cancer [73], and Heart Disease [74]. The future of health care may strongly rely on using data mining to reduce health care expenditures, determine treatment strategies and best practices, evaluate efficiency, detect fraudulent insurance and medical claims, and ultimately, improve the standard of patient care.

1.4 Statement of the Problem

Heart disease is the most widespread chronic condition leading to a tremendous rate of mortality all over the world; however, its early prediction can benefit in recouping patients'

health and decrease the fatality percentage. Based on the death rates, disability, and cost, there is an essential requirement for an accurate systematic tool for the diagnosis of heart disease. To diagnose heart disease at its early stages with low-cost, the screening tools are used, but those tools require prior blood sampling, which is an invasive and costly process. Consequently, there is a requirement to rationalize the risk features and build an accurate data mining model that can be utilized for public-standard screening to recognize patients at a high risk of heart disease and produce knowledge to facilitate initial intervention and enhance patient's health.

1.5 Objectives

The objectives of this research are as follows:

i. Review the literature about heart disease prediction using data mining techniques to fill the identified research gaps.

ii. To study and analyze the non-invasive heart disease risk features and their importance in predicting the disease.

iii. Develop the heart disease risk evaluation model to recognize patients at elevated risk of heart disease and provide information to enable early intervention.

iv. To analyze the heart disease risk evaluation model's efficiency using various metrics.

v. To validate the proposed heart disease risk evaluation model through benchmark heart disease datasets, with domain experts' knowledge.

1.6 Thesis Outline

The thesis is divided into six chapters:

Chapter 1 is an introductory chapter that discusses the background of the research work. It starts with an overview of heart disease, its mortality rates, and the global burden. This chapter discusses the recognition of heart disease, heart disease diagnosis and risk evaluation.

This chapter also explains data mining applications in healthcare systems, statement of the problem, and objectives of the research work.

Chapter 2 discusses a detailed systematic literature review of heart disease prediction and detection using different data mining techniques.

Chapter 3 discusses the feature selection methods which are applied to search the significant non-invasive subset of risk attributes for the early prediction of heart disease. This chapter applies different data mining techniques like Random Forest, Naive Bayes, K Nearest Neighbor, Support Vector Machine and Decision Tree to see whether these techniques will help medical professionals in early prediction of disease which would result in a reduction to severe and costly illness and complications. The developed risk models' performance is measured through medical and model metrics like the confusion matrix, AUROC Curve, model complexity, model comprehensibility, etc. Finally, the chapter is concluded by discussing the significance of non-invasive heart disease risk attributes.

Chapter 4 describes the data mining tasks and techniques. In this chapter, the data mining classification techniques like Random Forest, Decision Tree, K Nearest Neighbor, Support Vector Machine and Naive Bayes techniques are discussed to predict and detect heart disease at its initial stages. This chapter discusses the procedure of how the heart disease risk evaluation model is developed on the Jupyter Notebook Web Application. The performance measurement of the developed model is estimated through different evaluation techniques. The experimental results of the proposed techniques are explained, and the comparison among them is discussed. Experimental outcomes show the Random Forest heart disease model outperforms other proposed and prevailing models. Finally, the chapter is concluded by discussing the bias and variance errors of the predictive model.

Chapter 5 presents an introduction to hyperparameter optimization and its techniques. It discusses how to optimize the proposed model to improve model accuracy. In this chapter, the

comparison among hyperparameter optimized models is also performed, and the performance evaluation comparison between heart disease risk models with the default and hyperparameters is also discussed. The optimal sets of rules generated for heart disease risk assessment are explained. This chapter describes the significance of different combinations of risk attributes for cardiac disorder diagnosis. Finally, the evaluation components of the Heart disease expert system are discussed, and the developed risk evaluation model is presented.

Chapter 6 describes the conclusion and future work of the research. This chapter also summarizes the key questions like important attributes in the heart disease risk assessment, the importance of non-invasive risk features for the evaluation of cardiac disease. Finally, the research limitations are discussed along with future work.

CHAPTER 2

Literature Review

This chapter outlines the seminal contributions made by various researchers for the development of heart disease risk assessment models using different data mining techniques. This chapter also highlights the importance of early prognosis and identification of heart disease. Finally, the research gaps found in the prevailing literature are discussed.

2.1 Predicting Heart Disease Using Different Data Mining Techniques

In recent times, researchers made decisive contributions to heart disease disorder identification using various data mining tasks and techniques. Many researchers build heart disease risk models using the divergent data sets, different machine learning algorithms, various data mining approaches, and numerous tools which are explained as follows:

2.1.1 Predicting Heart Disease Using Supervised Tasks

In predictive tasks, the objective is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the *target or dependent variable*, while the attributes used for making the prediction are known as the *explanatory or independent variables* [59]. Following researchers use the supervised data mining techniques to predict the cardiac disorder:

Colombet et al. (2000) used Classification and Regression Tree (CART) and Multilayer Perceptron (MLP) data mining algorithms to diagnose cardiovascular risk from a real dataset consisting of 15,444 instances [75]. The dataset is randomly divided into a training set of (10,296) instances and a test set of (5,148) instances. Researchers' evaluate the developed risk model's performance using different performance measurements based on the test dataset. The implementation criteria, explicative criteria, and discriminative performance criteria are

considered based on ROC analysis. Researchers analyze the ROC curve using the ROCKIT software. The experimental results demonstrate that the MLP predictive ability in diagnosing heart disease is higher than the CART.

Yan et al. (2003) designed a predictive cardiac disorder risk decision support system using a Multilayer Perceptron Neural Network [76]. The proposed decision support system is trained using a back-propagation algorithm augmented with the momentum term, the adaptive learning rate, and the forgetting mechanics. A total of 352 medical records have been used to train and test the system. Researchers evaluate the performance of the developed risk evaluation model using three different evaluation methods like cross-validation, holdout and bootstrapping. The experimental outcomes demonstrate that the heart disease risk evaluation model based on MLP has great capability to classify five different types of cardiac disorders with strong predictive approximation.

Noh et al. (2006) used data mining classification techniques to diagnose Coronary Artery Disease (CAD) under the framework of ECG patterns and clinical investigations by analyzing Heart Rate Valve (HRV) from ECG [77]. The proposed method is an associative classifier based on the efficient FP-growth method by using a cohesion measure for pruning redundant rules. The dataset of 670 patients is used for the associative classifier, which utilizes multiple rules and pruning, and biased confidence (or cohesion measure). Using medical domain experts' knowledge researchers categorize patients as affected by CAD or as normal based on the stenosis of the luminal narrowing. Researchers used the stratified 10 fold cross-validation on the dataset and evaluated the performance of the developed model using different performance measurements like precision, F-measure, Recall, and Root Mean Square Error (RMSE). Results demonstrate that the proposed classifier outperformed other classification algorithms for the diagnosis of CAD.

Palaniappan and Awang (2008) developed a risk evaluation model using Decision Tree, Neural Network, and Naive Bayes data mining techniques [78]. The developed risk evaluation

model can extract interesting hidden patterns related to cardiac disorder and can answer intricate questions in which existing risk assessment tools fail. Researchers' obtain a small dataset from the Cleveland heart disease database consisting of a total of 909 instances with 15 medical risk features. The risk evaluation model is developed on the .NET platform, and for communication, the Data Mining Extension (DME) query language and functions are used. The performance of the developed risk evaluation model is checked through, the lift chart and classification matrix are used to check which model gave the highest percentage of correct predictions for diagnosing heart disease patients. From experimental results, it is found that the Naive Bayes risk evaluation model outperforms the Neural Network and Decision Tree models.

Shouman, Turner, and Stocker (2011) developed a novel classification model for the early prediction of heart disease patients using the decision tree technique [79]. While developing the risk evaluation model researchers integrate multiple classifiers voting technique with different multi-interval discretization methods like (equal frequency, chi-merge, equal width, and entropy) using different decision tree variants like (Gini Index, Gain Ratio and Information Gain). The heart disease decision rules are extracted, and then the efficient sets of rules are selected by using the reduced error pruning technique. The developed model obtained the highest accuracy of 79.1% on the configuration of equal width discretization information gain decision tree without voting. After applying the voting technique, the equal frequency discretization gain ratio achieved the highest accuracy of 84.1%.

Rani (2011) developed a robotic and reliable risk assessment model using the Neural Network algorithm [80]. The risk model extracts the proficient and reliable classification rules from the Cleveland heart disease dataset. The model is trained using Feed Forward Neural Network and Back-Propagation learning algorithm with momentum and variable learning rate. The performance of the network is analyzed through, huge test data, which is given as input to the network. Parallelism is implemented at each neuron in all the hidden and output layers to

speed up the learning process. The experimental results proved that the Neural Network technique provides satisfactory results for the classification task.

Kumari and Godara (2011) analyzed RIPPER, Decision Tree, Artificial Neural Networks, and Support Vector Machine algorithms on cardiovascular disease datasets [81]. Researchers use the Cleveland heart disease dataset having 303 records with 14 attributes and test it on the WEKA tool. An attribute selection algorithm is applied to pre-process the dataset, which resulted in 296 total records for data mining classification technique. Researchers use various performance measures like Accuracy, Specificity, Sensitivity, Precision, Error Rate, and AUROC to check the performance of the model. Results show that SVM model outperforms other classification algorithms in all parameters Sensitivity, Specificity, Accuracy, and Error Rate. Also on ROC space point, an SVM model is closer to perfect point (0.1) than other models, which shows SVM to be the best predictor of Heart disease.

Shouman, Turner, and Stocker (2012) developed a K Nearest Neighbor risk evaluation model using the Cleveland heart disease dataset to detect cardiac disorder patients well in advance with optimal accuracy [82]. Initially, the value of K is set to 1 and then iteratively incremented till the upper limit of 13 and when k=7 the highest accuracy and specificity of 97.4% and 99% are achieved respectively. In this work, researchers discovered that applying the voting technique did not show any progress in the precision even after estimating different values of parameter k.

Chaurasia (2013) developed a novel cardiac disorder detection model using different classification algorithms ID3, CART, and Decision Table [83]. The developed predictive risk model is trained and tested on the standard Cleveland Heart Disease dataset that consists of 11 significant disease risk features. After analyzing the heart disease dataset, it is executed on the WEKA tool. The performance of the risk evaluation model is checked through various performance measures. Researchers' used 10-fold cross-validation to minimize bias and to improve efficiency. To better understand the importance of each individual input variable for

heart disease prediction, Chi-Square, Info-Gain, and Gain Ratio tests were conducted. Experimental results demonstrate that CART outperforms Decision Table and ID3 classifiers with the highest accuracy and minimum error rate; however, the developed models time complexity increases.

Al-Milli (2013) developed a cardiac disorder risk prediction model using the back-Propagation Neural Network algorithm [84]. The researcher uses the Cleveland benchmark dataset consisting of 13 medical attributes and MATLAB tool to build the risk model. After parameter settings, the experiments were run 10000 iterations. In the MATLAB tool, the risk evaluation model is executed 11 times; however, each run provided varying results. From the experimental results, it is found that when the model was executed 10^{th} time, the highest variance from the training and testing process was achieved. The researcher used the box plot representation to illustrate the distribution of solution quality for training and testing datasets. In both cases, there is less dispersion of the output data, which demonstrates that it is a robust algorithm. The experiments conducted showed optimal performance compared to similar approaches of state of the art.

Masethe Hlaudi and Masethe Mosima (2014) designed a model to predict and classify heart attacks by using J48, Naive Bayes, Simple CART, REPTREE, and Bayes Net data mining algorithms [85]. The patient dataset used to build the heart disease model is collected from the health care professionals in South Africa that have 490 instances and 11 attributes. They use the WEKA tool for the prediction of heart disease. Researchers applied the stratified 10-fold cross-validation on the dataset for estimating the unbiased results. From the experimental results, it is found that the results did not provide any remarkable differences in heart disease prediction when different classification algorithms were applied.

Ngueilbaye, Lei, and Wang (2016) used Naive Bayes and Support Vector Machine classification algorithms for the initial prediction of cardiac disorder patients [86]. To check the performance of the applied classifiers, researchers used various measures like probability

and classification accuracy. Experimental results show that the Naive Bayes algorithm outperforms the SVM model. The small dataset of 315 instances was collected from different hospital databases.

2.1.2 Predicting Heart Disease Using Unsupervised Tasks

In Descriptive Tasks, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory and frequently require post-processing techniques to validate and explain the results [59]. The following researchers used different unsupervised data mining techniques to predict heart disease at its earliest:

Nguyen and Davis (2007) proposed KMIX (an improved form of K-Means Clustering) algorithm for the early prediction of cardiovascular disease [87]. The noisy heart disease dataset is cleaned through pre-processing techniques. The cleaned dataset having 341 instances and 19 vital heart disease features are then used to develop the model. The continuous-valued numerical attributes are transformed into the range [0,1] using the linear transformation method, and Boolean data is transformed into a discrete number text form. Using the WEKA tool, Sensitivity and Specificity are used to check the performance of the algorithm. Experimental results show that KMIX outperformed the K-Means algorithm with a sensitivity of 0.25 and a specificity of 0.89. Hence it can be seen that the performance of the KMIX clustering algorithm is appropriate for diagnosing the CVD patients.

Patil and Kumaraswamy (2009) developed an efficient approach for cardiac disorder prediction using a K-Means clustering algorithm [88]. The risk model is developed in Java on the heart disease dataset, which is obtained from the UCI machine learning repository. Researchers used the Maximal Frequent Itemset algorithm to mine the frequent patterns that are most appropriate for heart disease. After deriving the frequent heart disease risk patterns, the weights assigned to the patterns are calculated, and the pattern with a significant weight

higher than a predefined threshold value is used for the early detection of heart disease patients. The significant weighted patterns are pruned and verified by medical domain experts.

Shouman, Turner, and Stocker (2012a) demonstrated the efficiency of the K- Means Clustering technique in improving Naïve Bayes for the initial prediction of cardiac disorder patients [89]. Due to the inbuilt limitation of the Naive Bayes algorithm to deal with continuous attributes, the equal frequency discretization method is used to convert them into the discrete ones. To get the unbiased results, researchers divided the dataset into training and testing sets using the stratified 10 fold cross-validation. Researchers use the Cleveland heart disease dataset that consists of 297 instances with 13 medical risk attributes. Different methods of Initial Centroid Selection like Range, Inlier, Outlier, Random Attribute Values, and Random Row Methods are applied for the prediction of heart disease patients.

Experimental results demonstrate that integrating K-Means Clustering with Naïve Bayes using different Initial Centroid Selection enhanced the Naïve Bayes accuracy in predicting heart disease patients. Results show that the Random Attribute and Random Row Methods achieved higher accuracy than Inlier, Outlier, and Range methods with two clusters. The best accuracy achieved is by two clusters Random Row Initial Centroid selection method. However, increasing the number of clusters of the Random attributes and Random Row Initial Centroid selection methods did not show any enhancement in their accuracy in the diagnosis of heart disease patients.

2.1.3 Predicting Heart Disease Using Hybrid Data Mining Techniques

A combination of two or more methodologies within a design of single system results in a hybrid system. Hybrid systems extract the best from all methodologies and provide an optimal solution for the disease diagnosis. The following researchers use different hybrid data mining techniques to predict the heart disease at its initial stages:

Parthiban and Subramanian (2007) developed an intelligent heart disease prediction model on the Cleveland heart disease dataset using Coactive Neuro-Fuzzy Inference System (CANFIS) [90]. The proposed CANFIS model integrates adaptable fuzzy inputs with a modular Neural Network to rapidly and accurately approximate complex functions. To improve the learning capability of the CANFIS model, the Genetic algorithm is used to search for the best number of Member Function for each input and optimization of control parameters. The genetic algorithm combines selection, crossover, and mutation operators to find the best solution to a problem by searching until the specified criterion is met. The developed heart disease model acquired very less MSE (Mean Square Error).

Polat, Sahan, and Gunes (2007) proposed a novel system for the early prediction of cardiac disorder using the Artificial Immune Recognition System (AIRS) classifier with a fuzzy resource allocation mechanism [91]. Researchers firstly applied the K-NN based weighting process to the heart disease dataset and scaled the weights in the range of (0 and 1). Once this preprocessing step is completed, the Fuzzy-AIRS algorithm is applied to the weighted heart disease dataset. Researchers obtain the heart disease dataset (containing 13 attributes and 270 instances) from the UCI Machine Learning Database. In applications of the system, the heart disease dataset is classified for different values of k like 10, 15, and 20, which is used as the K-NN pre-processing step. The highest classification accuracy is reached when the value of k is 15. The obtained classification accuracy result of the proposed system is 87%, and it is very promising concerning the other classification applications. The results strongly suggest that K-NN weighted pre-processing and fuzzy resource allocation mechanism with AIRS can assist in the prediction of cardiac arrhythmias.

Tsipouras et al. (2008) predicted the CAD by developing the Fuzzy rule-based model [92]. Researchers used the heart disease dataset, which consists of demographic, historical, and laboratory data of 199 instances and 19 attributes. The random stratified 10-fold cross-validation is applied to the dataset to get the unbiased results. The performance of the developed model is checked through different performance evaluation measures.

Experimental results show the Sensitivity and Specificity of 62% and 54%, on decision tree rules. The average Sensitivity and Specificity increase to 80% and 65%, respectively, when the Fuzzification and optimization stages are used.

Tu, Shin, and Shin (2009) developed a predictive cardiac disorder risk model using the Bagging with Naive Bayes, C4.5, and Bagging with C4.5 classifiers. Researchers use live datasets collected from patients with heart disease. The bagging algorithm tries to neutralize the instability of learning techniques by simulating the process using a given training set [93]. Instead of sampling a new training dataset each time, the original training data is modified by deleting some instances and replicating others. Researchers carried out three different experiments on the WEKA tool. Experiment1 used the Decision tree algorithm, experiment2 used the Bagging with decision tree with a reduced error pruning option, and experiment3 used the Bagging with Naive Bayes algorithm. For each experiment, 10-fold cross-validation is used to minimize the bias produced by random sampling of the training and test data samples. Experimental results demonstrate that the Precision, Recall, and F-measure of bagging with Naïve Bayes, showed the optimal performance among the tested methods.

Das, Turkoglu, and Sengur (2009) developed an efficient Neural Networks ensemble model to predict heart disease at its earliest [94]. Researchers used the ensemble component to create new models by combining the posterior probabilities from multiple predecessor models. This newly developed risk model is then used to score new unseen data. Researchers use the SAS enterprise miner 5.2 to create a Neural Networks ensemble-based methodology for early prediction of heart disease. On the Cleveland, heart disease database having 297 instances and 14 attributes, 89.01% classification accuracy is obtained from the experiments. Researchers use the variable selection component to decrease the number of inputs by configuring the condition of the input variables as rejected.

Anbarasi, Anupriya, and Iyengar (2010) developed a risk model to accurately predict the existence of heart disease with a fewer number of attributes [95]. The genetic algorithm is

incorporated to determine the attributes which contribute more towards the diagnosis of heart ailments. Three classifiers like Naive Bayes, Classification by Clustering and Decision Tree, are used to diagnose heart disease patients. Observations exhibit that the Decision Tree technique outperforms the other two data mining techniques after incorporating feature subset-selection with relatively high model construction time. Experiments were conducted with the WEKA tool on a dataset of 909 instances. The genetic search for an optimal set of attributes starts with zero attributes, an initial population, and randomly generated rules. The generation of the new population continues until it evolves a population where every rule is satisfied by the population. All attributes are made categorical, and inconsistencies are resolved for simplicity. With 0.6 cross over probability and 0.033 mutation probabilities, the Genetic Search resulted in 6 attributes that contribute more towards the diagnosis of the cardiac disease.

Adeli and Neshat (2010) developed a model on benchmark Cleveland heart disease dataset consisting of 303 instances having 12 attributes using a fuzzy expert system [96]. Membership function of all the 11 input variables and one output variable is designed using an inference mechanism. Researchers use the Mamdani approach for Fuzzification, and the defuzzification process Centroid method was incorporated. In the fuzzy inference system, the quality of results depends on the fuzzy rules. The proposed system generated 44 rules and is best in comparison with the results of the other rule bases. The validity degree (k) for each rule is generated, and for aggregation of rules, the maximum validity degree is calculated with K=max (k1, K2... k44). The fuzzy expert system dealing with the diagnosis has been implemented, and the experimental results showed that the system did quite better than non-experts.

Aqueel and Hannan (2012) built a heart disease diagnosis model using Support Vector Machine (SVM), Genetic Algorithm, Rough Set Theory, Association Rules, and Neural Network algorithms. Researchers conducted numerous experiments on a dataset consisting of

909 records and 13 attributes [97]. To check the performance of these algorithms, researchers conducted experiments on the WEKA tool. Experimental results show that the developed decision tree risk model showed high predictive capability compared to other techniques after feature selection techniques are applied; however, the time complexity of the risk model increases.

Alizadehsani et al. (2013) used classification data mining algorithms to predict CAD. The present study employed C4.5 classification and Bagging classifiers to investigate the Lab and ECG data to identify the stenosis of each artery Left Anterior Descending (LAD), Left Circumflex (LCX), and Right Coronary Artery (RCA) separately [98]. The data were gathered from 303 random visitors to Rajaie Cardiovascular, Medical, and Research Center, Tehran, Iran. The accuracy in predicting the LAD stenosis was attained via Feature Selection. This research used the default setting of the Rapid Miner tool and obtained the Accuracy, Sensitivity, and Specificity of the algorithms. The Gini index and information gain were used to select the most important features. Furthermore, the use of features selected based on information gain enhanced the accuracy of the LAD stenosis diagnosis to 79.54%. The results indicate that EF (Ejection Fraction), Age, lymph, and HTN were among the ten most effective features on the stenosis of all of the arteries.

Jabbar, Deekshatulu, and Chandra (2015) proposed a new approach that combines K-Nearest Neighbour and the Genetic classifier for effective classification to predict cardiac disorder victims [99]. Genetic search is applied as a goodness measure to prune redundant and irrelevant attributes and to grade the features which contribute more towards classification. Least graded features are excluded, and the classifier is designed based on the classified features. The performance of the developed method is verified with six medical and one non-medical dataset. Among these seven heart disease datasets, one dataset is collected from different hospitals in Andhra Pradesh, INDIA, and the rest of the sex datasets are obtained from the UCI machine learning repository. Experimental results demonstrate that the classifier increases the efficiency of heart disease diagnosis.

Amin, Agarwal, and Beg (2013) developed a hybrid heart disease assessment model for its initial prediction based on significant risk attributes using the neural network and genetic algorithm [100]. The survey data collected by the American Heart Association is used, which consists of 12 important risk factors and 50 instances. After data pre-processing, neural network weights were initialized with the 'configure' function available in MATLAB. Then these configured weights were passed to the Genetic algorithm for optimization according to the fitness function. The ensemble risk model built-in MATLAB tool attained an optimal accuracy on training as well as on validation dataset with Least Mean Square Error of 0.034683 after 12 epochs.

Chaurasia and Pal (2014) developed a risk model on the Hungarian dataset using Naive Bayes, Decision Tree, and Bagging algorithms to predict heart disease with optimal accuracy [101]. Researchers used the 10 fold cross-validation to measure the unbiased estimate of the prediction models. The heart disease model is trained and tested on the WEKA tool. The experimental results show that the bagging algorithm does better than Naive Bayes and J48 algorithms with the highest accuracy of 85.03%.

Srinivas, Rao, and Govardhan (2014) proposed a new classifier by combining rough set theory with the fuzzy set for heart disease diagnosis [102]. Fuzzy Base Rules are generated using rough set theory, and the prediction is carried out by a fuzzy classifier. To get valid fuzzy rules; core analysis is done to identify the relevant attributes from the rough set theory after forming the indiscernibility matrix. Then, the fuzzy system is designed with the help of fuzzy rules and membership functions so that the prediction can be carried out within the fuzzy system designed. The proposed system is implemented using MATLAB 7.11, and the presence of heart disease is identified by inputting the data to the fuzzy system. The classifier experiments with the three widely applied datasets, namely, Cleveland, Hungarian, and Switzerland, which are downloaded from the UCI machine learning repository site. From the results, researchers' ensure that the proposed rough-fuzzy classifier outperformed the previous

approaches by achieving the accuracy of 80% on Switzerland's heart disease dataset and 42% on Hungarian heart disease dataset.

Dewan and Sharma (2015) developed a competent ensemble risk prediction model using a Genetic algorithm with the Back Propagation method [103]. The developed model is executed on a dataset consisting of 303 records for training and 270 for testing purposes. In pre-processing, the most common technique of WEKA tool, i.e., Replace missing value filter is used. To solve the drawback of being stuck in local minima, the best optimizer, i.e., Genetic Algorithm which uses the phenomena of mutation and crossover above various generations is embedded in the model. The weights which are used for Back Propagation are optimized first and then given as input to the network to get better results. Experimental results show that Neural Network is best among all the classification techniques to predict or classify non-linear data.

Sumana and Santhanam (2015) proposed a hybrid model for early heart disease diagnosis. After data cleaning, best-first-search and feature selection techniques were incorporated in cascaded fashion to get the relevant features for heart disease [104]. The resultant dataset is clustered using the K-Means algorithm, and the correctly clustered samples are trained with 12 distinct classifiers to develop the final model using stratified 10-fold cross-validation. The proposed model is evaluated using the WEKA tool on five other binary class medical datasets collected from the UCI machine learning repository to test the accuracy and time complexity of the classifiers. The observed outcomes demonstrate that regardless of the datasets and algorithms, the developed ensemble model enhanced classification accuracy and above on five different medical datasets with all 12 classifiers.

Beena, Rajinikanth, and Viswanadha (2016) selected the significant heart disease attributes by a combination of Computerized Feature Selection methods and Medical Features to increase the prediction accuracy and decision making for cardiac disorder diagnosis [105]. The default multi-class classification mode of the Cleveland heart disease dataset has been

converted into a binary classification form. Researchers used the Sequential Minimal Optimization algorithm to develop the risk model using MATLAB tool. It is found that the accuracy of the Feature Selection Method increases by controlling the discrete features; however, the model time complexity increases.

Bialy et al. (2016) presented a heart disease prediction model using Naive Bayes, Bayesian Net, Multi-Layer Perceptron (MLP), Sequential Minimal Optimization, C4.5, and Decision Tree [106]. The system combines the prediction results of each classifier in an ensemble model using a weighted average. The system is trained with two different disease datasets such as CAD dataset consisting of 920 cases with 14 attributes and Heart Valve Diseases (HVD) dataset consisting of 103 instances. The outliers and extreme values were detected and removed by using the inter-quartile range technique. The experimental results show that the Naive Bayes algorithm obtained the highest accuracy on both datasets. The obtained results were analyzed using the WEKA tool, and the classification performance was measured using the 10-fold cross-validation.

Arabasadi et al. (2017) proposed a hybrid model for the diagnosis of heart disease based on clinical data without the need for invasive diagnostic methods. Researchers use feature selection techniques like the Gini index, weight by SVM, Information Gain, and Principal Component Analysis (PCA) to train networks and modify weights to achieve minimum error [107]. They use the Error Back Propagation algorithm in Artificial Neural Network with MLP structure and sigmoid exponential function to build the heart disease model. The proposed risk model enhances the performance of Neural Network by increasing its initial weights using a Genetic algorithm. The model achieves an optimal accuracy, sensitivity, and specificity on the Z-Alizadeh Sani dataset, which are higher compared to the existing systems.

2.2 Research Gaps

Heart diseases can be predicted using several methods; however, the most cost-effective and reliable methods are based on the assessment of cardiac non-invasive risk attributes. Various

researchers predicted heart disease on risk factors using different data mining techniques, but a closer look at the reviewed literature reveals several shortcomings which are described as:

i. Most of the developed heart disease evaluation models lack generalization capability.

ii. The derived risk rules from heart disease data are complex and large, which makes the system slow and leads to inaccurate decisions.

iii. Different tools (like WEKA, RapidMiner, Orange; etc.) are used for experiments and simulation purposes; however, each tool has complications accompanied by it. There are many mechanisms for prediction but all have limitations like documentation for GUI is limited, scaling is a problem, Big Data cannot be handled, etc.

iv. Medical domain performance measures like sensitivity, specificity, accuracy, precision, etc. are used; however, the model measures like computational complexity, scalability, robustness, and comprehensibility are not used by the researchers.

v. The automatic splitting condition of the decision tree algorithm on numerical medical variables leads to the wrong diagnosis for medical professionals. The medical society has standard splitting criterions that are universally acknowledged (high blood pressure, high cholesterol, etc); hence decision tree algorithms should be trained on such cut-off values before applied on the medical dataset.

vi. The existing risk evaluation tools help in classifying victims at risk of heart disease; however, there is not known performance accuracy for them.

vii. Most of the researchers used clinical attributes in their contributions, which reduces their usability other than medical settings. However, none of the prevailing heart disease risk tools is based on purely non-invasive risk features.

viii. Most researchers use only a single feature selection technique to get the significant attributes; however, there are no investigations of using multiple feature selection techniques to derive the significant non-invasive attributes with their mean values for early heart disease risk evaluation.

To overcome these research limitations, we develop an effective, low-cost heart disease evaluation model using significant non-invasive risk attributes. The developmental procedure for the risk model is discussed in the subsequent chapters.

2.3 Chapter Summary

This chapter provides a detailed literature review of heart disease prediction and diagnosis using different data mining techniques. Researchers made decisive contributions to cardiac disorder identification using data mining methods to find out the factors that drag the world to this lethal disease. They found that behavioral risk factors are the primary causes of heart diseases. Many researchers build risk models using the divergent data sets, different machine learning algorithms, various data mining approaches, and numerous tools. Researchers found that there is no single algorithm that produces the best results for every dataset; however, hybridization and ensemble methods show optimal results. Researchers used cross-validation and error rates for experiments and simulation purposes using different tools, but each tool has complications accompanied with it. To improve health care systems and build effective models, we need to use the most appropriate and novel data demanding techniques on real-time datasets to get the accurate diagnosis well in advance. To predict and detect heart disease through data mining techniques at its earliest, it demands additional research to understand the novel discoveries about it.

CHAPTER 3

Applying Data Mining Techniques in Heart Disease Prediction

Medical industries are overwhelmed with the incomplete and noisy data and to extract the hidden information in an explicit structure from these large datasets the data mining techniques are applied. The reason to introduce data mining techniques in health care is not to take over specialists, but to give assistance where they struggle. This chapter describes the feature selection techniques which are used to find the significant non-invasive subset of risk attributes for the early prediction of heart disease. This chapter discusses the data mining techniques that are used to develop a risk evaluation model. In this chapter, the performance of the risk models is evaluated using various performance measures. The chapter also presents the significance of non-invasive risk features for the initial identification and treatment of cardiac patients.

3.1 Feature Selection Techniques

The feature selection techniques are used to discover the subsets of features that produce accurate and compact prediction models [134]. In this research work, the combination of Filter, Wrapper and Embedded feature selection methods are examined to reduce the complexity of the model and to get the preeminent non-invasive subset of risk attributes for heart disease prediction.

3.1.1 Extra Tree Classifier

The extra tree classifier also called extremely randomized trees is an ensemble learning method that builds multiple trees without replacement. The nodes of the decision tree are split based on random splits which lead to increased accuracy and extensively decrease computational load linked to the determination of optimal cut-points in standard trees and random forests [108].

3.1.2 Gradient Boosting Classifier

Gradient boosting is used for both classification and regression problems. It involves a loss function to be optimized using decision trees which are constructed through a greedy manner, and finally, these trees are added one at a time to minimize the loss function [109].

3.1.3 Random Forests

Random Forests are created from decision tree predictors that are used both for regression and classification tasks. The random forest is created using a number of decision trees from the randomly selected training set to surpass the overfitting problem of the individual decision tree [110]. The random forest classifier is explained exhaustively in the data mining techniques section.

3.1.4 Recursive Feature Elimination

The Recursive Feature Elimination (RFE) is a greedy optimization method that tries to get the finest performing feature subset. It recursively constructs models and puts apart the finest or the lowest-performing feature at each iteration. It builds up the next model using the remaining features until all the features are exhausted and then positions the features based on the rank of their elimination [111].

3.1.5 XG Boost Classifier

Extreme Gradient Boosting is an ensemble algorithm that uses an optimized gradient boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting and bias. Because of the XG Boost classifier's scalability, it learns fast and gets efficient memory usage [112].

3.2 Data Mining Tasks

The main goal of data mining is to learn from the data. Data mining tasks are utilized to determine the type of patterns found in the data mining process. Data mining tasks are

generally divided into two major types: Predictive Tasks and Descriptive Tasks, as shown in the below-given figure3.1 [113] [114]. In predictive tasks, the goal is to predict the value of a dependent (target) attribute based on the values of independent (exploratory) attributes. In Descriptive Tasks, the purpose is to extract patterns that describe the underlying relationships in data. Descriptive tasks are often exploratory and often need post-processing methods to explain and validate the results [59].

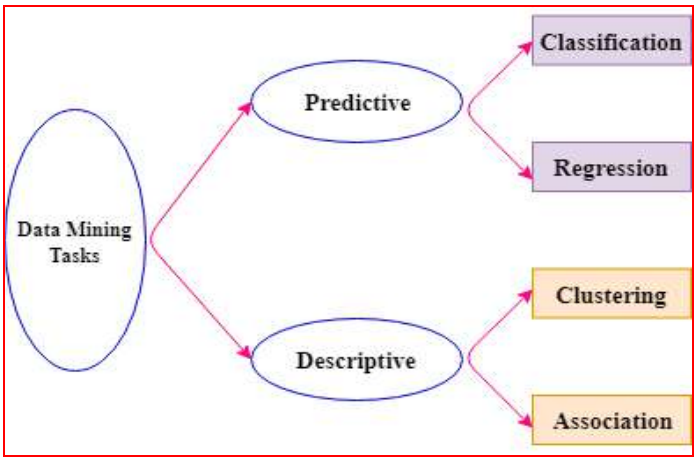

Figure3.1 Categorization of Data Mining Tasks

3.2.1 Predictive Data Mining Tasks

Predictive modeling means to build a model for the target variable as a function of the explanatory variables. Predictive modeling tasks are of two categories: Classification and Regression [59].

3.2.1.1 Classification: Classification data mining applications are predictive (Supervised) learning data mining tasks that predict discrete values of the target attribute [59]. If the target attribute values are two values (*such as yes and no*), then it is called as the *binary classification*. But, if the target attribute has multiple possible values (for example, drug A, drug B, and drug C) for any disease cure, then that is called as the *multi-class classification.*

3.2.1.2 Regression: Regression builds a predictive model for the continuous target variables as a function of the explanatory variable. e.g., predicting the cost of a stock three

months into the future, because the cost is a continuous-valued attribute. The objective of both tasks is to learn a model that reduces the error between the predicted and true values of the target variable [59].

3.2.2 Descriptive Data Mining Tasks

Descriptive modeling refers to the task of deriving the patterns that summarize the basic relationships in the data. There are two types of descriptive modeling tasks: Clustering and Association.

3.2.2.1 Clustering: The clustering task seeks to discover groups of strongly interrelated instances so that instances which fit into the same cluster are tightly interrelated to each other than the instances that belong to separate cluster [59].

3.2.2.2 Association: Association analysis is applied to identify patterns that characterize strongly associated features in the data. The discovered patterns are frequently expressed in the form of implication rules or features subsets [60].

3.3 Data Mining Techniques

Predicting heart condition from different symptoms is a stratified problem that is bound to erroneous assumptions and has impulsive effects. We use various data mining methods to extract knowledge from the heart disease dataset. The purpose of blending data mining methods in health care is not to take over specialists or assistants, but to give support to where they struggle [113] [115]. There are several core data mining techniques available, but the focus of this research is the application of classification techniques that would support medical professionals to identify the victims at high risk of heart disease.

3.3.1 Decision Tree

A Decision Tree is a non-parametric technique, which is most often used for classification; however, it can also be used for regression tasks [116]. Decision trees adopt greedy (i.e., non-

backtracking) approach and are constructed in a top-down recursive divide-and-conquer manner [60]. The algorithm begins with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built [59]. When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, to improve classification accuracy on unseen data [117]. In the decision tree, each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

Different types of decision trees are available, the distinction between them is the mathematical model used to select the splitting attribute in extracting the decision tree rules. Popular measures of attribute selection are Information Gain, Gain Ratio, and Gini Index [118]. The Information Gain attribute selection measure is used to select the attribute that best partitions the tuples into distinct classes. The Information Gain approach selects the splitting attribute that minimizes the value of entropy, thus maximizing the Information Gain. The Information Gain for each attribute is calculated using Equation (3.1).

$$Gain(A) = Info(D) - Info_A(D) \qquad (3.1)$$

Where Info (D) is the average amount of information needed to identify the class label of a tuple in D and is calculated using *Equation (3.2)*. InfoA (D) is the expected information required to classify a tuple from D based on the partitioning by A *and is* calculated in Equation (3.3).

$$Info(D) = -\sum_{i=1}^{m} pi \log_2 (pi) \qquad (3.2)$$

Where pi is the non-zero probability that an arbitrary tuple in D belongs to class Ci and is estimated by |Ci, D| / |D|. A log function to the base2 is used because the information is encoded in bits.

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * Info(D_j) \qquad (3.3)$$

The term | Dj | / | D| acts as the weight of the jth partition [59].

Below-given figure3.2 shows how the decision tree works to build the heart disease evaluation model. The principle reason to apply the decision tree is to develop a risk assessment model that can forecast the heart disease victims by learning decision rules from the training dataset. The experimental results obtained from the decision tree are explained in Chapter 4.

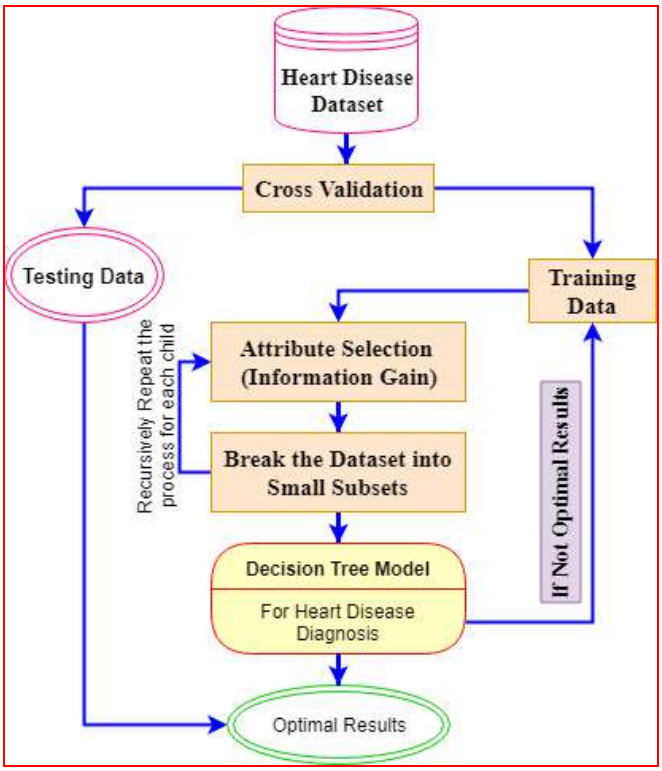

Figure 3.2 Decision Tree Model Working for Heart Disease Prediction

3.3.2 K Nearest Neighbour (K NN)

K Nearest Neighbour (K NN) is the basic, non-parametric, and instance-based data mining technique [119]. K NN uses learning by analogy, which compares the new unclassified record with the existing records using the distance metric. The closest existing record is used to assign the class to the newly unclassified record [59]. Below given figure3.3 shows the example of K NN classification.

The good value of k can only be determined experimentally by setting the value of k to 1 and then increment k to allow for new more neighbors. The k value that gives the minimum error rate is selected. The test set is used to estimate the error rate of the classifier. In the K NN algorithm, a new instance is classified by a closeness to the neighbors, which is defined in terms of the distance function. Many distance measures can be used, such as (Euclidean, Manhattan, and Minkowski) but in this research Euclidean measure is used because of the properties of the heart disease data [60] [120].

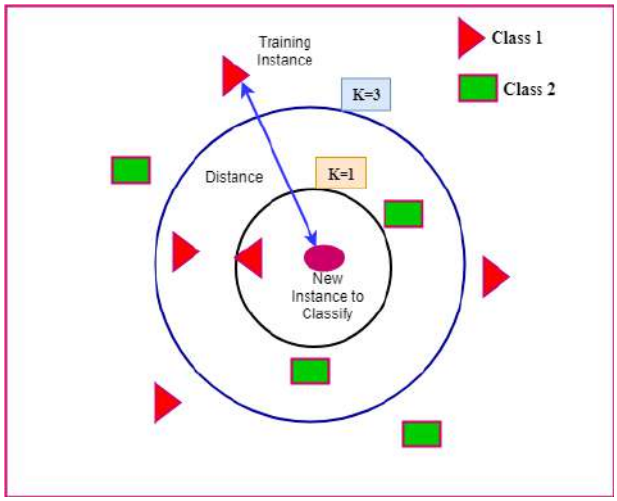

Figure3.3 K Nearest Neighbour classification Example

The Euclidean distance between two points is the length of the path connecting them. Euclidean distance is calculated as the square root of the sum of the squared differences between i= (xi1, xi2 .... xip) and j= (xj1, xj2 .... xjp) across all input attributes p.

$$d(i,\ j) = \sqrt{(xi1\text{-}xj1)^2+(xi2\text{-}xj2)^2+\ldots+(xip\text{-}xjp)^2} \qquad (3.4)$$

Before using Euclidean distance measure the values of attributes are normalized so that the attributes with the highest values will not overshadow the lowest valued attributes. In this research work, Min-Max normalization technique is applied to transform a value **P** of numeric attribute **Z** to $\mathbf{P}^{|}$ in the range [0, 1] by computing

$$P^{\backslash} = \frac{P - \min Z}{\max Z - \min Z} \qquad (3.5)$$

Where $min_{\mathbf{Z}}$ and $max_{\mathbf{Z}}$ are the minimum and the maximum values of attribute Z

In this research, the K NN technique is used to predict heart disease patients at its earliest. The experimental results achieved from the classifier are discussed in Chapter4.

3.3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised data mining technique that is used for both classification and regression purposes. SVM uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the optimal separating hyperplane. The *optimal* hyperplane for an SVM means the one with the largest *margin* between the two classes. The SVM finds this hyperplane using support vectors and margins [60] [121]. The support vectors are the data points that are closest to the separating hyperplane and are considered critical elements of the dataset and the margin is the maximal width of the slab parallel to the hyperplane that has no interior data points. The discriminant function *f* (*T*) for a test sample T is a linear combination of support vectors and is constructed as:

$$f(T) = \sum_{i} \alpha_i y_i \, (X_i \, . \, T) + b \qquad (3.6)$$

Where the vectors Xi is the support vectors, Yi is the class labels of Xi; vector T represents a test sample and (Xi.T) is the dot product of T with one of the support vectors Xi. αi and b are numeric parameters to be determined by the learning algorithm.

The following figure3.4 illustrates the linear support vector machine where light green circles represent data points of class x1 and red, indicating data points of x2. The purpose of SVM is to choose a hyperplane with the greatest possible margin between the hyperplane and any data point with the training set, giving a greater possible chance of new data being classified correctly. However, if there is no clear hyperplane, it is necessary to move to a higher

dimension view called kernelling in SVM. The idea is that the data will continue to be mapped into higher dimensions until a hyperplane can be formed to segregate it [59] [61].

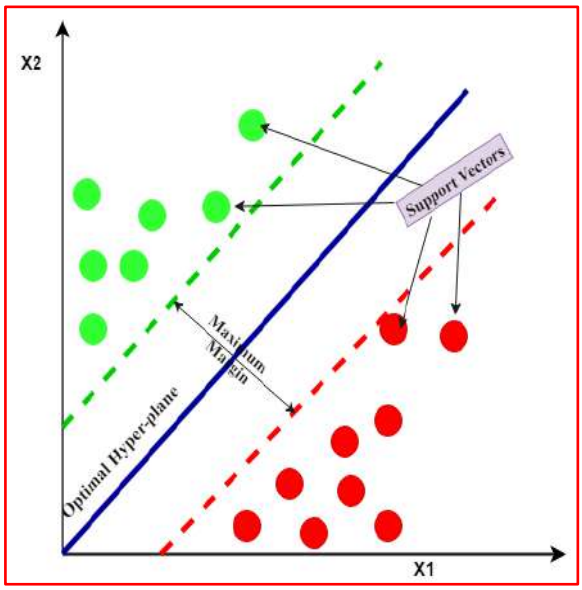

Figure3.4 Linear SVM Classifier for Two-Class Representation

Hence in the case of non-linear separation, the training data will be mapped into a higher-dimensional space *H*, and an optimal hyperplane will be constructed there. The mapping is performed by the kernel function ***K,*** which defines an inner product in *H*. When there is a mapping, the discriminant function is given like:

$$f(T) = \sum_{i} \partial_i y_i K(x_i, T) + b \qquad (3.7)$$

SVM is largely characterized by choice of its kernel function used, *e.g.*, polynomial, kernel and Gaussian radial basis kernel function. However, besides these kernel functions, there are other kernel functions. To determine parameters ∂ and y in the above equation, the construction of the discriminant function finally turns out to be a constrained quadratic problem on maximizing the Lagrangian dual objective function:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j) \qquad (3.8)$$

under constraints

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i >= 0, (i=1,2,\ldots,n) \qquad (3.9)$$

Where n is the number of samples in training data, however, the quadratic programming problem in equation (3.8) cannot be solved easily via standard techniques since it involves a matrix that has several elements equal to the square of the number of training samples [60]. The SVM technique is used in heart disease prediction because comprehensive and correct rules are crucial to help biologists and doctors. The heart disease prediction results obtained from the SVM model are discussed in Chapter 4.

3.3.4 Random Forests

Random Forests are an ensemble of simple decision trees that are used for both regression and classification problems. The random forest algorithm creates the forest with a number of decision trees from the randomly selected training set, with the goal of overcoming the over-fitting problem of the individual decision tree. In random forest classification, each decision tree votes and the aggregated votes decide the final classes of the test object; however, in the regression, the means prediction or regression of the individual trees is calculated [122].

Random forests are a substantial modification of bagging that builds a large collection of de-correlated trees and then averages them. Below given figure3.5 shows the working of the random forest algorithm in which each tree is grown on a different sample of original data. There is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error because, in each of k iterations, about 1/3rd of the samples are left out of the new bootstrap training set and are not used in the construction of the tree. In this way, a test set classification is obtained for each sample in about $1/3^{rd}$ of the constructed trees. The final classification for a sample is the class having the most votes from the trees in the forest [60] [61]. In this research work, the random forest algorithm is used in heart disease prediction, and diagnosis, and the results obtained are discussed in Chapter 4.

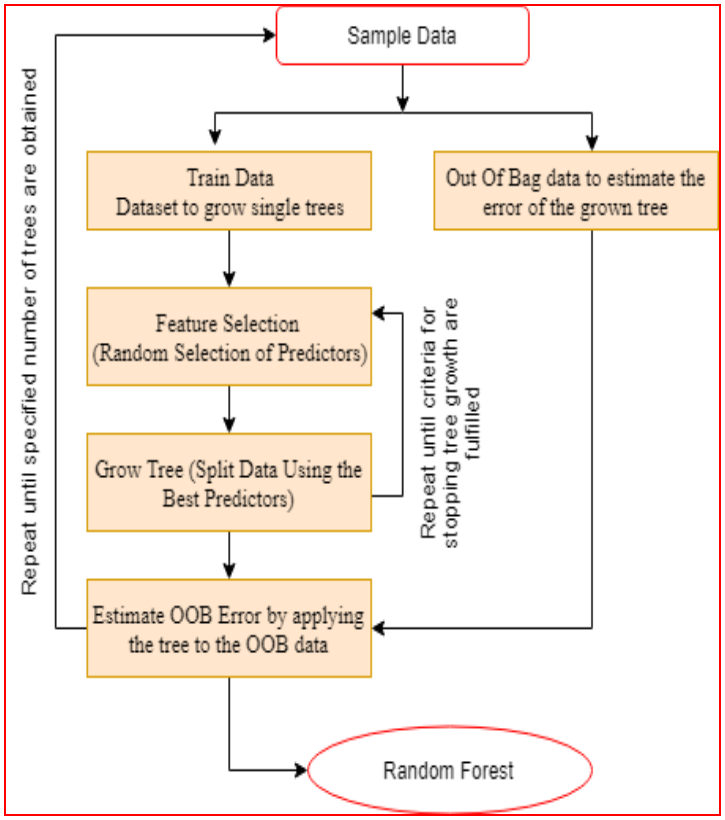

Figure 3.5 Random Forest Algorithms' Working

3.3.5 Naive Bayes

Naive Bayes predict class membership based on statistical probabilities. Naive Bayes classification is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. The algorithm assumes that the outcome of an attribute value on a given class is independent of the values of other attributes. This assumption is called *class conditional independence.* This assumption is made to simplify the computations involved and, in this sense, is considered *"Naive*" [59].

Naive Bayes classifier can handle an arbitrary number of independent variables, whether continuous or categorical. Naive Bayes classification calculates the *prior probability* of the target attribute and the *conditional probability* of the remaining attributes. 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 value with the largest probability is then selected [123] [124]. The probability of the testing instance for the target attribute value is calculated using the below-given equation:

$$P(H / X) = \frac{P(X/H)P(H)}{P(X)} \qquad (3.10)$$

Where *P (H/X)* is the *posterior probability,* of *H,* conditioned on *X and P (X/H) is the posterior probability of X conditioned on H. Similarly P(X) is the prior probability of X and P (H) is the prior probability of H.*

Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes and is a strong assumption which results in a fast and effective method [60] [61]. In this research, the Naive Bayes algorithm is used to predict and diagnose the heart disease at its earliest using the non-invasive risk attributes. The heart disease prediction results obtained from the Naive Bayes model are discussed in Chapter 4.

3.4 Model Evaluation Techniques

Model evaluation is the solution to making practical development in data mining. There are numerous methods of understanding structured patterns from the given dataset. However, to find out which method to apply to a specific problem, we need systematic methods to estimate how data mining techniques work and to compare one with another. In classification problems, the performance of an algorithm is measured in terms of the confusion matrix, cross-validation, error rate, sensitivity, specificity, accuracy, precision, and ROC curves, which are discussed as follows [59] [125]:

3.4.1 Confusion Matrix

The confusion matrix also called the contingency matrix, or error matrix is a principal source of performance measurement in classification problems. Below given table3.1 shows the two-

class confusion matrix, which provides insights into the types of errors being made by a classifier [59].

Table 3.1 Contingency Matrix for Two-Class Classification

		Expected Values	
Observed Values		**Positive**	**Negative**
	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Where True positives refer to positive tuples that are correctly labeled by the classifier. True Negatives refer to the negative tuples that are correctly labeled by the classifier. False Positives (also called a Type I Error) are the negative tuples that are incorrectly labeled as positives, and False Negative (also called Type II Error) are the positive tuples that are mislabelled as negative.

i. **Sensitivity** (also known as True Positive Rate or Recognition or Recall) is the proportion of positive tuples that are correctly classified as Positive [126].

$$Sensitivity = \frac{True\ Positive}{Positive} \qquad (3.11)$$

ii. **Specificity** (also known as True Negative Rate) is the proportion of negative tuples that are correctly classified as Negative [126].

$$Specificity = \frac{True\ Negative}{Negative} \qquad (3.12)$$

iii. **Accuracy** is the total percentage of cases that are correctly classified by an algorithm [127].

$$Accuracy = \frac{True\ Negative + True\ Positive}{TP + TN + FP + FN} \qquad (3.13)$$

iv. **Precision** is a measure of exactness (i.e., what percentage of entities categorized as positive are actually positive) [127].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (3.14)$$

v. **Error Rate** (**Misclassification Rate**) is the proportion of errors made over a whole set of instances. The error rate is a combination of training and generalization errors. Training errors are the number of misclassification errors committed to training data, whereas generalization error is the expected error of the model on previously unseen records. The best classification model has low training and generalization error. [59] [60] [61].

$$Error\ Rate = \frac{False\ Positive + False\ Negative}{Positive + Negative} \qquad (3.15)$$

3.4.2 Cross-Validation

The cross-validation technique measures the error rate of a learning model on a specific dataset. In cross-validation, the complete dataset is randomly split into mutually exclusive subsets of approximately equal size, and each record is used the same number of times for training and exactly once for testing. The training dataset allows data mining techniques to learn from this data. The testing dataset is used to evaluate the performance of the data mining technique about what is learned from the training dataset [59] [60] [61].

3.4.3 AUROC (Area Under the Receiver Operating Characteristics)

AUROC is a performance measure graph that demonstrates the performance of a classification model at different threshold settings. AUROC depicts how a greatly model is skilled in distinguishing between the classes. The ROC curve is plotted with True Positive Rate on the y-axis against the False Positive Rate on the x-axis, as shown in figure 3.6 [60] [61].

An outstanding model has AUROC value equivalent or close to 1, which means it has a fine measure of separability. A poor model has AUROC value equivalent or near to 0, which

means it reciprocates the result and predicts 0s as 1s and 1s as 0s. When the AUROC value is approximately 0.5, then the model cannot distinguish between positive and negative classes.

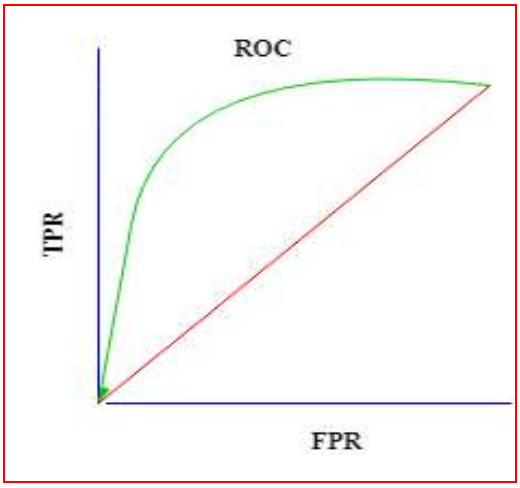

Figure 3.6 AUROC Representation

3.5 Data Collection and Research Methods for Risk Evaluation Model Development

The considerable challenge to collect a dataset is to get the quality and relevant data. The primary data is collected from different heterogeneous data sources like district hospitals and private clinics of Kashmir (INDIA) through quantitative data collection methods (interviews). This heart disease dataset comprises 5776 patient records accompanied by fourteen (14) non-invasive risk attributes as described in the below-given table3.2. The descriptive research methodology is followed to develop the risk evaluation model. The heart disease risk evaluation model is developed in the python programming language, and data cleaning, statistical modeling, data visualization, and machine learning operations are done in the Jupyter web application.

The Exploratory Data Analysis (EDA) is carried out on the Kashmir heart disease dataset to get different insights from the data. It is found that the heart disease dataset is noisy and includes numerous missing attribute values signified with the interrogation mark (?). Data imputation is performed (for *numeric* missing attribute values mean imputation data cleaning technique is applied to fill in the missing attribute values, and for the *categorical attributes*,

the *Mode* method is used for filling in the missing values). The heart disease dataset attributes are organized into two types nominal and numeric. For example, the Gender attribute values as "male" and "female" represent the nominal attribute, and the "age" attribute values as 70 years represent a numeric attribute. Furthermore, the nominal data attributes correspond to binary and ordinal variables, and continuous data attributes correspond to an integer, interval- scaled, and ratio-scaled variables [128].

Table 3.2 Description of the Heart Disease Dataset

Features	**Data Type**	**Features with subsequent values and explanation**
Age	Numeric	Represents the age of a patient in the number of years
Sex	Nominal	Represents sex of a patient, 0= Female, and 1= Male
Height	Numeric	Represents the height of the patient in Centimeters
Weight	Numeric	Represents the bodyweight of the patient in Kilograms
Systolic BP	Numeric	Represents the Systolic BP of the patient in mmHg
Diastolic BP	Numeric	Represents the Diastolic BP of the patient in mmHg
Hereditary	Nominal	Means if the patient had inherited heart disease, where 1=Yes, and 0= No
Healthy Diet	Nominal	Is the patient taking a nutritious diet? It is represented as 0=Following, 1=Occasionally and 2= Not Following
Physical Activity	Nominal	Whether the patient is doing exercise or not. 0= No Exercise, 1= Regular Exercise and 2= Occasionally
Alcohol Consumption	Nominal	How often the patient drinks alcohol, and it is represented as 0= Non-Alcoholic, 1= Occasionally and 2= Alcoholic
Smoking	Nominal	Whether the patient is smoking or not 0= Non-Smoker, 1= Regular and 2= Occasionally smoking
Socio-Economic Level	Nominal	Represents the economic level of the patient like 0=Poor, 1= Middle Class and 2= High Class
Diagnosis	Nominal	0= No and 1= Yes

3.6 Significance of Non-Invasive Heart Disease Attributes in Risk Evaluation

The cardiac disorder can be identified using numerous tests; however; these tests are expensive and cannot be used as public level screening tests. The existing risk assessment techniques use invasive heart disease risk features that require information from various blood tests before use. To overcome this problem, there is a requirement to streamline the cardiac disorder risk features with the goal that reasonable risk recognition strategies can be actualized [11]. The non-invasive heart disease risk features like age, height, weight, smoking habits, sex, and blood pressure are recognized effortlessly with no complex machines and instruments required [129] [130]. Although body weight and blood pressure need some instruments of measurement; however, these instruments can be accessible at home or in a drug store and do not require a hospital to take physical samples.

Chapter 4 and chapter 5 apply various algorithms to identify the non-invasive attribute combinations, which may demonstrate the most excellent performance in predicting heart disease patients. The non-invasive attributes are beneficial because they are low-cost attributes (Low-cost attributes means it costs nothing to determine a value for the attribute when diagnosing a patient). The effect of using different combinations of non-invasive attributes from the Kashmir heart disease dataset is investigated as input to Naive Bayes, Decision Tree, K Nearest Neighbor, Support Vector Machine and Random Forest data mining classifiers for performance evaluation. Equations using various non-invasive features are developed and checked to find out if they improve performance in predicting heart disease patients.

The existing risk calculators use invasive heart disease risk attributes like Total Cholesterol, HDL cholesterol, and Diabetes (usually measured through blood sugar or insulin levels) as well as non-invasive features such as age, smoking, sex, and resting blood pressure to discover patients at risk of heart disease [131]. However, the performance using only non-invasive features has not yet been measured. How correctly can data mining techniques (using only non-invasive attributes) classify patients would be examined in this research. Success

here would give an extraordinary chance for application to public level screening tests, thus enabling initial involvement in patients at high risk of heart disease and deciding reasonable treatment systems for those patients.

3.7 Chapter Summary

This research investigates developing a heart disease evaluation model using different data mining techniques. The primary data is collected from various heterogeneous data sources through quantitative data collection methods. This research work follows the descriptive research methodology and uses the Python programming language and Jupyter web application to build the risk evaluation model. In this research work, different feature selection techniques are applied over the Kashmir heart disease dataset to select the significant subset of non-invasive risk features for the initial prediction of heart disease. This research work applies different data mining techniques like Decision Tree, K Nearest Neighbor, Support Vector Machine, Random Forest and Naive Bayes to see whether these techniques will help medical practitioners in early prediction which would result in a reduction to severe and costly illness and complications. The developed risk evaluation model's performance is checked through various model evaluation techniques. Finally, the chapter is concluded by discussing the significance of non-invasive risk attributes.

CHAPTER 4

Discovering Knowledge in Heart Disease Data Using Data Mining Techniques

The Kashmir heart disease dataset is mined to derive knowledge for the early prediction and identification of the disease. This chapter describes Davis's data mining methodology, which is followed in this research for the development of heart disease model. In this chapter, the research design is formulated to simplify the research activities and make the research productive by the statement of objectives. Various feature selection techniques are applied to choose the significant non-invasive heart disease risk attributes. This research identifies a significant subset of risk attributes for data mining techniques for the initial prognosis of heart disorder patients. Finally, a risk evaluation model is developed to help medical specialists in classifying victims at elevated risk of cardiac disease.

4.1 Data Mining Methodology for Heart Disease Prediction

A data mining methodology is a technique for applying alternative methods to take raw data to a transformed dataset to generate knowledge for users. There are two eminent prevailing data mining methodologies for the "Knowledge Discovery from Data" process: CRISP-DM [132] [133] and SEMMA [134]. The industry-led consortium developed the CRISP-DM (Cross-Industry Standard Process Model for Data Mining), and the SEMMA (Sample Explore Modify Model Assess) is a data mining methodology derived from the Statistical Analysis Software Institute (SAS, 2008). These two methodologies are not suitable for our research work because they are too big and too complicated to use. Hence for this research work, Davis's data mining methodology is followed, as shown in figure4.1 [135]. The reason to apply this specific methodology is that it demonstrates our research objectives. This methodology contains the following phases:

i. **Data Selection:** In this phase, the relevant heart disease data from various heterogeneous sources is selected and then stored in the standard database.

ii. **Data Preparation:** In the data preparation step, the heart disease dataset is analyzed and prepared into an appropriate form for the data mining algorithms to derive meaningful insights from it and to get the optimal output.

iii. **Data Task Filter:** In this step, the heuristic decision rules are employed to determine expected results for heart disease prognosis in later steps. The selected dataset is then stored in the "Data Mining Task Warehouse.

iv. **Data Mining Techniques:** In this step, an appropriate algorithm is selected with a suitable dataset for the task requested in step3.

v. **Comparison and Evaluation**: In this phase, the classified outcomes are contrasted and estimated based on different data mining evaluation measures.

vi. **Building New Models:** In this phase, the developed supervised classification models are stored in the data mine warehouse for the next prediction problems. For new prediction tasks, the process is repeated from step 3 to step5.

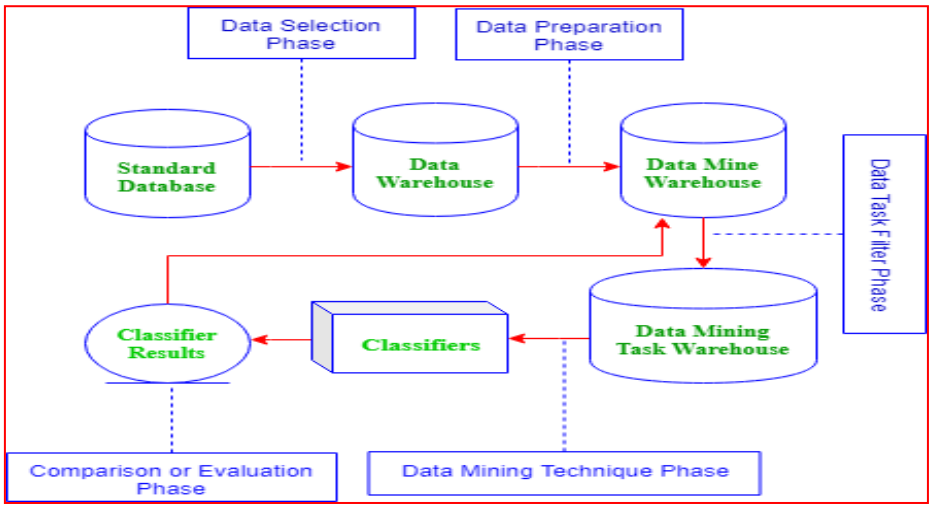

Figure4.1 Heart Disease Risk Evaluation Model Methodology

4.2 Research Design for Heart Disease Risk Evaluation Model

The research design is followed to simplify the research activities and make research productive by the statement of objectives. It provides information about the data inputs

required, the methods of analysis used, and a statement of objectives of the study to solve the research problem. The proposed research design (in figure4.2) is followed step by step to build the heart disease risk evaluation model.

The research design consists of three main phases having eight steps. The phases of the research design with their respective steps are explained as follows:

i. **Data Phase:** The data phase contains the whole process from data collection to feature engineering. This phase includes the qualitative data collection step, the pre-processing subsystem step, the cleaned data set storage step, and finally, the feature selection step.
ii. **Data Mining Phase:** The data mining phase includes classifiers Naive Bayes, K Nearest Neighbor, Decision Tree, Support Vector Machine, and Random Forest, which would be employed to develop the heart disease risk model.
iii. **Model Evaluation and Validation Phase:** The model evaluation and validation phase calculate and endorse the risk evaluation model using different data mining techniques. The risk evaluation model would be evaluated through the train-test split of the heart disease data using 10-fold cross-validation. The model would be then validated using different performance measures by comparing the results with the existing models and various model metrics (Sensitivity, Specificity, Accuracy, Misclassification Rate, and ROC Score).
iv. **Knowledge-Base Phase:** The Knowledge-base phase includes the steps to store and retrieve knowledge about heart disease. The generated heart disease risk rules would be stored in the knowledge base and cross-checked as per medical guidelines and through domain expertise.

In this chapter first phase (qualitative data collection, pre-processing subsystem, and feature selection) is discussed and the rest of the phases will be discussed as demanded by the research for the development of the heart disease risk evaluation model.

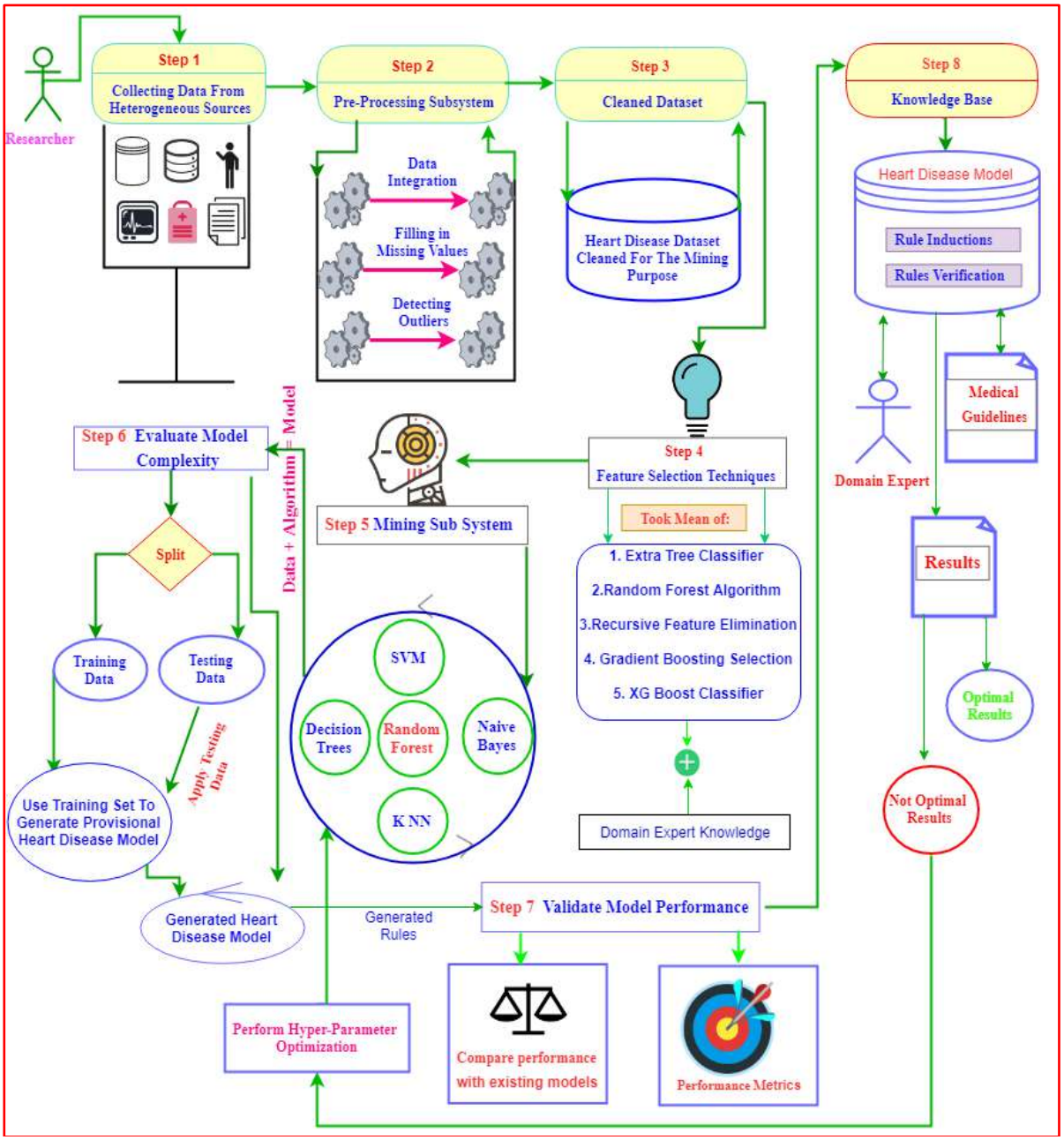

Figure 4.2 Detailed Steps of Research Design

4.3 Exploratory Data Analysis (EDA) Process

The basic statistical description is performed to learn about each attribute value of the Kashmir heart disease dataset. Knowing such fundamental statistics about each attribute helps to smooth noisy values, spot outliers, and fill in the missing values. The heart disease dataset consists of a combination of nominal and numeric risk attributes. The missing numeric values are removed through the simple mean imputation method, and categorical missing values are filled by mode imputation technique [136] [137].

4.3.1 Checking Class Imbalance and Data Distribution Problems in Dataset

Before performing any operations on the heart disease dataset, it is required to test class balance because highly imbalanced data makes the machine learning algorithms biased [138] [139]. To check the class balance, the skewness, and kurtosis statistical operations on the data are performed [140]. Skewness estimates the symmetry to see whether data distribution is same to the left and right of the center point and Kurtosis measures whether the data are light-tailed or heavy-tailed to a normal distribution [140] [141]. After the Skewness and Kurtosis statistical measure tests, it is found that the collected heart disease dataset is balanced and has a skewness of (-0.03065287) value and Kurtosis of (-2.000136) value which means the Kashmir heart disease dataset is normally distributed.

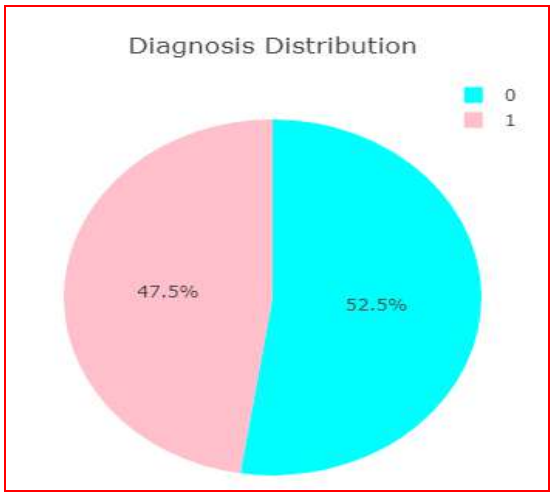

Figure 4.3 Heart Disease Distribution Based On Sex Attribute

The dataset contains 5776 records, of which 2760 are females, and 3016 are males. Among these 5776 instances, 2745 (47.5%) have heart disease, and 3031 (52.5%) are healthy. Below given figure 4.3 shows the pie chart representations of heart disease data distribution. Heart disease affects both men and women approximately in the same proportion with substantial death rates and disabilities. Predicting and detecting heart disease accurately in advance constitutes several basic causes like social, commercial, and cultural transition. The long-term disclosure to these risk attributes affects the hardest and ends up in death. Health reports

suggest that if there is not a reversal of behavioral risk factors, the disease will carry on to rise and would lead to human and economic loss.

4.3.2 Finding Correlation among Different Heart Disease Risk Attributes

In any dataset, there could be multifaceted and strange relationships among the variables; hence, it is imperative to determine and measure the degree to which attributes in the dataset are related to each other. This process of finding the degree of relationship between the dataset attributes is known as correlation [142]. The knowledge of the correlation between the attributes helps to prepare the data to meet the expectations of machine learning algorithms. Pearson's correlation is applied to check the mutual relationship among the heart disease attributes [140] [141]. A correlation could be positive (which means all the related attributes move in the same direction), negative (which means all the related attributes move in opposite directions) or neutral (which means that the attributes are not related to each other). The result of the applied Pearson's correlation coefficients among the heart disease variables is shown in the below-given figure 4.4 in the form of heatmap representation.

The heatmap grid represents the correlation between the heart disease attributes with their corresponding coefficients. The Symmetrical heatmap matrix represents all attributes across the top and down the side, to give a correlation between all pairs of features. The diagonal line across the matrix from the bottom-right corner to the top-left represents a perfect correlation of each attribute with itself. The value 1 means a perfect positive correlation among the attributes and value -1 means a perfect negative correlation among the heart disease attributes; a correlation coefficient close to zero indicates weak dependency among the heart disease attributes [143].

After analyzing the heatmap correlation results, it is found that the independent attributes of the Kashmir heart disease dataset are loosely correlated with one another. This loose correlation among independent heart disease attributes is a good sign to improve the performance of the model. However, if the attributes in a dataset are tightly correlated (called

multicollinearity), then change in one variable can lead to change to another variable that can deteriorate the performance of an algorithm [144]. Correlation among the attributes does not mean causation hence, the strong relationship among attributes should be evaluated significantly. Mostly, a relationship among attributes may look causal through strong correlation because of some overlooked factors.

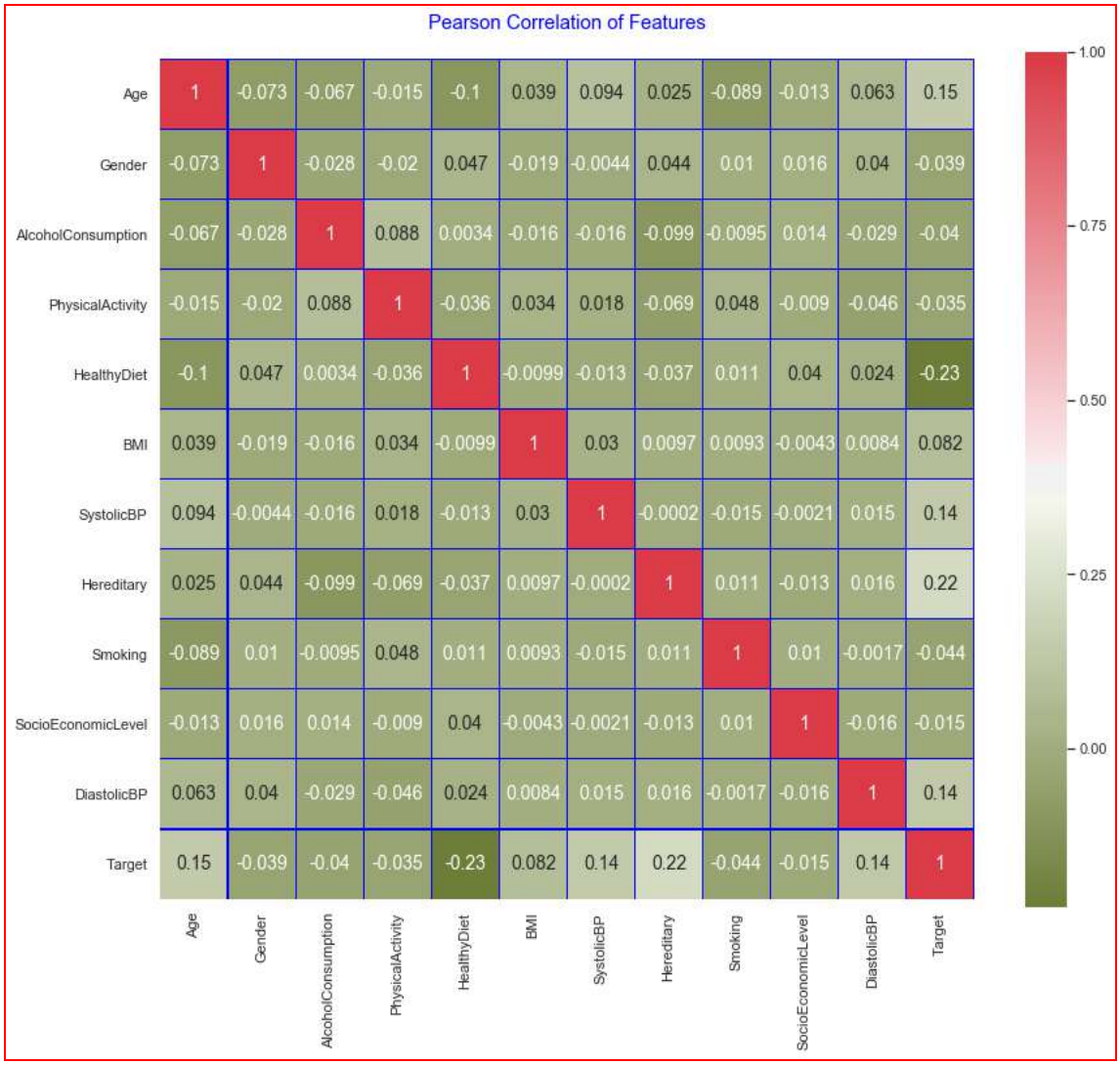

Figure 4.4 Correlation in Risk Attributes Through Heatmap Representation

4.4 Feature Selection Techniques for Heart Disease Risk Assessment

Even though heart disease attained an epidemic extent, yet it is controllable by reducing the modifiable risk factors. Diagnosing heart conditions effectively in advance is not free from erroneous statements and constitutes the number of undetermined factors [145]. The feature

selection methods are applied to select the significant and most appropriate subset of risk attributes for the initial prognosis of heart disease.

Feature selection helps to reduce the inappropriate and redundant attributes, which often decrease the performance of classifiers [146]. In this research (filter, wrapper, and embedded) feature selection methods are applied to get an appropriate feature subset for heart disease risk evaluation [146]. The five different Feature Elimination techniques (Extra Tree Classifier, Gradient Boosting Classifier, Random Forest, Recursive Feature Elimination, and XG Boost Classifier) are used on the heart disease dataset (as mentioned in table 4.1) to get the best non-invasive feature subset of risk attributes for heart disease prediction [147]. Each risk attribute is weighted by these feature selection techniques as per their role in disease prediction.

The applied feature selection techniques provide weight in between the scale of 0 to 1 to each heart disease risk attribute. After weight assignment to every risk attribute by the separate feature selection technique, the overall mean of all the applied weights to every attribute by these feature selection techniques is considered the final weight. The risk feature with the mean value near to 1 is considered important in predicting heart disease, and those risk attributes whose associated values are near to 0 are considered less significant in predicting heart disease.

Below given table 4.1 shows the different heart disease attributes with their respective weights assigned by different feature selection techniques and the last column in table shows the overall Mean of all the techniques. These heart disease risk attributes were identified by professional cardiologists *Sophia Airhart (Assistant Professor)*, and many other general physicians who are working in the cardiology department at various hospitals across, INDIA. The assigned weights to each heart disease risk attribute are validated and approved by different medical experts like *Dr. V. J. Sadhana (MD., FAHA Medical Director), Dr. Syed Aijaz Nasir (Cardiologist at AIIMS), Dr. Mohd. Khutubuddin Ansari (Medical Officer at MANUU)* etc. meanwhile these medical domain experts gave their respective opinions to

include some vital attributes also like (chest pain, and asthma) for early heart disease prediction and identification.

Table 4.1 Feature Selection Techniques Providing Weight to Each Risk Attribute

Attributes	Feature Selection Techniques with their results and Mean Values					
	ETC	**GBC**	**RF**	**RFE**	**XGB**	**MEAN**
Age	0.92	0.92	0.87	0.25	0.92	0.78
Sex	0.0	0.0	0.11	0.83	0.0	0.19
Alcohol Consumption	0.09	0.09	0.09	0.75	0.09	0.22
Physical Activity	0.25	0.25	0.08	0.67	0.25	0.30
Healthy Diet	0.71	0.71	0.52	1.0	0.71	0.73
BMI	0.74	0.74	0.79	0.0	0.74	0.60
Hereditary	0.38	0.38	0.4	0.92	0.38	0.49
Smoking	0.17	0.17	0.09	0.5	0.17	0.22
Systolic BP	1.0	1.0	1.0	0.08	1.0	0.82
Diastolic BP	0.88	0.88	0.78	0.33	0.88	0.75
Socio-Economic Level	0.17	0.17	0.11	0.42	0.17	0.21

After analyzing the results it is derived that the attributes [Systolic BP, Diastolic BP, Age, BMI, Hereditary, Healthy Diet, and Physical Activity] are the most important features for the early prediction of the heart disease because their corresponding numeric values are high and are also validated and approved by medical domain experts. The pictorial representation of attribute hierarchy with their respective weights is shown in figure 4.5.

Below given table 4.2 shows the descending order of heart disease attributes according to their mean values assigned by five different feature selection techniques. The attributes with the highest weight are crucial, and the attributes with lower values are less significant in

predicting heart disease at its earliest. The highly weighted significant subset of risk features is used to develop the heart disease risk model.

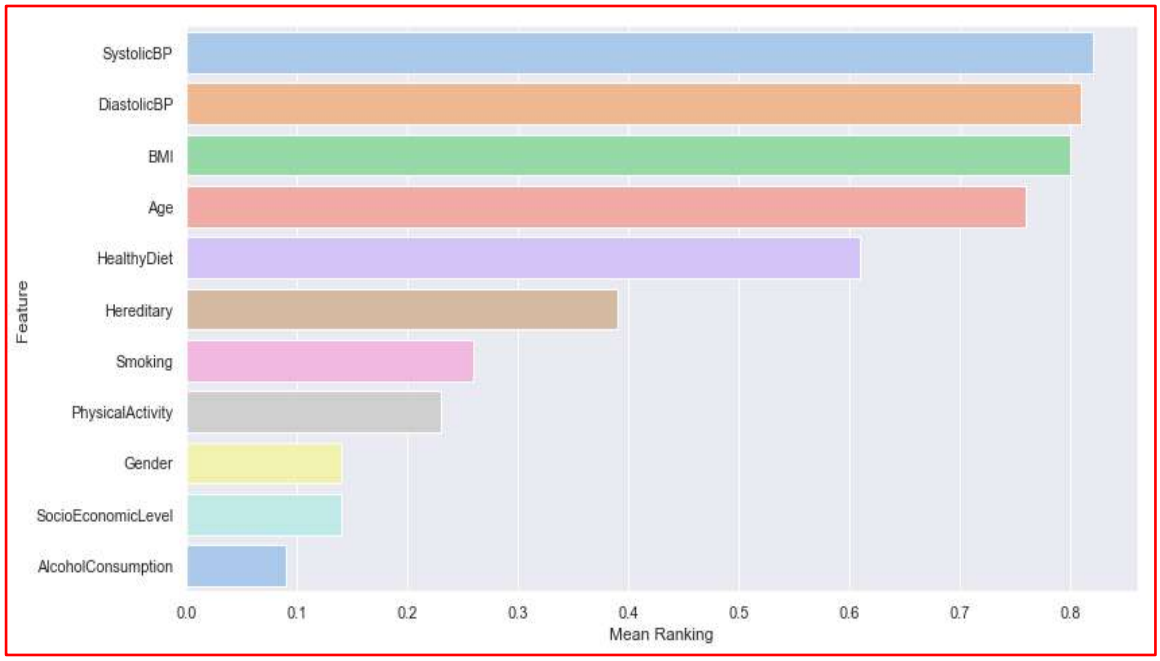

Figure 4.5 Risk Attribute Hierarchy by Feature Selection Techniques

Table 4.2 Mean Ranking of Risk Attributes by Feature Selection Techniques

Sr. No	Attributes	Mean Ranking of Attributes
1	Systolic BP	0.82
2	Diastolic BP	0.80
3	BMI	0.78
4	Age	0.76
5	Healthy Diet	0.54
6	Hereditary	0.42
7	Smoking	0.28
8	Physical Activity	0.24
9	Socio-Economic Level	0.16
10	Sex	0.14
11	Alcohol Consumption	0.12

4.5 Experimental Results of the Proposed Data Mining Techniques

It is found that the prevailing heart disease risk models are not free from limitations because they show varying results across different datasets that greatly reduces the effectiveness of the system. In this research, the heart disease dataset is mined through Decision Tree, K Nearest Neighbor, Random Forest, Support Vector Machine, and Naive Bayes techniques using 10-fold cross-validation to get unbiased results. Various medical domain performance metrics sensitivity, specificity, accuracy, precision, AUROC score, and misclassification rates, and the model measures like computational complexity and comprehensibility are calculated to obtain the optimal and accurate results. Below given sub-sections explain the experimental results obtained by different heart disease risk evaluation models.

4.5.1 Decision Tree Model Experimental Results

The rationale to apply a decision tree is to develop a heart disease risk evaluation model that can predict a class (diseased or healthy) by learning simple decision rules deduced from training data. The cross-validation on the training dataset is used to get the unbiased results [148] [149]. The performance results of the decision tree model are shown in the confusion matrix figure 4.6. From the decision tree model's confusion matrix (figure 4.6) the sensitivity, specificity, accuracy, precision, and error rates are derived that are described as follows:

The percentage of patients that were recognized accurately to have the heart disease (i.e., True Positive) upon the total number of patients who actually have the heart disease is Sensitivity. Putting the derived sensitivity values of the confusion matrix figure 4.6 in equation (3.11) the sensitivity of 82% is obtained. The closer the value for this measure is to 1, the better the rules are at identifying those patients who have heart disease.

Similarly, the percentage of victims that were recognized correctly to not have the heart disease (i.e., True Negative) upon the total number of patients who do not have the heart disease is Specificity. Putting the derived specificity values of confusion matrix figure 4.6 in

equation (3.12) the specificity of 0.8092% is obtained which means the decision tree model can recognize the healthy cases with an accuracy of 80%. The nearer the value for this measure is to 1 the best the rules are at identifying those patients without the disease. The overall accuracy of the decision tree model is obtained by using the equation (3.13) in figure 4.6 which is equivalent to 0.8185% which represents that the decision tree heart disease model's overall performance (in diagnosing both the diseased and non-diseased heart disease cases) , the higher the accuracy percentage, the more accurate the model is.

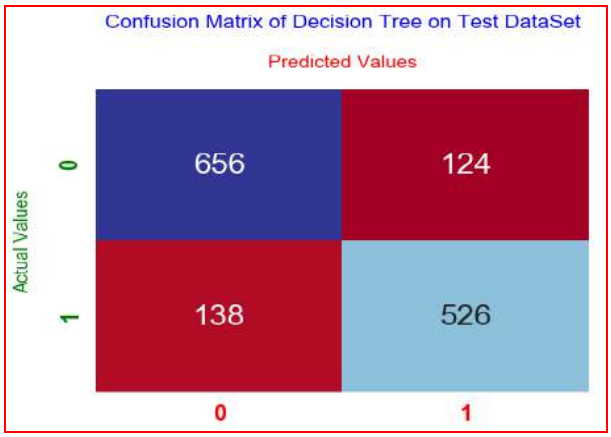

Figure 4.6 Decision Tree Model Confusion Matrix

Similarly, putting the values of the confusion matrix figure 4.6 in equation (3.14), a precision of 0.8410% is obtained. The closer the value for this measurement is to 1, the greater the chance that those with a positive outcome will actually have a disease. If a high precision rate of the decision tree model is obtained, then it means that the model will obtain a low false-positive rate. The error rate of the developed decision tree model is obtained by putting the values of the confusion matrix figure 4.6 in equation (3.15), which is equivalent to 0.1814%. The lower the percentage of misclassification rate of the model, the more accurate the model is in identifying the diseased and healthy cases.

The AUROC performance measurement is used to check the probability curve and measure of separability obtained by decision tree algorithm. AUROC demonstrates how efficiently the model can differentiate among the diseased and non-diseased patients. AUROC curve is the plot of the true positive rate (Y-axis) against the false positive rate (X-axis) for a number

of different candidate threshold values between 0.0 and 1.0. Below given figure 4.7 is the AUROC of the decision tree algorithm with an AUROC score of =0.817%. The area under a correlation curve plotting true positive against false positive is higher for models best able to correctly identify positive and negative cases.

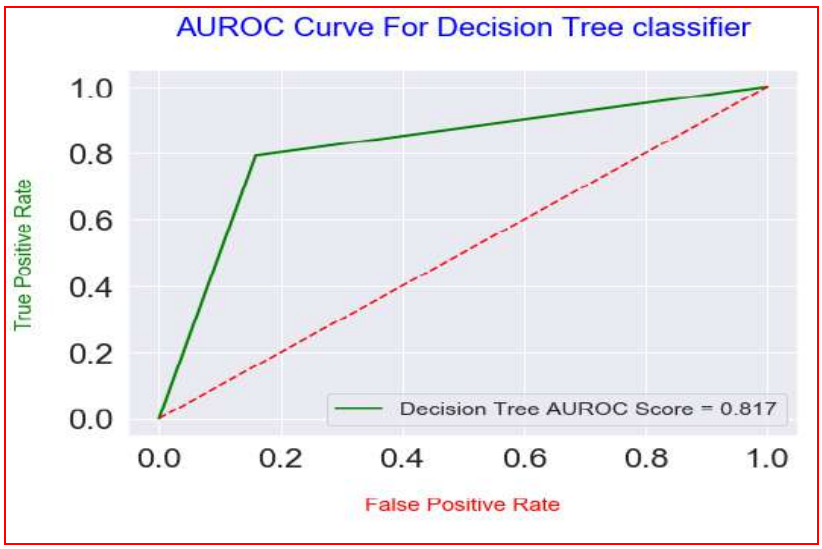

Figure 4.7 AUROC by the Decision Tree Model

We simulate the accomplished experimental results of the developed decision tree heart disease model with the prevailing research; the results obtained are to the best of our knowledge greater than the published results in the literature. However, the derived decision tree rules from heart disease data are complex and large, which increases the time complexity of the risk evaluation model and makes the system slow.

4.5.2 K Nearest Neighbor Model Experimental Results

The purpose of using the K Nearest Neighbor algorithm is to build a risk evaluation model that can predict heart disease at its earliest. The 10-fold cross-validation is used on the training data to get optimal and unbiased results. The performance results like sensitivity, specificity, accuracy, precision, and the error rates of the KNN classifier are derived from the confusion matrix figure 4.8.

The sensitivity of 0.7318% is obtained by using equation (3.11) in figure 4.8, which means the K Nearest Neighbor model can recognize the positive heart disease cases with an accuracy of

73%. Similarly, the amount of patients that were correctly recognized healthy is 0.66% by using the specificity equation (3.12) in figure 4.8; this means the K Nearest Neighbor model can recognize the healthy cases with an accuracy of 66%.

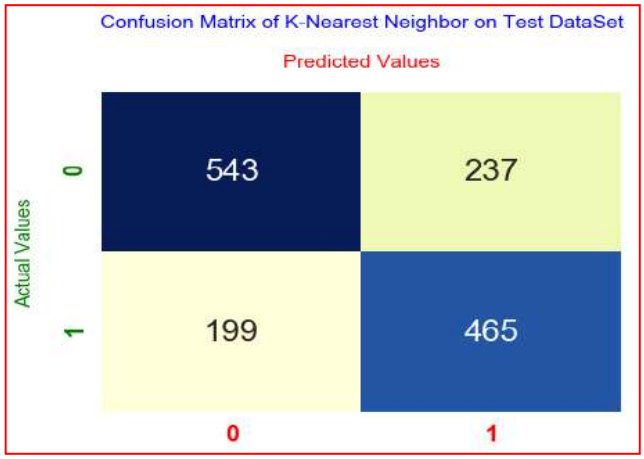

Figure4.8 K Nearest Neighbor Confusion Matrix on the Test Dataset

The overall accuracy of the K Nearest Neighbor model is obtained by putting the confusion matrix figure 4.8 into the equation (3.13) which is equivalent to 0.6980% this means that the K Nearest Neighbor heart disease model's overall performance accuracy (in diagnosing both the diseased and non-diseased heart disease cases) is 69%. Similarly, using equation (3.14) precision of the K Nearest Neighbor model is calculated that is equivalent to 0.696%. If a high precision rate of the K Nearest Neighbor model is obtained, it means the model will obtain the low false-positive rate. The error rate of the developed K Nearest Neighbor model is obtained using the equation (3.15) in the confusion matrix figure 4.8, which is equivalent to 0.3019%.

The AUROC performance measurement is used to see whether the predictive classification model can accurately differentiate between the diseased and healthy cases; however, the poor models will have difficulties in distinguishing between the two classes. Below given figure 4.9 shows the AUROC curve obtained from the K Nearest Neighbor algorithm with an AUROC score of =0.70%.

We simulate the accomplished experimental results of the developed K Nearest Neighbor heart disease risk model with the prevailing research; results show that the K Nearest Neighbor model is not optimal for heart disease prediction because the misclassification rate

is high. Apart from medical domain performance measures, the model performance measures like computational complexity and the comprehensibility of the developed heart disease risk model are calculated, which are also high. The higher values of misclassification rate and model complexity factors restrain its applications because medical prediction models must satisfy greater prediction accuracy and a single misdiagnose can lead to severe consequences like death.

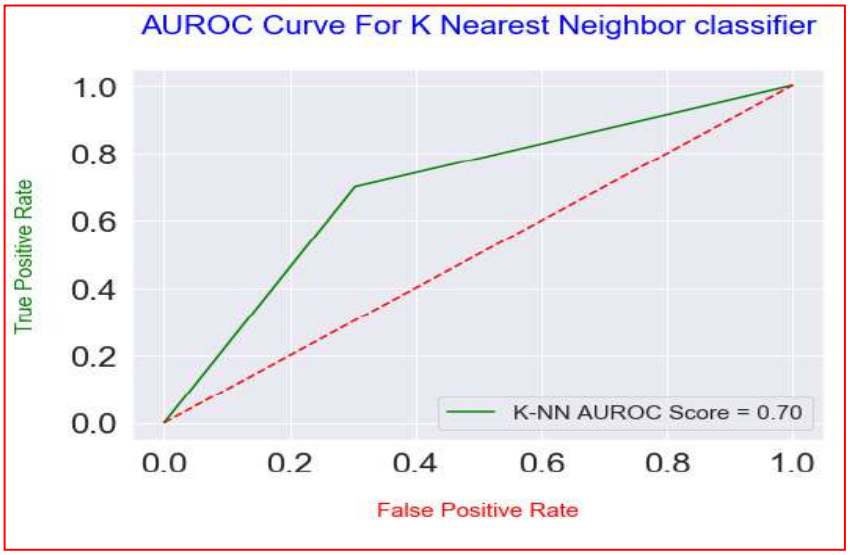

Figure4.9 AUROC by K Nearest Neighbor Model

4.5.3 Support Vector Machine Model Experimental Results

Support Vector Machine separates data into classes based on a hyperplane with maximum margin and searches for the optimal hyperplane between classes to create the support vectors [60] [150]. In this research, the Support Vector Machine is used to develop a risk model that can predict heart disease in its early stages. The performance results of the Support Vector Machine model on the Kashmir heart disease dataset are shown in confusion matrix figure4.10, and the sensitivity, specificity, accuracy, precision and error rates are derived from it which is described as follows:

Using equation (3.11) in confusion matrix figure 4.10 the sensitivity of the Support Vector Machine (SVM) model is calculated as 0.825%; hence the SVM model can recognize the positive heart disease cases with an accuracy of 82%. Similarly, by using equation (3.12) in figure 4.10 the specificity of 0.8152% is obtained, which means the SVM model can

recognize the non-heart disease cases with an accuracy of 81%. Similarly, the overall accuracy of 0.8213%, is obtained by putting the values of the figure 4.10 in equation (3.13), this means that the SVM heart disease model's overall performance (in diagnosing both the diseased and non-diseased heart disease cases) is 82%. Similarly, the precision of 0.8473% is obtained from figure 4.10 using equation (3.14), which means that the SVM model obtained the low false-positive rate. The misclassification rate of the developed SVM model is obtained using the equation (3.15) in figure4.10, which is equivalent to 0.1786%.

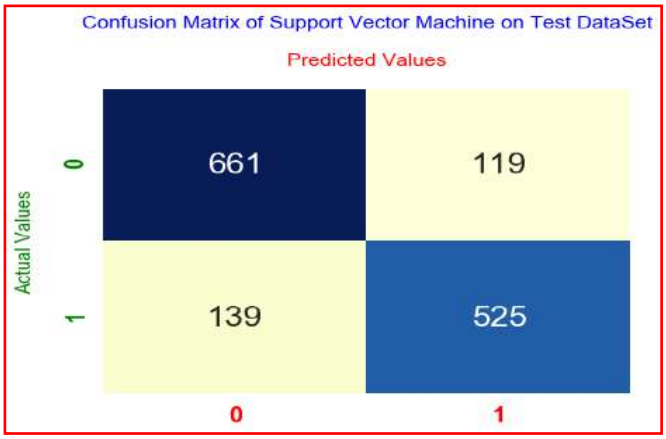

Figure 4.10 SVM Confusion Matrix on Test Dataset

The AUROC performance measurement is used to check the probability curve and measure of separability obtained by the SVM model. Below given figure4.11 shows AUROC obtained from the SVM model with an AUROC score of =0.82%.

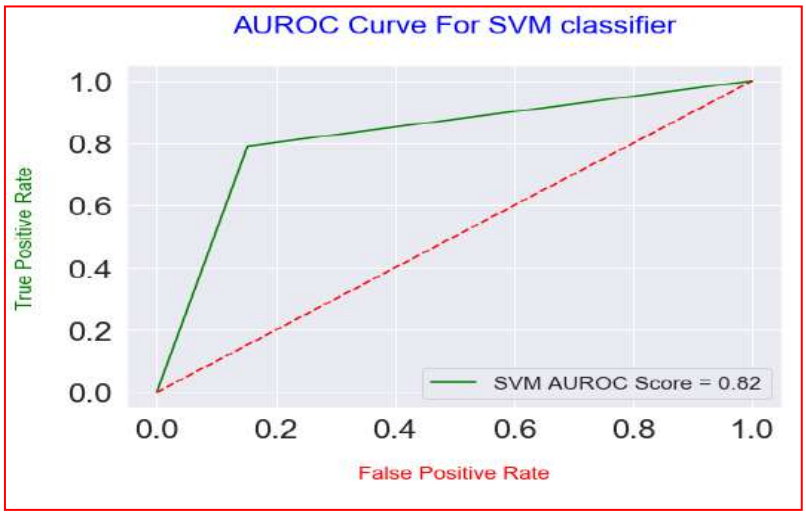

Figure 4.11 AUROC by Support Vector Machine Model

We simulate the accomplished experimental results of the developed Support Vector Machine heart disease model with the prevailing research; the results obtained are to best of our

knowledge greater than the published results in the literature; hence the developed SVM model is used for its practical implementation.

4.5.4 Random Forest Model Experimental Results

The predictive results of the Random Forest model on the Kashmir heart disease dataset are shown in the confusion matrix figure4.12. The sensitivity, specificity, accuracy, precision, and the error rates derived from the figure are explained as follows:

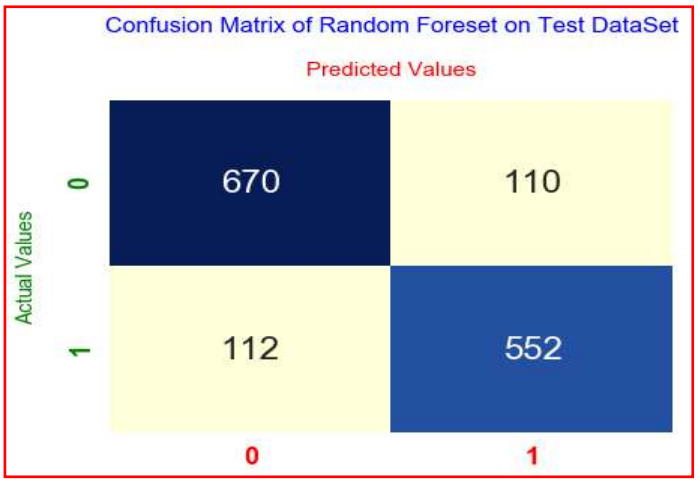

Figure 4.12 Random Forest Model Confusion Matrix on Test Dataset

The random forest model can recognize the positive heart disease cases with a sensitivity of 85% by putting values of figure 4.12 in equation (3.11). Similarly, the number of patients that were diagnosed healthy is equivalent to 0.8338% that is obtained by using equation (3.12) in figure 4.12. The closer the value for this measure is to 1, the better the rules are at identifying healthy patients.

The total accuracy of 0.8462% is achieved by using the equation (3.13) in figure 4.12 this means that the random forest heart disease model's overall accuracy (in diagnosing both the diseased and non-diseased heart disease cases) is 84%, the greater the accuracy percentage, the excellent the model is. Similarly, the precision of 0.8589% is obtained using equation (3.14) in figure 4.12. The closer the value for this measurement is to 1, the greater the chance that those with a positive result will have the disease. The error rate of 0.15% is obtained by putting the values of figure 4.12 in equation (3.15). The lower the percentage of

misclassification rate of the model, the more accurate the model is in identifying the diseased and healthy cases.

The AUROC score is calculated to check the probability curve and measure of separability obtained by the random forest model. The AUROC tells how good the model can differentiate among a diseased and healthy patient. Below given figure4.13 shows AUROC obtained from the random forest model with an AUROC score of =0.85%.

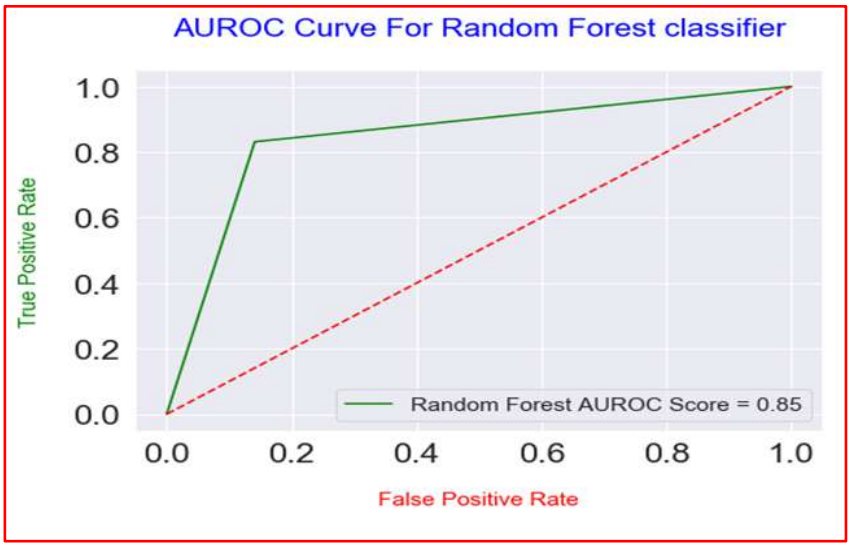

Figure4.13 AUROC by Random Forest Model

We simulate the accomplished experimental results of the developed random forest heart disease model with the prevailing research; the results obtained are to the best of our knowledge greater than the published results in the literature. Hence the proposed random forest model is used for the initial prediction of the heart disease patients.

4.5.5 Naive Bayes Model Experimental Results

Naive Bayes algorithms are a group of basic probabilistic algorithms based on applying Bayes' theorem with strong (naive) independence assumptions between the features [59] [61]. The 10-fold cross-validation is used on the Kashmir heart disease dataset to achieve the maximum accuracy and unbiased results. The performance results of the Gaussian Naive Bayes algorithm are shown in confusion matrix figure4.14, and the sensitivity, specificity, accuracy, precision, and error rates are calculated as follows:

Using equation (3.11) in the confusion matrix figure 4.14 the sensitivity of the Gaussian Naive Bayes model is obtained which is equivalent to 0.7273%; hence the Gaussian Naive Bayes model can recognize the positive heart disease cases with the sensitivity of 72%. Similarly, by using equation (3.12) in the confusion matrix figure 4.14 the specificity of the Gaussian Naive Bayes model is calculated that is equivalent to 0.663% which means the Naive Bayes model can recognize the non-heart disease with 66%.

The overall accuracy of the Gaussian Naive Bayes model is obtained by using the equation (3.13) in the confusion matrix figure 4.14 which is equivalent to 0.696%; this means that the Naive Bayes heart disease model's overall performance (in diagnosing both the diseased and non-diseased heart disease cases) is 69%. Likewise using equation (3.14) in the confusion matrix figure 4.14, the precision of the Gaussian Naive Bayes model is calculated, which is equal to 0.7012%. The error rate of the developed Gaussian Naive Bayes model is obtained by using the equation (3.15) in the confusion matrix figure 4.14, which is equivalent to 0.30%.

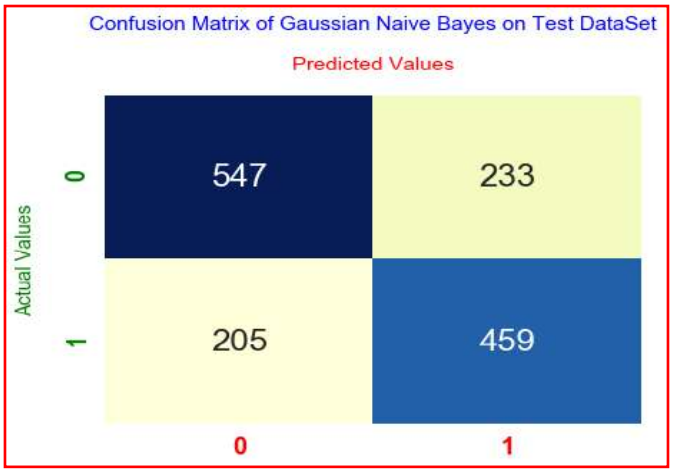

Figure 4.14 Naive Bayes Model Confusion Matrix on Test Dataset

The AUROC score is calculated to check how good the model can distinguish between diseased and healthy patients. Below given figure 4.15 shows the AUROC curve obtained from the Naive Bayes model with an AUROC score of =0.70%.

As per our knowledge, these experimental results of the developed Naive Bayes model are not the best for predicting heart disease because the misclassification rates are higher than the

existing proposed models in the literature. Apart from medical domain performance measures the computational complexity and the comprehensibility of the developed Naive Bayes model are high. The higher values of misclassification rate and model complexity factors restrain its applications because medical prediction models must satisfy greater prediction accuracy and a single misdiagnose can lead to severe consequences.

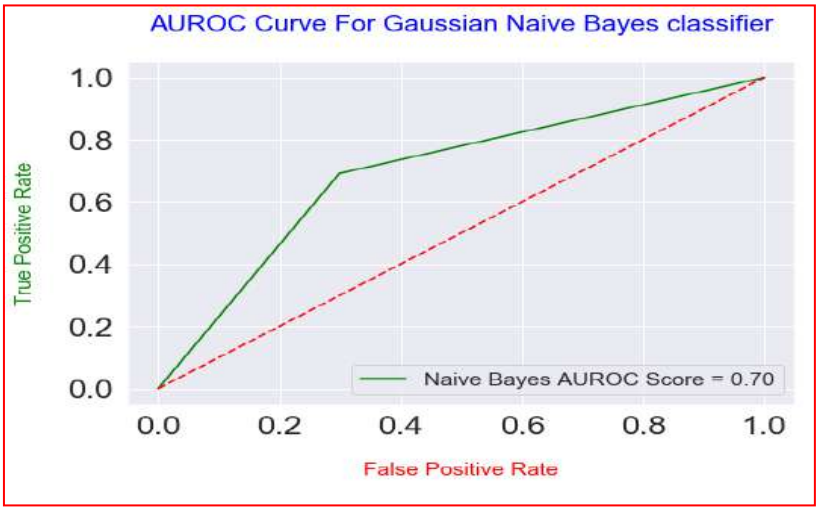

Figure 4.15 AUROC Curve by Gaussian Naive Bayes Model

4.6 Performance Comparison of The Developed Heart Disease Models

This section presents performance and comparison of the Decision Tree, K Nearest Neighbor, Support Vector Machine, Random Forest and the Naive Bayes risk prediction models through different measures as described in the following table 4.3.

Experimental results demonstrate that the random forest model performs most excellent in comparison to other risk models. The performance of the developed heart disease risk evaluation model is tested with the prevailing risk tools which demonstrate that the results are exceptionally encouraging with outstanding predictive accuracy. After the basic assessment of experimental results, it is imperative to cautiously check and assess the data to extract important knowledge, develop best models, and determine optimal factor settings. The results show that the random forest model outperforms other risk evaluation models with an optimal accuracy of 85%, the specificity of 83%, the sensitivity of 85%, precision of 85%, AUROC

score of 85% and with less misclassification rate of only 13%. The accuracy obtained by the random forest is highest for diagnosing heart disease and is not achieved by previous studies.

Table 4.3 Performance Measures of Developed Heart Disease Models

	Performance Measures					
Models	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Error Rate**	**AUROC**
Decision Tree	82%	80%	81%	84%	18%	81%
K Nearest Neighbor	73%	66%	70%	69%	30%	70%
Support Vector Machine	82%	81%	82%	84%	17%	82%
Random Forest	85%	83%	84%	85%	15%	85%
Naive Bayes	72%	66%	69%	70%	30%	70%

The above given figure 4.15 shows the combined AUROC curves of different developed heart disease risk evaluation models. From figure4.16, it is clear that the random forest risk evaluation model has the highest AUROC score of 0.85%, which means the model is highly skillful in predicting the diseased and non-diseased patients.

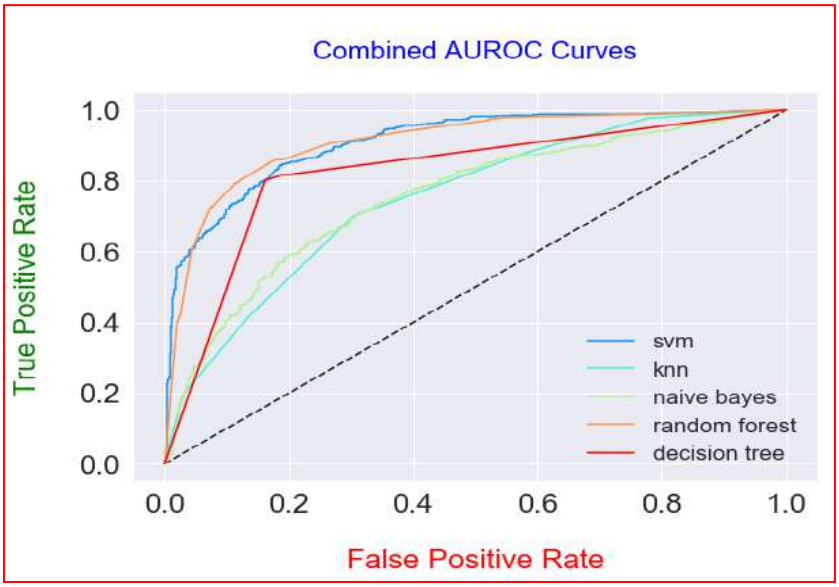

Figure 4.16 Combined AUROCs of the Developed Risk Evaluation Models

4.7 Developing an Accurate Heart Disease Data Mining Model

While developing the predictive data mining model, it is essential to locate the right balance of bias and variance to effectively minimize the test set error. It is necessary to adjust bias and

variance such that we won't get into under-fitted and overfitted problems [151]. If we managed to reduce the bias and variance, then the accurate model can be built [140]. Bias is an error from a faulty assumption in the learning algorithm, and when it is high, the model won't be able to accurately define the relationship between the attributes and the target outputs. However, the variance is an error resulting from fluctuations in the training dataset, and when its value is high; the model will not generalize to capture new data points [151].

The procedure to minimize the error is to reduce the bias and variance, but, it is not easy to reduce them concurrently because minimizing one term leads to an increase in the other term [152]. So it is important to find the balance between the bias and variance (bias-variance trade-off) so that to reduce the total error and have a perfect fit model [154].

$$Total\ Error = Bias\hat{}2 + Variance + Irreducible\ Error \qquad (4.1)$$

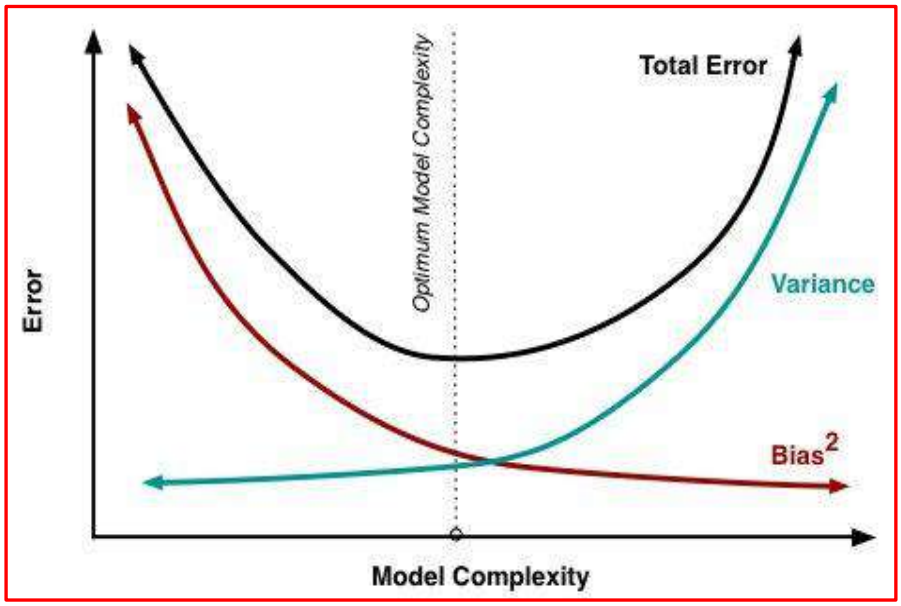

Figure 4.17 Bias and Variance Contributing to the Total Error

In practice, there is not an analytical procedure to find this location; instead, we use the hyperparameter optimization (discussed in chapter V) that helps to manage the behavior of machine learning classifiers when optimizing for performance and finding the right balance between Bias and Variance. Below given figure4.16 shows the total error representation of the supervised learning models. The total error is calculated using equation (4.1).

4.8 Chapter Summary

In this research work, Davis's data mining methodology is followed for the development of a heart disease risk assessment model. In this chapter, the research design is formulated to simplify the research activities and make research productive by the statement of objectives. The Kashmir heart disease dataset is mined using different feature selection techniques to choose considerable heart disease risk attributes. These significant risk attributes are used for data mining techniques for the early prediction of heart disease patients. The data mining classification techniques like Decision Tree, Support Vector Machine, K Nearest Neighbor, Random Forest, and Naive Bayes are applied for the early prediction and identification of heart disease patients. Experimental results demonstrate that the random forest model outperforms other models with the highest model accuracy, low misclassification rate, and low time complexity. The heart disease risk evaluation model will help medical practitioners in early prediction and diagnosis hence would reduce progression to severe complications.

CHAPTER 5

Hyperparameter Optimization of the Heart Disease Risk Evaluation Models

Heart Disease prediction and identification is often challenging due to its underlying complications. To predict heart disease more accurately and efficiently with minimum misdiagnosis, we optimize the hyperparameters of the developed models. Hyperparameter optimization is the process of tuning the optimal hyperparameters for a learning model [155]. It essentially means searching through an enormous universe of possible combinations of hyperparameters for the set that optimizes the desired figure of merit [156]. The model parameters are learned during the training phase; however, the hyperparameters are not directly learned from the training data but are tuned independently [157] [158]. This chapter describes how to optimize the hyperparameters of the developed models build in chapter 4. In this chapter, we also describe the different categories of hyperparameter optimization techniques and the comparison of the heart disease risk evaluation models with and without hyperparameters. Finally, we also discuss different combinations of risk features for cardiac disorder prediction, generated heart disease rules, heart disease expert system evaluation model components, and the developed heart disease risk evaluation model and end up with the chapter summary and conclusion.

5.1 Hyperparameter Optimization Techniques

Hyperparameters are the handles and levels that we draw and turn when developing a machine learning model. Hyperparameters are optimized by searching for different settings to check which values give maximum accuracy [159]. Model optimization is one of the hardest challenges in the implementation because of continuously modifying the code of the model to

decrease the testing error [160] [161]. The pictorial representation of hyperparameter optimization is shown in the below-given figure5.1.

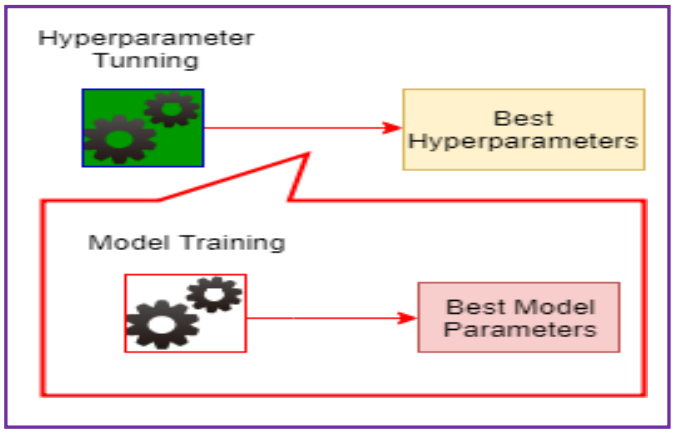

Figure5.1 Hyperparameter Optimization Representation

Hyperparameter tuning is an art to improve the performance of the model, and choosing appropriate hyperparameters will generate the most accurate outcomes and will give highly valuable insights into the data. The hyperparameters help to direct the behavior of the machine learning models when tuned for the performance and when finding the exact stability between bias and variance. Hyperparameter optimization is represented in equation form as:

$$x^{*} = \arg\min_{x \varepsilon X} f(x) \qquad (5.1)$$

Here f(x) represents an objective score to minimize the misclassification ratio evaluated on the validation set; x* is the set of hyperparameters that yield the lowest value of the score, and x can take on any value in the domain X. In simple terms, we want to find the hyperparameters that yield the best score on the validation set metric using different hyperparameter techniques.

5.1.1 Grid Search Hyperparameter Optimization

The conventional procedure of conducting hyperparameter optimization has been Grid search, which is a comprehensive search of candidate parameter values over all feasible values in the defined search space. After all conceivable parameter combinations of the model are examined; the finest combination will be retained. Grid search trains the algorithm for every

single permutation by utilizing the two set of hyperparameters (learning rate and a number of layers) and determines the accuracy using the cross-validation technique [162]. This validation technique gives assurance that the trained model obtains a maximum of the patterns from the dataset. Below given figure5.2 shows the grid search method layout.

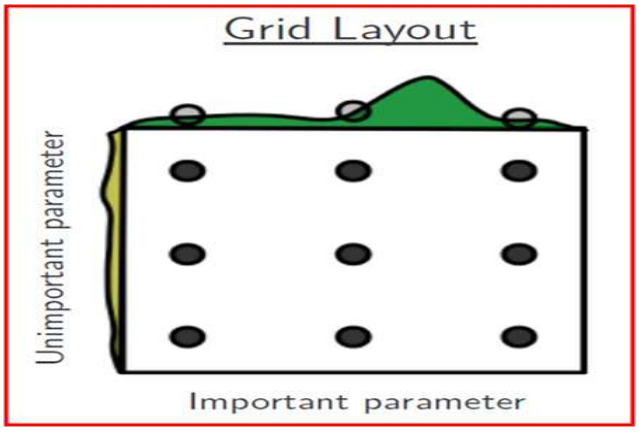

Figure5.2 Grid Search Layout

The Grid search method is a simpler method to use, but it is a costly approach and suffers if data have high dimensional space (the *curse of dimensionality*) [163].

5.1.2 Random Search Hyperparameter Optimization

In random search optimization, hyperparameter values are randomly selected from the defined search space. Random sampling allows the search space to include both discrete and continuous hyperparameters [164]. Random Search optimization is a parallel approach that permits the inclusion of previous knowledge by determining the distribution from which to sample. Below given figure5.3 shows the layout of random search optimization [165].

Grid and Random search hyperparameter optimizations are completely uninformed by past evaluations, and as a result, often spend a significant amount of time evaluating “bad” hyperparameters [166]. The Grid and Random search optimization techniques are based on experienced machine learning practitioners and are heuristic techniques. In these techniques, human expertise won’t attain a near-optimum setting of hyperparameters due to the mishandling of high dimensional data and can easily misinterpret when trying to tune multiple

hyperparameters [167]. Due to these drawbacks associated with the Grid and Random search hyperparameter optimization techniques, we prefer to use the Bayesian hyperparameter Optimization for the development of the heart disease risk evaluation model.

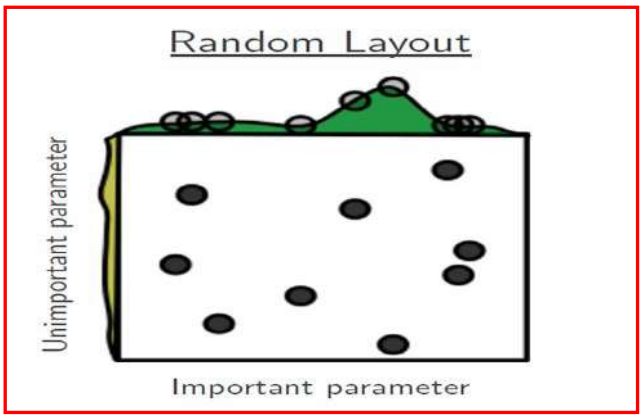

Figure 5.3 Random Search Layout

5.1.3 Bayesian Hyperparameter Optimization

In Grid and Random Search hyperparameter optimization, the model parameters are tuned randomly, and the new parameter combination is independent of all the previous trails executed [168]. However, the automatic hyperparameter optimization technique uses previous trail knowledge to make the best choice for the next parameter settings. Bayesian optimization first gathers the performance at various configurations then makes some conclusion and decides what configuration to try next. The reason is to decrease the number of checks while finding the best optimum [169].

In this research work for Bayesian Optimization, the Single Cross-validation (S-CV) methodology shown in Figure5.4 is used with each technique. The heart disease dataset is divided into k stratified folds when hyperparameters are optimized and executed. The Decision Tree, Support Vector Machine, and K Nearest Neighbor classifiers are trained on training dataset for every single solution found by the techniques. One part of the dataset is used for the validation of the model, and the remaining parts of the dataset are utilized for testing purposes. The validation and test performances are measured through the model

induced with the training dataset and the values of the hyperparameters found by the optimization technique. This process is reiterated for all k combinations in single cross-validation. The average validation accuracy is then used as the fitness value, which directs the search process. Finally, the individual with the maximum validation accuracy is returned (with its hyperparameters values), and the technical performance is considered the average test accuracy of the individual.

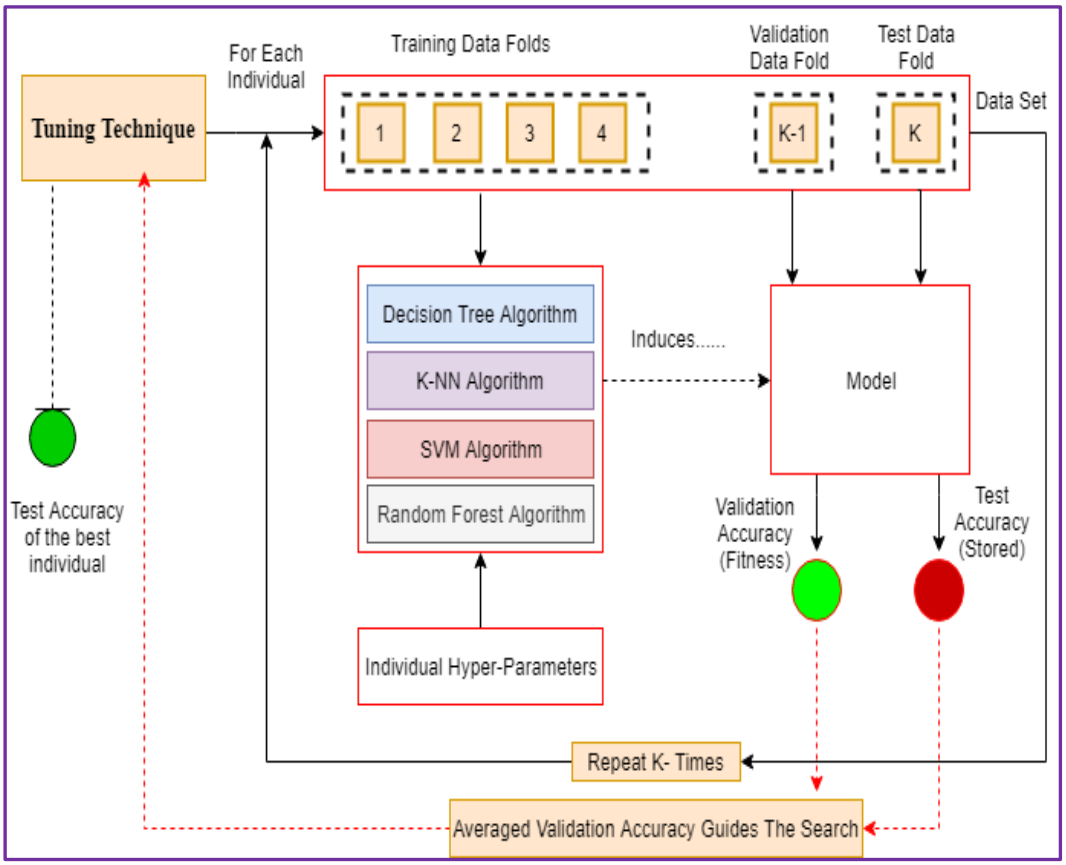

Figure5.4 Single Cross-Validation Methodology for Hyperparameter Optimization

5.2 Optimizing the Heart Disease Risk Evaluation Models

Data mining classifiers include numerous hyperparameters whose values influence the predictive performance of the induced models in complicated ways. Because of the large number of possibilities for the hyperparameter settings, we lack insight into how to professionally search this immense space of configurations. There are tons of potential parameters to tune on every model, and all the parameters are valuable, but it is required to

select the significant subset of parameters. Below given subsections describe how various developed risk evaluation models are optimized.

5.2.1 Decision Tree Hyperparameter Optimization Model

The most significant hyperparameters of the decision tree model are tuned to obtain the optimal accuracy; however, we should be careful to validate them on test data fold to avoid overfitting. The significant hyperparameters which improve the performance of the decision tree model that need to be tuned are as follows [170]:

i. **Max Depth:** The max depth hyperparameter of the decision tree algorithm represents how deep the tree can be. A decision tree with maximum depth captures more information from the data. In this research work, we fit a decision tree algorithm with a max depth of 100 and also plot the training and test AUROC scores. It is found that the developed model overfits if max depth values are set high.
ii. **Min Samples Split:** This hyperparameter of the decision tree algorithm represents the least amount of samples needed to divide an internal node. The min samples split vary from one sample at each node to considering all of the samples at each node. If the value of the min samples split parameter is increased, the tree becomes more constrained as it has to consider more samples at each node. However, if all of the samples at each node are considered, then the model underfits.
iii. **Min Samples Leaf:** This hyperparameter of the decision tree algorithm signifies the minimum number of samples needed to be at a leaf node. If the parametric value of min samples leaf is increased, then it leads to an underfitting problem.
iv. **Max Features:** The max features hyperparameter of the decision tree algorithm signify the number of features to include when searching for the best split; however if a high number of max features are selected then it causes an overfitting problem.

v. **Criteria:** The Criteria hyperparameter feature of the decision tree decides how the impurity of a split will be measured. The default value for criteria hyperparameter is "Gini" however; it can be set as "Entropy."

After tuning the hyperparameters of the decision tree model, we obtain the results as shown in the below-given table 5.1. The permutations and combinations of the optimized decision tree model show different results; however, we describe only those combinations which provide the highest accuracies.

Table5.1 Experimental Results of the Optimized Decision Tree Model

Max Depth	***Min Samples Split***	***Min Samples Leaf***	***Max Features***	**Criterion**	**Accuracy**
10	18	15	Auto	Entropy	81%
15	25	12	Auto	Gini	82%
18	11	10	Sqrt	Gini	72%
20	15	20	Sqrt	Gini	83%
25	10	50	Auto	Entropy	71%
30	12	30	Auto	Entropy	74%
35	22	25	Sqrt	Entropy	75%
40	8	14	Sqrt	Gini	78%
45	5	16	Auto	Entropy	73%
50	14	Not Used	Auto	Gini	78%
70	17	18	Auto	Entropy	75%
80	13	Not used	Auto	Entropy	84%
100	20	Not used	Sqrt	Entropy	84%

We applied the developed optimized decision tree model for the initial prediction and identification of cardiac disorder patients. The performance results of the model are shown in the confusion matrix figure 5.5.

From figure 5.5, we derive the True Positive Rate (Recall), True Negative Rate (Specificity), Accuracy, Precision, and Misclassification Rate, which are described as follows:

Using equation (3.11) in figure 5.5 the True Positive Rate of 0.833% is achieved which means the developed optimized decision tree model can recognize the positive heart disease cases with an efficiency of 83%. Similarly, using equation (3.12) in figure 5.5, the True Negative Rate of 0.80% is achieved which means the model can recognize the non-diseased cases with an efficiency of 80%. The accuracy of the optimized decision tree model is obtained using equation (3.13) in confusion matrix figure 5.5 that is calculated to 0.8185%; this describes that the decision tree model's overall accuracy in diagnosing both the diseased and healthy cases is 81%. Similarly the Precision of 0.8294% is obtained using the equation (3.14) in the confusion matrix figure 5.5; this means the model has a low false-positive rate. The misclassification rate of the proposed decision tree model is obtained using the equation (3.15) in the confusion matrix figure 5.5 which is equivalent to 0.18%.

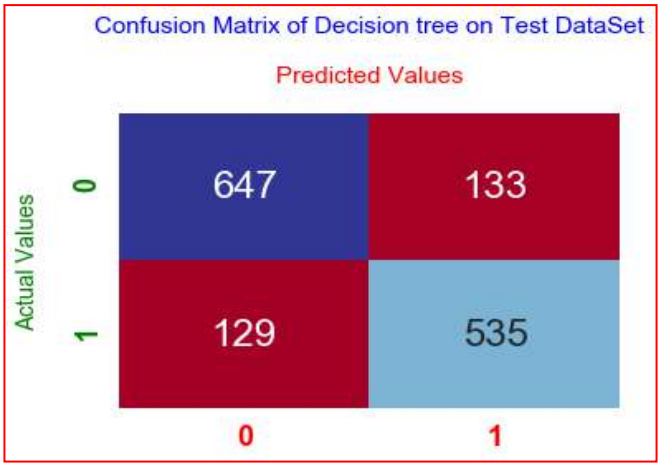

Figure5.5 Confusion Matrix of the Optimized Decision Tree Model

The AUROC performance measurement is used to see how good the model can differentiate among diseased and healthy patients. Better models perfectly differentiate among the diseased and non-diseased victims; however, the poor models get difficulties in distinguishing them. The below-given figure 5.6 shows AUROC obtained from the optimized decision tree model with an AUROC score of =0.82%. We simulate the accomplished experimental results of the

optimized decision tree heart disease model with the prevailing research; the results obtained are to the finest of our observation more than the published results in the literature. Hence we use the developed, optimized decision tree model for predicting heart disease patients; however, further improvement in model performance is required.

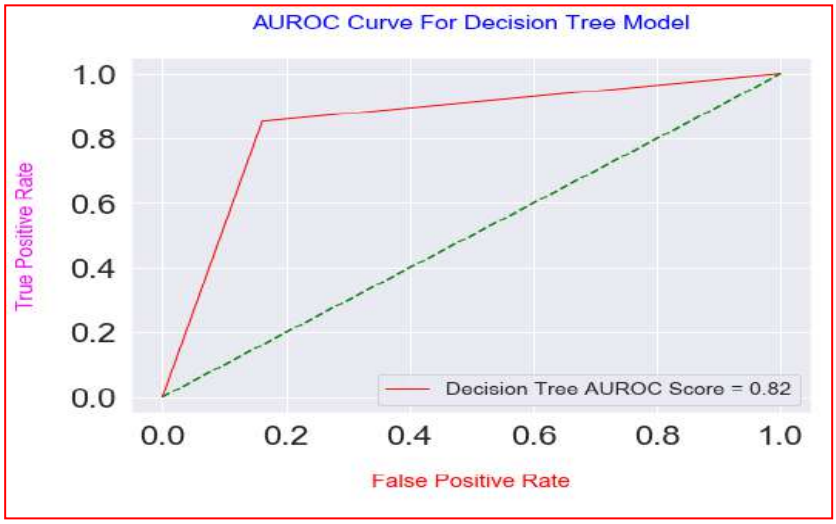

Figure 5.6 AUROC of the Optimized Decision Tree Model

5.2.2 K Nearest Neighbor Hyperparameter Optimization Model

K Nearest Neighbor classifies an unidentified neighbor based on majority votes. Each neighbor can be given equal weight, or the vote can be based on the distance. The hyperparameter optimization of the K Nearest Neighbor algorithm is performed to search the most excellent model that shows the maximum accuracy and the lowest error on the test dataset. The most important hyperparameters of the K Nearest Neighbor classifier are explored to check how they influence the model in terms of overfitting and underfitting. The primary hyperparameters of the K Nearest Neighbor model that are tuned are described as follows:

i. The number of neighbors' ***k***.

ii. The similarity function or the distance metric.

These two hyperparameters of the K Nearest Neighbor algorithm significantly influence the accuracy of the classifier. In the K Nearest Neighbor algorithm, the best K is the one that

provides the lowest test error rate; hence, the repeated calculations of the test error for other values of K are carried out. The test error rate is measured by reserving a subset of the training data from the fitting process. This subset (validation set) is used to choose the correct level of flexibility of the algorithm.

The experimental results of the optimized K Nearest Neighbor model are shown in the below-given table5.2. We use the different permutations and combinations of the K Nearest Neighbor model to attain the maximum accuracy. The parameter combinations which result in the highest accuracies are described in the table. We can see that when the **metric** attribute is configured as "Minkowski" and the **weight** attribute is configured as "Uniform" the performance of the K Nearest Neighbor model degrades to 67%. The 'best score' function is used to check the accuracy of the model because the 'best score' outputs the mean accuracy of the scores obtained through cross-validation.

Table5.2 Experimental Results of the Optimized K NN Model

Leaf Size	**Metric**	**Neighbors**	**Weights**	**Accuracy**
5	Euclidean	5	Distance	82%
10	Minkowski	11	Uniform	67%
30	City Block	13	Distance	85%
25	Euclidean	9	Distance	70%
15	Minkowski	7	Uniform	72%
20	City Block	11	Uniform	68%
12	Euclidean	15	Distance	75%
16	Minkowski	13	Uniform	77%
18	Minkowski	7	Uniform	80%
28	Euclidean	9	Distance	82%

When the value of k is set small, the low bias and high variance are obtained; however, when k is large, then high bias and low variance is achieved. So we configure the value of the k parameter in a sweet position. Through the Optimization search, the model accuracy is

improvised up to 85%. The experimental results show that when the hyperparameter combinations are set to [Leaf Size= 30, Metric= City Block, Weights=13], then the highest accuracy of 85% is achieved. The performance results of the proposed K Nearest Neighbor model are shown in the confusion matrix figure5.7.

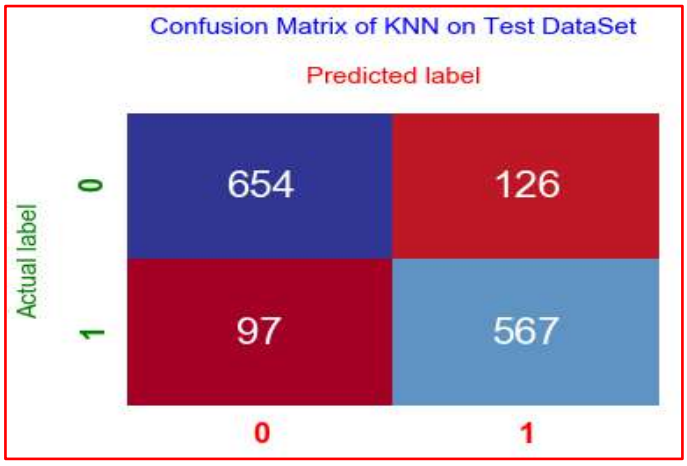

Figure 5.7 Confusion Matrix of the Optimized K NN Model

From the figure5.7, we derive the sensitivity, specificity, Accuracy, Precision, and misclassification rates, which are described as follows:

Using equation (3.11) in figure 5.7, we obtain the True Positive Rate of 0.87% hence K Nearest Neighbor model can recognize the positive heart disease cases with a sensitivity of 87%. Similarly, the True Negative Rate of 0.81% is achieved by putting values of figure 5.7 in equation (3.12) the result means that the K Nearest Neighbor model can recognize the non-diseased cases with the efficiency of 81%. The overall accuracy of 0.84% is obtained by putting the values of figure 5.7 in equation (3.13) these values means the K Nearest Neighbor model's overall performance in predicting both the diseased and healthy cases. To obtain the Precision of the optimized K Nearest Neighbor model we put the values of the confusion matrix figure 5.7 in equation (3.14) which is 0.83%; this means that the optimized K Nearest Neighbor model has the low false-positive rate. The misclassification rate of the developed K Nearest Neighbor model is obtained by putting the values of figure 5.5 in the equation (3.15), which is equivalent to 0.15%.

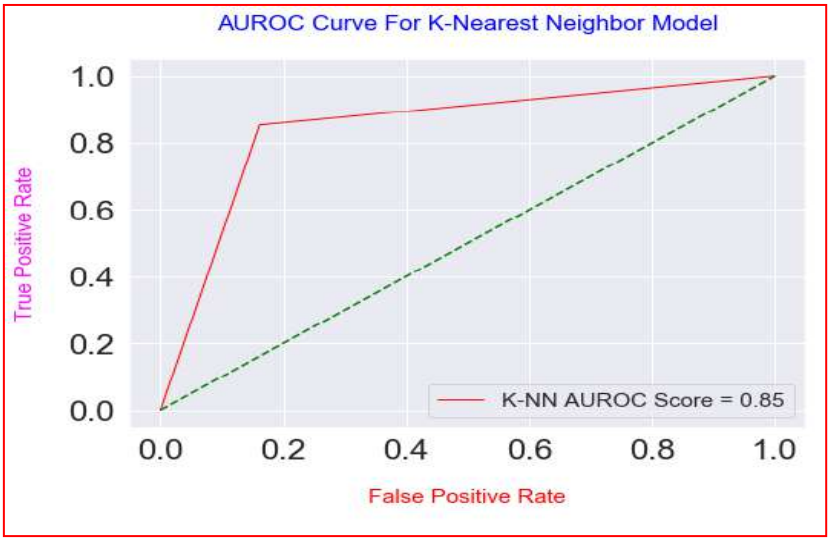

Figure 5.8 AUROC of the Optimized K NN Model

The AUROC performance measurement is used to check the probability curve and measure of separability obtained by optimized K Nearest Neighbor model, below given figure5.8 shows AUROC curve obtained from the K Nearest Neighbor model with an AUROC score of =0.85%. We simulate the accomplished experimental results of the optimized K Nearest Neighbor heart disease model with the prevailing research; the results obtained are to the best of our knowledge greater than the published results in the literature. Hence we can use the developed, optimized K Nearest Neighbor model for predicting the heart disease patients; however, further improvement in model performance is required.

5.2.3 Support Vector Machine Hyperparameter Optimization Model

The most significant hyperparameters of the Support Vector Machine model are tuned to obtain the highest accuracy for the initial identification of heart disease; however, we should be careful to validate them on the test dataset. The hyperparameters of the Support Vector Machine model that are optimized are as follows:

i. **Kernel:** The Kernel hyperparameter chooses the type of hyperplane to separate the data. The kernel hyperparameter changes the given input data into the requisite form. In the Support Vector Machine, various categories of kernel functions are available like

Polynomial, Radial Basis Function (RBF), Linear and Sigmoid. In this research, all types of kernel functions are used; however, the RBF kernel function provided significant results.

ii. **Regularization:** The Regularization hyperparameter of the Support Vector Machine is the penalty parameter, which represents misclassification or error term. The misclassification means how much error is bearable. If a regularization hyperparameter is set with a smaller value, then it creates a small-margin hyperplane, and if it is set with a larger value, then it creates a larger-margin hyperplane.

iii. **Gamma:** The Gamma hyperparameter of the Support Vector Machine classifier is used for non-linear hyperplanes. If Gamma hyperparameter is set with a smaller value, then it loosely fits the training dataset; however, a higher value exactly fits the training dataset, which causes overfitting.

The behavior of the developed Support Vector Machine risk assessment model is extremely sensitive to the gamma hyperparameter. The different accuracies achieved after tuning various hyperparameters of the Support Vector Machine model are shown in the below-given table 5.3. The “Kernel” and “Regularization” hyperparameters of the Support Vector Machine model are configured with the best set of permutations and combinations to obtain the optimal accuracy.

It is found that when the Kernel hyperparameter values are set to (linear or sigmoid or Sqrt), the time complexity of the risk model increases. The experimental results show that when the Support Vector Machine hyperparameter combinations are set to [Kernel=rbf, Gamma=0.1, Regularization=1.0], the highest accuracy of 81% is achieved. The results of the model are shown in the confusion matrix figure 5.9.

From the figure5.9, we derive the sensitivity, specificity, Accuracy, Precision, and misclassification rate, which are described as follows:

Table5.3 Experimental Results of the Optimized SVM Model

Kernel	***Gamma***	***Regularization***	**Accuracy**
Linear	0.001	0.11	71%
Sigmoid	0.1	1.0	70%
Sqrt	0.00001	0.001	68%
rbf	0.1	1.0	81%
Linear	0.001	0.001	72%
rbf	0.0001	0.1	80%
Linear	0.01	0.10	73%
rbf	0.0011	0.0001	78%
Sqrt	0.0001	0.010	75%
Sqrt	0.1	0.11	76%
Sigmoid	0.01	1.0	74%
Linear	0.0001	1.0	71%
Sigmoid	0.010	0.11	77%
Rbf	0.11	0.0001	69%
Sqrt	0.10	0.001	73%

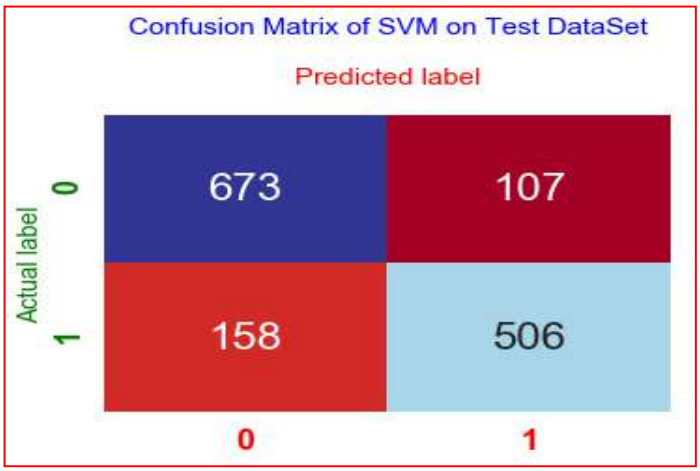

Figure 5.9 Confusion Matrix of the Optimized SVM Model

Putting the values of confusion matrix figure 5.9 into the equation (3.11) the True Positive Rate of 0.80% is achieved; which means the Support Vector Machine model can recognize the

positive heart disease cases with 80% efficiency. Similarly, the True Negative Rate of 0.82% is achieved by putting the values of confusion matrix figure 5.9 into the equation (3.12), these results mean the optimized Support Vector Machine model can recognize the non-diseased cases with an efficiency of 82%.

The overall accuracy of 0.81% is obtained by putting the values of confusion matrix figure 5.9 into the equation (3.13) this means that the Support Vector Machine model's overall performance in predicting both the diseased and healthy cases is optimal. The Precision of the optimized Support Vector Machine model is equivalent to 0.86%; this means that the model has a low false-positive rate. The misclassification rate of the developed optimized SVM model is 0.18%.

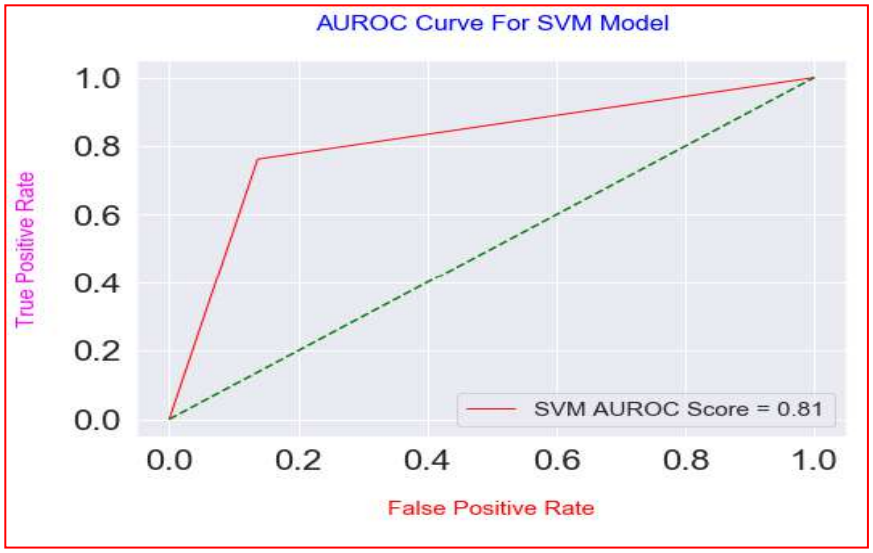

Figure 5.10 AUROC of the Optimized SVM Model

The AUROC performance measurement is used to see how exact the model can distinguish between diseased and non-diseased patients. Below given figure5.10 shows the AUROC curve obtained from the optimized Support Vector Machine with an AUROC score of =0.81%. The accomplished experimental results of the optimized Support Vector Machine heart disease risk model are simulated with the prevailing research; the experimental results of the optimized Support Vector Machine model are not best for predicting heart disease. Its usage is restrained for the practical implementation because the Support Vector Machine heart

disease risk evaluation model's time complexity is high, and it causes an overfitting problem that leads to heart disease misdiagnosis.

5.2.4 Random Forest Hyperparameter Optimization Model

A random forest classifier fits various decision trees on a variety of sub-samples of the dataset and uses averaging to improve the predictive accuracy and control the overfitting problem. In this research for the risk model development, the most influential hyperparameters of the random forest classifier are explored and configured, which are discussed below:

i. **N Estimators:** This hyperparameter of the random forest algorithm represent the total number of trees in the forest. The larger the numbers of trees are in the random forest, the more accurately it can learn from the data. In this research work, the execution process is stopped at 32 trees because adding more number of trees reduces the model performance.

ii. **Max Depth:** The Max Depth hyperparameter represents how deep each tree in the forest can be. A tree with maximum depth captures more information from the data. In this research work, we fit each tree with max depth hyperparameter value ranging in between 1 to 32 and draw the training and test errors; however, the developed random forest model overfits for large values.

iii. **Min Samples Split:** This hyperparameter of the random forest algorithm represents the minimum number of samples required to split an internal node. The min samples split vary from one sample at each node to considering all of the samples at each node. If the value of the min samples split parameter is increased, the tree becomes more constrained because then it has to include more samples at every node. Experimental results show that when all of the samples at each node are used, the model does not learn enough from the data and leads to the underfitting problem.

iv. **Min Samples Leaf:** This hyperparameter represents the minimum number of samples needed to be at a leaf node. If the parametric value of this hyperparameter is set high, then it leads to underfitting.

v. **Max Features:** This hyperparameter signifies the number of features to include when searching for the best split; however setting the value of max features large causes an overfitting problem.

After tuning the hyperparameters of the random forest model, we obtain the performance results which are described in the below-given table 5.4.

Table5.4 Experimental Results of the Optimized Random Forest Model

Criterion	**Max Depth**	**Max Features**	***N Estimators***	**Min Samples Leaf**	**Accuracy**
Gini	70	Not Used	Not Used	Not Used	85%
Entropy	60	Auto	Not Used	Not Used	86%
Gini	50	Auto	100	Not Used	87%
Entropy	80	Auto	100	100	73%
Gini	100	Auto	100	50	76%
Entropy	30	Not Used	80	60	80%
Gini	40	Not Used	90	40	78%
Gini	25	Auto	70	30	75%
Entropy	20	Auto	40	25	82%
Entropy	35	Auto	30	20	81%
Gini	45	Not Used	60	35	80%

The permutations and combinations of the optimized random forest model show different results; however, we describe only those parametric combinations which provide the highest accuracies. The experimental results show that when the hyperparameter combinations are

configured as [Criterion= Gini, Max Depth= 50, Max Features= Auto, N Estimators=100] the highest accuracy of 87% is achieved.

The optimized random forest model is applied for the initial prediction and identification of heart disease. The performance metrics results of the model are shown in the confusion matrix figure 5.11.

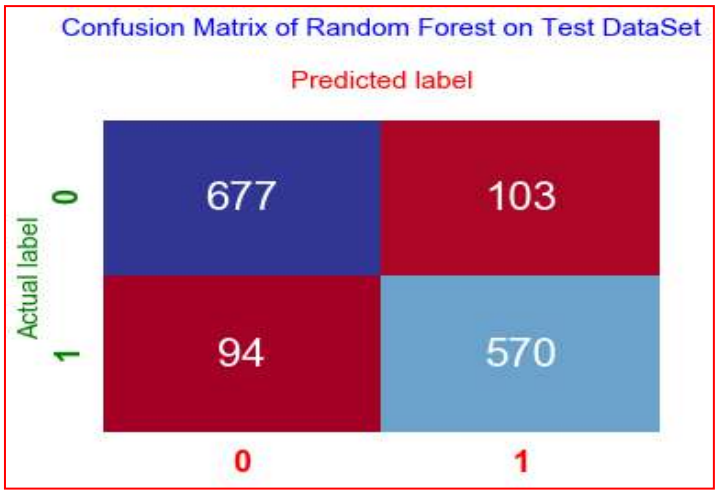

Figure 5.11 Confusion Matrix of the Optimized Random Forest Model

From the figure5.11, we derive the sensitivity, specificity, Accuracy, Precision, and misclassification rate, which are described as follows:

Putting the values of optimized random forest confusion matrix figure 5.11 into the equation (3.11) we obtain the True Positive Rate of 0.87%; hence the optimized random forest model can recognize the positive heart disease cases with the efficiency of 87%. Similarly, by putting the values of figure 5.11 into the equation (3.12), we get the True Negative Rate of 0.84% that means the optimized random forest model can recognize the non-heart disease cases with the efficiency of 84%. The accuracy of the optimized random forest model is obtained by using the equation (3.13) into the confusion matrix figure5.11 which after calculations is equivalent to 0.86% which is the overall performance in diagnosing both the diseased and healthy heart disease cases. The Precision of 0.86% is achieved; this means that the optimized random forest model has a low false-positive rate. The misclassification rate of

the developed, optimized random forest model is obtained by putting the values of confusion matrix figure 5.11 into the equation (3.15), which is equivalent to 0.13%.

We also use the AUROC performance metric to check the probability curve and measure of separability achieved by an optimized random forest model. Below given figure5.12 shows AUROC obtained from the random forest model with an AUROC score of =0.86%.

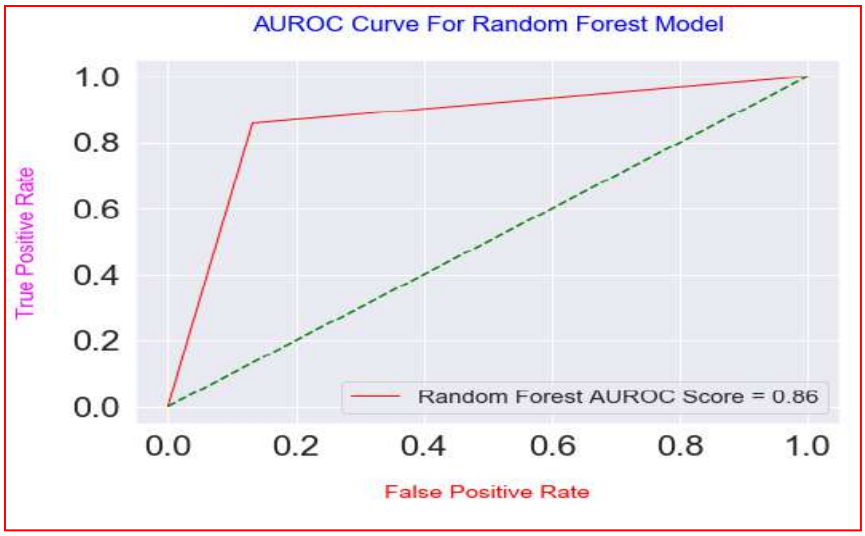

Figure 5.12 AUROC of the Optimized Random Forest Model

We simulate the accomplished experimental results of the optimized random forest heart disease model with the prevailing research; the results obtained are to the best of our knowledge greater than the available results in the literature; however further improvement in model performance is required. Hence we use the developed, optimized random forest model for the initial prediction of heart disease victims.

5.3 Performance Comparison among Developed Optimized Risk Models

In this section, an assessment and comparison of the optimized models are described and to test the performance of developed, optimized heart disease risk models different measures are used which are described in the below-given table 5.5.

The performance outcome demonstrates that the optimized random forest model outclasses other risk models. The performance of the developed model is verified with prevailing designs which demonstrate that the outcomes are promising with magnificent predictive function.

After a precarious assessment of experimental results, we recognize that it is essential to precisely examine and calculate the data to extricate precious knowledge and develop the models.

Table5.5 Performance Measures of Developed Optimized Heart Disease Models

	Performance Measures of the Models					
Models	**TPR**	**TNR**	**Accuracy**	**Precision**	**Error Rate**	**AUROC**
Decision Tree	0.83%	0.80%	0.82%	0.82%	0.5%	0.82%
K Nearest Neighbor	0.87%	0.81%	0.84%	0.83%	0.15%	0.85%
Support Vector Machine	0.80%	0.82%	0.82%	0.86%	0.18%	0.82%
Random Forest	0.87%	0.84%	0.87%	0.86%	0.13%	0.86%

The outcomes show that the random forest model obtains the optimal accuracy of 87%, with a minimum misclassification rate of only 0.13%. Figure 5.13 shows the combined AUROC curves of different optimized heart disease risk evaluation models. Random forest heart disease risk evaluation model has the highest AUROC score of 0.87 %, which means the model has the best capability to differentiate among the diseased and non-diseased heart victims; however, further, improvement is needed.

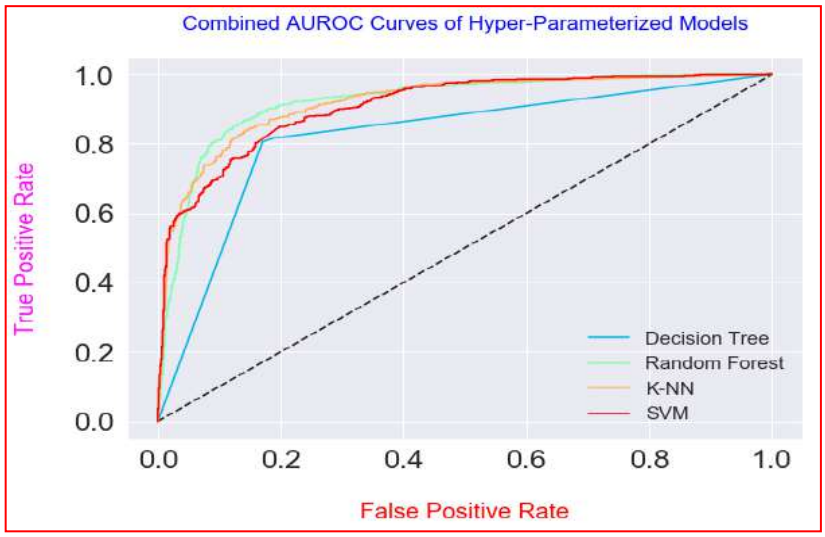

Figure 5.13 Combined AUROC of the Optimized Risk Evaluation Models

5.4 Performance Comparison between the Default and Optimized Risk Evaluation Models

Although heart disorder is the principal source of mortality across the globe; however, it has also been identified as among the most preventable and controllable disease. At least 80% of heart disease could be avoided by a healthy diet, regular bodily exercise, and evasion of tobacco products. The early detection and treatment are aimed to decrease progression to the expensive and costly illness of heart disease. Keeping these things under consideration, the heart disease risk evaluation model is developed using different data mining methods for the initial prognosis of the disease with high predictive power. To improve the performance and minimize the misclassification rate of the heart disease risk model, we optimize the hyperparameters of the models. Below given table5.6 describes the comparison among the default and the optimized heart disease risk models.

Table 5.6 Performance Comparison of Developed Heart Disease Risk Models

Performance Measures	Comparison among Different Heart Disease Risk Evaluation Models								
	Decision Tree		**K NN**		**SVM**		**Random Forest**		**Naive Bayes**
	DMP	***HPT***	***DMP***	***HPT***	***DMP***	***HPT***	***DMP***	***HPT***	***DMP***
TPR	0.82	0.83	0.73	0.87	0.82	0.80	0.85	0.87	0.72
TNR	0.80	0.80	0.66	0.81	0.81	0.82	0.83	0.84	0.66
Accuracy	0.81	0.82	0.70	0.84	0.82	0.82	0.84	0.87	0.69
Precision	0.84	0.82	0.69	0.83	0.84	0.86	0.85	0.86	0.70
Error Rate	0.18	0.05	0.30	0.15	0.17	0.18	0.5	0.13	0.30
AUROC	0.81	0.82	0.70	0.85	0.82	0.82	0.85	0.86	0.70

TPR means (True Positive Rate), TNR means (True Negative Rate), DMP means (Default Model Parameters), and HPT means (Hyperparameter Tuning). The experimental results illustrate that the random forest model outperforms other models both on default and hyperparameter settings. The performance of the developed cardiac disorder risk evaluation model is verified with prevailing designs which demonstrate that the outcomes are promising

with magnificent predictive function. After a precarious assessment of experimental results, we recognize that it is essential to precisely examine and calculate the data to extricate valuable knowledge and develop the model.

5.5 Different Combinations of Risk Features for Early Heart Disease Prediction and Identification

In this section, the results of different permutations of risk features for the untimely prognosis and identification of heart disease are presented. Below given table5.7 demonstrates the performance of various combinations of the non-invasive features in early heart disease predictions. The combinations of Systolic BP, Diastolic BP, Heredity, and Age show the best accuracy of 77.3% obtained by the decision tree model. We also measure the sensitivity and specificity of all the attribute combinations. Here sensitivity is most effective in diagnosing sick cases to provide proper care. By adding BMI (Height and Weight) attribute with the combination of [Age, Systolic BP, Diastolic BP, and Heredity], risk features accuracy is increased up to 78.9% by the random forest model. However, further combinations of the risk attributes with different permutations and combinations decrease the accuracy.

An optimal set of predictive risk rules are generated using the above-derived attribute combinations, which help in the initial prognosis and recognition of heart disease victims. The generated heart disease risk evaluation rules are pruned, evaluated and validated by different medical domain experts; however, their use is restricted as the extracted rules are inductive because they are based on the specific ethnic heart disease dataset.

Table5.7 Integrating Different Non-Invasive Heart Disease Risk Factors

Techniques	Risk Attributes	Sensitivity	Specificity	Accuracy
Decision Tree	Systolic BP, Diastolic BP, Age, Heredity	78%	80%	77.3%
	Systolic BP, Diastolic BP, Age, BMI	72%	70%	70.9%
	Age, Healthy Diet, BMI	68%	61%	63.3%
	Systolic BP, Diastolic BP, Age, Physical Activity	53%	60%	58.6%

	Healthy Diet, BMI, Physical Activity, Age	58%	41%	50.9%
	Healthy Diet, Physical Activity, Age, Systolic BP, Diastolic BP	45%	43%	42.5
	Physical Activity, Age, Healthy Diet, BMI, Systolic BP, Diastolic BP	38%	30%	38.2%
	Age, Physical Activity, Smoking, Systolic BP, Diastolic BP, Healthy Diet, Alcohol Consumption, BMI	30%	28%	42.7%
K Nearest Neighbor (KNN)	Age, Healthy Diet, Alcohol Consumption, Smoking	42%	45%	38.2%
	Age, BMI, Healthy Diet	70%	60%	67.9%
	Age, BMI, Alcohol Consumption, Smoking, Sex	52%	50%	48.9%
	BMI, Systolic BP, Diastolic BP, Age, Physical Activity	38%	35%	42.7%
	BMI, Systolic BP, Diastolic BP, Age	68%	74%	72.5%
	Age, Systolic BP, BMI, Diastolic BP, Heredity	68%	70%	72.8%
Random Forest	Systolic BP, Diastolic BP, Age, Healthy Diet, Smoking	51%	48%	45.4%
	BMI, Age, Systolic BP, Diastolic BP, Heredity	72%	78%	78.9%
	Alcohol Consumption, Physical Activity, Age, Systolic BP, Diastolic BP, BMI, Smoking, Healthy Diet	35%	45%	58.7%
	Age, Sex, Physical Activity, BMI,	32%	34%	40.8%
	Age, Sex, Physical Activity, BMI, Systolic BP, Diastolic BP	39%	45%	42.6%
Support Vector Machine (SVM)	Systolic BP, Diastolic BP, Age	72%	62%	76.1%
	Systolic BP, Diastolic BP, Age, BMI, Heredity	70%	78%	75.2%
	Healthy Diet, Age, BMI	41%	53%	50.9%
	Systolic BP, Diastolic BP, Age, BMI, Physical Activity	50%	44%	51.6%
	BMI, Physical Activity, Alcohol	49%	50%	52.4%

	Consumption, Age			
	Age, Alcohol Consumption, BMI, Healthy Diet	41%	59%	52.2%
Naive Bayes	Systolic BP, Diastolic BP, Age	74%	78%	75.1%
	Age, Alcohol Consumption, Healthy Diet, Sex, BMI	40%	44%	48.8%
	Systolic BP, Diastolic BP, Age, BMI, Heredity	68%	75%	77.2%
	Systolic BP, Diastolic BP, Alcohol Consumption, Heredity, Age, BMI, Smoking, Healthy Diet, Sex, Physical Activity,	46%	51%	50.6%

5.6 Heart Disease Expert System Evaluation Model Components

The developed heart disease risk evaluation model is innovative because it identifies the degree of risk of heart disease patients using only the non-invasive data features, thus supporting its application as a public screening test. For simplicity, we have called this model as HDREM (Heart Disease Risk Evaluation Model). Below given figure 5.14 shows three main components of HDREM and their working: the knowledge base; inference engine; and the interface.

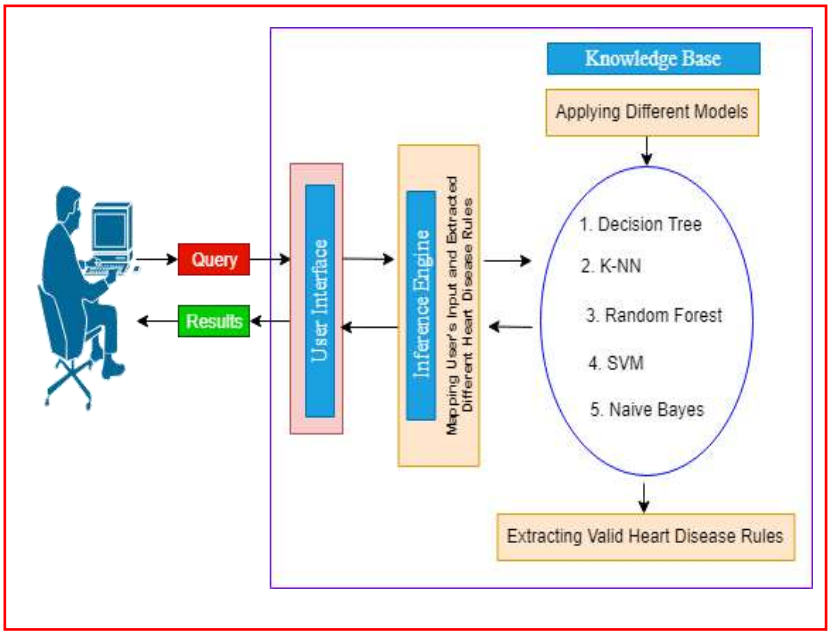

Figure 5.14 Heart Disease Expert System Evaluation Tool Components

The knowledge base component applies the proposed models on non-invasive heart disease data attributes to extract the expert system rules. The inference engine uses the extracted rules, and the users' input component draws 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.

5.7 Heart Disease Risk Evaluation Model (HDREM)

The heart disease risk evaluation model is developed that can be applied for public scanning to predict and diagnose patients at serious, elevated risk disease and give information to facilitate immediate intervention. In Chapter 4 and Chapter 5, the developed models use various risk attributes in predicting heart disorder victims. The results demonstrate that the combination of Age, Systolic BP, Diastolic BP, BMI, Healthy Diet, Hereditary, and Physical Activity provides the best results. These results seem sufficiently high that the Decision Tree, Random Forest, K Nearest Neighbor, Support Vector Machine and Naive Bayes methods could be used to create a screening test for the evaluation of heart disease risk for use at a public level. The rules are extracted to create a chart as community screening tests to support health care experts diagnose the degree of risk of heart disease patients.

The HDREM development plan consists of two major phases. The first phase includes loading the attributes and applying the proposed models to the non-invasive features of the Kashmir heart disease dataset, and then the diagnostic rules are extracted and stored. In the second phase, the user enters his/her data; 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. The HDREM is implemented using the Python Jupyter Notebook web application. Figure 5.15 shows the start screen of HDREM, where the user enters his/her data, and the degree of heart disease risk is calculated and displayed.

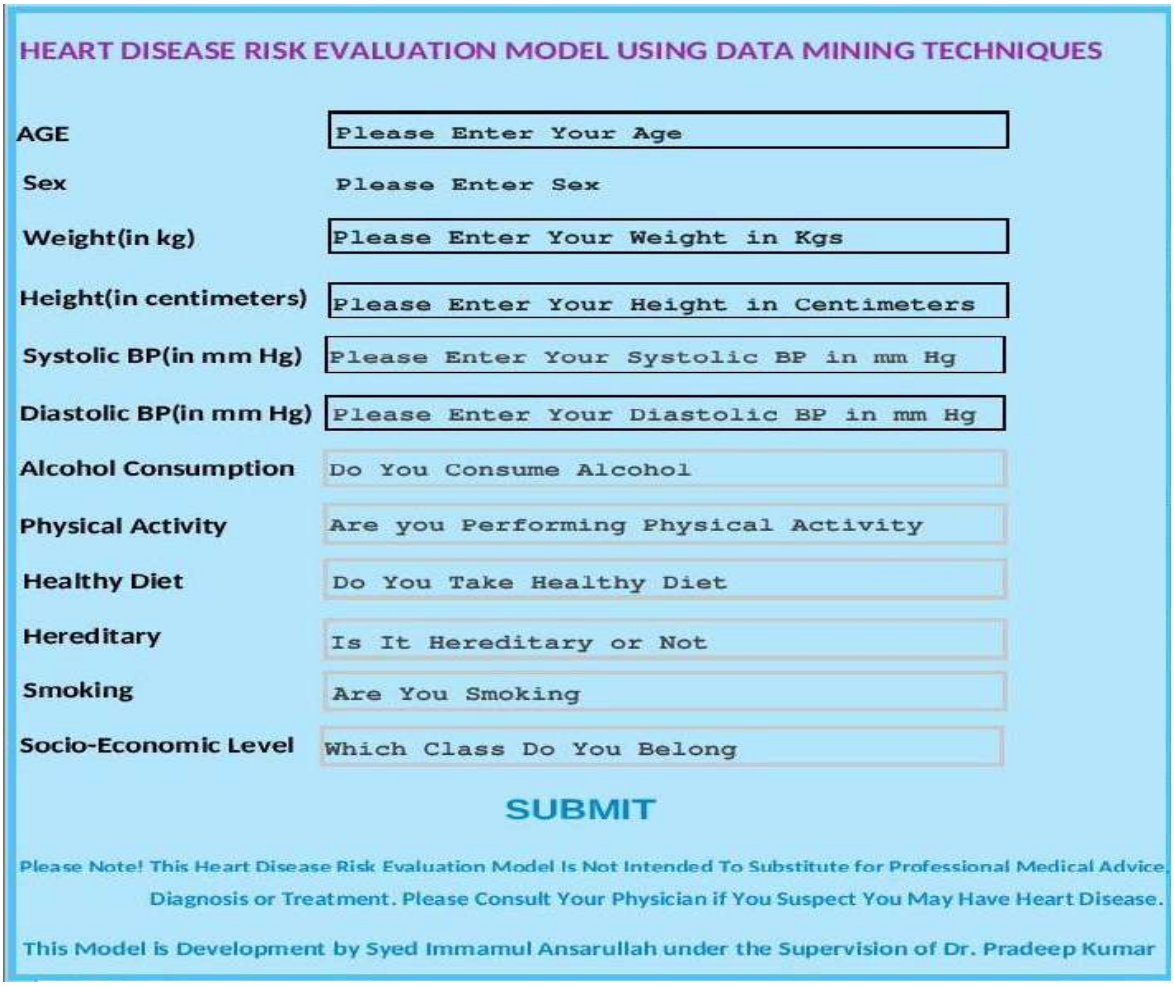

Figure 5.15 Heart Disease Risk Evaluation Model Interface

Figure 5.16 is an output result of the data entered by the user into different corresponding attribute values of the start screen; here, HDREM calculates the high degree of heart disease risk for the entered data. The threshold value greater or equal to 50 is diseased and below it not diseased.

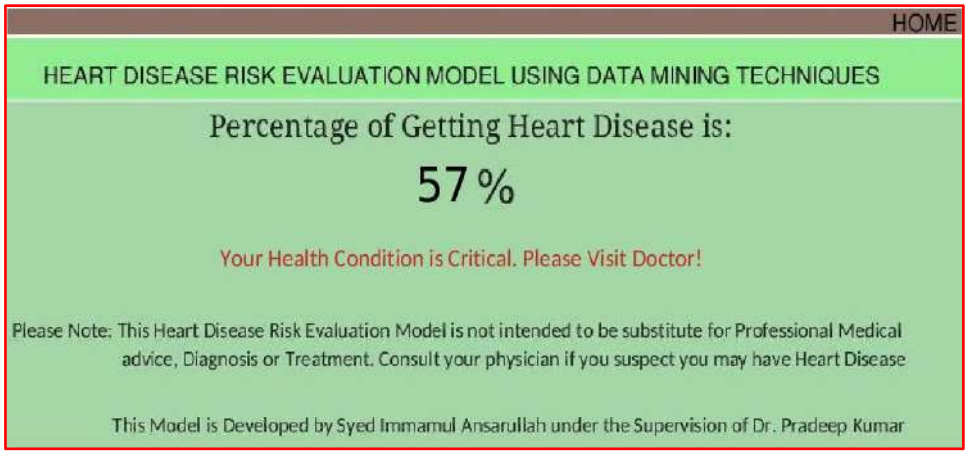

Figure 5.16 High-Risk Heart Disease Evaluation Example

Figure 5.17 is the second example of the HDREM model; in this case, a low risk of heart disease is calculated based on the entered data at the start screen.

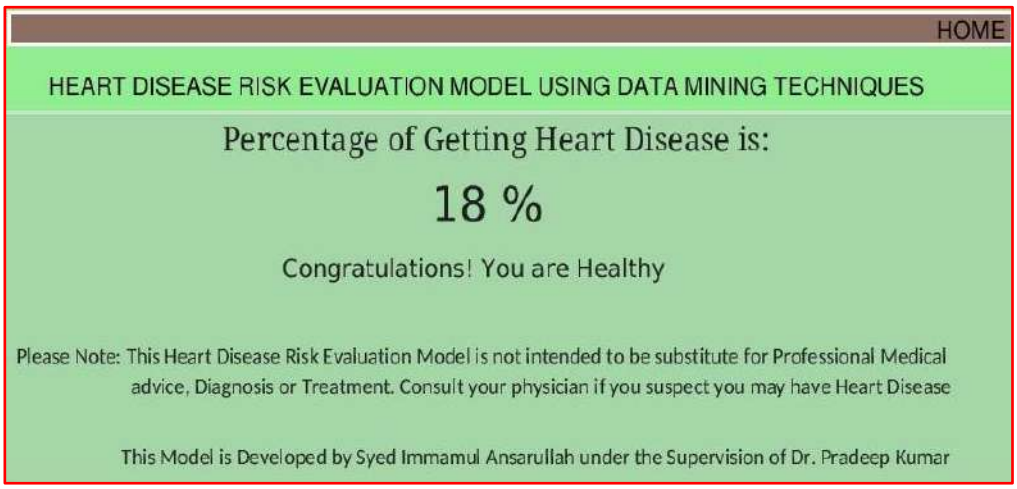

Figure 5.17 Low-Risk Heart Disease Evaluation Example

These examples demonstrate that HDREM can act as a public level screening test. The simplicity of the user interface allows health care practitioners to identify patients at high risk of heart disease using very low-cost non-invasive attributes. The HDREM is implemented on mobile as well as desktop applications.

5.8 Chapter Summary

In this chapter, we introduce hyperparameter optimization techniques and their various types. We optimize the developed models to help health care professionals in the initial prediction and detection of heart disease patients that would decrease growth to critical and exorbitant diseases and complications. The developed heart disease risk evaluation model is built on the Jupyter Notebook web application, and its performance is computed through various measures like TPR, TNR, Accuracy, Precision, Misclassification Rate, AUROC, and cross-validation technique to compute the unbiased measurements. Experimental results show random forest heart disease model surpasses other proposed models. This chapter also describes components of the Heart Disease Risk Evaluation Model and the extracted set of risk rules for the initial prediction of disease.

CHAPTER 6

Conclusion and Future Work

This chapter outlines the research conclusions, discusses the research limitations, and illustrates the future research aspects. Predicting heart disease from diverse features is a multifaceted approach which is often followed by unrepresented effects. The escalated health care costs, repeated hospitalizations, and premature mortality have transformed heart disease into an epidemic worldwide [151]. Heart disease has the greatest disease burden, topping the charts of DALY (Disability Adjusted Life Years) and YLL (Years of Life Lost) as reported by the WHO [152]. Although heart disease is among the most widespread chronic condition leading to a tremendous rate of mortality across the globe, it is recognized as among the most preventable and controllable diseases [154]. Although newer diagnostic innovations have now turned into the standard of consideration, yet these modalities are exorbitant and operationally complex that limits their utilization in rural areas, in primary health care set-ups, and at public-level screening evaluations. By knowing the limitations of the existing systems, we developed a risk evaluation model that helps in the initial recognition and identification of heart disease victims and to achieve this objective; the research poses the question

Can data mining support medical specialists in the early identification and examination of heart disease in a public setting?

Researchers made decisive contributions to heart disease disorder identification using diverse data mining approaches, divergent datasets, different machine learning algorithms, and many tools; however, each design has limitations in it. The fundamental aim of this research is, in answering the above question;

To provide the medical practitioners with a public-level screening model for the untimely risk assessment and identification of heart disease patients with great predictive efficiency.

This research focuses on the following key questions to attain the objective of early prediction of heart disease:

i. Can significant features in the prediction of heart disease patients be determined?

a. Chapter 3 explains that the substantial attributes for heart disease risk evaluation are identifiable.

ii. Can data mining techniques on non-invasive attributes be generously applied for the timely diagnosis of heart disease patients?

b. Chapter 4 discusses that we can effectively apply the data mining techniques to the non-invasive attributes to diagnose heart disease at its initial stages with significant predictive accuracy.

iii. Could hyperparameter optimization methods be emphatically utilized to enhance the performance of non-invasive data features for the early prediction of heart disorder victims?

c. Chapter 5 discusses we can successfully apply optimization techniques to the significant heart disease attributes to enhance the accuracy of the heart disease risk evaluation model.

iv. Can a heart disease risk evaluation model be developed, using non-invasive heart disease data attributes?

d. Chapter 6 describes that a reliable heart disease risk evaluation model, using non-invasive data attributes can be developed.

The results from Chapters 3 to 6 that support and confirm answers to the key questions posed in this research work also show support for the principal question that data mining can assist medical practitioners in the early detection of heart disease in a public screening environment.

6.1 Research Conclusion

The existing risk tools predict heart disease using the clinical datasets having attributes and inputs from the multifaceted examinations conducted in the medical labs. However, none of the tools predicts cardiac disease based on visible risk features that can be used efficiently for the diagnosis. A model dependent on such risk features would support medical practitioners as well as it would provide victims with a message about the possible existence of cardiac disorder even before he/she visits a clinic or does exorbitant health inspections. These risk models are applicable where people lack the facilities of the integrated primary medical care technologies for untimely prediction and cure.

In this research work, the cardiac disorder risk evaluation model is developed based on non-invasive attributes using the Jupyter notebook web application. The Kashmir heart disease dataset is mined using the Random Forest, K Nearest Neighbor, Support Vector Machine, Decision Tree and Naive Bayes classifiers to discover if an individual possessing certain modifiable risk features, will have the cardiac disorder or not. The specificity, sensitivity, precision, accuracy, misclassification rate, and AUROC score are calculated for each method using out-of-sample testing to check how accurately the risk evaluation model performs. The hyperparameter optimization is performed to obtain the optimal results out of these techniques. Experimental results show that the random forest model outperforms other models with the highest sensitivity, specificity, precision, accuracy, AUROC score, and with minimum misclassification rate.

We simulate the accomplished outcomes against the prevailing research; the results obtained are, to the best of our perception, greater than published values in the literature. The research investigated whether we can apply the data mining techniques with reliable accuracy to non-invasive attributes to create diagnostic rules to build a public-level heart disease risk evaluation model. We discuss the details of the contributions of this work in below subsections.

6.1.1 Significant Attributes in Heart Disease Risk Evaluation

Chapter 3 identifies the significant non-invasive risk features which are used to predict heart disease patients at its earliest. In this chapter, each heart disease risk attribute is weighted as per their importance in disease prediction by five different feature selection techniques. After the assignment of weight to every risk attribute by different feature selection techniques, the overall mean of all weights is considered for risk model development. The higher numeric weight value means that an attribute is significant and plays a crucial role in predicting heart disease patients at its initial stage. The assigned numeric weights to risk attributes are validated and approved by various medical domain experts. Finally, data mining techniques use weighted risk attributes in predicting and diagnosing heart disease patients.

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

Chapters 4 and 5 recognize the importance of different non-invasive features in the prediction of cardiac disorder victims. The main importance of non-invasive features is that they are low-cost attributes and can be used in public-level evaluation tests to predict victims at elevated risk of cardiac disorder. Chapter 5 evaluated the non-invasive attribute combinations to see which attribute combinations show the best performance in identifying cardiac disease patients at its initial stages. Table 5.7 summarizes the results of applying the proposed data mining techniques on different non-invasive heart disease attribute combinations. When investigating the effect of different combined non-invasive features in the prediction of heart disorder victims, the combination of (Age, Sex, Systolic BP, Diastolic BP, BMI, and Heredity) attribute results are very interesting and create a Public level screening test for the assessment of heart disease risk.

6.1.3 Building the Heart Disease Risk Evaluation Model (HDREM)

Although the cardiac disorder can be diagnosed by numerous investigations like Stress Test,

an electrocardiogram, and cardiac angiograms; however, these modalities are costly and operationally complex that restrains their use in rural areas and at public-level screening evaluations. To overcome the drawbacks of the prevailing risk systems, we developed a risk evaluation model for the initial prediction and detection of heart disease is developed. In this research work, five different feature selection methods are applied to the Kashmir heart disease dataset, and their mean values are calculated to select the most favorable attributes for initial identification of heart disease. The five different standard classification algorithms like Decision Tree, Naive Bayes, Random Forest, K Nearest Neighbor, and Support Vector Machine are applied to carry out the experiments. The main contribution of this risk evaluation model is its ability to identify the risk of heart disorder victims using only low-cost non-invasive features. The random forest and decision tree risk rules are extracted to identify heart disease using only non-invasive data attributes.

6.2 Research Limitations

The limitations associated with this research work are described as follows:

i. Using only the basic data mining techniques: Random Forest, K Nearest Neighbor, Support Vector Machine, Decision Tree, and Naive Bayes in heart disease prediction.
ii. Using some non-invasive attributes only while not using other non-invasive attributes such as Depression Level, Low Educational Status, and Ethnicity, etc.
iii. No usability testing of the proof-of-concept evaluation tools with community-level screening health care providers (e.g., pharmacists).
iv. Testing the data mining methods on the Kashmir heart disease dataset, which contains a small number of records and is based on specific ethnicity?
v. Not diagnosing the different categories of heart disease like Arrhythmia and Cardiac arrest etc.

6.3 Future Work

In the future, the proposed research can be enhanced in the following directions:

i. Investigating the performance of other robust data mining methods like Genetic Algorithm, Neural Networks, and Ensembling of techniques to achieve comparative performance results.
ii. Studying the significance of adding other non-invasive attributes (socioeconomic level, Depression Level, and Ethnicity) on the performance of different data mining methods for the early identification of heart disease risk patients.
iii. Identifying the significance of controlled non-invasive attributes such as weight and smoking on different age and sex groups in the risk estimation of heart disease.
iv. Using heterogeneous real-world datasets having a different number of attributes, diverse population groups, and a huge number of records.
v. Testing the realistic usability and acceptability of the HDREM among pharmacists and other community health care providers.
vi. To develop a one-size-fits-all heart disease risk model using data mining techniques that could successfully prescribe a treatment plan for the disease also.

References

1. Omran AR (2005). The epidemiologic transition: A Theory of the Epidemiology of Population Change. Milbank Mem Fund Q, volume 49 on page 509.

2. World Health Organization (2010). Global status report on noncommunicable diseases 2010. https://www.who.int/nmh/publications/ncd_report_full_en.pdf

3. World Health Organization (2011a). The Top Ten Causes Of Death. Accessed 18 August 2017, from http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

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

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

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

7. Heart Disease and Stroke Statistics (2019): A Report from the American Heart Association. DOI: 10.1161/CIR.0000000000000659.

8. Centers for Disease Control and Prevention (2013). Chronic Disease Prevention and Health Promotion. Accessed 27 September 2017, from http://www.cdc.gov/nccdphp/

9. World Health Organization (2011b). Burden: Mortality, Morbidity, and Risk Factors. Accessed 3 April 2017 http://www.who.int/nmh/publications/ncd_report_chapter1.pdf

10. Mathers Colin D., Boerma Ties and Fat Doris Ma (2009). Global and Regional Causes of Death. British Medical Bulletin, Volume 92, Issue 1, Pages 7–32, https://doi.org/10.1093/bmb/ldp028

11. Paladugu, S. (2010). Temporal mining framework for risk reduction and early detection of chronic diseases. A Thesis Submitted to the University of Missouri-Columbia.

12. Scully Jackie Leach (2004). What is Disease? EMBO Rep. 5(7):650–653. https://doi.org/10.1038/sj.embor.7400195

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

14. Gaziano et al. (2009). Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries. Curr Probl Cardiol. 35(2): 72–115. DOI: 10.1016/j.cpcardiol.2009.10.002

15. American Heart Association Annual Report for 2017-2018. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease

16. American Heart Association (2013). What is Cardiovascular Disease (Heart Disease)? http://www.heart.org/HEARTORG/Caregiver/Resources/WhatisCardiovascularDis ease/What-is-Cardiovascular-Disease_UCM_301852_Article.jsp

17. Atherosclerosis. National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/health-topics/atherosclerosis

18. Atherosclerosis and Cholesterol. https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis

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

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

21. European Public Health Alliance (2013). Cardiovascular Health Takes Center Stage in Brussels. Accessed 12 March 2016, from http://www.epha.org/a/5899

22. Heart and Circulatory Disease Statistics (2019). British Heart Foundation. https://www.bhf.org.uk

23. Australian Bureau of Statistics (2013). Accessed 12 March 2016, from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3303.0Media%20Release12011?opendocument&tabname=Summary&prodno=3303.0&issue=2011&num=&view=

24. Australia's Leading Causes of Death (2017). https://www.abs.gov.au

25. Shahwan-Akl, L. (2010). Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 1(1), 1-7.

26. Simons L.A., Simons J., Friedlander Y., McCallum J., and Palaniappan L. (2003). Risk Functions for Prediction of Cardiovascular Disease in Elderly Australians: The Dubbo Study. Medical Journal of Australia, 178(3), 113-116.

27. Economic and Social Survey of Asia and the Pacific (2010). http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp

28. G L Khor (2001). Cardiovascular Epidemiology in the Asia-Pacific Region. Asia Pacific Journal of Clinical Nutrition 2001; 10 (2):76-80.

29. Huang Yanzhong, Moser Patricia, and Roth Susann (2015). Health in the Post-2015 Development Agenda for Asia and the Pacific.

30. World Health Organization (2013c). Deaths from Coronary Heart Disease. http://www.who.int/cardiovascular_diseases/en/cvd_atlas_14_deathHD.pdf

31. Francesco Paolo Cappuccio and Michelle Avril Miller (2016). Cardiovascular Disease and Hypertension in Sub-Saharan Africa: Burden, Risk, and Interventions. Intern Emerg Med.; 11: 299–305.

32. Gregory A. Roth (2017). Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology.DOI: 10.1016/j.jacc.2017.04.052

33. Hansson K. Goran. (2005). Inflammation, Atherosclerosis, and Coronary Artery Disease. The New England Journal of Medicine. DOI: 10.1056/NEJMra043430

34. Thomas A. Gaziano (2008). Reducing the Growing Burden of Cardiovascular Disease in the Developing World: Disease burden can be lowered with cost-effective interventions, especially by reducing the use of tobacco around the world. Health Aff (Millwood). 26(1): 13-24.

35. Becker, Roberto, Silvi, John, Ma Fat, Doris, L Hours, Andre and Laurenti Ruy. (2006). A Method for Deriving the Leading Causes Of Death. Bulletin of the World Health Organization, 84 (4) , 297 - 304. World Health Organization. https://apps.who.int/iris/handle/10665/269620

36. American Heart Association (2011). American Heart Association Live and Learn. http://www.americanheart.org/presenter.jhtml?identifier=4478

37. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, Finkelstein EA, Hong Y, Johnston SC, Khera A, Lloyd-Jones DM, Nelson SA, Nichol G, Orenstein D, Wilson PW, and Woo YJ (2011). Forecasting the Future of Cardiovascular Disease. Circulation March 1, Vol. 123, Issue 8.

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

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

40. Ensminger, M. E., and Ensminger, A. H. (1993). Foods & nutrition encyclopaedia (Second Edition), Volume 1. CRC Press, ISBN 9780849389818

41. Cupples L. and D Agostino (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.

42. Department of Health & Aging, A. G. (2012). Seniors and Aged Care Australia websites have been replaced. Accessed 16 August 2016, from http://www.agedcareaustralia.gov.au/internet/agedcare/publishing.nsf/Content/Prevention+and+awareness-2

43. Colin D. Mathers and Dejan Loncar (2006). Updated Projections of global mortality and burden of disease, from 2002 to 2030. Published online November 28. DOI: 10.1371/journal.pmed.0030442

44. Pietrzak E, Cotea C, and Pullman S (2014). Primary and Secondary Prevention of Cardiovascular Disease: is there a place for Internet-based interventions. Journal of Cardiopulmonary Rehabilitation and Prevention. Sep-Oct; 34(5):303-17. DOI: 10.1097/HCR.0000000000000063.

45. World Health Organization Press (2014). Global Status Report on Non-Communicable Diseases. https://www.who.int/nmh/publications/ncd-status-report-2014/en/

46. Nag Tanmay and Ghosh Arnab (2013). Cardiovascular disease risk factors in Asian Indian population: A systematic review. Journal of Cardiovascular Disease Research 4 (2013) 222 – 228.

47. Din, S., Rabbi, F., Qadir, F., and 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.

48. World Health Organization, Regional Office for the Eastern Mediterranean. (2005). Clinical guidelines for the management of hypertension. https://apps.who.int/iris/handle/10665/119738

49. Porter, Thomas, and Green, Barbara (2009). Identifying Diabetic Patients: A Data Mining Approach. AMCIS 2009 Proceedings. Paper 500. http://aisel.aisnet.org/amcis2009/500

50. Thompson G. R. (2004). Management of dyslipidemia. Heart (British Cardiac Society), 90(8), 949–955. DOI:10.1136/hrt.2003.021287

51. Hubert, H. B., Feinleib, M., McNamara, P. M., and 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.

52. Panzarasa, S., Quaglini, S., Sacchi, L., Cavallini, A., Micieli, G., and Stefanelli, M. (2010). Data Mining Techniques for Analyzing Stroke Care Processes. In the Proc. of the 13th World Congress on Medical Informatics.

53. Reynolds Risk Score (2015). About the Reynolds Risk Score. Accessed 10 November 2016, from http://www.reynoldsriskscore.org/home.aspx

54. Framingham Heart Study (2013). About the Framingham Heart Study. Accessed 5 October 2017, from http://www.framinghamheartstudy.org/about/history.html

55. Bitton, A., and Gaziano, T. (2010). The Framingham Heart Study's Impact on Global Risk Assessment. Progress in cardiovascular diseases, 53(1), 68-78.

56. National Vascular Disease Prevention Alliance (2012). Guidelines for the management of absolute cardiovascular disease risk. ISBN: 978-0-9872830-1-6

57. Daniel T. Larose and Chantal D. Larose (2005). Discovering Knowledge in Data-An Introduction to Data Mining (Second Edition). Wiley-Interscience Publications.

58. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). Discovering Data Mining. From Concept to Implementation. Upper Saddle River, NJ: Prentice-Hall.

59. Jiawei Han, Micheline Kamber, and Jian Pei (2006). Data Mining Concepts and Techniques (Third Edition): Morgan Kaufmann Publishers.

60. Ian H. Witten, Eibe Frank, and Mark A. Hall (2011). Data Mining Practical Machine Learning Tools and Techniques (Third Edition): Morgan Kaufmann Publishers.

61. Max Bramer (2007). Principles of data mining (Second Edition). eBook ISBN: 978-1-84628-766-4. DOI: 10.1007/978-1-84628-766-4. Publisher: Springer-Verlag London

62. Fayyad, U. (1997). Data mining and Knowledge Discovery in Databases: Implications for Scientific Databases, Ninth International Conference on Scientific and Statistical Database Management.

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

64. Ruben D. Canlas Jr. (2009). Data Mining in Healthcare: Current Applications and Issues.

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

66. Pang Ning Tan, Michael Steinbach, and Vipin Kumar (2010). Introduction to Data Mining (First Edition). Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©2005. ISBN: 0321321367.

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

68. I.N. Lee, S.C.Lio and M. Embrechts (2009). Data mining techniques applied to medical information. Informatics for Health and Social Care, 25(2), 81-102.

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

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

71. L.M. Renard, V. Bocquet, G. Vidal-Trecan, M.-L. Lair, S. Couffignal, C. Blum-Boisgard (2011). An algorithm to identify patients with treated type 2 diabetes using medico-administrative data. BMC Med Inform Decis Mak, 11 (2011), p. 23.

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

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

74. Das, R., Turkoglu, I., and Sengur, A. (2009). Effective Diagnosis Of Heart Disease Through Neural Networks Ensembles. Expert Systems with Applications, Elsevier, 36 (2009), 7675–7680.

75. I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M. C. Jaulent (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron, and logistic regression. Proc. AMIA Symp. PP. 156–60.

76. H. Yan (2003). Development of a Decision Support System for Heart Disease Diagnosis Using Multilayer Perceptron. IEEE Int. Symp. Circuits Syst., vol. 5, pp. 709–712.

77. Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu (2006). Associative Classification Approach for Diagnosing Cardiovascular Disease. ICIC, 2006, LNCIS 345, pp. 721 – 727, 2006.

78. S. Palaniappan and R. Awang (2008). Intelligent heart disease prediction system using data mining techniques. IEEE/ACS Int. Conf. Comput. Syst. Appl., vol. 8, no. 8, pp. 343–350.

79. M. Shouman, T. Turner, and R. Stocker (2011). Using Decision Tree For Diagnosing Heart Disease Patients. Proc. Ninth Australia's. Data Min. Conf., pp. 23–30.

80. K. U. Rani (2011). Analysis of Heart Diseases Dataset using Neural Network Approach. Int. J. Data Min. Knowl. Manag. Process, vol. 1, no. 5, pp. 1–8.

81. M. Kumari and S. Godara (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. Int. J. Comput. Sci. Trends Technol., vol. 2, no. 2, pp. 304–308.

82. M. Shouman, T. Turner, and R. Stocker (2012). Using data mining techniques in heart disease diagnosis and treatment. Japan-Egypt Conf. Electron. Commun. Comput., pp. 173–177, 2012.

83. V. Chaurasia (2013). Early Prediction of Heart Diseases Using Data Mining. Caribb. J. Sci. Technol., vol. 1, no. December, pp. 208–217, 2013.

84. N. Al-Milli (2013). Back-propagation Neural Network for Prediction of Heart Disease. J. Theory. Appl. Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

85. D. Masethe Hlaudi and A. Masethe Mosima (2014). Prediction of Heart Disease using Classification Algorithms. Proc. World Congr. Eng. Comput. Sci., vol. II, pp. 22–24.

86. A. Ngueilbaye, L. Lei, and H. Wang (2016). Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the Republic of Chad. Int. J. Sci. Res., vol. 5, no. 5, pp. 1564–1571, 2016.

87. Thuy Nguyen Thi Thu, and Darryl.N. Davis (2007). A Clustering Algorithm for Predicting Cardiovascular Risk. World Congress on Engineering 2007: 354-357

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

89. M. Shouman, T. Turner, and R. Stocker (2012a). Integrating Naive Bayes and K Means Clustering with different Initial Centroid Selection methods in the diagnosis of heart disease patients. airccj.org, pp. 431–436, 2012.

90. Latha Parthiban and R. Subramanian (2007).Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm. World Academy of Science, Engineering, and Technology International Journal of Medical and Health Sciences Vol: 1, No: 5, 2007.

91. Polat, K., Sahan, S., and Gunes, S. (2007). Automatic detection of heart disease using an artificial immune recognition system with fuzzy resource allocation mechanism and k-nearest neighbor based weighting preprocessing. Expert Systems with Applications, 32(2), 625-631.

92. Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 4, pp. 447–58, 2008.

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

94. George Edward Pelham Box (1979). Robustness in the strategy of scientific model building. Robustness in Statistics, Pages 201-236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

95. M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," Int. J. Eng. Sci. Technol., vol. 2, no. 10, pp. 5370–5376, 2010.

96. A. Adeli and M. Neshat (2010). A Fuzzy Expert System for Heart Disease Diagnosis. Proc. Int. Multi-Conference Engineers Comput. Sci., vol. I, pp. 1–6, 2010.

97. A. Aqueel and S. A. Hannan (2012). Data Mining Techniques to find out Heart Diseases : An Overview. Int. J. Innov. Technol. Explore. Eng., vol. 1, no. 4, pp. 18–23, 2012.

98. R. Alizadehsani, J. Habibi, M. Javad, B. Bahadorian, and Z. Alizadeh (2013). A data mining approach for diagnosis of coronary. Comput. Methods Programs Biomed. PP. 1–10, 2013.

99. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra (2015). Computational Intelligence Technique for Early Diagnosis of Heart Disease. IEEE Int. Conf. Eng. Technol., Vol No, March, pp. 1–7, 2015.

100. S. U. Amin, K. Agarwal, and R. Beg (2013). Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors. ICT 2013 - Proc. 2013 IEEE Conf. Inf. Commun. Technol., no. ICT, pp. 1227–1231, 2013.

101. V. Chaurasia and S. Pal (2014). Data Mining Approach to Detect Heart Diseases. Int. J. Adv. Comput. Sci. Inf. Technol. Vol. 2, no. 4, pp. 56–66, 2014.

102. K. Srinivas, G. R. Rao, and A. Govardhan (2014). Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories. Arab. J. Sci. Eng., vol. 39, no. 4, pp. 2857–2868, 2014.

103. A. Dewan and M. Sharma (2015). Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification. International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 704–706.

104. B. V. Sumana and T. Santhanam (2014). Prediction of diseases by cascading clustering and classification. International Conference on Advances in Electronics, Computers, and Communications, ICAECC 2014, 2014, p. 8.

105. Beena G N Bethel, T V Rajinikanth, and Viswanadha S Raju (2016). A Knowledge-Driven Approach for Efficient Analysis of Heart Disease Dataset. International Journal of Computer Applications 147(9):39-46, August 2016

106. Omar Karam, Mostafa A. Salama and Randa El Bialy (2016). An ensemble model for Heart disease data sets : a generalized model. ACM, pp. 191–196, 2016.

107. Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, and Yarifard AA (2017). Computer-Aided Decision Making For Heart Disease Detection Using Hybrid Neural Network-Genetic Algorithm. Computer Methods and Programs in Biomedicine 141 (2017) 19-26. https://doi.org/10.1016/j.cmpb.2017.01.004

108. Geurts. Pierre, Ernst. Damien and Wehenkel. Louis (2006). Extremely Randomized Trees. Machine Learn 63: 3-42. DOI 10.1007/s10994-006-6226-1

109. Natekin. Alexey and Knoll. Alois (2013). Gradient Boosting Machines- A Tutorial. Frontiers in Neuro Robotics Volume7| Article21.

110. Biau Gerard (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research 1063-1095.

111. Khaing T. Kyaw (2010). Enhanced Features Ranking and Selection using Recursive Feature Elimination (RFE) and K- Nearest Neighbor Algorithms in Support Vector Machine for Intrusion Detection System. International Journal of Network and Mobile Technologies VOL 1/ ISSUE1/ JUNE.

112. Chen. Tianqi and Guestrin. Carlos (2016). XG BOOST: A Scalable Tree Boosting System. KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785-794 at San Francisco, California, USA.

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

114. Richard J. Roiger (2017). Data Mining: A Tutorial - Based Primer (Second Edition). CRC Press Taylor & Francis Group New York.

115. Jure Leskovec, Anand Rajaraman, and Jeffrey Ullman (2014). Mining of Massive Datasets (Second Edition). Cambridge University Press. ISBN-10: 1107077230.

116. J.R. Quinlan. Machine Learning 1: 81-106 (1986). © Kluwer Academic Publishers, Boston – Manufactured in the Netherlands. https://doi.org/10.1007/BF00116251

117. Esposito, F., Malerba, D., Semeraro, G., and Kay, J. (1997). A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5), 476-491.

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

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

120. I.Ketut Agung Enriko, Muhammad Suryanegara, and Dadang Gunawan (2018). Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records. Published in: Proceeding ICFET '18 Proceedings of the 4th International Conference on Frontiers of Educational Technologies Pages 127-131.

121. Purnami, S., Zain, J., and Embong, A. (2010). A New Expert System for Diabetes Disease Diagnosis Using Modified Spline Smooth Support Vector Machine. Computational Science and Its Applications – ICCSA 2010 (Vol. 6019, pp. 83-92): Springer Berlin Heidelberg.

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

123. Mudasir M Kirmani and Syed Immamul Ansarullah (2016) Classification models on cardiovascular disease detection using Neural Networks, Naive Bayes and J48 Data Mining Techniques". International Journal of Advanced Research in Computer Science. Volume 7, No. 5, September-October 2016.

124. Vipin Kumar (2010). The Top Ten Algorithms in Data Mining (First Edition). CRC Press Taylor & Francies Group Boca Raton London New York. ISBN-13: 978-1-4200-8964-6.

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

126. Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar and Ravi Thomas (2008). Understanding and using Sensitivity, Specificity, and Predictive Values. Indian Journal of Ophthalmology, Jan- Feb; 56(1): 45-50.

127. Mudasir Manzoor Kirmani and Syed Immamul Ansarullah (2016). Prediction of Heart Disease using Decision Tree a Data Mining Technique. International Journal of Computer Science and Network, Volume 5, Issue 6.

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

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

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

131. Shillabeer, A., and Roddick, J. F. (2006). Towards Role Based Hypothesis Evaluation For Health Data Mining. Electronic Journal of Health Informatics, 1(1), 1-9.

132. Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). The CRSIP –DM Process Model. Available online at website: http://www.industrieldatamining.dk/images/crisp-dm.pdf

133. Shearer, C. (2000). The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data warehousing. Vol.5. Pp: 13-22.

134. SAS (2008). Available online at web site: http://www.sas.com/technologies/analytics/datamining/miner/semma.html

135. Davis, D. N. (2007). Data Mining and Decision Systems. Lecture notes. Available online at website: http://intra.net.dcs.hull.ac.uk.

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

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

138. Charles X. Ling and Victor S. Sheng (2011). Class Imbalance Problem. Encyclopedia of Machine Learning, 171–171.DOI:10.1007/978-0-387-30164-8_110

139. Irene Y. Chen, Fredrik D. Johansson, and David Sontag (2018). Why Is My Classifier Discriminatory? 32nd Conference on Neural Information Processing Systems, Montreal, Canada.

140. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, date.

141. MM Mukaka (2007). Statistics Corner: A Guide To Appropriate Use Of Correlation Coefficient In Medical Research. Malawi Medical Journal; 24(3): 69-71 September.

142. Kathleen F. Weaver, Vanessa C. Morales, Sarah L. Dunn, Kanya Godde, and Pablo F. Weaver (2018). An Introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences. Published by John Wiley & Sons, Inc. Companion.

143. Shilin Zhao, Yan Guo, Quanhu Sheng, and Yu Shyr (2016). Advanced Heat Map and Clustering Analysis Using Heatmap3. Hindawi Publishing Corporation BioMed Research International Volume 2014, http://dx.doi.org/10.1155/2014/986048

144. Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua, and James W. Lillard, Jr.(2015). A Study of Effects of MultiCollinearity in the Multivariable Analysis. Int J Appl Sci Technol.

145. Hensher Martin, Price Max, and Adomakoh Sarah (2006). Disease Control Priorities in Developing Countries. https://www.who.int/management/referralhospitals.pdf

146. Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang (2018). Feature selection in machine learning: A new perspective. Neurocomputing 300 (2018) 70–79. https://www.journals.elsevier.com/neurocomputing

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

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

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

150. Amin Ul Haq, Jian Ping Li,Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun(2018). A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mobile Information Systems. Volume 2018. https://doi.org/10.1155/2018/3860146

151. Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R.Day, Clint Richardson, Charles K. Fisher and David J. Schwab (2019). A high-bias, low-variance introduction to Machine Learning for physicists, Physics Reports. https://doi.org/10.1016/j.physrep.2019.03.001

152. Vinod Chandra S.S and Anand Hareendran S. (2014). Artificial Intelligence and Machine Learning. Publisher: Prentice Hall India Learning Private Limited. ISBN-13: 978-8120349346.

153. Alex Smola and S.V.N. Vishwanathan (2008). Introduction to Machine Learning (First Edition). Cambridge University Press. ISBN: 0521825830.

154. Chris Albon (2018). Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (First Edition). Publisher: Shroff/ O' Reilly. ISBN-13: 978-9352137305

155. Andreas C. Muller and Sarah Guido (2016). Introduction to Machine Learning with Python (First Edition). O'Reilly Media. ISBN-13: 978-1449369415.

156. Aurelien Geron (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow (First Edition). O'Reilly. ISBN-13: 978-1491962299.

157. Jeremy Jordan (2017). Hyperparameter Tuning for Machine Learning Models. https://www.jeremyjordan.me/hyperparameter-tuning/

158. Hazan, E., Klivans, A., and Yuan, Y (2018). Hyperparameter optimization: A spectral approach. Proceedings of the International Conference on Learning Representations (ICLR'18).

159. Falkner, S., Klein, A., and Hutter, F (2018). BOHB: Robust and Efficient Hyperparameter Optimization at Scale. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden.

160. Noemi DeCastro-Garcia, Angel Luis Munoz Castaneda, David Escudero Garcia, and Miguel V. Carriegos (2019). Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm. https://doi.org/10.1155/2019/6278908.

161. Cathy O'Neil and Rachel Schutt (2013). Doing Data Science: Straight Talk from the Frontline (First Edition). Publisher: O'Reilly Media. ISBN-13: 978-1449358655.

162. Dernoncourt F., Nemati S., Kassis E.B., and Ghassemi M.M. (2016). Hyperparameter Selection. Secondary Analysis of Electronic Health Records. Springer, Cham.

163. Umit Mert Cakmak and Sibanjan Das (2018). Hands-On Automated Machine Learning. Published by Packt Publishing Ltd. ISBN: 9781788629898.

164. Schilling N., Wistuba M., Drumond L., and Schmidt-Thieme L. (2015) Hyperparameter Optimization with Factorized Multilayer Perceptrons. Machine Learning and Knowledge Discovery in Databases. ECML PKDD, 2015. Lecture Notes in Computer Science, VOL 9285. Springer, Cham.

165. James Bergstra and Yoshua Bengio (2012). Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research. Volume 13, Issue 1, January 2012.

166. Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. DOI: 10.1016/j.ecolmodel.2019.06.002

167. Sebastian Raschka (2017). Python Machine Learning (Second Edition). Published by Packt Publishing Ltd. ISBN-13: 978-1787125933.

168. Jake VanderPlas (2017). Python Data Science Handbook (First Edition). Publisher: Shroff Publishers & Distributors Pvt. Ltd. ISBN: 9789352134915

169. David Barber (2014). Bayesian Reasoning and Machine Learning. Cambridge University Press. Online ISBN: 9780511804779. DOI: https://doi.org/10.1017/CBO9780511804779

170. Mantovani, R., Horvath, T., Cerri, R., Vanschoren, J., and Carvalho, A (2016). Hyper-Parameter Tuning of a Decision Tree Induction Algorithm. 5th Brazilian Conference on Intelligent Systems (BRACIS). Pp. 37-42.IEEE Computer Society Press.

171. Rafael Gomes Mantovani, Tomas Horvath, Ricardo Cerri, S Barbon Junior, Joaquin Vanschoren and Andre de Carvalho (2019). An empirical study on hyperparameter tuning of decision trees. https://arxiv.org/pdf/1812.02207.pdf

172. Tammo Krueger, Danny Panknin, and Mikio Braun (2015). Fast cross-validation via sequential testing. Journal of Machine Learning Research 16 (2015) 1103-1155.

173. Wojciech M Czarnecki, Sabina Podlewska, and Andrzej J Bojarski (2015). Robust optimization of SVM hyperparameters in the classification of bioactive compounds. Journal of Cheminformatics. DOI: 10.1186/s13321-015-0088-0

174. Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix (2019). Hyperparameters and Tuning Strategies for Random Forest. arXiv:1804.03515v2 [stat.ML] 26 Feb 2019

175. Tinu Theckel Joy, Santu Rana, Sunil Gupta, and Svetha Venkatesh(2019). Fast Hyperparameter Tuning using Bayesian Optimization with Directional Derivatives. https://arxiv.org/pdf/1902.02416.pdf

176. Jasper Snoek, Hugo Larochelle, and Ryan P. Adams (2019). Practical Bayesian Optimization of Machine Learning Algorithms. Neural Information Processing Systems (NIPS) Conference 2019.

177. Willi Richert and Luis Pedro Coelho (2013). Building Machine Learning Systems with Python (First Edition). Published by Packt Publishing Ltd, Livery Place Birmingham B3 2PB, UK. ISBN 978-1-78216-140-0.

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