

**THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION  
INSTITUTE OF SCIENCE AND TECHNOLOGY**

**HEART DISEASE SYSTEM PREDICTION USING DATA MINING  
TECHNIQUES**

**MASTER THESIS**

**Mohammed Ibrahim Mahdi AL-AZZAWI**

**THE DEPARTMENT OF INFORMATION TECHNOLOGY  
THE PROGRAM OF INFORMATION TECHNOLOGY**

**MARCH 2017**

**THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION**  
**INSTITUTE OF SCIENCE AND TECHNOLOGY**

**HEART DISEASE SYSTEM PREDICTION USING DATA MINING  
TECHNIQUES**

**MASTER THESIS**

**Mohammed Ibrahim Mahdi AL-AZZAWI**

**ID: 1406050035**

**THE DEPARTMENT OF INFORMATION TECHNOLOGY**  
**THE PROGRAM OF INFORMATION TECHNOLOGY**

**Assist. Prof. Dr. Yuriy ALYEKSYEYENKOV**

**MARCH 2017**

Türk Hava Kurumu Üniversitesi Fen Bilimleri Enstitüsü'nün 1406050035 numaralı Yüksek Lisans öğrencisi "Mohammed AL-AZZAWI" ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "HEART DISEASE SYSTEM PREDICTION USING DATA MINING TECHNIQUES" başlıklı tezini aşağıda imzaları bulunan jüri önünde başarı ile sunmuştur.

**Tez Danışmanı : Yrd. Doç. Dr. Yuriy ALYEKSYEYENKOV**  
Türk Hava Kurumu Üniversitesi



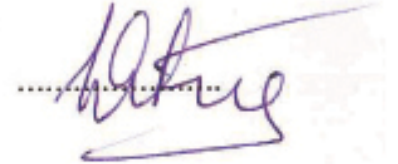
**Jüri Üyeleri : Doç. Dr. Fahd JARAD**  
Çankaya Üniversitesi



**: Yrd. Doç. Dr. Shadi AL SHEHABI**  
Türk Hava Kurumu Üniversitesi



**: Yrd. Doç. Dr. Yuriy ALYEKSYEYENKOV**  
Türk Hava Kurumu Üniversitesi



**Tez Savunma Tarihi: 31.03.2017**

## STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

A handwritten signature in purple ink, consisting of several loops and a long horizontal stroke at the bottom.

31.03.2017

Mohammed Ibrahim

## **ACKNOWLEDGEMENTS**

I would first like to thank my thesis advisor, Assis. Prof. Dr. Yuriy ALYEKSYEYENKOV of the Computer Engineering Department at THK University, without whose helpful advice, valuable comments and guidance this thesis could not be completed. His door was always open for me whenever I needed his help. To whom who lighted up my way through darkness, to whom who made me what I'm today and yet didn't have the chance to see what a great man he left behind, my dear Brother God bless his soul. I owe you my past and I owe you my future, with my best regards and love to you Brother Rest in Peace. And to my other brother and my sisters who always gave me advice and guidance, god bless you wish you many years ahead by my side. I would like to thank my who Owns My Heart H, friends and teachers for everything.

## TABLE OF CONTENTS

<b>STATEMENT OF NON-PLAGIARISM PAGE</b> .....	iii
<b>ACKNOWLEDGEMENTS</b> .....	iv
<b>TABLE OF CONTENTS</b> .....	v
<b>LIST OF FIGURES</b> .....	viii
<b>LIST OF TABLES</b> .....	ix
<b>LIST OF ABBRIVIATION</b> .....	x
<b>ABSTRACT</b> .....	xi
<b>ÖZET</b> .....	xiii
<b>CHAPTER ONE</b> .....	1
<b>1. INTRODUCTION</b> .....	1
1.1 Introduction .....	1
1.2 Research Background.....	1
1.3 Problem Statement .....	4
1.4 Purpose of the research.....	4
1.5 Research Aim and Objectives .....	5
1.6 Contribution of Research.....	6
1.7 Research Methodology.....	6
1.8 Structure of the Research .....	7
<b>CHAPTER TWO</b> .....	8
<b>2 . BACKGROUND THEORY - LITERATURE REVIEW</b> .....	8
2.1 Introduction .....	8
2.2 Data Mining: Meaning, Origin and Application .....	8
2.3 Data Mining and Heart Disease.....	12
2.4 Heart Disease, LOS, and Data Mining .....	13
2.4.1 Types of Heart Diseases.....	14
2.4.2 Symptoms of Heart Disease.....	16
2.4.3 Length of Stay- Concerned situation of heart disease.....	18

2.5	Data Mining Techniques and Healthcare Needs .....	19
<b>CHAPTER THREE .....</b>		<b>21</b>
<b>3. DATA-MINING ALGORITHMS TO APPLY AND COMPARE .....</b>		<b>21</b>
3.1	Introduction .....	21
3.2	Naive Bayes.....	21
3.3	WAC (Weighted Associative Classifier) .....	24
3.4	Decision Tree .....	27
3.5	Logistic Model .....	30
3.6	Two-Step Clustering.....	31
3.7	Algorithm Accuracy Measurement .....	31
3.8	Data Presentation and Preparation.....	35
3.9	Data Description.....	35
3.9.1	Input Attributes .....	36
3.9.2	Key Attribute.....	37
3.9.3	Predictable Attribute .....	38
3.10	Descriptive Analysis.....	38
3.11	Correlation Coefficient Analysis (CCA).....	38
<b>CHAPTER FOUR .....</b>		<b>39</b>
<b>4. GENERAL DISCUSSION AND COMPARISON .....</b>		<b>39</b>
4.1	Overview of Algorithm and Technology .....	39
4.2	Existing and Proposed System .....	40
4.2.1	Existing system .....	40
4.2.2	Proposed System .....	41
4.3	Comparison of data mining techniques used.....	43
4.4	Research Feasibility .....	45
4.5	Testing Criteria.....	46
4.6	Summary of Data Mining Techniques .....	48
4.7	Limitations.....	51
<b>CHAPTER FIVE .....</b>		<b>53</b>
<b>5. CONCLUSION AND FUTURE WORK .....</b>		<b>53</b>
5.1	Findings .....	53
5.2	Conclusion.....	55

5.2	Future Work and Recommendation.....	57
<b>REFERENCES</b>	.....	58
<b>CURRICULUM VITAE</b>	.....	70



## LIST OF FIGURES

<b>Figure 1</b> : The process of Data mining.....	9
<b>Figure 2</b> : Traditional decision making approach if the practitioners .....	11
<b>Figure 3</b> : Decision Tree - Basic Illustration.....	32
<b>Figure 4</b> : Explorer window of Weka Workbench.....	32
<b>Figure 5</b> : Knowledge Flow and Data Mining .....	34
<b>Figure 6</b> : Experiment Environment.....	35
<b>Figure 7</b> : User_Module .....	42
<b>Figure 8</b> : Admin Modulet .....	43

## LIST OF TABLES

<b>Table 1</b> : Input Attributes determining the quality og healthcare services.....	36
<b>Table 2</b> : Comparison of Naïve Bayes and Weighted Associative Classifier .....	44
<b>Table 3</b> : Pros and Cons of Data Mining Techniques .....	49

## LIST OF ABBRIVIATION

CHD	Coronary heart disease
CACS	Coronary Artery Calcium Score
WAC	Weighted Associative Classifier
MIS	Management Information Science
KDD	Knowledge Discovery in Databases
WWW	World Wide Web
OLAP	Online Analytical Processing
CVD	Chemical vapor deposition
OLTP	Online Transaction Processing
LOS	Length of Stay
IHDPS	Institute for Health and Disability Policy Studies
CMAR	City Managers' Association Rajasthan
SVM	Support Vector Machines
HRV	Heart rate variability
NCC2	Naive Credal Classifier 2
ODANB	One Dependency Augmented Naive Bayes
CCI	Class Conditional Independence
GUI	Graphical User Interface
ID3	Iterative Dichotomiser 3
CF	Cluster Feature
SQL	Structured Query Language
CSV	Comma-Separated Values
ARFF	Attribute-Relation File Format
CCA	Correlation Coefficient Analysis
SQL	Structured Query Language

## **ABSTRACT**

# **HEART DISEASE SYSTEM PREDICTION USING DATA MINING TECHNIQUES**

AL-AZZAWI, Mohammed Ibrahim

Master, Department of Information Technology

Thesis Supervisor: Assist. Prof. Dr. Yuriy ALYEKSYEYENKOV

March 2017, 70 page

The Quality of Services (QoS) has always been of utmost significance in the healthcare sector since the services entail accountabilities to both the ethical and social perspectives. In today's world of technology, the traditional approaches to treatment based on trials or doctors' experiences have been denied, as proficient decision-making systems have taken place even in the diagnosis system. In this regard, the adoption of data mining techniques for managing patients' data has been facilitating the data management needs of the healthcare sector. Nonetheless, different data mining techniques entail certain challenges associated with the respective proficiency in accordance with the nature of the data. In order to assess the effectiveness of data mining techniques in the healthcare sector, the study has focused the implications in managing the patients' data to predict the likelihood of heart disease in future. Consequently, the researcher has developed a prototype having two different data mining techniques of Naïve Bayes and WAC - Weighted Associative Classifier. Comparing the performance efficacy of both the techniques through the workbench of Weka 3.6.6, and focusing on the symptoms of heart disease, the performance of the developed prototype is affirmed to be a success with

respect to the intended purpose; where multiple input attributes have favoured the classification of the prediction system. Accordingly, the efficacy of prediction and classification techniques of data mining has been approved for diagnostic purposes. The study also presents certain recommendations for developing an intelligent trusted automation system, having the prediction mechanism of vessel stenosis and attributes, and feature reduction for accurate prediction of heart disease. Besides, the adoption of fuzzy approaches to data mining has also been proposed.

## ÖZET

### VERİ MADENCİLİĞİ TEKNİKLERİ KULLANARAK KALP HASTALIĞI SİSTEMİ TAHMİNİ

AL-AZZAWI, Mohammed Ibrahim

Yüksek Lisans, Bilişim Teknolojileri Anabilim Dalı

Tez Danışmanı: Doç. Dr. Yuriy ALYEKSYEYENKOV

Mart 2017, 70 sayfa

Verilen hizmetlerin hem etik hem de sosyal açılardan bir sorumluluk ortaya çıkarması nedeniyle, Hizmet Kalitesi (QoS) sağlık sektöründe en önemli hususlardan biri olmuştur. Günümüz teknoloji dünyasında ehliyetli karar verme sistemleri tanı sisteminde bile yer aldığı için, deneme veya doktor tecrübesine dayalı geleneksel tedavi yaklaşımları hep reddedilmiştir. Bu açıdan, hastaların verilerini yönetmek için veri madenciliği tekniğinin kullanılması sağlık sektöründe veri yönetimini kolaylaştırmaktadır. Yine de, farklı veri madenciliği teknikleri verinin doğasına bağlı olarak yetkinlik ile ilgili bazı zorluklar ortaya çıkarmaktadır. Sağlık sektöründe veri madenciliği tekniklerinin etkililiğini değerlendirmek amacıyla, çalışma hastaların gelecekteki olası kalp hastalıklarını tahmin etmek için hasta verileri yönetimindeki sonuçlara odaklanmaktadır. Sonuç olarak araştırmacı Naïve Bayes ve WAC - Weighted Associative Classifier (Ağırlıklı İlişkisel Sınıflandırıcı) isimli iki farklı veri madenciliği prototipi geliştirmiştir. Weka 3.6.6 iş tezgâhı ile her iki tekniğinde performans etkililiği karşılaştırılarak ve kalp hastalıkları semptomlarına odaklanılarak, geliştirilen prototipin performansı çoklu girdi özelliklerinin tahmin sisteminin sınıflandırmasını desteklemesiyle hedeflenen amaç açısından başarı olarak

onaylanmıřtır. Buna gre, veri madencilięinin tahmin ve sınıflandırma tekniklerini etkililięi tanısal amaçlar için onaylanmıřtır. Çalıřma ayrıca kalp hastalıklarının doęru ve tam tahmini için damar stenozu (darlıęı) ve özellikleri ile özellik azaltma tahmin mekanizmasına sahip yetenek güvenilen otomasyon sistemi geliřtirmek için bazı öneriler sunmaktadır. Bunun yanında, veri madencilięine bulanık yaklařımların uyarlanması amaçlanmıřtır.

## **CHAPTER ONE**

### **INTRODUCTION**

#### **1.1 Introduction**

This section serves as an introduction of the main area of the research. The entire section entails the background information of the effectiveness of the data mining techniques, in relation to the challenges posed by the healthcare system. Consequently, the problem statement, contribution of the research, the purpose of the research, and the adopted methodology for the accomplishment of the objectives of the research are also described in the proceeding section.

#### **1.2 Research Background**

Quality of services in the healthcare sector is a definite prospect to be considered, as its avoidance is not acceptable from ethical and social perspectives. With the increasing awareness of human rights among the society individuals, the healthcare sector has received increasing pressure to deliver quality services, rather than delaying the process of treatment through trial-based tests or judgments based on experiences. It has been a practice to carry out the process of diagnosis of patients' health, based on doctors' experiences and the relevant expertise. Even though, the significance of expertise and experience cannot be denied; yet the resulting consequences are not favourable in all the cases. Patients have to go through numerous tests, based on the anticipated assertions of the doctors. Later on, most of the tests are found to be irrelevant or waste of time and money as well, since the resulting disease comes out to be totally different [1-3]. Most importantly, this particular phase of trial-based diagnosis also leads to delayed treatment that might cause adverse outcomes as well. Therefore, it turns out to be the most significant need to recognize the worth of human life. Patients need adequate and effective treatment, rather than being treated as a source of gaining experience. The entire



system of making decisions regarding patient care must be mechanized or technologically revolutionized at all the levels [4-6].

In this regard, multiple efforts of proficiently improving the management of patients, data have been undertaken. Among certain other approaches of managing the huge amount of data, significant amount of research has been carried out on the data mining techniques [7-12]. It is associated with the innate potential of data mining techniques to manage the entire dataset, regardless of the size or the volume of the data. The complex nature of data mining has been affirmed to facilitate the data management and decision-making needs, based on the inculcation of defined algorithms and past experiences within the existing packages, and offered software solutions [13]. Besides other industries, the field of medical or healthcare has been regarded as the most critical with respect to the needs of effectively and efficiently managing the datasets. The aforementioned assertion is based on the fact that the medical diagnosis involves human life that cannot be taken carelessly. Therefore, the automation of this particular field of interest would be a definite act of kindness towards humanity [2, 3].

Data mining has been facilitating the management of the data in terms of yielding identified patterns from the datasets, which are left unresponsive or ignored if dealt with traditional measures of data management. As a result, a considerable amount of challenging situation has been managed; however, certain issues still prevail. It is asserted with respect to the potential of different techniques of data mining with respect to the nature of the data, and the associated efficacy required [10-12]. The incessant advancements in the technology have been the main reasons of this challenging situation, among the techniques available. Besides, the potential or competence of the personnel involved is also significant in utilizing the benefits offered by the data mining techniques. The implications of data mining techniques have been conquering diverse range of patients' needs [8-12].

However, the current study is entirely focused on the prediction of heart disease threats among the patients, based on the analysis of the patients' health profile. Heart diseases or cardiovascular diseases are characterized by "Stroke (cerebrovascular disease)", "congenital heart disease", "Hypertensive heart disease", "peripheral artery disease", "inflammatory heart disease", and "rheumatic heart

disease". It has been noted that the physical inactivity of the patients, tobacco and alcohol consumption, and unhealthy diet are the main reasons of the occurrence of heart disease among the individuals, regardless of their age. Numerous studies have been carried out in this regard, which have adopted certain statistical measures and data mining techniques of facilitating the professionals of healthcare organisations to [14-20].

Based on the similar needs of efficiently managing the clinical data [15] have proposed an integrated system, having the computer-based management of the patients' data. The system was intended to mitigate the issues caused by the medical errors, as the resulting impacts were observed to adversely affect the patient safety. Therefore, preventive and development measures were crucially required to acquire enhanced patient outcome. In this regard, it was recognised that there is a need of a dedicated system of data modelling and analysing for potentially generating the environment enriched with relevant knowledge of improving the quality delivery of the services and the associated clinical decisions as well. Have adopted the three data mining classifiers of *Naive Bayes* [15], *Clustering Classifier*, and *Decision Tree*, for predicting the diagnosis of the heart patients, with respect to the likelihood of heart related diseases. It was comprehended that the performance outcome of *Decision Tree* was exceptional as compared to the other two techniques. Considering the performance of other two techniques, it has been established that the performance of Naïve Bayes was constant with respect to the nature of the system, while Classification through Clustering could not perform adequately.

Have also strived to predict the likelihood of the events of coronary heart disease (CHD) [7]. In this regard, the authors have considered the *Coronary Artery Calcium Score* (CACS), as a significant factor towards the improvement of CHD prediction. The results have been beneficial towards the intended objectives of mining the data of CACS into the system of CHD prediction. However, the need of a definite data mining technique was still required to make the outcomes effective in the long-run. In addition to this, [20] has also proposed the clinical decision making to be based on the multi-layered processes of evaluating the available data of the patients, since the patients' data has enough information that would direct effective diagnosis results if adequately indentified. It has been affirmed that the poor quality

of the decision making process of the clinicians has been costing even to the life of the patients, which is not desirable in any circumstances. In this regard, the data mining techniques of *Naive Bayes or Bayes' Rule* and *Decision Tree* have been evaluated towards the attainment of well-identified patterns from the patients' data, so that the process of diagnosis could be carried out in an efficient manner [20].

The current study is also a result of acknowledged need of making the process of data management in the healthcare organisations. The researcher has developed a prototype that entails two different techniques of mining the data (*Naive Bayes* and *WAC - Weighted Associative Classifier*); in order to better recommend the most effective approach of predicting the likelihood of heart diseases among the patients.

### **1.3 Problem Statement**

The increasing health issues have been resulting in increasing the concerns of healthcare organisations, regarding the management of the huge amount of heterogeneous data. With the traditional systems of decision making process and the inadequacy of managing the data proficiently, healthcare organisations have been encountering adverse situations [21]. In this regard, the approach of mining the entire dataset has been recognized as an effective one. However, still there lie certain concerns of adopting the most adequate technique to acquire cross-domain knowledge effectively. It is asserted based on the fact that the technology is going through rapid advancements that makes the preceding techniques less effective, as innovations keep on enhancing the potential of the new techniques [21, 22]. Therefore, the adopted methodology of mining the data needs to be potentially proficient in classifying, clustering, and identifying the hidden patterns in the heterogeneous data. If the issues of depersonalisation, heterogeneity, multi-relational aspects of the data are not resolved, the implication of data mining would not be feasible at all.

### **1.4 Purpose of the research**

The healthcare organisations (medical centres and hospitals) have been facing the most critical issue of facilitating the patients with quality services in an

affordable manner. With respect to the element of quality services, it asserts the assurance of correct diagnosis of the patients, along with the effective administration of the treatments. Entailing the aspect of human life, there seems the most critical aspect of making legitimate decisions in the healthcare sector, as the poor decisions are affirmed to lead towards adverse consequences. The entire healthcare system needs to better manage the huge amount of data to be utilized while making decisions. In this regard, the entire data management needs to be computerised, based on the implications of decision-support systems. More specifically, the efficacy of the data towards adequate decision-making needs to be ensured by mining the data [23]. As a result, the hidden information within the mined data is easily discovered that directs the decision making to be effective. Accordingly, the approach of data mining has been adopted within the healthcare sector; however, the effectiveness of the outcomes may vary based on the used techniques' proficiency. Consequently, the healthcare organisations need to explore the variety of available data mining techniques to acquire the best solution to the management of data. This particular research presents a prototype of managing the healthcare data, using data mining techniques of *Naïve Bayes* and *WAC* (Weighted Associative Classifier).

### 1.5 Research Aim and Objectives

The aim of a research is crucial that needs to be determined precisely, so that the relevant objectives of the study could be formulated, accordingly. This particular study is aimed at *the instigation of a system to deal with the complex queries of the healthcare management that are created due to the incompetence of the traditional decision-support systems*. In this regard, the system is expected to predict the likelihood of heart disease rate among the patients, by means of constantly monitoring the factors of blood sugar, blood pressure in relation to the sex and age of the patients. The system is ensured to inculcate significant knowledge in terms of identifying the relationships and patterns in between the factors medically related to the probability of heart disease. The tool or the system can facilitate adequate and effective diagnosis system of the heart patients, with its web-based user-friendly environment. The current study has analyzed the efficacy of the prototype based on two techniques of data mining, with respect to different performance measures.

Accordingly, following objectives have been formulated that would support the accomplishment of the aim of the study, as the development of the prototype needs to be a success:

- To explore the concept of data mining, and understand the techniques of mining the data.
- To identify the needs of healthcare organisations with respect to the implications of data mining techniques.
- To evaluate the effectiveness of data mining techniques as a solution to the data management challenges of healthcare organisations.
- To facilitate the instigation of the system to improve the decision-making process of the healthcare activities and diagnosis system of heart related diseases.

#### **1.6 Contribution of Research**

The information systems in healthcare organisations are designed to deal with the billing needs, and inventory management, along with certain statistical needs as well. Based on the critical nature of data management, some healthcare institutions have deployed decision support systems, but the efficacy of the systems seems limited to some extent. It is asserted based on the fact that the currently deployed systems are capable of dealing with the queries regarding the average age of heart patients, average stay of the patients at hospital, and the gender-based segregation of the patients receiving services, as well. However, the systems are not aligned with the needs of complex queries of prediction-based scenarios, regarding the needs of hospital services, including the stay, diagnosis, and treatment needs. Therefore, the development of a web-based user-friendly system is anticipated to serve the mitigation of the challenges faced by the healthcare organisations, with respect to the management of the heterogeneous data.

#### **1.7 Research Methodology**

In order to develop the system of data mining within the management needs of data warehouses, the current study has evaluated the implications of two techniques of data mining, including *Naïve Bayes* and *WAC* (Weighted Associative

Classifier). The classified, clustered, and identified patterns as a result of the mining of the data have been observed to mitigate the inadequacies of the traditional decision-support systems of healthcare organisations. In this regard, the researcher has acquired the assistance of a variety of materials, including a Personal Computer (CPU: Core i7, RAM: 4 GB), expertise in .Net Programming Language (C#, Visual basic, and others), proficient knowledge of SQL Server 2008, and Visual studio 2010. Besides, the workbench of Weka 3.6.6 has been used for the analysis of the data, in terms of comparing the performance outcomes of the two data mining techniques.

## 1.8 Structure of the Research

The entire thesis completion is structured as:

**Chapter 1:** This chapter has presented the essentials reflecting the needs of conducting this particular study. Among the essentials, the background information representing the healthcare needs of an innovative system deployment has been described. Besides, the research purpose, problem statement, research contribution, the formulated aim and objectives, and the adopted methodology for the accomplishment of those objectives have also been described in this section.

**Chapter 2:** It is basically the detailed literary section, based on the topic under consideration. Previous research attempts and the associated literary findings supporting the intended objectives of this particular research are also presented.

**Chapter 3:** This section is the detailed demonstration of the adopted methodology. The areas focused, materials used, and techniques deployed are discussed in this chapter.

**Chapter 4:** This section contains the generated results or findings as a result of the instigated system of mining the heterogeneous data of the hospitals.

**Chapter 5:** Based on the results or findings, the concluding remarks are presented in this particular chapter. Moreover, certain recommendations have also been provided to enhance the credibility of the research, and the potential of the deployed system within the field of interest.

## **CHAPTER TWO**

### **BACKGROUND THEORY - LITERATURE REVIEW**

#### **2.1 Introduction**

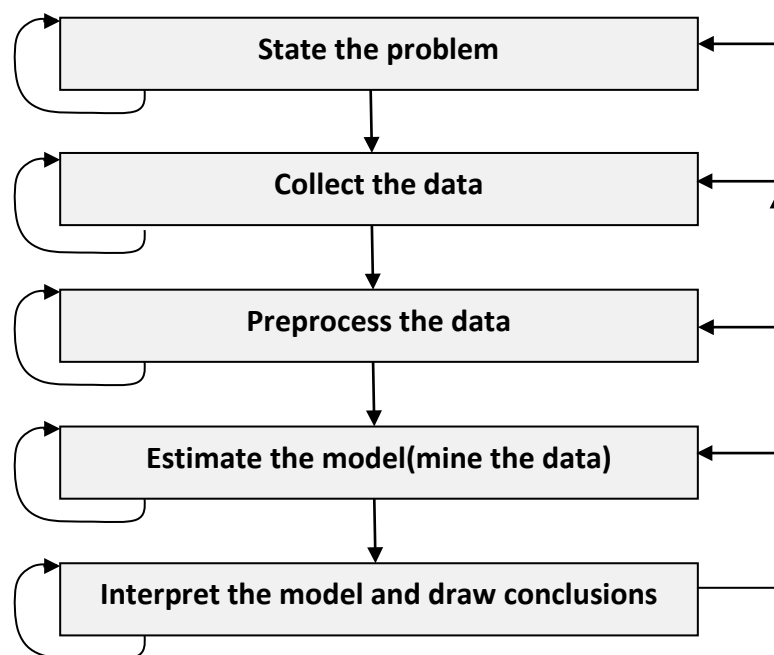
This section presents the literary findings of the data mining needs in the management of organisational data. The objectives of evaluating the implications of data mining techniques in the prediction of the likelihood of heart diseases among the patients have been assessed in accordance with the past studies conducted in this regard. The proceeding section presents the basics of data mining, along with the variety of techniques that are efficient in dealing with the complex queries of managing the data. Moreover, the healthcare needs of efficient and accurate management of the patients' data are also discussed in this section. Accordingly, the literature intends to assert the value of the most efficient method to be deployed within the healthcare sector.

#### **2.2 Data Mining: Meaning, Origin and Application**

The process of knowledge discovery has been of critical significance in the databases; thus, it needs to be well-defined, entailing various types of knowledge acquisitions to attain identified datasets. Data mining is regarded as the core step in the successful discovery of the valuable information from the bulk of databases. Formally, data mining is defined as the process that facilitates the extraction of embedded data from the huge amount of databases in a non-trivial manner. Moreover, it is ensured that the extracted data is characterized as being formerly unidentified, but potentially valuable to the involved organisation [5, 24, 25]. The technology of data mining offers the personnel involved a "user-centred approach" to the unique and hidden attributes within the bulk amount of data. It is asserted that the acquired

knowledge discovery would assist the administration of healthcare organisations in improving the service quality.

Besides, the discovered knowledge is also affirmed to be facilitating the medical practitioners or clinicians, in order to decline or mitigate the adversities caused by a particular drug effect, along with the recommendation of cost-effective or affordable measures of treatment or certain other alternatives [17, 26, 27]. More specifically, the implications of data mining algorithms are also contended to be potentially advantageous towards the anticipation of the future behavior of the patients, based on the evaluation of the patients' history.



**Figure 1:** The process of Data mining

Historically, the notion of identifying valued patterns from a huge amount of data has been named in different ways, including data mining, information discovery, knowledge extraction, information harvesting, processing of data pattern, and data archaeology as well. With respect to the recent trends of MIS "Management Information Science" in the data warehouses or databases, the terms "data mining" or "KDD - Knowledge Discovery in Databases" are observed to prevail. Data mining then becomes the core area of the entire KDD process that involves the inference of algorithms that are aligned to discover the patterns that have been unidentified or

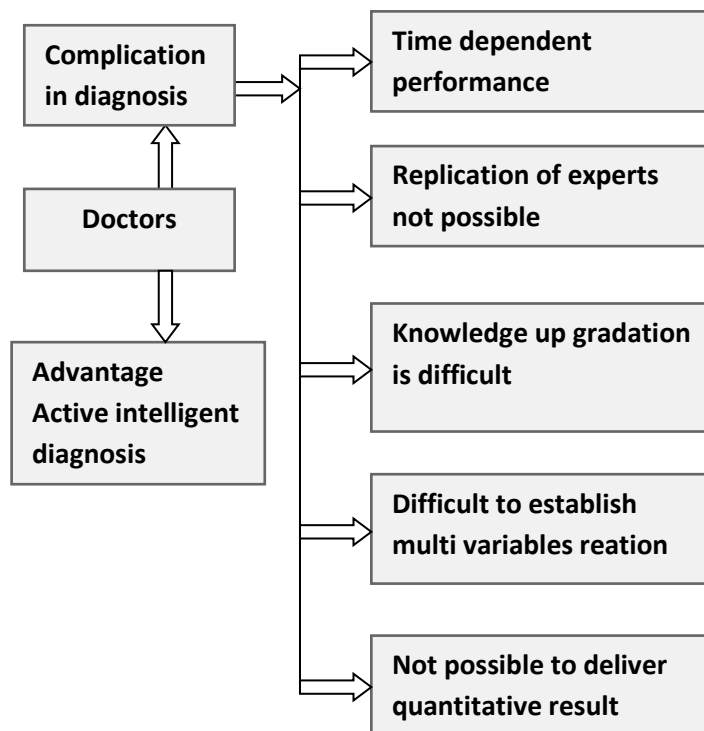


unknown, followed by the exploration of the entire datasets, and the development of the model [12, 28, 29]. This model facilitates the understanding of the entire process of analyzing the data to head towards prediction.

Previously, the concept of data mining dates back to the period of 1960s. It was the time when the systematic evolution of information technology and database had taken place. The data management acquired a shift from primitive processing of files to the powerful and sophisticated systems of databases. Later in the period of 1970s, it was observed that the database systems transformed from network and early hierarchical structure to the relational nature of database systems. It was then practiced to employ the tools of data modelling, and the methods of indexing the data to be accessed [12, 30]. Besides, in this period, the integration of query languages, optimized processing of queries, user interfaces and the management of transition provided the users flexible and convenient access to the data. It was the mid-1980s period, when the relational technologies were adopted, in reaction to the increasing surge of research and development needs on the sophistication of the database systems.

It was followed by the extensive exploration of the features of distributed, diversified, and shared data. It led towards the recognition of WWW (World Wide Web) as a notation of internet-based information systems or heterogeneous systems of global level, within the industry of information technology. It headed towards the convergence of the areas of Information Retrieval, Artificial Intelligence, and statistical database systems, which carried out the integration of fast speed microcomputers to better retrieve and analyse the data. Finally, it was in 1990s, when data warehouses were developed, as a collaboration of the transactional and the operational databases. Later on, the advanced systems of OLAP, Decision support systems, staging or scrubbing of data, and algorithms of association rule were developed. It was the time, when data mining acquired the position within the business practices, rather than just being an innovative measure of technology. It was accomplished due to the increasing datasets, along with the associated needs of effective and efficient management of the data [1]. Besides, the acquisition of new customers and the attainment of revenue growth have also been the major causes of integrating data mining within the businesses.

According to the study of Mosadeghrad [31], the provision of quality service has become the most challenging situation for the healthcare organisations, due to the inadequate decision making systems. The situation gets even worse, when affordability of the patients is considered, towards the medical tests or the treatment plans. By means of delivering quality services, the healthcare organisations are accountable to facilitate the patients with effective and correct measures of diagnosis and treatment as well. Certainly, the poor decision making of the clinicians would definitely lead towards adversities that cannot be regarded as acceptable or somewhat reasonable (figure below represents traditional decision making approach if the practitioners). The need of deploying an efficient information management system cannot be denied within the massive databases of healthcare [18, 30]. The data may include patient-related data, the transformed data, or even the data of resource management. At all the levels, the healthcare organisations are expected to fulfil the data management needs of all the areas. With the integration of data mining algorithms, it is anticipated that the record maintenance of millions of the patients would eventually assist the practitioners in dealing with the critical or complex queries of healthcare.



**Figure 2:** Traditional decision making approach if the practitioners

In accordance with the study results [32], it has been established that even the perceptions and the associated needs of the practitioners, patients, or clients have shifted from qualitative to quantitatively visualized assessments, based on the facilitation of integrated information handling of the patients. Moreover, the quantitative integration of the information is also demanded to be carried out with the inclusion of all the clinical data and imagery information as well. For instance, the practitioners would be capable of carrying out comparison of the diagnostic details of multiple patients characterized by identical conditions. Besides the comparison of the patients' data, the practitioners can even take assistance of other practitioners, dealing with the same findings, regardless of their location [32].

It is noteworthy to mention here that the prospects of comparison and assistance from the practitioners having identical experiences are based on the ultimate need of making precise and efficient decisions of medical diagnosis; thus, automation is the most efficacious solution in this regard [16, 19, 30] has regarded data mining essential to the organisational needs of knowledge discovery. This process involves an iterative series of data clean-up, integration of data, selection of the most relevant data, and the pattern recognition and presentation of knowledge. It highlights that data mining accomplishment through the stages of description of class, its association, along with the classification and clustering as well. Afterwards, the patterns are identified that lead towards the stages of prediction by means of time series analysis. As a result, data mining is defined as a process that is discovery driven, contrasting to the traditional decision support systems.

### **2.3 Data Mining and Heart Disease**

By means of Heart Disease, it is comprehended that the diverse nature of diseases that have influence over the heart or its activity are regarded as heart diseases. Globally, the death rate due to cardiovascular or heart disease is affirmed to be 31% of all the deaths (17.5 million) [33]. In the United States, at least 610,000 people annually encounter death due to heart disease [34]. Within the European region, the heart disease causes almost 2 million deaths on annual basis [35]. Multiple conditions are included in the aspect of heart disease that are influential to the blood vessels, and thus, to the entire heart activity of pumping the blood across

the body. Heart disease or cardiovascular disease (CVD) has adverse impacts, which might lead towards severe sickness, disabilities, or even death. If the coronary arteries are narrowed, there is a definite reduction in the supply of oxygen and blood to the heart, which eventually causes CHD "Coronary Heart Disease". In case of an abrupt blockage of a coronary artery, there are increased chances of heart attack that is mainly because of the blood clotting. Besides, there could be certain other symptoms, as if the patient might have chest pains that would lead towards heart attacks. Therefore, the practitioners and the patients are expected to consider such particular signs as critical, as the inadequate blood reception of the heart muscles is undoubtedly a threatening sign.

It is observed that most of the healthcare organisations have the deployment of OLTP [18] (Online Transaction Processing) as a hallmark of effective utilisation of the data. However, the system is not aligned with the needs of decision-making process, and the associated needs of pattern analysis to acquire the most relevant datasets. It is contended that the organisations must have efficient data management mechanism that would empower the staff and the entire management, based on the tools of knowledge discovery and critical thinking. As a result of such a strategic decision making process, the success of the organisation turns out to be ascertained. There could be the deployment of data mart, data mining algorithms or OLAP "Online Analytical Processing", as a source of supporting the decision making process of the organisation.

#### **2.4 Heart Disease, LOS, and Data Mining**

Basically, heart is the most vital organ of the body as it keeps the body functional. The Performance activity of the heart needs not to be affected or disturbed in any means, as it directly affects the health of the individuals. Even the resulting impacts are utmost adverse that the individuals may encounter death. Any disorder or abrupt condition in the flow of blood causes heart disease to occur. Therefore, the practitioners and the patients are accountable to critically notice and report the symptoms, in order to mitigate or control the likelihood of adverse situations [36]. With respect to the symptoms of heart disease, chest pain is regarded as the basic and the most critical one. However, it cannot be practiced to just focus the chest pains, as

there might be certain conditions that are not characterized by any specific symptom condition, yet are life-threatening. For instance, atherosclerosis has no notable symptom, but excessive complications are associated with its prevalence. Based on this assertion, it has been established that the patients' health profile or the data must be evaluated in terms of predicting the likelihood of heart disease. In this regard, certain risk factors must be focused, including age and sex of the patients, blood pressure and blood sugar levels, obesity, along with the consumption patterns of alcohol and smoking as well.

#### **2.4.1 Types of Heart Diseases**

The patients' data may help the practitioners to better predict the possibility of being affected by any heart disease. However, the entire datasets cannot be reviewed; thus, effective ways of identifying the valued datasets need to be utilized. It is asserted based on the fact that the data of patients would direct the practitioners to recommend the most appropriate or required tests and treatments [37]. Among types of heart disease, five most common diseases are described in the proceeding section:

**Congestive Heart Disease:** Congestive refers to the blocked way of the pumping of blood from heart to other parts of the body. This congestive performance of heart if left unnoticed or untreated, results in periodically reducing the efficiency of heart. The patients going through congestive arteries are observed to face shortness of breath, edema, swelling, and even kidney problems. As a result, even the weight of the patient might increase as well. Besides these symptoms, the root cause of this congestive condition might be alcohol abuse or the patients' records of high levels of blood pressure [38]. The practitioners may predict the likelihood of this particular heart disease by reviewing the patients' history, in terms of being alcoholic, suffered from a heart problem, or even family history of heart problem. Therefore, the existence of any of the symptoms must direct the process of treatment without having delays of diagnosis tests. The diet of the patients should be changed, exercise must be recommended, and alcoholic substances must be limited as well, rather than recommending useless tests or medical evaluations based on mere judgments or experiences.

**Congenital Heart Disease:** Congenital refers hereditary aspects of a particular disease. Based on the prospect of associated with the family, this particular disease is regarded as being inevitable as the development of the heart is defected. Even though this particular disease cannot be treated, yet the practitioners may guide the patients accordingly, rather than recommending inadequate tests or preventive measures. Undoubtedly, it requires adequate management of patients' history to guide the diagnosis process [39].

**Coronary Heart Disease:** Being the most common disease of heart, this particular disorder leads the patient to encounter heart attacks. It happens due to the reduced blood supply to the heart that results in damaging the heart by means of depositing fats on the blood vessels' linings. As a result, the vessels are narrowed that eventually minimizes the blood flow. The patients then suffer from pain that is regarded as the pain of angina. If the prevalence of this particular disease of heart is intended to be mitigated or controlled, the practitioners must consider the symptoms, critically [36]. Among the most common symptom, cholesterol level is the main determiner, as the higher levels of cholesterol direct the increasing proportion of fatty deposits. Additionally, the consumption of tobacco and cigarettes may directly assert the probability of the heart disease. Therefore, the evaluation of the patients' health profile is crucial, if the increasing rate of heart diseases is expected to be controlled.

**Rheumatic Heart Disease:** This particular disease is the derivative of strep infections of throat. Although throat infections seem to be the of least concern, with respect to the probability of heart disease, yet the avoidance of sore throats lead towards certain infections that eventually affect the performance activity of heart. However, throat infections can only be avoided by the patients, but should not be disregarded by the practitioners, if prediction of the heart diseases is intended [40]. The patients' history of prolonged throat infections must trigger the likelihood of heart disease, if the valued patterns are adequately identified and highlighted.

**Pulmonary Heart Disease:** Having a history of lung problems, the patients need to have awareness of its association with heart disease. If lungs are in

complicated condition of blocked or slowed blood flow, the resulting increase in pressure over the lungs affects the blood supplier as well. The shortness of breath, chest pain, dyspnoea, and syncope must be promptly comprehended as the directives of pulmonary heart disease, rather than triggering the disease further, by delaying or adopting useless diagnosis measures [41]. If left untreated or avoided, the disease ultimately causes death; thus, the practitioners must have access to the identified patterns of the patients' data.

#### **2.4.2 Symptoms of Heart Disease**

The principal symptoms that may direct the prediction of the likelihood of heart disease need to be recognized. It is asserted based on the fact that the data mining would yield the identified patterns from the huge amount of datasets, which would be focused on the most concerned areas of interest. Therefore, the algorithms need to be focused on the indications of chest pain, dyspnoea, heart palpitations, discomfort, fatigue, syncope, daytime sleepiness, and lethargy as the warning signs.

**Chest pain or Discomfort:** Feeling discomfort in berating or experiencing chest pain is directly associated with the probability of heart disease. Although chest pain is usually ignored, its repeated occurrence might direct the individuals towards heart disease. The individuals experiencing chest pain might wrongly interpret the feeling or the origination of the pain, as it could be the feel of pain, squeezing, pressure, numbness, choking, or any discomfort associated with the upper abdomen, neck, chest, arms, head, or even jaw [42]. If the pain prevails for long or the individuals encounter the pain on frequent basis, the practitioners must promptly regard it as a symptom of heart disease, and recommend the most appropriate treatments.

**Dyspnoea:** Dyspnoea refers to the abnormal breathing, which is eventually a symptom of pulmonary and cardiac disease. The exhausting routine of even the healthy individuals or moderate experiences of the individuals that are not accustomed to physical activities might result in dyspnoea. However, its repeated

appearance should not be avoided as it may lead towards the likelihood of heart disease among the patients [43].

**Syncope:** It refers to the state of unconsciousness or lost consciousness, which is caused by the reduction of perfusion in the brain. The patients encounter a sudden fainting experience that lasts for a shorter period of time. In most of the cases, this could be the outcome of exhausted lifestyle. However, it needs to be figured out promptly, as the converse outcomes could be life-threatening if left untreated. The lost consciousness could be due to metabolic, neurologic, cardiac, and vasomotor factors, which ultimately results in death [44]. Therefore, the history of the patients must be evaluated based on this factor as well, if the prediction of heart disease is expected to be provided.

**Fatigue/Daytime Sleepiness/Lethargy:** These symptoms seem to be the most common, as almost everyone experience fatigue in daily life. However, if the condition gets prolonged, and the individuals are unable to perform the daily activities at normal pace, the situation then demands intense consideration. The situation of feeling drowsy all the time or even falling asleep suddenly is regarded as Narcolepsy, which requires prompt treatment. If it carried the added impacts of fatigue as well, the probability of heart disease increases [45]. Therefore, any abnormality or repeated occurrence of even normal symptoms must be medically evaluated.

**Heart Palpitations:** If an individual experiences increased heart rate, it somehow, is related to the cardiac problem. However, the associated reason would determine the intensity of the likelihood of heart disease. People experience increased heart rate or even observe skipping or irregular heartbeat; thus, it is referred as abnormal rhythm of heart. Therefore, abnormal rhythm of heart may lead towards heart disease that needs to be predicted promptly [46]. Any abnormal activity related to heart should not be avoided as its frequent occurrence makes the individual suffer from critical conditions of heart disease.



### 2.4.3 Length of Stay- Concerned situation of heart disease

With respect to the disease of heart, another most important element to be considered is of "Length of Stay - LOS". LOS is the number of days for which an individual remains hospitalized. Heart disease makes the LOS prolonged; thus, it has been observed that the practitioners or the hospital administration try to make the LOS shorter, in order to control the costs of hospitals. However, the length of stay cannot be decided based on mere assertions or previous experiences. There is an intense need of carrying out proper evaluation of the patients' data to predict the severity of the sickness, in relation to the identified cost of health, and the utilisation of the resources of health care. Therefore, LOS becomes a major element of determining the cost of facilitating the patients, in terms of allocating the hospital resources [47]. Adequate and accurate prediction is essential based on the fact that the hospitals' numbers of beds for the patients are limited. Most importantly, the administration is observed to have limited financial resources as well, which substantially increases the financial pressure. Thus, reduction of costs needs to be achieved at all the levels; however, the approach of ignoring the health status of the patients, and minimizing tier LOS cannot be regarded as acceptable [48].

Data mining has been contended as the best technological solution to determine the adequate LOS for a patient in a hospital. Has asserted that the success of a hospital administration is governed by the hospital's prediction and evaluation of the LOS [49], regardless of the concerns of being laborious. If the prediction of LOS and heart disease is precisely carried out, the decisions regarding the resource allocation could be efficiently made. LOS has been the most common, but important aspect of predicting heart disease. However, certain other critical aspects are also there that would enhance the credibility of the prediction of the likelihood of the heart disease among the patients. It is also observed that the inadequacy of predicting heart disease also result in failure of the measurement of the admission requests of the patients in future [48]. In this regard, it has been contended that the entire process of decision-making, related to the patients' care require a proper system, rather than carrying out based on the mere judgments.

## 2.5 Data Mining Techniques and Healthcare Needs

A number of studies have been conducted to present the most effective and efficient data mining tool for dealing with the prediction needs of the healthcare organisations. Accordingly, different probabilities have been acquired as the research outcomes, with respect to different algorithms of data mining. Has developed an intelligent system for predicting the heart disease (IHDPS) [18]. The system entailed the data mining techniques of Neural Network, Naive Bayes, and Decision Tree as well. As a result, it was observed that all the three techniques had their respective strengths. It was observed from the identified relationships and patterns within the datasets, based on the user-friendly, web-based system [50]. Likewise [18], has adopted multiple classifiers of data mining. By means of using Bayesian Classifiers, Decision Tree (C4.5), Multiple Association Rules (CMAR), and Support Vector Machine (SVM), the authors have developed an accurate and proficient system of measuring the "*Heart rate variability*" (HRV).

By means of just using Neural Networks, [51] have efficiently predicted the likelihood of blood pressure, heart disease and even diabetes. The entire execution of the involved activities was carried out by the supervising algorithm of back propagation, in order to train and test the gathered data. In the same manner, [52], has focused on the issues of identifying the association rules that are constrained towards the prediction of heart disease. The identified patterns were reduced based on certain constraints associated with the attributes. As a result, two groups were created, characterized by the absence and presence of heart disease. Has evaluated the issues of medical area [18], based on the data mining algorithm of decision trees. Adopted the classification approach to extract data regarding the prevalence of HRV [18].

Have conducted a survey of multiple data mining techniques as a source of predicting heart disease [16]. The techniques included Decision Tree, Naive Bayes, Neural Networks, KNN and others. It has been observed that Decision Trees and Bayesian classification have been efficacious as compared to the techniques of Neural Networks and Clustering based Classification and even KNN. More specifically, the integration of genetic algorithm has further enhanced the credibility of the two algorithms of Decision Trees and Bayesian Classification. Has yielded the

results that among multiple techniques of data mining [17]; Neural Networks holds greater significance, as the outcomes have been exceptional towards the prediction of heart disease among the patients. Besides, the efficacy of Decision Tree has also been supported, as the accuracy of the prediction seemed high as well.

Have evaluated the effectiveness of the data mining techniques of Decision Tree [14], Artificial Neural Network, Naive Bayes, and Rule Based on the huge amount of healthcare data. In order to make the process of decision making to be effective, the algorithms of NCC2 "Naive Credal Classifier 2" and ODANB "One Dependency Augmented Naive Bayes" Classifiers have been used. As a result, even the incomplete datasets have been evaluated to provide the information. Besides, the even the factors of sex, age, blood sugar level, and blood pressure level were also considered while predicting the likelihood of heart disease among the patients. Cumulatively, it has been established that the data mining algorithms facilitate the prediction of the likelihood of heart disease in a considerably efficient manner [18-20]. The current study is based on the same assertions, as the factors undertaken by [14], have significant contribution to the successful prediction of the likelihood of heart disease among the patients.

## **CHAPTER THREE**

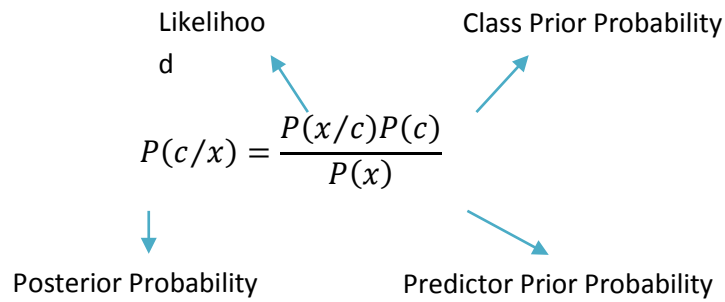
### **DATA-MINING ALGORITHMS**

#### **3.1 Introduction**

This section presents the detailed description of the data-mining algorithms that have been selected for applying and comparing in the case of healthcare needs of predicting the likelihood of heart diseases among the patients. Even though, the research has compared two algorithm techniques of Naive Bayes and WAC (Weighted Associative Classifier), this section has reviewed multiple other techniques as well.

#### **3.2 Naive Bayes**

With respect to the classifiers of Bayesian method, it is affirmed that all require Weka workbench for the implementation. The classifiers include Naive Bayes, Complement Naive Bayes, ADOE, and Multinomial Naive Bayes, as the mostly preferred ones. Basically, Bayes' Theorem is the origin of Naive Bayes that represents that the relationship in between the predictors of the datasets is characterized by independent assumptions. There is no prospect of complexities or impediments with respect to the iteration; thus, the algorithm is affirmed to be feasible for comparatively larger datasets. Even though, this particular classifier represents simple facets, the performance outcomes are remarkable that makes it comparable and potentially competent with multiple other classifiers [53, 54].



$$P(c/x) = P(x_1|c) \times P(x_{21}|c) \times \dots \times P(x_n|c) \times P(c)$$

Where,

$P(x/c)$  = Predictor's probability, ensure the class is given

$P(c)$  = Previous class's probability

$P(c/x)$  = The target's probability (sub-sequent class), ensuring that the given attribute represents the predictor.

The Naive Bayes algorithm facilitates the target users by means of providing the chance of subsequent probability computation that is always based on the predictors. As a result, the entire process is also regarded as "*Class Conditional Independence*".

#### **ALGORITHM Naive Bayes Classifier**

Let the training dataset  $D$  consist of  $n$  points  $x_i$  in a  $d$ -dimensional space, and let  $y_i$  denote the class for each point, with  $y_i \in \{c_1, c_2, \dots, c_k\}$ . The Bayes classifier directly uses the Bayes theorem to predict the class for a new test instance,  $x$ . It estimates the posterior probability  $P(c_i | x)$  for each class  $c_i$ , and chooses the class that has the largest :

Where:

$D_i$  = training dataset

$n_i$  = instance points size

$X_j$  =  $d$ -dimensional space (test point)

$y_i$  = class for each instance point

$c_i$  = the class

$\hat{P}$  = posterior probability (Likelihood)

$\hat{\mu}_i$  = The sample mean

$\widehat{\Sigma}_i$  = covariance matrix for class

$Z_i$  = centered data matrix for class

$\widehat{y}$  = The predicted class

Then the bayes classifie for agiven training dataset ( $\mathbf{D} = \{(\mathbf{X}_j, y_j)\}_{j=1}^n$ ): **It** can be calculated using the following algorithm

**A. for  $i = 1, \dots, k$  do**

To classify points, we have to estimate the likelihood and prior probabilities directly from the training dataset  $\mathbf{D}$ . Let  $\mathbf{D}_i$  denote the subset of points in  $\mathbf{D}$  that are labelled with class  $c_i$  as following :

**B.  $\mathbf{D}_i \leftarrow \{ \mathbf{X}_j | y_j = c_i, j = 1, \dots, n \}$  training dataset**

**C.  $n_i \leftarrow |\mathbf{D}_i|$  let  $n_i$  = instance points size**

Let the size of the dataset  $\mathbf{D}$  be given as  $|\mathbf{D}| = n$ , and let the size of each class-specific subset  $\mathbf{D}_i$  be given as  $|\mathbf{D}_i| = n_i$ . The prior probability for class  $c_i$  can be estimated as follows:

**D.  $\widehat{P}(c_i) \leftarrow n_i/n$  posterior probability (Likelihood)**

To classify a numeric test point  $x$ , the Bayes classifier estimates the parameters via the sample mean and sample covariance matrix. The sample mean for the class  $c_i$  can be estimated as

**E.  $\widehat{\mu}_i \leftarrow \frac{1}{n_i} \sum_{x \in \mathbf{D}_i} \mathbf{X}_j$  The sample mean**

the sample covariance matrix for each class can be estimated using as follows, where  $Z_i$  is the centered data matrix for class  $c_i$ .

**F.  $Z_i \leftarrow \mathbf{D}_i - \mathbf{1}_{n_i} \widehat{\mu}_i^T$  centered data matrix for class**

**G.  $\widehat{\Sigma}_i \leftarrow \frac{1}{n_i} Z_i^T Z_i$  covariance matrix for each class**

The posterior probability is then given as following :

**H. return  $\widehat{P}(c_i), \widehat{\mu}_i, \widehat{\Sigma}_i$  for all  $i = 1, \dots, k$**

Further, because ( $\mathbf{x}$  and  $\widehat{P}(c_i), \widehat{\mu}_i, \widehat{\Sigma}_i$ , for all  $i \in [1, k]$ ) remains fixed for  $\mathbf{x}$ , we can predict the class for  $x$  by following :

$$\widehat{y} \leftarrow \operatorname{argmax}_{c_i} \{ f(x | \widehat{\mu}_i, \widehat{\Sigma}_i) \cdot P(c_i) \}$$

Having a set of objects that is associated with an identified class with recognised vector of variables, there is an observed need of constructing a rule

capable of assigning approaching objects to the class, in a manner that the future objects are described by the variables' vectors only. Such problem cases have been regarded as the issues of supervised classification due to being ubiquitous; thus, resolved through different rules, which include the most credible contribution of naive Bayes [55]. The importance or significance of this particular rule of classification lies in its easy-to-construct attribute, as there is no need of complicated or typical estimations of iterative parameter schemes. Accordingly, it turns out to be efficient in dealing with huge amount of datasets. The algorithm is easy in terms of interpretation aspects as well; thus, offers flexibility and feasibility even to the unskilled users as well. Consequently, it is referred as potentially reliable in terms of being robust in performance [55, 56].

Being a statistical method, Naive Bayes is capable of predicting the probabilities of class membership, considering the fact that the attributed values of the classes are independent of the influence of others [56]. In this regard, it has been contended that the classifier considers the presence or even absence of a particular feature totally discrete, regardless of the presence or absence of any other specific feature of the class [57]. For instance, a problem case of a series of  $n$  attributes turns out to yield  $2^n$  assumptions that are independent, if Naive Bayes classifier is employed. These prospects based on the probabilistic approach makes Naive Bayes potentially effective even in complex situations of real-world [57]. On the other hand, certain flaws or errors are also reported with respect to the performance of Naive Bayes Classifier. These errors may include the following factors:

**Training Data Noise:** The selection of efficient training data would minimize the impacts of noise [57].

**Variance:** It occurs if the datasets are too small; thus, associated with the training data noise, and can be mitigated if appropriated datasets are selected [57].

**Bias:** The datasets of training are first divided into multiple groups. If the machine learning algorithm divides the data into larger groups, **Bias** error occurs [57].

### 3.3 WAC (Weighted Associative Classifier)

In this particular technique, associate rules are used for the classification of the discovered patterns from the databases. The selected items or datasets are

evaluated based on the analysis approaches of correlations or associations that are facilitated significantly by WAC. While exploring the association rule, there is no need of class attribute; thus, representing unsupervised approach of learning. Conversely, classification demands supervision as the class attributes are the essentials of the classifier construction, which is followed by the prediction of the data [50]. Having an integrated impact of both the association and classification aspects, WAC becomes increasingly accurate algorithm to determine the predictive values. Therefore, WAC algorithm is feasible for the prediction needs of the environment requiring intensive accuracy.

$$WSP (A \rightarrow Class_{label}) = \frac{\sum_{x=1}^{|w_i|} weight(w_i)}{\sum_{y=1}^{|n|} weight(w_i)}$$

Has referred WAC as a new concept that integrates the use of Confidence Framework and Weighted Support for the purpose of extracting the rules of association from the repository of data [58]. The WAC reflects exceptional performance outcomes that are ascertained with its potential attribute of having no insignificant relation among the variables. For instance, the implications of WAC within the heart-related datasets are remarkable, since the data is pre -processed for being encountered with the mining technique. Accordingly, the classifier assigns a particular *weight* range of 0 to 1 that adds the attribute of being important throughout the prediction model. Within the allocated range for *weight*, the more influential attributes are weights as closer to **0.9**, while the less effective attributes are weighted as almost **0.1** [16, 58].

Afterwards, the pre -processed data is ready to undergo the implications of mining rules of WAC algorithm. Once these rules are generated through the weighted element of *Support and Confidence* rather than the traditional ones, these are gathered in **Rule Base**. Whenever, there is a new entry into the record management, the *Class Association Rule* activates for predicting the label of the class that is accessed right from the *rule base*. There are training datasets that are represented as {attribute ( $a_i$ ), value of attribute ( $v_i$ ), weighted value ( $w_i$ )}. Accordingly, the integrated element of weight illustrates the importance of the particular attribute throughout the processing [58]. Based on these specific measures



deployed over the datasets, studies have affirmed the efficient performance of WAC, in terms of extracting the patterns in the most accurate manner. Undoubtedly, the prospect of *weighted attributes* serves as the mark of identification for WAC [16].

- 1) **Attribut Weight:** Attribute weight is assigned depending upon the domain.
- 2) **Attribute set weight:** Weight of attribute set X is denoted by W(X) and is calculated as the average of weights of enclosing attribute. And is given by

$$W(X) = \frac{\sum_{i=1}^X \text{weight}(a_i)}{\text{Number of attributes in } X}$$

- 3) **Record weight/Tuple Weight:** Consider the data in relational table, the tuple weight or record weight can be defined as type of attribute weight. It is average weight of attributes in the tuple. If the relational table is having n number of attribute then Record weight is denoted by W(r<sub>k</sub>) and given by

$$W(r_k) = \frac{\sum_{i=1}^{r_k} \text{weight}(a_i)}{\text{Number of attributes in a record}}$$

- 4) **Weighted Support:** In associative classification rule mining, the association rules are not of the form  $X \rightarrow Y$  rather they are subset of these rules where Y is the class label. Weighted support WSP of rule  $X \rightarrow \text{Class\_label}$ , where X is set of non empty subsets of attribute-value set, is fraction of weight of the record that contain above attribute-value set relative to the weight of all transactions. This can be given as

$$WSP(X \rightarrow \text{Class\_label}) = \frac{\sum_{i=1}^{|X|} \text{weight}(r_i)}{\sum_{i=1}^K \text{weight}(r_i)}$$

Here n is the total number of records.

- 5) **Weighted Confidence:** Weighted Confidence of a rule  $X \rightarrow Y$  where  $Y$  represents the Class label can be defined as the ratio of Weighted Support of  $(X \cup Y)$  and the Weighted Support of  $(X)$ .

$$\text{Weighted Confidence} = \frac{\text{Weighted Support } (X \cup Y)}{\text{Weighted Support } (X)}$$

### 3.4 Decision Tree

Has regarded Decision tree as the most efficient and credible algorithm of data mining [59]. Decision tree is among the techniques of data mining, which is interlinked with the conventional statistical technique of linear regression. Besides, Decision tree also entails the prospects of cognitive sciences that is instilled in its neural networks. Since its inception, it used to be developed imitating human brain methods of detecting patterns and forming concept accordingly [60]. Accordingly, decision tree is affirmed to be simple but potentially effective in analytical needs of multiple variables. This particular classification of data mining techniques serves as a unique substitute of the typically complex analytical techniques of multiple linear regression, certain other neural network-related algorithms of data mining techniques, and multiple other analytical tools of business intelligence [59, 60].

Decision tree has been contended as a classifier that is characterised with its capability of expressing the instance spaces as recursive partitions. There are nodes forming *Rooted Tree* like appearance, making the classifier to be a *directed tree*, which has a node as "*root*" with no defined incoming-edges. The remaining nodes are characterised by one exact incoming-edge, while *test* node or *internal* node is the one having outgoing edges [59, 60].

#### **ALGORITHM Decision Tree Classifier**

can be calculated using the following algorithm

Where,

$D = \text{training dataset}$

$X = \text{Point}$

$n = \text{number of points (partition size)}$

$X_i = \text{attribute}$

$X_j = \text{Numeric decisions}$

$\pi = \text{purity threshold}$

$c_i = \text{class}$

$D_y = \text{subsets (corresponds to all points } x \in D \text{ that satisfy the split decision)}$

$D_N = \text{subsets (corresponds to all points that do not satisfy the split decision)}$

Let our training dataset is  $D$  then sort  $D$  on attribute  $X$ , so that  $1 \leq X_j \leq n - 1$

a)  $n \leftarrow |D|$  put partition size in  $n$

Different split points are evaluated for each attribute in  $D$ , and the size of class  $c_i$  can found and put in  $n_i$

b)  $n_i \leftarrow |X_j | X_j \in D, y_j = c_i \}$

calculated purity threshold by following :

c)  $\text{purity}(D) \leftarrow \max_i \left\{ \frac{n_i}{n} \right\}$

The stopping condition for given dataset, If the number of points  $n$  in  $D$  drops below the user-specified size threshold  $\eta$ , then we stop the partitioning process and make  $D$  a leaf. This condition prevents over-fitting the model to the training set, by avoiding to model very small subsets of the data. The stop point can be calculated using following steps.

d) if  $n \leq \eta$  or  $\text{purity}(D) \geq \pi$  then

1.  $c^* \leftarrow \text{argmax}_{c_i} \left\{ \frac{n_i}{n} \right\}$  // majority class

2. create leaf nod, and label it with class  $c^*$  return

Then we should initialize best split point or set of midpoints (The simplest condition is based on the size of the partition  $D$ )

e)  $\text{split point}^*, \text{score}^* \leftarrow (\emptyset, 0)$

The best split point is chosen to partition the data into partition  $D$  into  $D_y$  and  $D_N$   $\text{split point}^*$ , and call recursively by following steps :

f) foreach (attribute  $X_i$ ) do

1. if  $X_i$  is numeric then

a.  $(v, \text{score}) \leftarrow \text{EVALUATE} - \text{CATEGORICAL} - \text{NUMERIC}(D, X_j)$

b. if  $\text{score} > \text{score}^*$  then  $(\text{split point}^*, \text{score}^*) \leftarrow (X_i \leq v, \text{score})$

2. else if ( $X_j$  is categorical) then

a.  $(V, \text{score}) \leftarrow \text{EVALUATE} - \text{CATEGORICAL} - \text{ATTRIBUTE}(D, X_j)$

b. if  $\text{score} > \text{score}^*$  then  $(\text{split point}^*, \text{score}^*) \leftarrow (X_i \leq v, \text{score})$

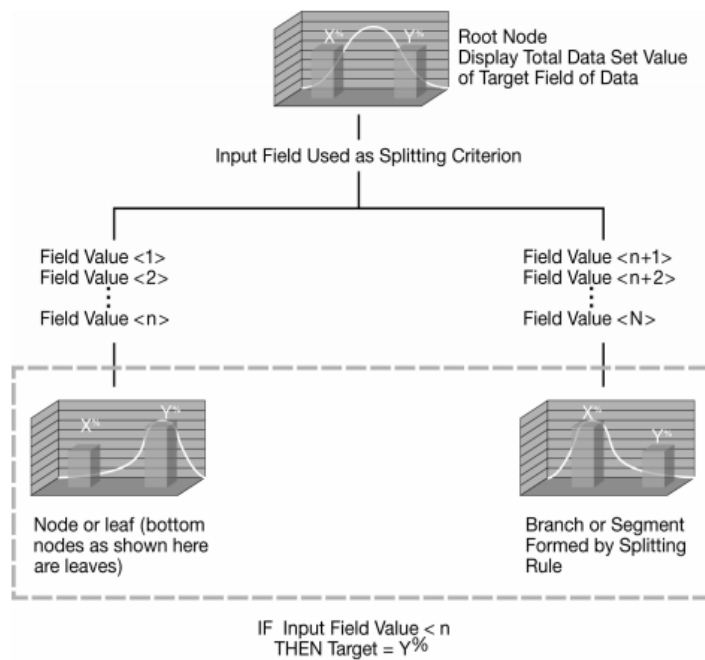
create internal node *split point\**, with two child nodes,  $D_Y$  and  $D_N$

- g)  $D_Y \leftarrow \{X \in D | X \text{ satisfies a } split\ point^*\}$
- h)  $D_N \leftarrow \{X \in D | X \text{ does not satisfy } split\ point^*\}$

Result we get : **DECISIONTREE**(  $D_Y$  ) **DECISIONTREE**(  $D_N$  )

Additionally, the remaining nodes are terms as *Leaves/Terminals*, whose count is dependent on the implications of some specific discrete function of the attributed values of the input. Each leaf is characterized with having an adequate target value or it may have a probability vector as well, having instilled value of the attributed target [59].

Generally, the decision makers are observed to prefer uncomplicated or basic decision trees due to the integrated comprehensibility, based on the respective impacts of the accuracy of the outcomes. It is also stated that rule induction is reflected in the induction of decision tree, since each leaf's functionality is conjoined with tests, considering the class prediction of the leaf as the class value. Below is the basic illustration of the decision tree, presenting the perspectives of both the categorical and continuous object of analysing [60].

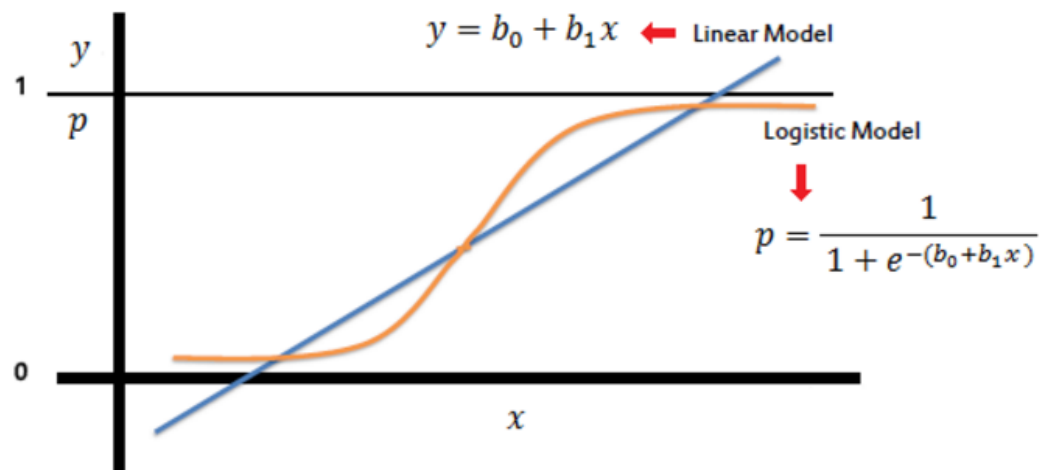


**Figure 3: Decision Tree - Basic Illustration**

The Graphical User Interface (GUI) of the workbench of Weka provides the users opportunity of directly formulating the decision trees, rather than the traditional essentials of source code for executing Decision Tree algorithm over the datasets. The classification panel front of the analysis tool has different learning algorithms of *Decision Tree* that may include C4.5 clone (or J48), ID3, random trees, and certain other decision tree algorithms as well [59].

### 3.5 Logistic Model

Logistic regression reflects the statistical technique of estimation for the mapping of the object with respect to the real or authentic value of the value of prediction. The regression approach utilizes curve fitting, forecasting (prediction), causal relationships, hypotheses testing and multiple other approaches between the variables, whether linear or logistic regression models. Logistic regression outperforms linear regression as it potentially predicts binary values that are beyond the acceptable range of 0 to 1 [61]. Although logistic curve represents the values in between 0 to 1, the use of "odds" as the natural algorithm differentiates it from linear regression. In addition to this, there is no compulsion of the predictors to have equal variance or normal distribution.



- The constant ( $b_0$ ) is referred as the determiner of the movement of the curve to the left.
- Slope ( $b_1$ ) refers to the steepness of the curve of the logistic regression.

- Following is the equation of the logistic regression:

$$\frac{p}{1-p} = \exp(b_0 + b_1x)$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x$$

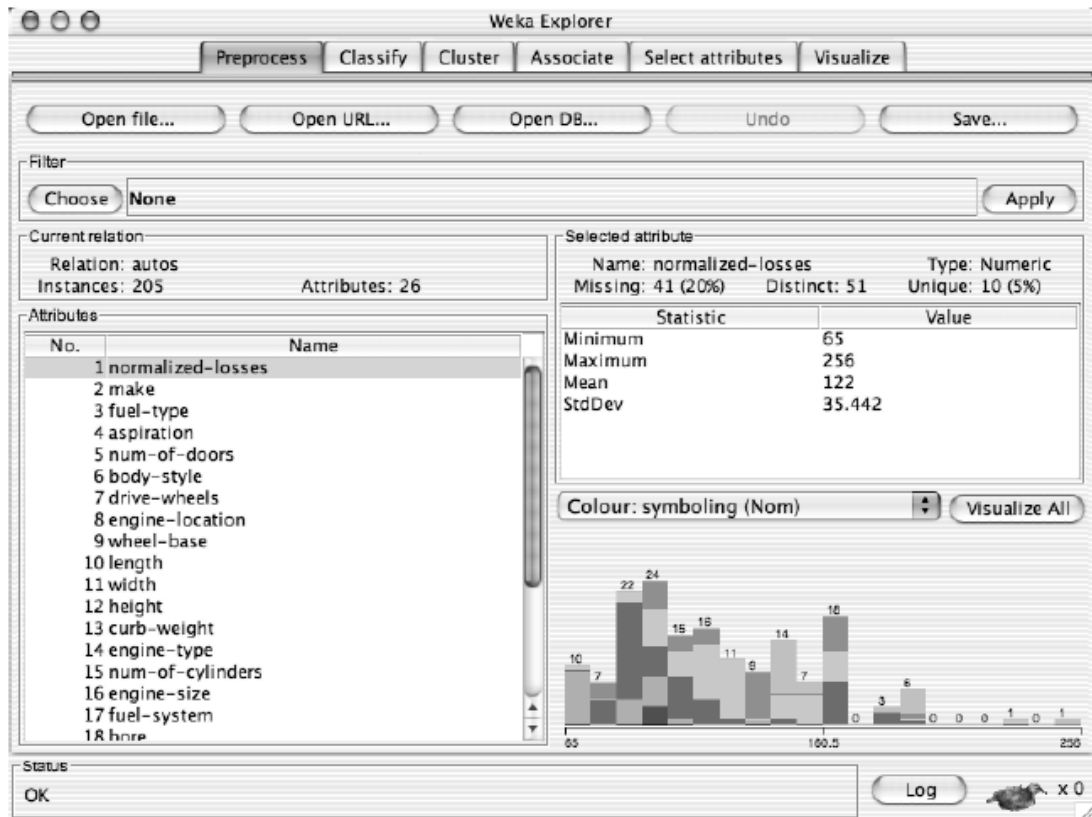
$$p = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p)}}$$

### 3.6 Two-Step Clustering

This particular algorithm serves as a tool that explores the entire dataset in order to reveal naturally occurring clusters or groups that are not apparently identifiable. Based on multiple features, this particular algorithm is differentiated from conventional approaches of clustering. The algorithm potentially handles even the continuous and categorical variables by means of regarding the variables as independent in terms of a joint collaborating of normal and multinomial distribution. Besides, the number of clusters is automatically selected, which is carried out by the comparison of the values of a particular criterion across multiple solutions of clustering [12]. Furthermore, even scalability is also facilitated greatly that is ensured by the CF "Cluster Feature" that is characterized with the summary of records.

### 3.7 Algorithm Accuracy Measurement

With respect to the measurement of the accuracy of the algorithms used, the research has adopted the tool of Weka workbench. Then entire process of data mining goes through the Weka workbench, including the setting of the entire input data, the analysis of the adopted algorithms in terms of statistical measures, and the overall process and the involved essentials of the input and the results, all are visualized as well. Regardless of the algorithm or the technique of data mining, Weka workbench performs exceptionally with all the clustering, regression, association rules, and classification aspects of data mining. However, it is noteworthy to mention here that all the algorithms are characterized with their respective relational tables, as a result of the response to the queries within the database [5, 62, 63].



**Figure 4:** Explorer window of Weka Workbench

Figure 3 represents that the main window or the GUI of the Weka workbench offers multiple panels to be used for different operations of data mining. The description of each panel is presented in the section below:

**First Panel:** At this stage, SQL query supports the loading or extraction of the data from a particular database or a file as well. The supported formats of the uploaded file are CSV, and ARFF. The access to the database for the purpose of approaching the SQL queries is ascertained by *Java Database Connectivity*. Afterwards, the retrieved datasets are allowed to undergo multiple operations of *Filtering*, *Normalizing* and others.

**Second Plan:** The processes or activities of classification and regression are offered at this panel of the Weka workbench. The process is carried out on the dataset that is already processed. Most importantly, the performance outcomes are visible both in terms of statistical and graphical manner.

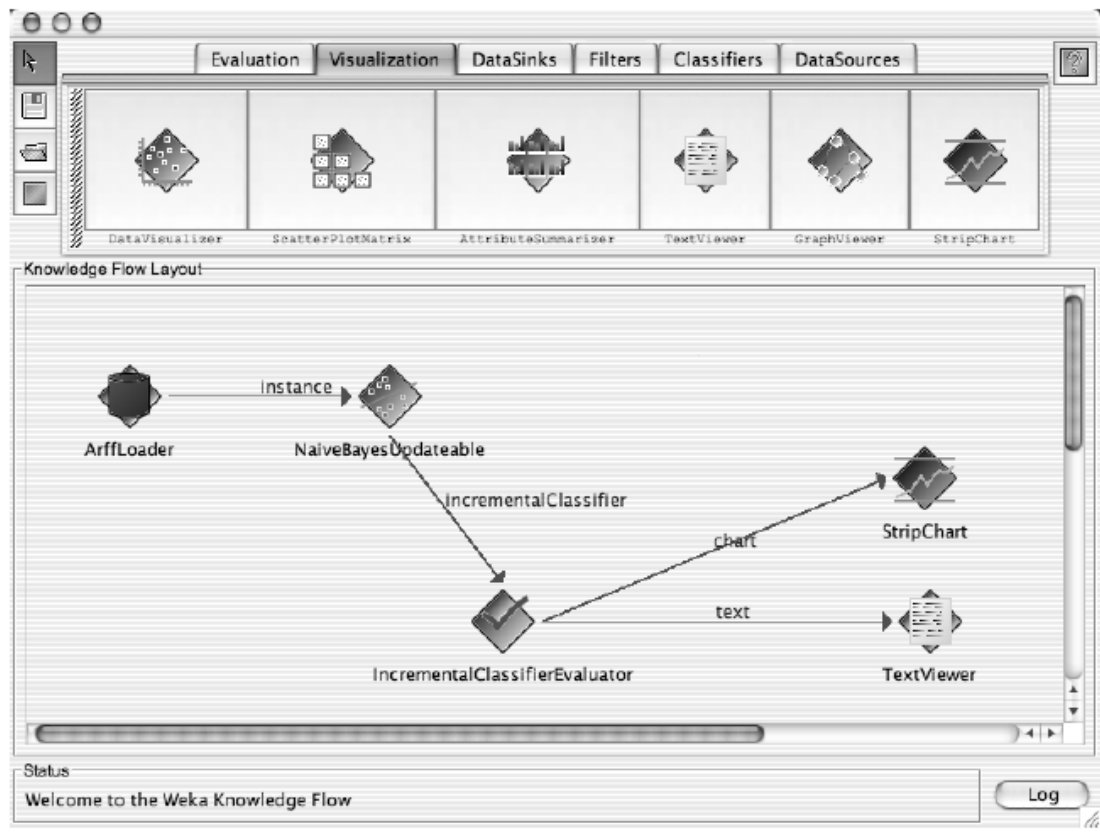
**Third Panel:** Clustering is carried out at this panel of the Weka workbench, entailing the element of visualizing the statistics of the entire process. However, it is crucial that the estimates of density are also represented.

**Fourth Panel:** This stage or the panel of the Weka workbench is based on the process of the rules of association. The rules are comprehended by means of accessing the algorithms.

**Fifth Panel:** Datasets that are mostly characterized based on their predictive nature are offered to the users for the implications of the data mining methods [63, 64].

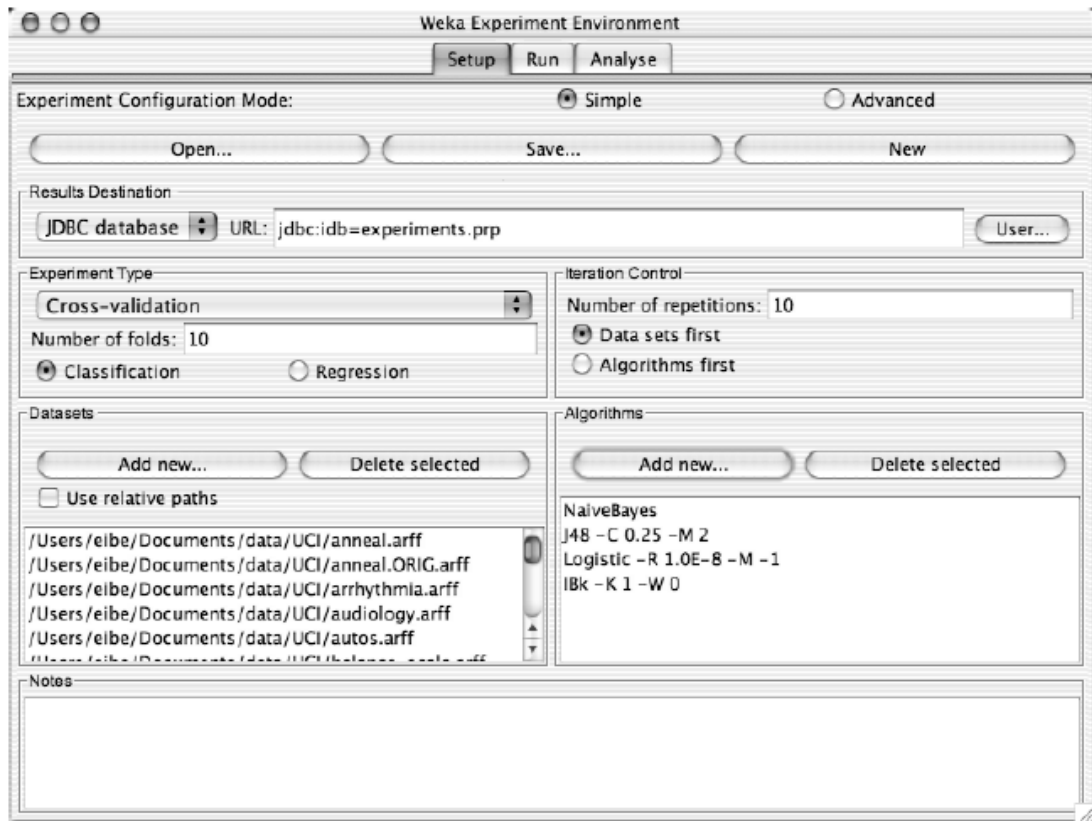
It is contended that this particular workbench is feasible for small to medium sized datasets. In this regard, the workbench offers another interface of "Knowledge Flow", represented in the figure below, facilitating the needs of database evaluation by means of the options of the data source, algorithm to be learned, tools of pre-processing the data, and the tools of visualizing and evaluating the entire datasets. Resultantly, incremental learning is offered that could not be attained by the *Explorer GUI*.





**Figure 5:** Knowledge Flow and Data Mining

Numerous data mining options are also provided to the users with the workbench of Weka, in order to discover solution to a particular issue or problem. In this manner, the users can adopt the best suited algorithm, as the outcomes of each algorithm are easily compared and contrasted. This interface of Weka workbench is regarded as "*The Experimenter*" (shown below). More specifically, the integrated element of "*Java Remote Method Invocation*" makes it possible for the user to perform all possible computations.



**Figure 6:** Experiment Environment

### 3.8 Data Presentation and Preparation

Data mining, being the process of exploring the massive datasets to generate the patterns to be utilized for predicting the likelihood of diseases, has acquired a notable position even in the business environment. With respect to the objectives of this particular study, the efficacy of data mining algorithms of *Naive Bayes* and *WAC* (Weighted Associative Classifier) has been evaluated within the healthcare industry. In this regard, the researcher has adopted 15 attributes, regarding the heart disease. Previously 13 attributes have been adopted, but two more attributes of smoking and obesity have been explored with the data mining algorithms. The data mining techniques have been *Naive Bayes*, and *WAC* (Weighted Associative Classifier).

### 3.9 Data Description

Since accurate and efficient healthcare is the need of general public, the researcher has used the public database of heart disease. The database records of Cleveland heart disease (303 records), and Statlog database (270 records) were used

in this particular research, to be evaluated in terms of the selected algorithms of data mining. The entire dataset was comprised of three distinctive demarcations of *Input Attributes*, *Key Attributes*, and *Predictable Attributes*. The detailed datasets are presented in the section below:

### 3.9.1 Input Attributes

**Table 1:** Input Attributes determining the quality of healthcare services

No.	Attribute	Values	Description
1.	Sex	Male is referred as 1 Female is referred as 0	Representation of Male or Female
2.	Age	Continuous	Representation of age in <b>years</b>
3.	thebstps	Continuous, value to be measured in mm of hg	Blood pressure characterized by resting condition
4.	cp	Four values scaled: Typical (Type 1) refers as 1 Typical (Type angina) refers as 2 Non-angina pain refers as 3 Asymptomatic condition refers to 4	Type reflecting pain in chest
5.	Restecg	From 0-2: Normal is referred by 0 Having abnormal wave of ST_T is referred as 1. The left ventricular hypertrophy is referred as 2.	Results of resting electrographic
6.	chol	Continuous, having the representation of mm per dl.	Level of serum cholesterol

7.	fb	Two scales have been designed: 1 $\geq$ 120 mg per dl 0 $\leq$ 120 mg per dl	Blood sugar while fasting
8.	slope	If unslopy, it is referred as 1. If flat, referred as 2. If downsloping, referred as 3.	Slope of the exercise at the peak of ST segment
9.	thalach	Values are continuous	The rate of heart that is maximum
10.	thal	Normal, referred as 3. Fixed, referred as 6. Reversible defect, referred as 7.	Type of defect
11.	exang	No means 0 Yes means 1	Angina induced by exercise
12.	ca	Values range in between 0-3	The vessels that are coloured by means of fluoroscopy
13.	oldpeak	Values are continuous	The depression of ST that is induced by the exercise that is relevant to the rest
14.	Smoke	Past, referred as 1 Current, referred as 2 Never, referred as 3	Smoking
15.	Obes	Yes means 1 No means 0	Obesity

### 3.9.2 Key Attribute

The key or main attribute has been selected to be the identification number of the patient: **PatientID**.

### 3.9.3 Predictable Attribute

The predictability of heart disease in terms of diagnosis is represented as:

No heart disease ( $1 = < 50\%$ )

Has Heart Disease ( $0 = > 50\%$ )

### 3.10 Descriptive Analysis

With respect to the analysis of the acquired data, the tool of Weka 3.6.6 has been utilized. It is noted that there were certain values that were missing at the initial stage. The values were identified based on the efficacy of the data mining tool, through its filter of *ReplaceMissingValues*. Multiple other techniques were also applied on the datasets, including *Confusion matrix*.

### 3.11 Correlation Coefficient Analysis (CCA)

Correlation deals in the areas of statistics and probability that are aligned to provide the solutions for the strength measurement of the dependency among the numeric variables under consideration. If the two variables (random variables) under consideration are independent, the relationship will surely be uncorrelated [65]. Coefficient of correlation is illustrated as,

$$P = P(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Where,

$\sigma_X$ , and  $\sigma_Y$  = standard deviation of the variables being compared.

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(X.Y) - \mu_X \mu_Y$$

It is established that the correlation in between the variables is the strongest, if the covariance has the highest absolute value. Besides, it is also noted that the coefficient of correlation entails certain other interesting aspects as well. At first,  $\rho$  takes the quantities that are dimensionless in the range of  $-1$ , and  $1$ ;  $\rho \in [-1, 1]$ . Moreover, there is no dependency of shifting or scaling [65].

$$P(a_X X + b_X, a_Y Y + b_Y) = P(X, Y) = P$$

## **CHAPTER FOUR**

### **GENERAL DISCUSSION AND COMPARISON**

#### **4.1 Overview of Algorithm and Technology**

Heart disease prediction system has established a highly distributed environment of the internet by focusing on the drawbacks of the existing system. Microsoft .NET framework has been used as a front end technology; whereas, Microsoft SQL server has been employed as a backend software. The instigation of .NET framework as a front end software has been used to endow an effective object-oriented programming environment, which reduces the versioning conflicts and software deployment. The core intention of applying this technology was to strengthen the communication process based on the executed object code. The additional and explicit features of .NET framework provides the development of third-party runtime hosts along with the runtime hosts of proposed technology. Thereby, .NET framework enables internet deployed software to be exceptionally considering effective by considering the security features of the runtime. .NET framework has been implemented to expand the productivity of developed as well as it enhances the performance of heart disease prediction system.

Microsoft SQL server has been employed in the heart disease prediction system for sustaining the association between information in the database. The instigation of SQL server provides effectiveness in recovering the consistency of complete information if system failure occurs. The feature of SQL server comprehensively managed the client/server system for organizing the data effectively. By using backend technology, individuals can connect easily with the database system to acquire all information from a server. It has been adhered that the server communication of SQL server further facilitates the communication among SQL server and application running on the server. Easy deployment and use of SQL features have provided management to manage the system overall the organization.

The database system of the Microsoft SQL server comprises of information that support processing activities of heart disease prediction system. To fulfil the requirements of scalable applications, ADO.NET framework has been instigated in heart disease prediction system to address the user requirements. The standalone features of ADO.NET framework allows management to employ this technology in the heart disease prediction system.

As mentioned earlier, naïve Bayes algorithm is implemented in prediction system by considering the main factors for prediction. The naïve bayes algorithm is used to identify the existing information from dataset when user access to the system. The naïve bayes algorithm is concerned to address the queries of patient in regards to heart disease factors. Naïve bayes algorithm predict the symptoms or disease at some stage after initializing all the queries and information of patient. The prediction of naïve bayes is based on the demographic variables, including age and gender.

## **4.2 Existing and Proposed System**

### **4.2.1 Existing system**

The provision of quality services at affordable costs is becoming the major concern for healthcare organizations. The instigation of decision support system and computer-based information systems in healthcare organizations have provided much support in endowing quality services. The core intention of these systems is to organize the information of patients regarding patient. However, the problem exists in this scenario is that this information is insufficiently used in supporting the decision-making. Another issue observed from the implementation of existing system is the restriction towards the use of decision support system. These systems are restricted in predicting the symptoms of diseases. The complex queries cannot be addressed effectively by these systems due to their limitations.

The existing system comprises of numerous hitches that causes complexity in maintaining the system. The drawbacks of existing system reveals probability of not providing accurate and appropriate consequences. In this regards, the symptoms or predictors cannot be diagnosed properly through this systems. Healthcare professionals require user friendliness systems for identifying the predictors or

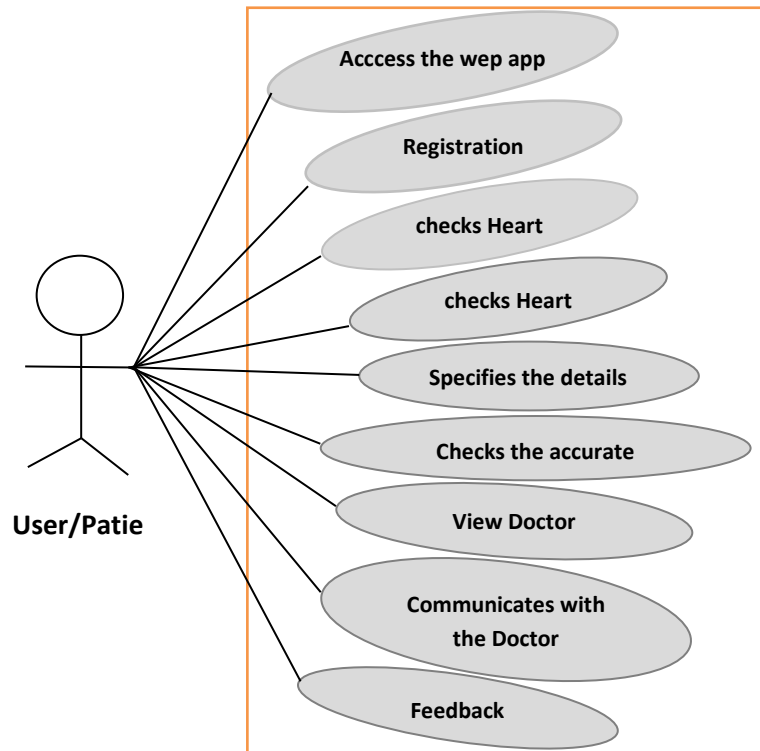
symptoms but the existing systems are not much efficient in this approach. Due to this negativity, these systems devour ample time for processing the activities.

#### **4.2.2 Proposed System**

The processing of activities in the system is based on the initial analysis, which lacks due to the drawbacks of the existing system. This study has proposed a heart disease prediction system to support healthcare professionals in predicting symptoms related to severe heart disease issues. The proposed healthcare system is instigated on the basis of online consultation project and it is an end user support system. By considering the drawbacks of existing system, the proposed system is much effective in providing instant guidance for users regarding their heart disease. The massive storehouse of the proposed system allows users to explore information about assorted symptoms that is associated with the backend information about patients. In this regards, intelligent data mining techniques have been used to support the proposed system and to ensure the information of patients are organized effectively. The proposed system comprises of 2 modules that effectively handles the activities of doctor and provides treatment to the patient.

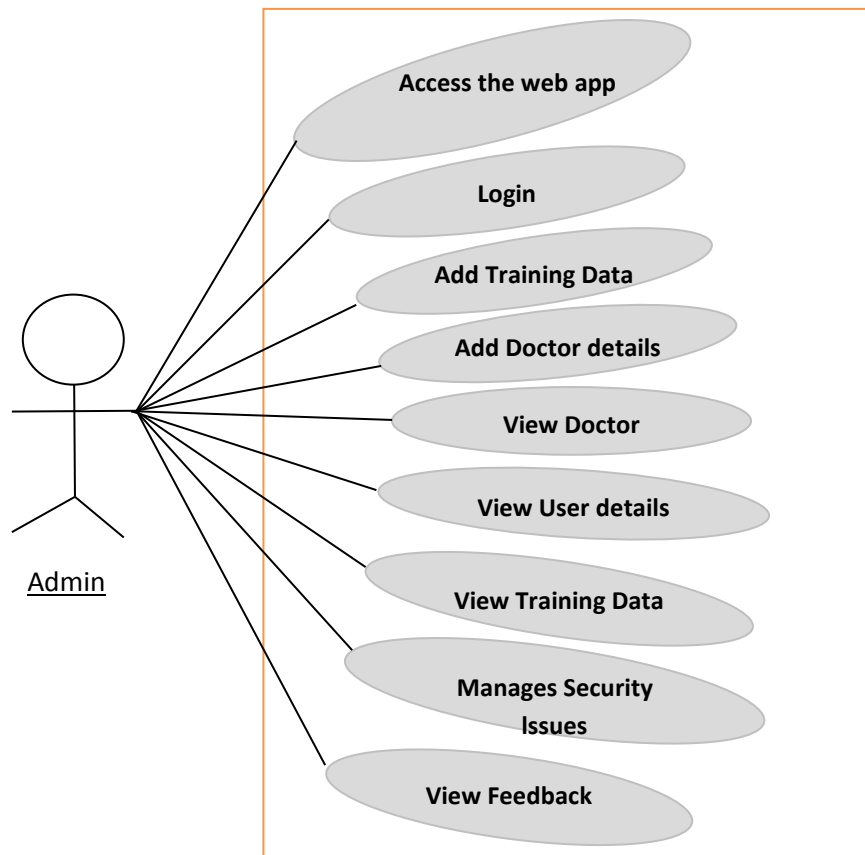
The modules of heart disease prediction system are further depicted through the case diagrams, which reveal the activities of patient and admin system. As illustrated in figure 7, the user module has allowed patients to share the information regarding their disease on the system in order to get accurate results and treatment for the doctor. Heart disease prediction system allows patients, interacting with doctors to get feedback about their disease. The proposed system allows patients to register online for acquiring guidance without visiting the consultant personally.





**Figure 7: User Module**

In addition, the admin module comprises of information regarding patient and doctor as illustrated in figure 8. Moreover, it is also used to store data into its database system for providing feedback to the management. The admin module of heart disease prediction system also assists in controlling the security issues regarding invalid data feed. The complete sequence of both modules are presented in the sequence diagram, which explicitly listed the steps used in the heart disease prediction system and the accessibility of user.



**Figure 8:** Admin Module

### 4.3 Comparison of data mining techniques used

Data mining techniques have been efficient in determining the diagnosis and prediction of different heart disease types. Prior studies have presented and implemented different data mining techniques for the prediction of heart disease [66, 67]. These techniques have been effective in managing the large databases of the patients as well as used symptoms for the prediction of risk factors. In this study, naïve bayes and weighted associative classifier are used to predict the heart disease. These two techniques have been appropriate in predicting the heart disease and health care data. The comparison of these two techniques have been essentially revealed in the prediction of heart disease [68].

Naïve bayes classifies a separate phenomenon whereby input features can be participated on the individual basis. This independence phenomenon has been described through a platform, which comprises of aspects such as the target variable,

the classification and the nodes [18]. Thereby, the naïve bayes classifier works sufficiently due to the underlying assumption of independence. It has been recognized that the independence assumption classifies effective predictive features, revealing the independence of the class. Simplicity and easy implementation are included in the advantages of the classifier, establishing it a prominent method to be executed for a new classification problem [69]. Pros and cons of naïve bayes and weighted associative classifier have been listed in the table below.

**Table 2:** Comparison of Naïve Bayes and Weighted Associative Classifier

Data mining technique	Pros	Cons
Naïve bayes classifier	<ul style="list-style-type: none"> <li>Less model complexity</li> <li>Easy to implement</li> <li>Work efficiently</li> </ul>	<ul style="list-style-type: none"> <li>Cannot be used on large dataset</li> <li>Poor when independence assumptions are not fulfilled.</li> <li>Super simple</li> </ul>
Weighted associative classifier (WAC)	<ul style="list-style-type: none"> <li>Better accuracy</li> <li>Robust and predictive.</li> <li>High accuracy</li> </ul>	<ul style="list-style-type: none"> <li>Minimum support</li> </ul>

The naïve bayes model is represented through a binary vector in the space of words. Binary variables describe the independent Boolean for the inputs. The instigation of naïve bayes classification determines the text documents in the form of word counts. The classifier develops a log linear decision rule assigning an independent parameter to measure the comparative degree of association among the class-word pairs [70]. According to the study, naïve bayes has been implemented to detect the risk factors of heart disease system among patients. The monitoring of heart disease is fundamental as it becomes the reason of fatality [71].

The accessibility of patient record is designed using a graphical user interface for the prediction of heart disease. The prediction of heart disease system is effective

in which records are excavated through a repository. The implementation of naïve bayes classifier has been proposed for the prediction of heart disease using greedy feature selection. The main notion behind the implementation was to identify the attributes in order to accomplish the accuracy of classification. The input and target attributes were categorized in accordance to the procedure or diagnosis codes [17].

Weighted associative classifier has been implemented on the basis of weighted association rule for the classification. Confidence framework and weighted support are used by weighted ARM for the extraction of associative rule acquired from data repository [72]. The implementation of weighted associative classifier has been effective in acquiring significant results as compared to other classification algorithms. Prior studies have significantly implemented weighted associative classifier on the basis of association rules [17, 73]. These associative rules has been used without the provision of supplied support values to predict the heart disease system. The data mining technique in the weighted associative classifier discovers the mining problem by allowing attributed weight to every mining issue. It has been investigated that weighted associative classifier comprises of linguistic terms in the monitoring of heart disease system [74].

The instigation of naïve bayesian approach has been designed to consider that particular data mining technique is important in the healthcare domain. The prediction of disease is not identified comprehensively from all the symptoms existed in the healthcare domain [75]. Thereby, weighted associative classifier has been used to classify diverse symptoms of particular disease. The assigning of each attribute's weight is based on the domain based weights. The issue of downward closure property invalidation has been addressed in weighted association rule mining by assigning a weight. The formulae was proposed for the weighted support and weighted confidence on the basis of quantitative and Boolean weighted items [76].

#### **4.4 Research Feasibility**

Preliminary studies have been used in this research to demonstrate the applicability and feasibility of data mining techniques for the prediction of heart disease system. The development of these studies has been specific in determining the appropriateness and visibility of data mining techniques in heart disease

prediction. Healthcare sector has been specifically mentioned for demonstrating the feasibility of data mining techniques [77]. Various limitations have been observed regarding data mining techniques in heart disease prediction even though it is recognized as an effective approach in numerous sectors. The theoretical demonstrations have been discussed in this study for the feasibility of implementing data mining technique. The comprehension of data mining techniques has been effective in accomplishing prediction of heart disease by evaluating the feasibility of the implementation [5].

The healthcare sector has faced massive augmentation towards collected and generated electronic data that are complex to organize. This approach has been successful in fulfilling requirements of data-warehousing technologies and database management systems. Thereby, the need for data mining techniques have been increased for quicker data analysis. The rapid expansion of neural networks have enabled the healthcare sector to investigate the massive and complex data [78].

Naïve bayes and weighed associative classifier have been dominant in comparing cost-estimating model by using three different models. These approaches have been significant in addressing the heart disease prediction. It has been explored that competitive results are required for the direct comparison of traditional statistical methods [51]. The applicability of naïve bayes and weighted associative classifier has been effective in indicating diverse symptoms of a disease. The feasibility of data mining technique lies in the minimization of time and effort of the automated user. Furthermore, the prediction of heart disease system is monitored effectively through these data mining techniques. The classification of these approaches are feasible in determining the input attributes [5].

#### **4.5 Testing Criteria**

One of the most important technique is cross validation for the estimation of generalizing performance of a model that has been predicted. The main notion behind the cross validation is to break or divide the data once or more than once for the estimation of risk of each algorithm. It selects the algorithm with very little amount of estimated risk, whereas, it can also be considered as an alternative of random subsampling. Hold-out is also a simple validation method that depends on a

single division of data. It is the simplest type of cross validation and the data is divided into two separate series called the testing or training set.

The training set is only used to fit the function approximate. Then the output values are asked to be predicted for the data in the testing set. The mean absolute error is provided by the overall errors it makes and evaluates the model. This method can be repeated various times to enhance the estimation of the performance of classifier. This approach can also be termed as a subsampling approach, but this technique encounter some possible problems associated with the hold out method. K-fold Cross-validation is one of the way to enhance the holdout method and the data is segregated into k sets.

The hold out method is repeated k times and every time the k sub-sets are put together to create a set of training. The average errors are then computed considering the k trails. When the number of k is increased, the variance of the resultant estimate decreases. Unlike all the methods that have been discussed in which sampling was applied without replacement, the bootstrapping method records the samples with replacement. The already selected data for training is put back to the original data set and it is equally possible to be redrawn.

Testing criteria is constructed to ensure that the activities of heart disease prediction system are accomplished successfully. The testing evaluation is based on the success of the project in order to verify that each component in the prediction system performs effectively. The object code of the system has been written comprehensively on the ASP.NET along with C# as a front-end designing interface. During the implementation of components, few applications were inaccurate which were taken into consideration for the correction. Testing system was merely based on four testing criteria, which includes acceptance testing, system testing, integration testing and unit testing.

The implementation of user acceptance testing has been done to investigate the user acceptance for the system in which requirements are needed at the time of development. The investigation of system is based on the criteria of system testing, which effectively analyzes the accuracy and efficacy of the system for live operation. The objective of the system are achieved when the entire components of the system performs accurately. The construction of program structure is tested in the

integration testing, which considers the unrevealed flaws exists in the interface. The module and software design are verified in the procedure of unit testing in which modules are tested independently.

#### **4.6 Summary of Data Mining Techniques**

Data mining techniques is potentially implemented in every sector even though it has been existed around for numerous years. Data mining techniques comprise of machine learning, database technology and statistical analysis for extracting secreted associations and pattern acquired from massive repositories [19]. Data mining technique is specifically termed as a nontrivial extraction process, which is used to extract unknown, potentially useful and implicit information from the database. In addition, it is also described as a process of exploring, selecting and modelling massive quantities of information in order to explore the associations or regularities. These regularities and associations are revealed with the intention to acquire valuable and apparent consequences of the repository [18].

According to the study, two strategies have been used in the data mining techniques, including supervised and unsupervised learning. Supervised learning comprises of a training set including learning model parameters [68]. On the other hand, no training set is included in the unsupervised learning, which probably entails k-means clustering. Data mining technique possesses a unique pattern to identify the modelling purpose. These patterns are associated directly with the intention of the model used [66]. Classification and prediction are the two dominant and common modelling objectives, possessed by data mining technique. The prediction of categorical labels such as unordered and discrete categories are entailed in the classification models whereas continuous-valued functions are predicted within the prediction models. Classification models are based on the neural networks and decision trees while prediction models comprise of clustering, regression and association rules [79].

The critical implementation of data mining is related with the investigation of data and data mining tools to discover the patterns from the specified data set. The most significant intention of data mining technique is the exploration of patterns independently without acquiring user efforts and input [71]. The influential

capabilities of data mining is surrounded on the intention of usage decision building as well as for forecasting anticipations of market trends. The functions of data mining techniques can be effective in several different sectors in different approaches. The implementation of data mining tool is efficiently observed in aggressive surroundings for data analysis. Various trends and market patterns are evaluated by using data mining tools and techniques. Furthermore, data mining technique is very essential in diagnosing diseases [67].

Data mining technique is specifically related to the diagnoses of diseases; however several applications are complex to be handled in healthcare sector. Data mining technique is also preferable for the detection of heart disease including diverse symptoms and factors, which are comprised of unpredictable effects [80]. Thereby, the diagnosis process and knowledge and experience utilization is dealt with the patient’s data. It has been evident that less cost spent in providing the services is considered to be the major constraint in the healthcare organizations [81]. According to the study, the accurate diagnosis of patients and effective treatment are denoted by valuable quality service. The accomplishment of clinical tests and diagnosis of risks can be handled appropriately by implementing machine learning or data mining technique [42]. Table 2 has listed the pros and cons of data mining techniques discussed in this section.

**Table 3: Pros and Cons of Data Mining Techniques**

No.	Data mining techniques	Pros	Cons
1.	Naïve bayes classifier	Less model complexity Easy to implement Work efficiently	Cannot be used on large dataset Poor when independence assumptions are not fulfilled. Super simple
2.	Weighted associative classifier (WAC)	Better accuracy Robust and predictive. High accuracy	Minimum support



3.	Decision tree	Simple Easy to interpret Fast and scalable Consider consequences	Over fit Inaccurate Expensive Less expected complex
----	---------------	---	---

The effective use of data mining technique is adhered in health care specifically in the prediction of coronary heart diseases. It has been revealed that numerous applications are included in predicting the risk factors associated to the heart disease. The features of data mining technique can use symptoms and patient records for predicting the risk factors in the health care sector [82]. A framework has been suggested, identifying the key aspects of medical data mining technique. The prediction of heart disease in the framework specifically commences with identifying the medical issues. The second step entails the pre-processing and cleaning of the data before heading towards the data mining tools. The evaluation of the expertise and risk prediction is entailed in the concluding step of the framework [83].

Classification, clustering, association and prediction are the most dominant techniques included in the data mining. Classification technique is a traditional data mining technique, which is relied on the machine learning approach. The intention of classification is used to attribute the set of information within a predefined set of groups or classes [19]. The use of mathematical techniques such as decision tress, neural network, statistics are included in the classification technique. Clustering is a most significant data mining technique, which entails equivalent cluster of substance involving mechanical technique features [69]. The classes are defined in the clustering technique, which is quite different from the classification approach. These classes are identified in classifying objects that are related to the predefined classes. For instance, the record of patients having similar risk factors can be identified in the form of clusters using clustering technique [17].

Association is regarded as a best data mining technique as compared to classification and clustering techniques. A relationship based on the specific item of similar operation is use to identify the patterns in the association technique. For

instance, heart disease prediction is used in the association technique to analyze the dissimilar categories of patients [68]. These categories are observed through the risk factors, which are essential for the identification of disease. The relationship between independent variables and dependent variables are discovered in the prediction technique. For instance, the prediction of profitability can be observed by using prediction analysis. In this approach, profit can be regarded as a dependent variable whereas sale could be regarded as an independent variable. A fixed regression curve can be drawn for profit prediction using the historical scale and profit data [67].

The employment of information systems and data mining techniques in hospital sector is used on the management of data and healthcare records. By applying these systems, large amount of data can be generated in the form of text, charts, numbers and images. On the contrary, clinical decision making is rarely supported by these systems [84]. The intuition and experience of doctors relies on the clinical decisions as compared to the knowledge-rich data hidden in the repository. The quality of service endowed to the patients are affected from the errors, excessive medical costs and unwanted biases. The reduction of medical errors, unwanted practices and improvement of patient safety are observed from the integration of clinical support system and computer patient records [81].

#### **4.7 Limitations**

- The widespread expansion of data mining techniques in different domains specifically in the healthcare sector has created complex issues and complex problem solving environment.
- The classification of these issues are based on the several factors, which includes distributed data, distributed operations, outliers, and data integrity and data privacy.
- The classification issues have been massively interacted with data mining techniques, restricting the capability for measuring large datasets.
- The major limitation in the data mining technique is the compilation of all the information in a comprehensive approach due to technical and organizational causes.

- The facilitation of distributed data and mining techniques is possible when algorithms and services are consequently developed.
- The paucity of distributed operations in data mining technique is also becoming a limitation for numerous sectors.
- The accessibility of data mining operations and algorithms is dependent on the problem solving approach.
- Innovative tools, grid services and algorithms are required to solve complex problems and analyzing large datasets.
- Massive data restricts the implementation of data mining techniques for the classification of diseases.
- The development of data mining algorithms is restricted due to low-dimensional and small datasets.
- Data privacy, governance and security are having serious issues in distributed environments due to the automated data mining approach.
- The implementation of grid-based data mining technology is preferable to overcome data privacy and security issues.
- The underlying assumptions of data mining technique usually causes over-fitting in a particular dataset.
- The minute size of training database can specifically cause over-fitting issue.
- The interpretation of results also cause issues to handle the queries and interfaces. Therefore, there is a need to recognize desired results and to handle data appropriately for interpretation of results.
- The data redundancy from diverse resources causes serious challenges for the implementation of data mining techniques. This limitation is vital as data integrity is a major issue in the prediction of heart disease detection system.
- Algorithms of data mining techniques are usually constructed for small datasets; therefore, the problems are created when algorithms are applied for the large datasets.
- Parallelization and sampling are the efficient tool to overcome this scalability.

## CHAPTER FIVE

### CONCLUSION AND FUTURE WORK

#### 5.1 Findings

The proposed System in this study have been specifically dominant in predicting heart disease system. Machine learning and statistical features of these data mining techniques have extracted significantly the symptoms of heart disease. By considering unique modelling objectives, naïve bayes and weighted associative classifier have possessed a modern approach for the prediction of heart diseases. The proposed framework has explicitly focused on the symptoms associated with heart diseases. The incorporation of naïve bayes and weighted associative classifier simply focuses on the input attributes of the model. Due to the underlying independence, these algorithms performed perfectly to predict the heart disease system.

It has been identified that prediction and classification are the two most dominant features of data mining technique. Diverse forms of databases can be handled through these two approaches. The specified patterns of data mining techniques merely focused on the exploration of input attributes and consequences. It has been adhered that naïve bayes and weighted classifier are significant in diagnoses purpose.

The most dominant techniques of data mining technique comprises of prediction, association, classification and clustering. These techniques are massively relied upon machine learning approach, which specifically intensifies a predefined set of classes. The classification of naïve bayes and weighted associative classifier is based on the simplicity and applicability of the incorporation. The log linear features of the naïve bayes classification assigns a separate parameter in order to identify the relative extent of class word pairs association.

Graphical user interface is incorporated in these algorithms for the effective prediction of heart disease. The main notion behind the implementation was to identify the attributes for accomplishing the accuracy of classification. The input and target attributes were categorized in accordance to the procedure or diagnosis codes. The notion of implementing weighted associative classifier is based on the association rules. The prediction of heart disease system can be supported through the provision of associative rules. The monitoring of heart disease system is evaluated through linguistic terms of weighted associative classifier. Thereby, weighted associative classifier has been used to classify diverse symptoms of particular disease. Furthermore, Boolean weighted items and quantitative weighted items have been used to support the input attributes of the weighted associative classifiers.

It has been investigated that environment factors contributed significantly in the progression of the disease because the occurrence of heart diseases are interacted with numerous symptoms and changes according to geographic region. The feasibility and accessibility of data mining techniques can be widely viewed to determine the symptoms of heart diseases. The algorithm used in this study has focused specifically and explicitly on the accuracy of datasets. Furthermore, the effectiveness of algorithm has generated association rules to possess extensive confidence. The focused demographic attributes has been sufficient in explaining the symptoms of heart diseases.

Heart disease prediction system have been classified by using KNN classification, J48 and naïve bayes including appropriate features. Another study has employed these algorithms for the prediction of heart disease. These approaches have been significantly used to detect the heart disease, possessing own capability to acquire appropriate results [82]. The construction of the system was entirely based on the hidden patterns and associations between decision tree, neural network and naïve bayes. It has been identified from the experimental results of the intelligent heart disease prediction system that employed data mining techniques are effective in expanding the reliability, scalability and web-based features of the system [42].

Naïve bayes and weighted associative classifier are used for the enhancement of heart disease prediction system. The study has included numerous input attributes

for the classification of prediction system. The information regarding the heart disease system specifically focused on the heart disease database [85]. The employment of weighted associative classifier has been significant to identify the intelligent and effective heart disease prediction system. Weighted input variables have been entailed in this approach to classify the weighted association in order to enhance the accuracy as compared with other present associative classifiers. The study has revealed that weighted associative classifier depicts 81% accuracy for the heart disease prediction system [81].

Naïve bayes has been employed in the study to explore the relationship between attributes and heart disease risk factors. Waikato environment for knowledge tool has been implemented to analyze the heart disease dataset [84].

The information about massive volumes of data was extracted by using naïve bayes approach. The analysis was completed on the basis of bayes theorem and input attributes. The Weka tool was preferred for the analysis as it is valuable machine learning algorithm, which is written in java [79]. The weka tool comprises of visualization, clustering, regression, pre-processing, classification and association tools. On the contrary, the mining and classification process was performed by using naïve bayes approach [86].

## **5.2 Conclusion**

The aim of this study was to investigate the data mining techniques used in the prediction of heart disease system. High occurrence rates of heart diseases have allow to focus on this dimension using the data mining techniques. The importance of data mining technique in healthcare sectors has been dominant in discussing specific, legal, ethical, social and heterogeneous restrictions. The massive datasets existed in the healthcare related studies have been encountered from diverse sources. On the contrary, data mining techniques are not appropriate in the validation and prediction of heart disease system. This study has employed naïve bayes and weighted associative classifier for the prediction of heart disease. Among data mining techniques, these techniques have been significant in determining the accuracy and evaluation of heart disease prediction. It has been analyzed that naïve bayes algorithm is effective in processing heart disease dataset as well as provides

better performance in predicting accuracy. The speed of the naïve bayes algorithm is much faster as compared to other algorithms. Thereby, naïve bayes has been effective in providing sufficient results in regards to heart disease prediction system.

Comparison between existing and proposed system has been discussed in this study to explore the significance of proposed techniques. The models of naïve bayes and weighted associative classifiers have significantly addresses the input attributes of the system. The comparison between algorithms in this study has provided great emphasis to comprehend the attributes and classification of these algorithms. It has been revealed that intelligent heart disease prediction system can also be expanded and improved on the basis of these classifications. Furthermore, medical attributes can be incorporated in the data mining techniques for significantly addressing the causal relationship between heart disease predictions.

It has been evident that numerous studies employed data mining techniques for the prediction of heart disease. Decision tree, Bayesian classifier, genetic algorithm and weighted associative classifier are the most dominant among data mining techniques. On the contrary, the important dilemma in these techniques is the classification of data mining tasks. Numerous issues have been discussed in the form of classification problems. It has been indicated that naïve bayes and weighted associative classifier are well-effective in solving issues due to their expanded applicability and feasibility of attributes. Therefore, naïve bayes and weighted associative classifier have been employed in this study to classify the attributes of heart disease prediction. The information are regarded as an input in the data mining system in order to analyze the performance of network. The results have indicated that naïve bayes and weighted associative classifier provides effective results for the classification of heart disease prediction.

The life of humans can be saved from heart attacks if heart diseases are diagnosed earlier. The feature reduction and optimization of proposed data mining techniques have been significant in classifying chronic heart disease prediction. It has been adhered that the experimental results of data mining techniques is sufficient in predicting the accurate decisions.

### **5.3 Future Work and Recommendation**

With widespread considerations of data mining technique in healthcare domains, future studies should focus on the enhancement of the prediction of heart disease using combination of vessel stenosis and attributes. In addition, the employment of feature reduction should be made in order to accomplish the effective accuracy of heart disease prediction. Furthermore, unique classifiers must be developed to explore the issues regarding cyanotic heart diseases, coronary microvascular diseases and pulmonary heart diseases. It is important to develop and effective intelligent trusted automated system for the prediction of heart disease. This intelligent trusted automated system must be focused on the symptoms and domain knowledge at the lower cost.

Intelligent heart disease prediction system can be expanded and improved appropriately. For instance, the specified system can utilize and implement medical attributes for the prediction of heart disease. Furthermore, incorporation of other data mining technique can be focused in future studies such as association rules, time series and clustering. Despite using categorical data, continuous data can be used for the purpose of heart disease prediction. It has been suggested that future studies should focus on the implementation of text mining in order to handle the huge amount of accessible data in healthcare databases. Moreover, the implementation of data mining and text mining in future projects will be a major concern.

Prediction of heart diseases can be significantly enhanced if fuzzy approaches are incorporated with data mining techniques. Numerous attributes can fulfil the prediction of heart diseases. It has been revealed that fuzzy K-NN classifier can be used along with unstructured information accessible in the healthcare sector. This implementation can focus directly on the structured heart disease information in order to endow effective accuracy of the system in diagnosing and indicating the symptoms of heart disease. Furthermore, the incorporation of fuzzy weighted association rule can be significant in the form of classification rule to enhance the accuracy of classification models. This algorithm can be used as a proposed strategy for identifying the accuracy of classification models.



## REFERENCES

1. Duan, L., Street, W. N. and Xu, E., 2011. Healthcare information systems: data mining methods in the creation of a clinical recommender system. *Enterprise Information Systems*, 5(2), pp.169-181.
2. Andersen, R. M., Rice, T. H. and Kominski, G. F., 2011. *Changing the US health care system: Key issues in health services policy and management*. John Wiley and Sons.
3. Kongstvedt, P. R., 2012. *Essentials of managed health care*. Jones and Bartlett Publishers
- 4] Massoud, M. R., Mensah-Abrampah, N., Barker, P., Leatherman, S.,
4. Kelley, E., Agins, B., Sax, S. and Heiby, J., 2012. Improving the delivery of safe and effective healthcare in low and middle income countries. *BMJ*, 344, p.e981.
5. Tomar, D. and Agarwal, S., 2013. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5), pp.241-266.
6. Koh, H. C. and Tan, G., 2011. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), p.65.
7. Polonsky, T. S., McClelland, R. L., Jorgensen, N. W., Bild, D. E., Burke, G. L., Guerci, A. D., and Greenland, P. 2010. Coronary artery calcium score and

- risk classification for coronary heart disease prediction. *Jama*, 303(16), 1610-1616.
8. Gupta, S., Kumar, D., and Sharma, A. 2011. Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(2), 188-195.
  9. Palaniappan, S., and Awang, R. 2008. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 108-115). IEEE.
  10. Delen, D., Walker, G., and Kadam, A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2), 113-127.
  11. Yeh, J. Y., Wu, T. H., and Tsao, C. W. 2011. Using data mining techniques to predict hospitalisation of hemodialysis patients. *Decision Support Systems*, 50(2), 439-448.
  12. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F. and Hua, L., 2012. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4), pp.2431-2448.
  13. Liu, H. and Motoda, H., 2012. *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer Science and Business Media.
  14. Srinivas, K., Rani, B. K., and Govrdhan, A. 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(02), 250-255.
  15. Anbarasi, M., Anupriya, E., and Iyengar, N. C. S. N. 2010. Enhanced prediction of heart disease with feature subset selection using genetic

- algorithm. *International Journal of Engineering Science and Technology*, 2(10), 5370-5376.
16. Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011a. Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*, 3(6), pp.2385-2392.
  17. Bhatla, N., and Jyoti, K. 2012. An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, 1(8), 1-4.
  18. Dangare, C. S. and Apte, S. S., 2012. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10), pp.44-48.
  19. Chaurasia, V., and Pal, S. 2013. Early prediction of heart diseases using data mining techniques. *Carib. j. SciTech*, 1, 208-217.
  20. Ishtake, S. H., and Sanap, S. A., 2013. Intelligent heart disease prediction system using data mining techniques. *International Journal of Healthcare and Biomedical Research*, 1, 94-101.
  21. Groves, P., Kayyali, B., Knott, D. and Van Kuiken, S., 2013. The 'big data' revolution in healthcare. *McKinsey Quarterly*, 2.
  22. Milovic, B. and Milovic, M., 2012. Prediction and decision making in health care using data mining. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, 1(12), p.126.

23. Rebuge, Á. and Ferreira, D. R., 2012. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2), pp.99-116.
24. Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P., 2013, May. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 48-55). IEEE.
25. Provost, F. and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), pp.51-59.
26. Kantardzic, M., 2011. *Data mining: concepts, models, methods, and algorithms*. John Wiley and Sons.
27. Ilayaraja, M. and Meyyappan, T., 2013, February. Mining medical data to identify frequent diseases using Apriori algorithm. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on* (pp. 194-199). IEEE.
28. Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.
29. Freitas, A. A., 2013. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science and Business Media.
30. Gupta, G. K., 2014. *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd..
31. Mosadeghrad, A. M., 2013. Healthcare service quality: Towards a broad definition. *International journal of health care quality assurance*, 26(3), pp.203-219.

32. Alrubaiee, L. and Alkaa'ida, F., 2011. The mediating effect of patient satisfaction in the patients' perceptions of healthcare quality-patient trust relationship. *International Journal of Marketing Studies*, 3(1), p.103.
33. Who, 2017. Cardiovascular diseases (CVDs) Fact sheet, *World Health Organisation*, Retrieved from, <http://www.who.int/mediacentre/factsheets/fs317/en/>
34. CDC, 2017. Heart Disease Facts, *CDC.Gov*, Retrieved from, <https://www.cdc.gov/HeartDisease/facts.htm>
35. Eurostat, 2016. Cardiovascular diseases statistics, *Eurostat, Statistics Explained*, Retrieved from, [http://ec.europa.eu/eurostat/statistics-explained/index.php/Cardiovascular\\_diseases\\_statistics#Deaths\\_from\\_cardiovascular\\_diseases](http://ec.europa.eu/eurostat/statistics-explained/index.php/Cardiovascular_diseases_statistics#Deaths_from_cardiovascular_diseases)
36. Heran, B. S., Chen, J. M., Ebrahim, S., Moxham, T., Oldridge, N., Rees, K., Thompson, D. R. and Taylor, R. S., 2011. Exercise - based cardiac rehabilitation for coronary heart disease. *The Cochrane Library*.
37. Mann, D. L., Zipes, D. P., Libby, P. and Bonow, R. O., 2014. *Braunwald's heart disease: a textbook of cardiovascular medicine*. Elsevier Health Sciences.
38. Bishop, S. P. and Altschuld, R. A., 2016. Increased glycolytic metabolism in cardiac hypertrophy and congestive failure. *American Journal of Physiology--Legacy Content*, 218(1), pp.153-159.
39. Perloff, J. K. and Marelli, A. J., 2012. *Clinical recognition of congenital heart disease*. Elsevier Health Sciences.

40. Mendis, S., Puska, P. and Norrving, B., 2011. *Global atlas on cardiovascular disease prevention and control*. World Health Organisation.
41. McMurray, J. J., Adamopoulos, S., Anker, S. D., Auricchio, A., Böhm, M., Dickstein, K., Falk, V., Filippatos, G., Fonseca, C., Gomez - Sanchez, M. A. and Jaarsma, T., 2012. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012. *European journal of heart failure*, 14(8), pp.803-869.
42. Nahar, J., Imam, T., Tickle, K. S. and Chen, Y. P. P., 2013. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4), pp.1086-1093.
43. Joshi, M., Joshi, A. and Bartter, T., 2012. Symptom burden in chronic obstructive pulmonary disease and cancer. *Current opinion in pulmonary medicine*, 18(2), pp.97-103.
44. Moya, A., García-Civera, R., Croci, F., Menozzi, C., Brugada, J., Ammirati, F., Del Rosso, A., Bellver-Navarro, A., Garcia-Sacristán, J., Bortnik, M. and Mont, L., 2011. Diagnosis, management, and outcomes of patients with syncope and bundle branch block. *European heart journal*, 32(12), pp.1535-1541.
45. Lin, E. H., Von Korff, M., Ciechanowski, P., Peterson, D., Ludman, E. J., Rutter, C. M., Oliver, M., Young, B. A., Gensichen, J., McGregor, M. and McCulloch, D. K., 2012. Treatment adjustment and medication adherence for complex patients with diabetes, heart disease, and depression: a randomized controlled trial. *The Annals of Family Medicine*, 10(1), pp.6-14.

46. Koyak, Z., Harris, L., de Groot, J. R., Silversides, C. K., Oechslin, E. N., Bouma, B. J., Budts, W., Zwinderman, A. H., Van Gelder, I. C. and Mulder, B. J., 2012. Sudden cardiac death in adult congenital heart disease. *Circulation*, pp.CIRCULATIONAHA-112.
47. Stefan, M. S., Rothberg, M. B., Priya, A., Pekow, P. S., Au, D. H. and Lindenauer, P. K., 2012. Association between  $\beta$ -blocker therapy and outcomes in patients hospitalised with acute exacerbations of chronic obstructive lung disease with underlying ischaemic heart disease, heart failure or hypertension. *Thorax*, pp.thoraxjnl-2012.
48. Hachesu, P. R., Ahmadi, M., Alizadeh, S. and Sadoughi, F., 2013. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare informatics research*, 19(2), pp.121-129.
49. Dudas, K., Lappas, G., Stewart, S. and Rosengren, A., 2011. Trends in out-of-hospital deaths due to coronary heart disease in Sweden (1991 to 2006). *Circulation*, 123(1), pp.46-52.
50. Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011b. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), pp.43-48.
51. Sudhakar, K., & Manimekalai, D. M. (2014). Study of heart disease prediction using data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1).
52. Ordonez, C. and Chen, Z., 2012. Horizontal aggregations in SQL to prepare data sets for data mining analysis. *IEEE transactions on knowledge and data engineering*, 24(4), pp.678-691.

53. Bhardwaj, B. K., and Pal, S. 2012. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.
54. Patil, T. R., and Sherekar, S. S. 2013. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2), 256-261.
55. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1), 1-37.
56. Doreswamy, H.K., 2012. Performance Evaluation of Predictive Classifiers for Knowledge Discovery from Engineering Materials Data Sets. *arXiv preprint arXiv:1209.2501*.
57. Mukherjee, S. and Sharma, N., 2012. Intrusion detection using naive Bayes classifier with feature reduction. *Procedia Technology*, 4, pp.119-128.
58. Sundar, N. A., Latha, P. P. and Chandra, M.R., 2012. Performance analysis of classification data mining techniques over heart disease database. *IJESAT/ International Journal of engineering science & advanced technology ISSN*, pp.2250-3676.
59. Rokach, L., and Maimon, O. 2014. *Data mining with decision trees: theory and applications*. World scientific.
60. Seema, Rathi, M., and Mamta. (2012). Decision Tree: Data Mining Techniques, *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Vol 1(3)



61. Koh, H. C. and Tan, G., 2011. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), p.65.
62. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., and Witten H. I., 2006. WEKA - A Machine Learning Workbench for Data Mining, *Chapter 1*
63. Moraru, A., Pesko, M., Porcius, M., Fortuna, C., and Mladenec, D., 2010. Using Machine Learning on Sensor Data, *Journal of Computing and Information Technology - CIT* 18(4), 341–347 .
64. Aggarwal, C. C. (Ed.). 2013. *Managing and mining sensor data*. Springer Science and Business Media.
65. Letkowski, J., and Gulati, A., 2013. A deceptive side of data mining .
66. Kaur, K., & Singh, L. M. (2016). HEART DISEASE PREDICTION SYSTEM USING ANOVA, PCA AND SVM CLASSIFICATION.
67. Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. *International journal on recent and innovation trends in computing and communication*, 2(10), 3003-3008.
68. Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937-1946.
69. Dai, W., Brisimi, T. S., Adams, W. G., Mela, T., Saligrama, V., & Paschalidis, I. C. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. *International journal of medical informatics*, 84(3), 189-197.
70. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases.

71. Chadha, R., Mayank, S., Vardhan, A., & Pradhan, T. (2016). Application of Data Mining Techniques on Heart Disease Prediction: A Survey. In *Emerging Research in Computing, Information, Communication and Applications* (pp. 413-426). Springer India.
72. Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, 24(1), 27-40.
73. Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Applications*, 67, 239-251.
74. Xu, S., Shi, H., Duan, X., Zhu, T., Wu, P., & Liu, D. (2016, March). Cardiovascular risk prediction method based on test analysis and data mining ensemble system. In *Big Data Analysis (ICBDA), 2016 IEEE International Conference on* (pp. 1-5). IEEE.
75. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1), 1-37.
76. Wang, L., Ji, P., Qi, J., Shan, S., Bi, Z., Deng, W., & Zhang, N. (2014). Feature weighted naïve Bayes algorithm for information retrieval of enterprise systems. *Enterprise Information Systems*, 8(1), 107-120.
77. Verma, L., Srivastava, S., & Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of medical systems*, 40(7), 1-7.

78. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2015). Heart disease prediction system using data mining technique by fuzzy K-NN approach. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1* (pp. 371-384). Springer International Publishing.
79. Khare, S., & Gupta, D. (2016, August). Association rule analysis in cardiovascular disease. In *Cognitive Computing and Information Processing (CCIP), 2016 Second International Conference on* (pp. 1-6). IEEE.
80. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.
81. Letian, W., Han, L., Zhang, L., & Guo, S. (2016). GW27-e0397 An analysis and diagnosis system of coronary heart disease based on big data platform. *Journal of the American College of Cardiology*, 68(16), C82.
82. Peter, T. J., & Somasundaram, K. (2012, March). An empirical study on prediction of heart disease using classification data mining techniques. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 514-518). IEEE.
83. Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on* (pp. 173-177). IEEE.
84. Kharya, S., & Soni, S. (2016). Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection. *International Journal of Computer Applications*, 133(9), 32-37.

85. Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In *Proceedings of the world congress on engineering and computer science* (Vol. 2, pp. 22-24).
86. Kausar, N., Palaniappan, S., Samir, B. B., Abdullah, A., & Dey, N. (2016). Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. In *Applications of intelligent optimization in biology and medicine* (pp. 217-231). Springer International Publishing.

## CURRICULUM VITAE

### PERSONAL INFORMATION

**Surname, Name:** Mohammed Ibrahim Mahdi AL-AZZAWI

**Date and Place of Birth:** 10 December 1988, Iraq/ Wasit

**Marital Status:** Single

**Phone:** +96477705984465

**Email:** [muhamad.alazzawi@gmail.com](mailto:muhamad.alazzawi@gmail.com)



### EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	THK University, Information Technology, Ankara, Turkey.	2017
B.Sc.	Software Engineering. Al-Rafidain University College, Baghdad, Iraq.	2011
High School	Al – Hussein School. Wasit/ Al-Hai, Iraq.	2007