Predictive Diagnosis Through Data Mining: A Secondary Review

By

Ritwik Chawla

Dissertation submitted for the partial fulfillment of the

MBA Hospital and Health Management

Academic year, 2015-2017

The IIHMR University, Jaipur

1, Prabhu Dayal Marg, Sanganer, Airport
Jaipur 302029, Rajasthan
Ph: 0141 3924700, Fax: 3924738
Website: **www.iihmr.edu.in**

Predictive Diagnosis Through Data Mining: A Secondary Review

By

Ritwik Chawla

Dissertation submitted for the partial fulfillment of the

MBA Hospital and Health Management

Academic year, 2015-2017

NTT DATA Information Processing Services Private Limited, Bengaluru

The IIHMR University, Jaipur

1, Prabhu Dayal Marg, Sanganer, Airport
Jaipur 302029, Rajasthan
Ph: 0141 3924700, Fax: 3924738
Website: www.iihmr.edu.in

INSTITUTE OF HEALTH
MANAGEMENT RESEARCH
(WHO Collaborating Centre for District Health System Based on Primary Health Care)

IIHMR UNIVERSITY

TO WHOMSOEVER MAY CONCERN

This is to certify that Dr. Ritwik Chawla student of MBA Hospital and Health Management from The IIHMR University, Jaipur has undergone internship training at NTT DATA Services, Whitefield, Bengaluru from 6th February to 5th May 2017.

The Candidate has successfully carried out the study designated to him during internship training and his approach to the study has been sincere, scientific and analytical.

The Internship is in fulfilment of the course requirements.

I wish him all success in all his future endeavours.

(Col.) Dr. Ashok Kaushik
Dean, Academics and Student Affairs
The IIHMR University, Jaipur

16.5.17
Dr. Sandesh Kumar Sharma
Associate Professor
The IIHMR University, Jaipur

NTT DATA Information Processing Services Private Limited
(Formerly known as Dell Business Process Solutions India Private Limited)
Plot 123, EPIP Phase II, Whitefield Industrial Area
Bengaluru, Karnataka 560 066, India

NTT DATA
Global IT Innovator

To whomsoever it may concern

This is to certify that **Dr. Ritwik Chawla** of **The IIHMR University, Jaipur** has been working with NTT DATA Services for his summer project.

Project Details:

Project Name : Predictive Diagnosis through Data Mining: A Secondary review
Duration : 90 days
Location : Bangalore
Guide Name : Gaurav Bhati & Shivang Sharma
Sponsor Name : Ajay Aiyar

He has successfully completed his project and his performance during the tenure of the internship has been found to be satisfactory.

His findings in course of the project has been found to be practical and relevant and some of the recommendations will be incorporated on the floor on approval from the business.

Thanking You,

Regards,

Dilipkumar Santhanam
Business Partner Director, Services HRBP Team
NTT DATA Services

THE IIHMR UNIVERSITY, JAIPUR

CERTIFICATE BY SCHOLAR

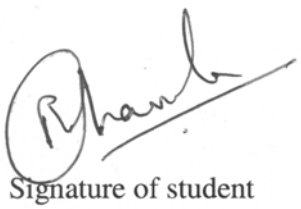
This is to certify that the dissertation titled **Predictive Diagnosis through Data Mining: A Secondary Review** and submitted by Dr. Ritwik Chawla Enrollment No. IIHMR-U/01/2015-17/0335 under the supervision of Dr. Sandesh Kumar Sharma (Associate Professor), for award of MBA in Hospital and Health Management of the University carried out during the period from February 6, 2017 to May 5, 2017 embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other institute or other similar institution of higher learning.

Signature

THE IIHMR UNIVERSITY, JAIPUR

CERTIFICATE BY SCHOLAR

This is to certify that all ethical issues have been considered while collecting data for my dissertation titled **"Predictive Diagnosis through Data Mining: A Secondary Review"**



Signature of student

1. Introduction

Today, in the world a large amount of data is blowing up. Virtually every country generates data in abundance in all known fields, including healthcare. The data supply of world has reached 2.8 zettabytes (ZB) or 2.8 trillion Gigabytes in 2012 as per the Gantz, 2012. "Data universe is made up of images and videos, security footage at airports and major events, banking data, subatomic collisions recorded by the Large Hadron Collider, transponders recording highway tolls, voice calls zipping through digital phone lines, and texting as a widespread means of communications." (37)

Huge amount of data was produced in the industry of healthcare, this is being taken in relation to healthcare associated processes in the form of "Electronic Health Records (EHR), Health insurance claims, medical imaging databases, disease registries and clinical trials." The data is dependent on a wide range of healthcare associated functions, which includes clinical patient records, medical images from laboratory, health behaviors, clinical and medical decision support system, surveillance of disease, genomic data and public health management systems, etc (36). This rapidly growing, incredible amount of data is getting collected and being stored in large and multiple data repositories which has by far surpassed the human ability for conception without having powerful and accurate data analysis tools. Reports suggests that global digital healthcare associated data was estimated to be at a level of 500 petabytes in 2012. Report further suggests that the data is expected to reach 25,000 petabytes or 25 Exabytes in 2020 (38). Unfortunately, this data is rarely used to support clinical decision making.

The questions arises, how can this huge data be turned into useful information to make intelligent clinical decisions for the benefit of the people. The answer lies with data analytics.

A lot of efforts have been put in to manage with the explosion of healthcare associated data produced on one hand, and to gain valuable and evocative knowledge from it on the other hand. This has provoked many researchers to put in all modern and technical innovations like big data analytics, machine learning, predictive analytics, and various other learning algorithms to excerpt useful and meaningful knowledge for making important decisions.

With a short span of history, the role of Data Mining extends to many domains such as:

A) Financial Data
B) Retail Industry
C) Banking
D) Telecommunications Industry
E) Biological Data Analysis
F) Health and Medicine
G) Communications
H) Credit handling
I) Market trend prediction
J) Education Industry
K) Fraud Detection

In the early 1990s, many businesses had started using data mining for many such activities like, fraud detection, maintenance scheduling, credit scoring. Now, even healthcare organizations are seeing great value and benefit in data mining.

Following is a case of failed diagnosis which would explain the need for data mining.

A 34-year-old male went to a primary care doctor and presented with sternal pain after lifting a tree in his backyard. The pain elevates when the patient raises his arms. An ECG was instructed, and the results came out to be as negative. The patient did not go for cardiac enzyme test as the doctor diagnosed that the muscle strain or spasm could be the cause of the patient's symptoms. The doctor told the patient to go on a vacation after some rest. After two days on his vacation, the patient suddenly died from myocardial infarction. (3)

Discussion of the Case

Such a case of failed diagnosis not only increases the disease burden on a patient, but can be catastrophic. Various factors may have contributed to diagnostic errors. One of the most prevalent is **clinical misjudgment**, which is a complex process that involves various cognitive functions and is potentially vulnerable to a variety of **cognitive biases**. The case above illustrates how an oversight in clinical misjudgment can lead to a diagnostic error with a profound outcome. In this case, a cognitive bias known as "**anchoring**" occurred.

Anchoring means a snap judgment or a tendency to make a diagnosis based on the first clinical sign or symptom or any relevant lab abnormality (4).

Today, in the medical domain hospitals are storing large amount of heterogeneous data in the form of text, numbers, images, patient records, and test results on a regular basis. This information is not only valuable for diagnosis and therapy but is rarely used after the actual treatment and the release of the patient for the improvement of medical procedures. A comprehensive analysis of this information, may help to come up with new medical hypotheses and to find statistical evidence for existing hypotheses. These new insights help us understand which factors may increase the risk of getting a disease, which tests can provide valuable information on the status of a disease or which treatment works best for a patient.

The knowledge obtained from data can support health care workers in treatment planning or diagnosis. Powerful methods from machine learning, association learning and mining of data have been adapted to meet the specific demands of medical domains.

Data Mining interconnects database technology, different modelling techniques, analysis of statistical data, recognition of patterns, and machine learning. It uses innovative tools for database management of large data and easy automatic and/or semiautomatic analysis to identify important trends as well as associations believed to be informative. This is because it is original, implicit to the data, and has a greater potential in providing support for prediction and decision making (1).

In simple terms data mining is a process of discovering patterns, trends, and knowledge from data which is being generated from various organizations stored in databases which are previously unknown. Using the data to make information and making use of the same information to build predictive models.

Discovering useful information from data usually takes place in two major forms: description and prediction. It is just used to make the data simplified and then summarize in a manner that one would be able to understand, and then allows to deduce information about specific cases which could be based on the patterns that had been observed. Pattern detection that are commonly used can be of several main types. Figure 2.1 shows the general form showing what data mining can help us to do (29).

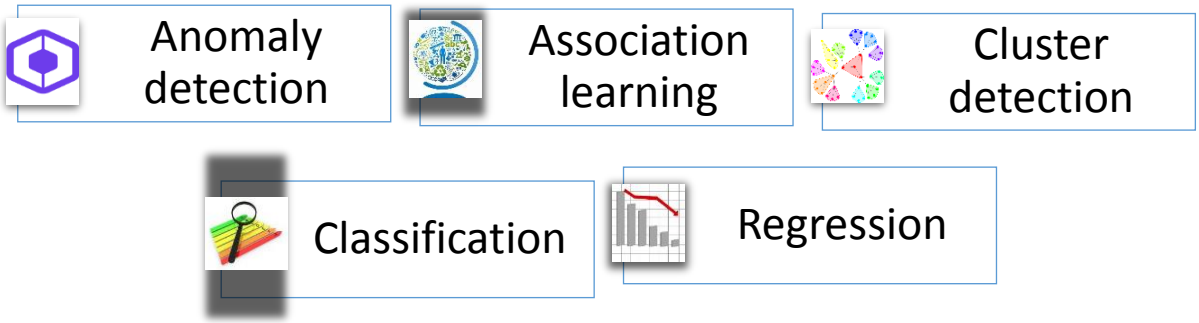


Figure 1.1: Types of Pattern detection in Data Mining

a) **Anomaly Detection** – It is used to determine if something or some data is particularly different from the pattern, at the time of review and audit.
b) **Association learning** – It is a rule-based machine learning method for discovering interesting relations between variables in large databases. These types of findings are often used for targeting coupons/deals or advertising.
c) **Cluster detection** – Data itself determine the groups to capture the relevant distinctions between apparent groups in the data.
d) **Classification** – Data mining can be used to categorize new cases into these pre-determined categories. Algorithms can distinguish persistent systemic differences between items in each group and then apply these existing rules to a new classification set of problems.
e) **Regression** – Data mining can be used to make predictive models based on many variables at the same point of time.

1.1 Data Mining History

Data Mining science is practiced everywhere, but its beginning had started many centuries earlier. A paper published by Thomas Bayes' in **1763** explained a theorem to relate current or hypothesized probability to prior or posterior probability which came out to be known as Bayes' Theorem. This was fundamental step in data mining and current probability. In **1805** Carl Friedrich Gauss and Adrien-Marie Legendre had applied regression analysis to

approximate the relationships between multiple variables using a precise method based on least squares. Regression analysis was one of an important tool used in data mining.

In the year **1943** Walter Pitts and Warren McCulloch in a published paper titled "A logical calculus of the ideas immanent in nervous activity" created an abstract model of a neural network for the very first time. Each neuron out of all those can perform three different tasks: accept inputs, process these inputs, and produce one or two outputs. In **1965** Fogel L. is credited for formation of an organization by the name Decision Science Inc. This new organization developed applications on evolutionary programming in various fields.

1970s brought sophisticated management systems databases and it became probable to store huge amounts of data in petabytes and query on it. Warehouses where the data was stored allowed multiple users to change from a transaction-oriented perspective of decision making to a much more logical way of reviewing data. During **1975**, John Holland had written an Adaptation on Artificial and Natural Systems, which came out as a ground-breaking book on genetic algorithms.

In **1980s** HNC trademarked the phrase "database mining." At that time, it was done as a general tool for constructing models on neural network. During the period, many sophisticated algorithms started "learning" relationships generated from the data which allowed SME's to make out what exactly these associations meant. In **1989** "Knowledge Discovery in Databases" (KDD) term was conjoined by Gregory Shapiro. During the same time Gregory co-founded the foremost data mining workshop called as "KDD."

"Data Mining" was introduced in **1989** by Fayaad. After which the term "data mining" started appearing in the community related to database. Many retail organizations and financial communities had started using data mining techniques to analyze and predict data and identify trends which would help to grow their consumer base, including prediction and variations in rates of interest, cost of stocks, and demand of customers. In **1992** Isabelle Guyon, Bernhard Boser, and Vladimir Vapnik had suggested greater improvements on "Support Vector Machine" (SVM) model which helped making some classifiers which were nonlinear and variable. Data mining as an independent discipline wasn't introduced until **2001,** William Cleveland is credited for it.

Today, "Data Mining" is virtually prevalent in various businesses, engineering science, and healthcare to name some. Data Mining of Credit card transactions, national security, actions

through stock market, experiments of human genome and pharmaceutical trials are just the start for applications in this science. Today, terminologies like “Big Data” are becoming common with gathering of data going inexpensive day by day and the explosion of smart devices proficient enough to gather massive amount of data.

The Figure: 2.1 shows the chief markers and “firsts” in the data mining history and how this has advanced and unified with Big Data and Data Science. (14)

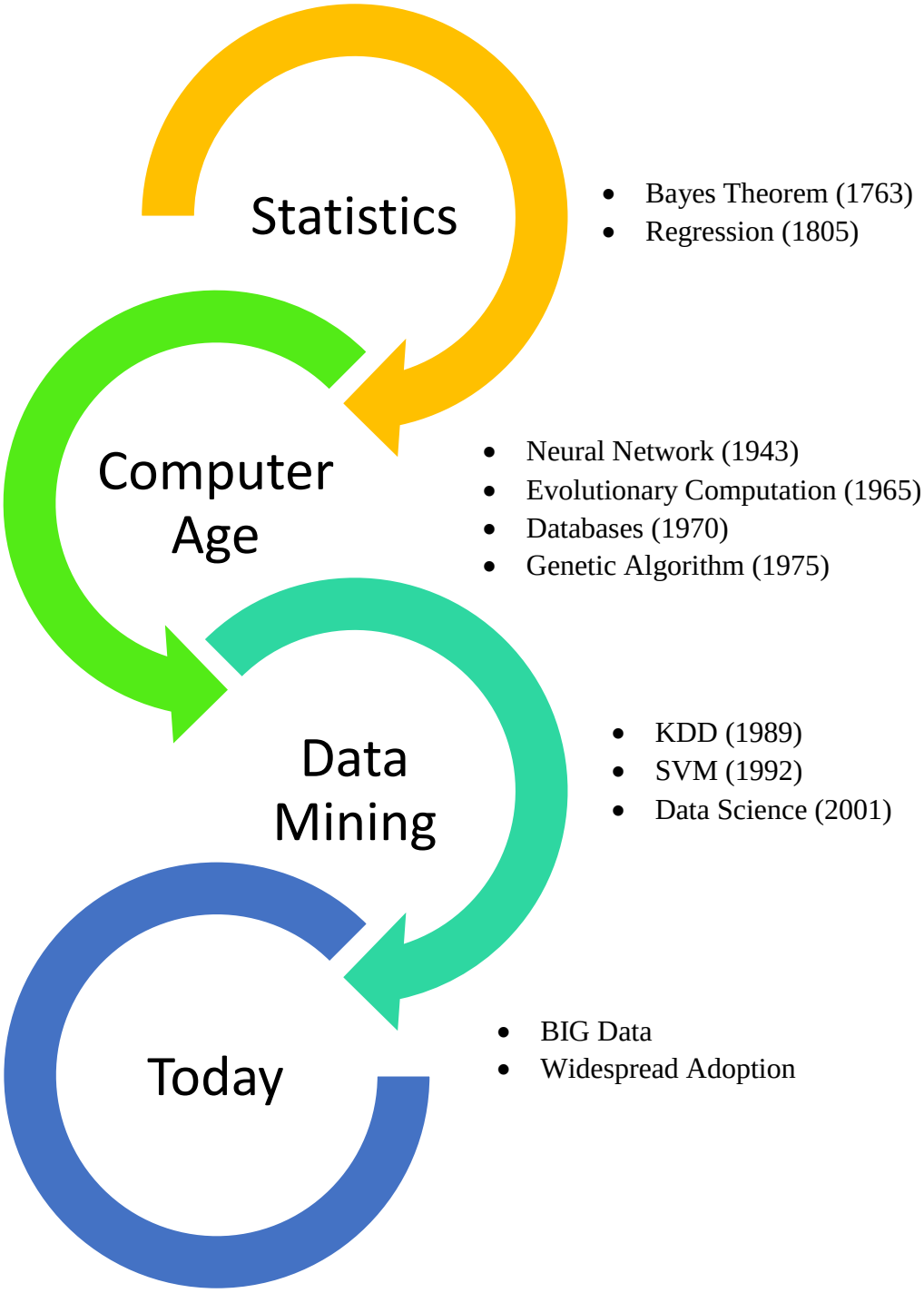


Figure 1.2: History of Data Mining

2. Objectives

2.1. To study the various Data mining techniques of diagnosis and prediction system.

2.2. To study the applications and compare the performances of three predictive data mining techniques of heart disease prediction.

2.3. To study the global market and prospects of predictive diagnosis through data mining.

To achieve these research objectives, this study has been done which will address the following research questions:

3. Research Questions

- Are there any assuring techniques to predict various diseases with high accuracy and limited resources?
- Do different data mining techniques really provide an effective way to predict heart disease?
- What are the plausible benefits of different technique of data mining in relation with heart disease?
- How different data mining techniques of heart disease prediction system performs?
- How can data mining in healthcare provide a novel predictive diagnosis system in the future?

4. Rationale

In present years, data-mining has raced to become one of the most treasured tools for mining, utilizing data and for establishing patterns and predictions to produce meaningful information which is useful for decision making and forecasting.

Many people are involved with this process and are probing for the right tools to solve predictive problems related to data-mining. This report will give a route to what can be done when we are confronted with few of such kind of problems. The report will explain the data processing techniques which can be practiced on a dataset to expand the understanding of the type and application of techniques used. Heart disease prediction has been reviewed to assess the performances of three different predictive data mining techniques. Finally, the results of

the performances of three techniques have been reviewed and each of these techniques have been compared to each other on the same data set.

5. Study Purpose

The purpose was to identify the data mining technique which performs best for a given dataset for prediction of heart disease and use the same directly in place of depending on the common hit-and-try method. The procedure when efficiently used, will help decrease the lead time to build models for forecasting and/or prediction of many medical diseases.

6. Study Methodology

A scientific literature review search was conducted using all the articles and journals available to know the various predictive diagnosis technique using data mining and their performances.

6.1 Study Design: Exploratory design with Secondary review

6.2 Study Settings: NTT Data Services, Bengaluru

6.3 Study Duration: Three months (February 2017 – April 2017)

6.4 Data Sources: Based on secondary literature review of predictive diagnosis through data mining techniques following the IMRAD format. Various sources like Library books, BioMed Central, Google Scholar, Medline, Elsevier, IEEE Explorer, Springer, and other search engines available at Google were searched from February to April 2017. These articles discussed various predictive diagnostic techniques of data mining in healthcare, its application, and comparison of performances, global market, and its future aspects. The MeSH terms which has been used to index the articles are: (1)"data mining", (2)" predictive diagnosis", (3)"data mining through predictive diagnosis", (4) "Naïve Bayes", (5) Decision Tree, (6) Neural Network. The search criteria did not include any limitation on publication date, and the earliest eligible article searched was published after 2007 only. The reference lists of included articles were also searched systematically.

6.5 Inclusion Criteria: Study has covered three major techniques of data mining i.e., "Naïve Bayes Classifier", "Decision Tree", and "Artificial Neural Networks" for

of Heart Disease prediction system. Various articles from leading international journals, web pages, published newspaper articles, published books, conference papers, thesis, etc. was reviewed for the study purpose.

6.6 Exclusion Criteria: The applicability of Data Mining in Healthcare extends to variety of diseases, like Diabetes, Cancer, Genetic disorders, and other major diseases. However not all diseases have been covered, except one covered in Inclusion Criteria. Articles written in language other than English was not included in the study. Unpublished study material and material from blogs was not referred for the study purpose.

7. Literature Review

7.1 The Knowledge Discovery Process Model

Data Mining (DM), Knowledge Discovery in Databases (KDD), and Knowledge Discovery (KD) are three different terms usually used in this reference. The explanation of these terminologies have been done with references to definitions from scientific literature published by leading Data scientists.

Fayyad explains "*Data Mining* is concerned with the application, under human control, of low-level DM methods and tools, which in turn are defined as algorithms designed to examine data, or to extract patterns in specific related categories from data" (8). Data Mining also recognized by many other terms including, "Information Discovery", "Knowledge Extraction", "Data Archaeology", "Data pattern processing" , and "Information Harvesting" (9).

Fayyad defined *"Knowledge Discovery* as a process that try to find new knowledge about an application field. It consists of many steps, with one of them being Data Mining, each step is aimed at accomplishment of a discovery task, and is accomplished by the application of different discovery methods" (8).

Knowledge Discovery in Databases is concerned with the discovery process of knowledge which is applied on databases. Also, it was defined by Fayyad as "an insignificant process of identifying valid, original, potentially useful, and ultimately comprehensible patterns in data" (9).

Knowledge Discovery and Data Mining is concerned with the Knowledge Discovery process applicable to any source of data. KDDM has been projected now as the best suitable name for the complete process of Knowledge Discovery (8). The process of KDD has been described in the Figure: 8.1. It has been divided into few steps as follows (15):

- **Application domain learning:** It is relevant to past knowledge and the aim of application.
- **Target dataset creation:** It is done to select Data.
- **Data preprocessing/cleaning:** Reduces noise and error which may take away 60% of the effort.
- **Transformation of Data and reduction:** Finding meaningful features, reduction in dimensionality and variability, and appropriate representation.
- **Data Mining Functions selection:** Data summarization and classification, clustering, association, regression, etc.
- **Selection of Data Mining algorithm(s):**, Naïve Bayes, Decision tree, Neural Network, etc.
- **Data Mining technique performed:** Searching patterns for greater interest
- **Evaluation of pattern and Presentation of Knowledge:** Data visualization, , eliminating redundant patterns, transformation, etc.
- **Usage of Knowledge discovered**

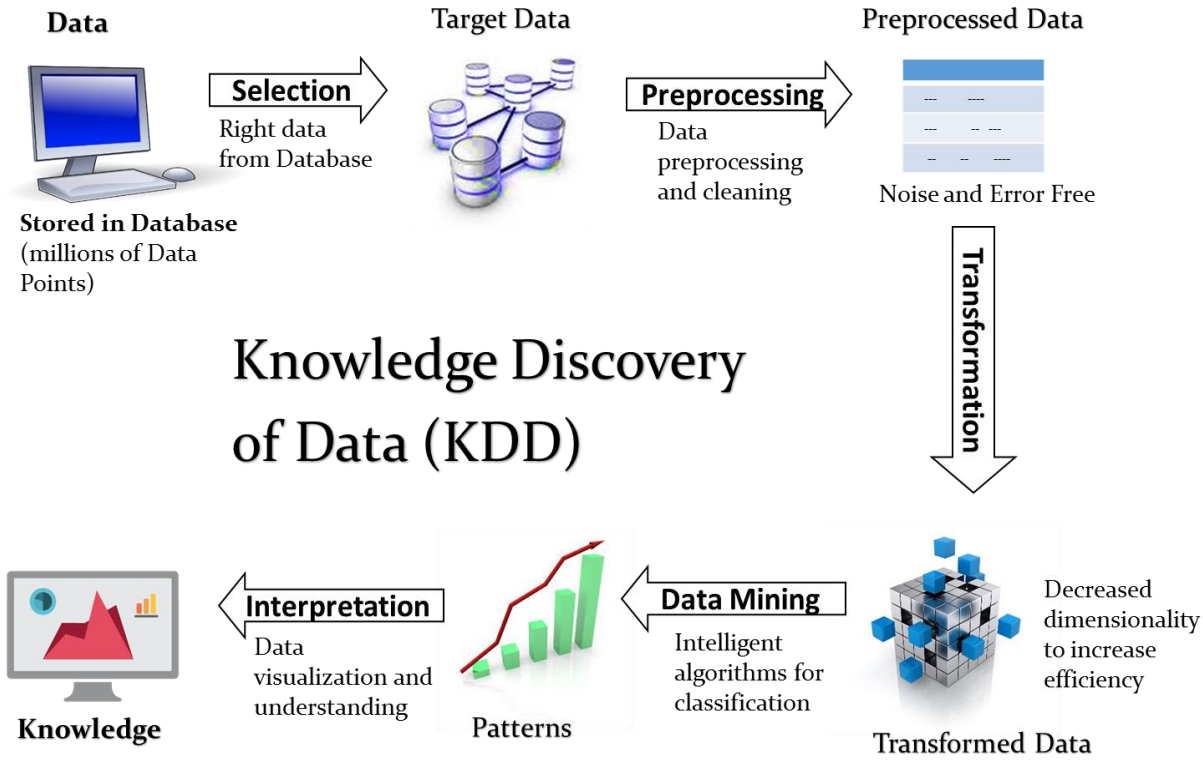


Figure 7.1: Showing the Steps of Knowledge Discovery of Database

The progress of the typical KDDM model started some three decades ago. The initial KDDM model consisted of 9 steps. Fayyad et. al developed this in the middle of 1990s. Cabena et al. made the next model which was made up of 5 steps and was presented in the year 1998. Next model was made up of eight steps and was established by Buchner and Anand during the same time. The fourth model was proposed in 1996 known as CRISP-DM (CRoss-Industry Standard Process for Data Mining) that included six steps initially established in 1996 by a association of four organizations: OHRA (an insurance organization), NCR (a database provider), Daimler Chrysler, and SPSS (a commercial provider of Data Mining solutions) and). The latter two organizations provided case studies and sources of data. Officially (as version 1.0) released in the year 2000 this model continues to relish a robust manufacturing provision (Table 8.1).

Table 7.1: Comparison of chief existing KDDM Models for Knowledge Discovery

KDDM Model	Fayyad et al.	Cabeena et al.	Buchner and Anand	CRISP-DM	Cious et al.	General Model
Application area	Academic	Industrial	Academic	Industrial	Academic	-
Total Steps	Nine	Five	Eight	Six	Six	Six
References	(Fayyad et al., 1996d)	(Cabena et al., 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios et al., 2000)	N/A

7.2 Data Mining Tasks

Data mining tasks are used to specify the kind of patterns to be found in data mining process. In general, data mining tasks can be classified into two categories: **Predictive** and **Descriptive** (12).

A Predictive model is one which makes a prediction about values of data using known results found from different data. The goal of this model is to identify strong links between variables of a data table (columns). Examples includes providing a diagnosis for a medical patient based on a set of test results or estimating the probability that customers will buy a specific product.

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

The major difference between the two models is that, a descriptive model serves to explore the properties of the data examined, not to predict new properties. On the contrary, a predictive model has a specific objective to predict the value of some target characteristic of an object based on observed values of other characteristics of the object.

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

7.3 Data Mining Steps

A Data mining project has a life cycle which consists of nine steps. The process starts with determining the KDD goals, and ends with the implementation of the discovered knowledge.

As many decisions are made by the user, the steps are interactive and iterative, this means moving back to previous steps may be required sometimes. So, based on the KDD methodology, the following nine steps are undertaken for any data mining project specific for predictive diagnosis of any disease (9).

Step 1: Understanding the Application Domain

Here, the researcher works closely with the Hospital's Department head, a real-time observation of the system is done to understand the business process of the hospital. The sub goal in this step is determining the data mining goals and their success criteria. These goals are obtained by translating the medical goals into data mining goals.

Step 2: Selecting and Creating the Target Dataset

The researcher use data collected by the Hospital which contains diagnostic information on various tests which were conducted on many patients over the years. To reduce the number of variables, the researcher selects only 15 as the most important variables for inclusion in the dataset. After recording, the new database now contains thousands of such instances, wherein each instance resembles a single file of a patient.

Step 3: Pre-processing and Cleansing

The selected data is checked for noise, inconsistency, missing values, and outlier detection using distribution frequency. Noises and inconsistencies are corrected manually, while missing values are replaced with the most probable value determined with regression. Outliers are replaced with a mean value of the attributes to reduce variability. Data cleaning is performed after addressing issues and requirements of the tools selected for the pre-processing phase.

Step 4: Data Transformation

In this stage, the researcher consults with the domain expert for a few transformations to implement on the dataset. This is done to make the data more suitable for the data mining algorithms.

To reduce the dimensionality, the dataset is encoded with different mechanisms. These are applied to obtain a reduced or compressed representation of the original data. To be able to reconstruct the original data from the compressed data without any loss of information a lossless reduction method is also adopted.

Attribute selection is also done as a part of transformation. This is necessary to reduce the number of features a classification algorithm will examine and reduce errors which can come from these irrelevant features. Here, the researcher uses the best first search method to select the best attributes from 15 attributes that are available.

Step 5: Choosing the appropriate Data Mining Task

This step is crucial as the major step is to decide what type of data modelling technique must be used. Selecting an appropriate modelling technique depends entirely on the goal of the study, for instance if the goal of this study is to detect heart disease using data mining techniques a classification technique can be adopted to develop predictive models for detection.

Step 6: Choosing the Data Mining Algorithm

To select the best algorithms various important steps are taken into consideration: the performance of the algorithms based on previous projects, the data mining goal set in the first place, the structure of the data available, and the algorithm's application in the medical field.

Here, in WEKA machine learning software, after checking through the available algorithms the most frequently used algorithms are Decision Tree, Neural Network, and Bayesian Classifier.

Step 7: Employing Data Mining Algorithm

The selected classification algorithms are employed. Experiments designed are conducted on a full training dataset containing instances. Usually researcher considers two scenarios, one containing all 15 attributes and the other containing only 8 selected attributes.

A 10-Fold Cross Validation is adopted for randomly sampling the training and test data sets. This entire process is done with the help of WEKA 3.6.4 machine learning software.

Step 8: Evaluation

Models must be evaluated to assess the degree to which they meet the data mining goals identified initially. This includes understanding the results, understanding if the new information obtained is novel and interesting, medical interpretation of the results obtained, and checking the impact on the project goal. The evaluation for the algorithms is based on different parameters such as, the classification accuracy, area under the ROC curve, and confusion matrix table.

Step 9: Using the Discovered Knowledge

This is the last step in the data mining project through knowledge discovery process. The success of this step determines the effectiveness of the entire KDD process. After all the steps are carried out successfully a final report is written to summarize the project outcome.

7.4 Classification

Classification is defined as the process of finding a model that describes and distinguishes data classes or concepts, for being able to use the model to predict the class of objects whose class label is unknown. The model which is derived is based on the analysis of a set of training data (i.e., data objects whose class label is known) (12).

Some of the major tools used for constructing a classification model include.

1. Naïve Bayes classifier
2. Decision Tree
3. Artificial neural network (ANN)
4. Bagging algorithm
5. K- nearest neighborhood (KNN)
6. Support vector machine (SVM)

7.4.1 Bayesian Classifiers / Naïve Bayes Technique

This Classification method is named after Thomas Bayes (1702-1761), who also proposed the Bayes Theorem. It captures uncertainty about a data model by determining probabilities of the outcomes. This method can be used to solve various diagnostic and predictive problems. It assumes a simple underlying probabilistic model in which, class membership probabilities are predicted statistically, in a way that the probability of a given sample belongs to a class. The occurrence (or nonoccurrence of a feature of a class is considered as independent to the presence (or absence) of any other feature.

For example, Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. A system discovers the concealed knowledge associated with diseases from historical records of the patients having heart disease. For each input, it gives the possibility of attribute for the expectable state. (Figure 8.2)

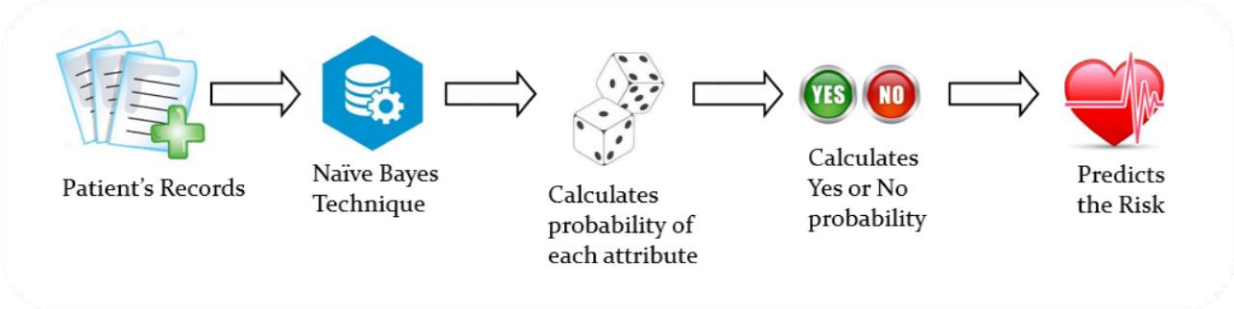


Figure 7.2: Steps of implementation of Naïve Bayes algorithm on patient data.

The mathematical expression for this method is explained below (10):

The Bayes Theorem:

$$\boldsymbol{P(h/D) = \frac{P(D/h)P(h)}{P(D)}}$$

Here,

- $\boldsymbol{P(h)}$, the probability that some hypothesis h is true, called the **prior probability of h.**
- $\boldsymbol{P(D)}$ Is the probability that some training data will be observed.
- $\boldsymbol{P(D/h)}$ Is the probability that some data value holds given the hypothesis.

- $\boldsymbol{P(h/D)}$ Is called the posterior probability of h. After we observe some data d what is the probability of h? It is also called **conditional probability.**

Usually in data mining this theorem is used to decide among alternative hypotheses$(h_1, h_2, h_3, h_4, \ldots\ h_n)$ (10)

Hence, for each hypothesis, h_1, h_2, h_3, $h_4 \ldots$ h_n, probability is computed to decide among different alternatives:

$$P(h/D) = \frac{P(D/h_1)P(h_1)}{P(D)}$$

$$P(h/D) = \frac{P(D/h_2)P(h_2)}{P(D)}$$

$$P(h/D) = \frac{P(D/h_n)P(h_n)}{P(D)}$$

Out of all the probabilities which have been computed, the hypothesis with the highest probability will be called the **maximum a posteriori hypothesis, or h_{MAP}.**

It can be translated into a following formula:

$$h_{MAP} = argmax_{h \in H} P(h/D)$$

Here, H is the set of all the hypotheses. So $h \in H$ means "for every hypothesis in the set of hypotheses." Hence for every hypothesis in the set of hypotheses $P(h/d)$ is computed and hypothesis with the largest probability is picked. Using Bayes Theorem, it can be converted to:

$$h_{MAP} = argmax_{h \in H} \frac{P(D/h)P(h)}{P(D)}$$

Here in these calculations, the denominators are identical i.e**.** $\boldsymbol{P(D)}$**.** Thus, they are independent of the hypotheses. If a specific hypothesis has the max probability, it will still be the largest if hypothesis is not divided by $\boldsymbol{P(D)}$ Hence the formula to find the most likely hypothesis is:

$$h_{MAP} = argmax_{h \in H} P(h/D)$$

An example is explained subsequently from (16) to better understand the technique

Considering a medical domain where it must be determined whether a patient has a kind of cancer or not.

It is known that only 0.8% of the people in the U.S. have this form of cancer. A simple blood test can be done that would determine whether someone has it or not. The test comes back with binary answers i.e. POS or NEG. It is also known that when the disease is present the test returns a correct POS result 98% of the time. It returns a correct NEG result 97% of the time in cases when the disease is not present.

Now, a patient undergoes a blood test for the cancer and the test result is POS.

Bayes Theorem can help determine whether the patient has cancer or not.

$$P(cancer) = 0.008$$

$$P(\ no\ cancer) = 0.992$$

$$P(POS/cancer) = 0.98$$

$$P(POS/no\ cancer) = 0.03$$

$$P(NEG/cancer) = 0.02$$

$$P(NEG/no\ cancer) = 0.97$$

Finding the **maximum a posteriori probability:**

$$P(POS/cancer)\ P(cancer) = 0.98(0.008) = 0.0078$$

$$P(POS/no\ cancer)\ P(no\ cancer) = 0.03(0.992) = 0.0298$$

Selecting h_{MAP} the patient can be classified as **not having cancer.**

To know the exact probability:

$$P(cancer/POS) =\ 0.0078/(0.0078 + 0.0298) = 0.21$$

Hence, the patient has a 21% chance of having cancer.

7.4.2 Decision Tree Classifier

Berry and Linoff defined decision tree as *"a structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become increasingly like one another." (11)*

These are simple and powerful form of multiple variable analysis, produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted tree that originates with a root node at the top. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. (Figure 7.3)

Attributes can be categorical and continuous. Finite and small set of values, such as sunny, overcast and rain, (for weather) are called **categorical attributes**. Attributes with numerical values, such as Temperature and Humidity, are known as **continuous**. A specially-designated categorical attribute is called ***the classification*****.**

Descriptions of all the object are held in tabular form in a training set. Each row of the table comprises of the (non-classifying) attribute values and the classification corresponding to one object. The aim is to develop classification rules from the data in the training set. A decision tree is created by a process known as splitting on the value of attributes i.e. testing the value of an attribute and then creating a branch for each of its possible values. This process is known as **recursive partitioning**. In the case of continuous attributes the test is normal whether the value is 'less than or equal to' or 'greater than' a given value known as the split value. The splitting process continues until each branch can be labelled with just one classification.

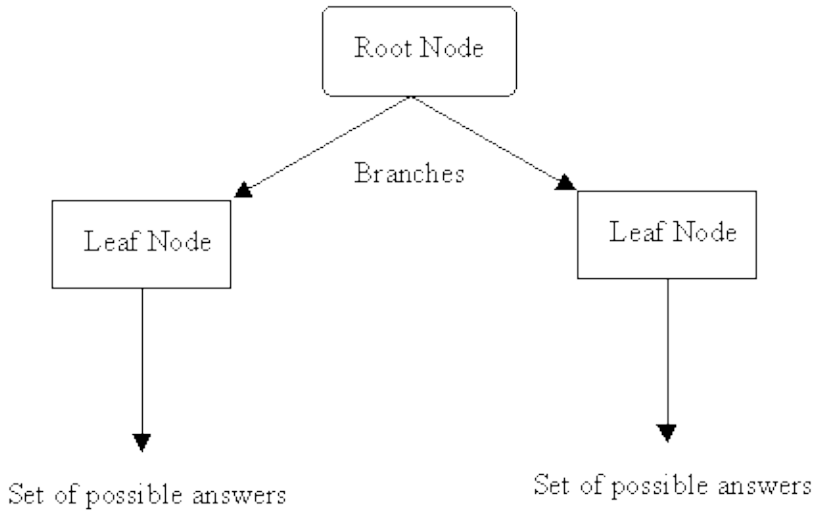


Figure 7.3: Decision Tree Representation

Hence, the decision tree can be used not only equivalent to the original training set but also generalization of it which can be used to predict the classification of other instances. These are often called as unseen instances and a collection of them are generally known as a test set or an unseen test set, contrary to the original training set. Decision tree usually has two functions: data compression and data prediction.

Through a powerful algorithm called TDIDT, these methods are widely used as a means of generating classification rules, TDIDT stands for **Top-Down Induction of Decision Trees** (Figure 8.4). This has been known since the mid-1960s and has formed the basis for many classification systems, as well as being used in many commercial data mining packages. Two of the best-known are ID3 and C4.5. ID3 and C4.5 is an algorithm for building decision trees.

IF all the instances in the training set belong to the same class

THEN return the value of the class

ELSE (a) Select an attribute A to split on+

(b) Sort the instances in the training set into subsets, one for each value of attribute A

(c) Return a tree with one branch for each non-empty subset, each branch having a descendant sub tree or a class value produced by applying the algorithm recursively

+ Never select an attribute twice in the same branch

Figure 7.4: TDIDT Basic Algorithm

7.4.2.1 Algorithm in Decision Tree

- **ID3** the word stands for Iterative Dichotomiser 3. It is one of the decision tree model that builds a decision tree from a fixed set of training instances.
- **C4.5** is the latest version of ID3 induction algorithm also an extension of ID3 algorithm. It builds a decision tree from training dataset using Information Entropy concept often called as Statistical Classifier. C4.5 is a widely used free data mining tool.
- **C5.0** model is an extension of C4.5 decision tree algorithm. Both C4.5 and C5.0 can produce classifiers expressed as either decision tree or rule sets. In many applications, rule set are preferred because they are simpler and easier to understand. The major differences are tree sizes and computation time. C5.0 is used to produce smaller trees and very fast than C4.5.
- **J48** decision tree is the implementation of ID3 algorithm developed by WEKA project team. J48 is a simple C4.5 decision tree for classification. With this technique, a tree is constructed to model the classification process. Once the tree is build, it is applied to each tuple in the database and the result in the classification for that tuple.

7.4.2.2 Types of Reasoning

Implicit form of Decision rules in a decision tree are often called *rule induction* or tree induction or decision tree induction.

Different types of reasoning are existent. The most common is *deduction*, where the conclusion is always followed from the truth of the premises (first two statements), this is 100% reliable. For example

All Men Are Mortal

Raj is a Man

Therefore, Raj is Mortal

A second type of reasoning is called *abduction*. For example,

All Men Are Mortal

Raj is Mortal

Therefore, Raj is Man

Here, even though the conclusion is consistent with the truth of the premises, but it may not necessarily be correct. Raj may not be man or perhaps name of some female. Reasoning of this kind is often successful in practice but can sometimes lead to incorrect conclusions.

A third type of reasoning is called *induction*. The process of generalization of such reasoning is based on repeated observations. For example,

If 1,000 people are seen with two legs, reasonably it can be concluded that "if x is a man then x has 2 legs" (or more simply "all people have two legs"). This is induction.

The decision trees derived from datasets are of this kind. They are generalized from repeated observations (the instances in the training sets) and they can be expected to be good enough to use for predicting the classification of unseen instances in variety of cases, but they may not be trustworthy.

Example:

A Training data can be obtained from a Hospital or a data repository. This dataset is converted into a tabular form representing data from each patient record. The selected attributes are titled above the table. When working on large dataset, a set is available which have 4 attributes namely, Sex (categorical), RestECG (categorical), Age (continuous), and Diagnosis (class). The training data has been arranged in a tabular form (Table 8.3).

Table 7.2: A Sample of training data set for Decision tree

S. No.	Sex	RestECG	Age	Diagnosis
1	Male	0	63	No
2	Female	1	67	Yes
3	Female	0	62	No
4	Male	1	56	No
5	Female	2	66	Yes
6	Female	1	69	Yes
7	Male	2	68	Yes
8	Female	0	64	No
9	Female	1	70	Yes
10	Female	0	65	No

A decision tree is generated out of this training data set (Figure 8.5).

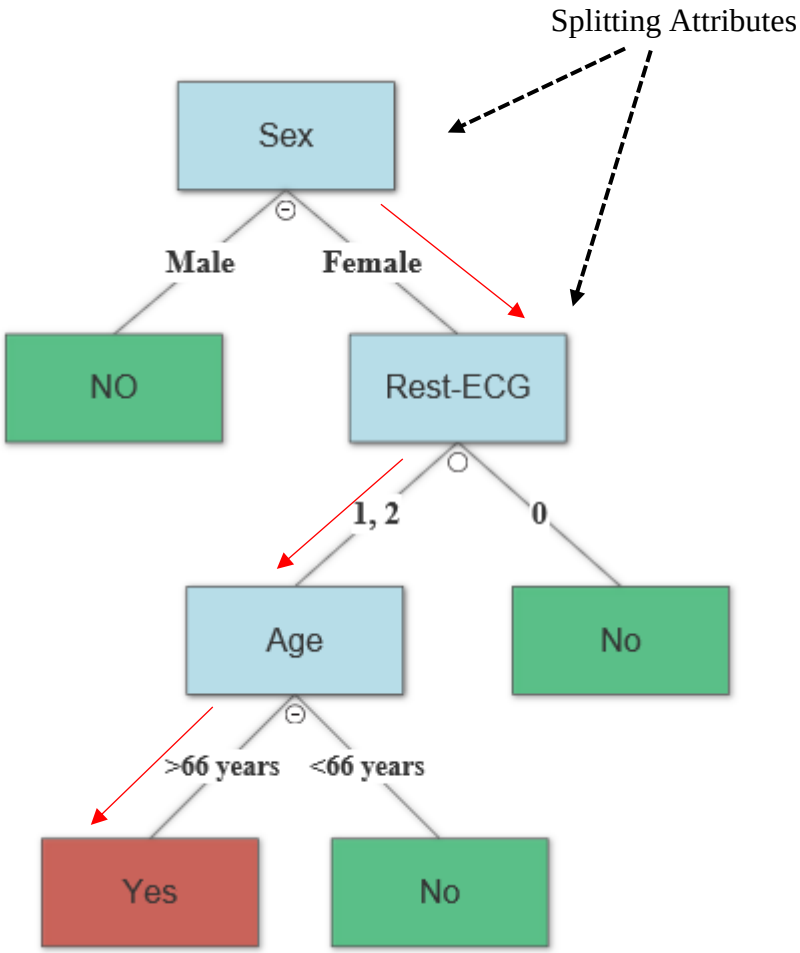


Figure 7.5: A Decision Tree of given Dataset

This technique when developed with an algorithm generate some significant rules that are useful for diagnosing heart disease. The researcher selects the most significant rules which cover most of the instances in the dataset. These selected rules are then discussed with domain expert for consideration.

For example, in the Figure 8.5 a rule can be generated following the Red Arrows.

IF Sex = Female **AND** RestECG = having ST_T wave abnormal **OR** left ventricular hypertrophy **AND** Age > 66 years **THEN** Diagnosis = YES

After rules are generated the domain expert either accepts or rejects the rule based on the results obtained from the algorithm. These rules help to decide how much the patient is likely to have heart disease.

7.4.3 Artificial Neural Network

Artificial neural networks have been inspired by knowledge of biological nervous system, the brain. The brain has a functional unit which performs all the actions called neurons (18).

The neuron has the cell body and fibers. One of these fibers - the axon - is responsible for transmitting information to other neurons. All other are dendrites, which carry information transmitted from other neurons. The synaptic button of other neurons surrounds dendrites. Neurons are highly interconnected and highly complicated. Building such an artificial network is very difficult. Hence statisticians view the brain models as inspiration to build artificial neural networks as the wings of a bird are the inspiration for the wings of an airplane (18).

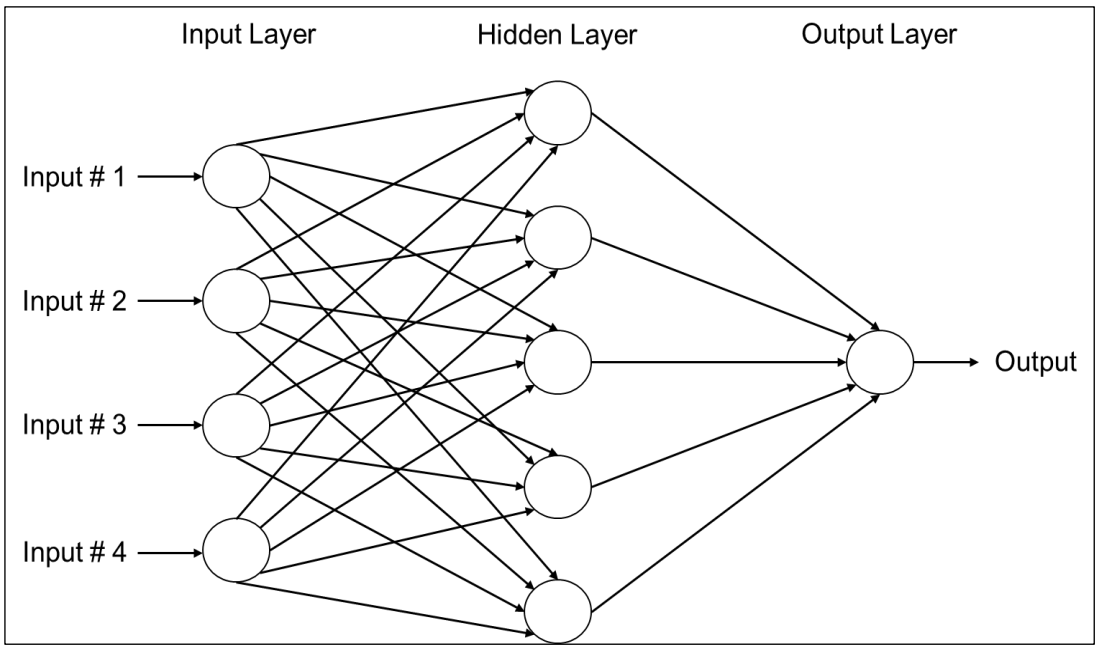


Figure 7.6: A Neural Network

- The first layer of an artificial neural network, called input layer, is constituted by the inputs x_p and the last layer is the output layer. The intermediary layers are called hidden layers.
- The number of layers and the quantity of neurons in each one of them is determined by the nature of the problem.

- A vector of values presented once to all the output units of an artificial neural network is called case, example, pattern, sample, etc. (Figure 8.6)

An artificial neuron can be identified by three basic elements:

1. A set of synapses, each characterized by a weight, that when positive means that the synapse is excited and when negative, it means inhibitory. A signal x_j in the input of synapse *j* connected to the neuron *k* is multiplied by a synaptic weight w_{jk}
2. Summation – to sum the input signs, weighted by respective synapses (weights).
3. Activation function – restrict the amplitude of the neuron output.

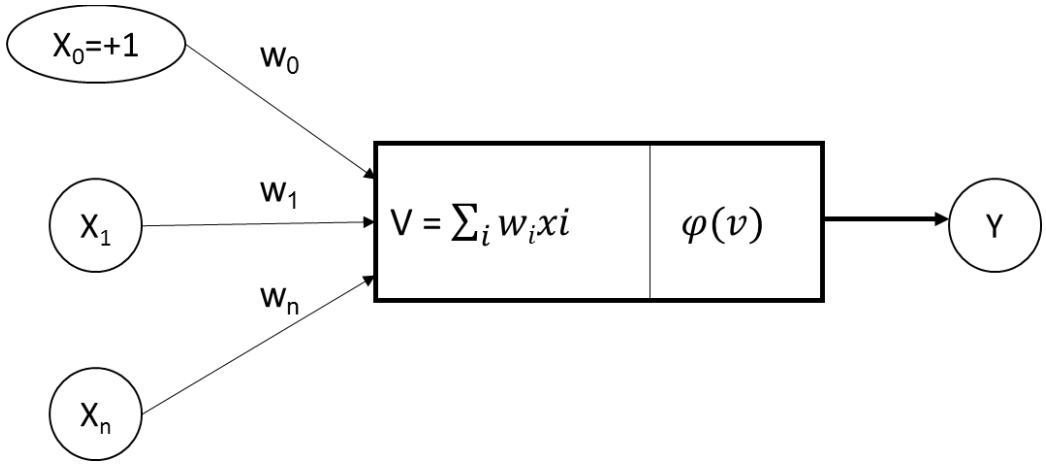


Figure 7.7: Model of an artificial neuron

Models will be displayed as networks diagrams. Neurons are represented by circles and boxes, while the connections between neurons are shown as arrows (Figure8.7):

- **Circles** represent observed **variables**, with the name shown inside the circle.
- **Boxes** represent **values** computed as a function of one or more arguments. The **symbol** inside the box indicates the **type of function**.
- **Arrows** indicate that the source of the arrow is an argument of the function computed at the destination of the arrow. Each arrow usually has a **weight** or parameter to be estimated.

- A **thick line** indicates that the values at each end are to be fitted by **same criterion** such as least squares, maximum likelihood etc.
- Several artificial neurons form an artificial neural network. Each artificial neuron is constituted by one or more inputs and one output.
- These inputs can be outputs of other neurons and the output can be the input to other neurons. The inputs are multiplied by weights and summed with one constant; this total then goes through the activation function. This function has the objective of activate or inhibit the next neuron.
- Mathematically, the k^{th} neuron can be described in the following form:

$$u_k = \sum_{j=1}^{p} w_{kj} x_j \ and \ y_k = \varphi(u_k - w_{k0})$$

Where $x_0, x_1, \ldots . x_p$ are the inputs; $w_{k1}, \ldots, w_{kp}$ are the synaptic weights; u_k is the output of the linear combination; w_{k0} is the bias; $\varphi(.)$ is the activation function and y_k is the neuron output.

Network Architecture

Architecture refers to the way neurons in a neural network are organized. There are three different classes of network architectures

1. **Single-Layer Feedforward Networks**

   In a layered neural network the neurons are organized in layers. One input layer of source nodes projects onto an output layer of neurons, but not vice versa. This network is strictly of a feedforward type.

2. **Multilayer Feedforward Networks**

   These are networks with one or more hidden layers. Neurons in each of the layer have as their inputs the output of the neurons of preceding layer only. This type of architecture covers most of the statistical applications. The neural network is said to be fully connected if every neuron in each layer is connected to each node in the adjacent forward layer, otherwise it is partially connected.

3. **Recurrent Network**

   Is a network where at least one neuron connects with one neuron of the preceding layer and creating a feedback loop. This type of architecture is mostly used in optimization problems.

8. Related Work

Heart disease or cardiovascular disease is the class of diseases that involve the heart or blood vessels (arteries and veins). Today most countries face high and increasing rates of heart disease and it has become a leading cause of morbidity and death worldwide in men and women over age sixty-five. In many countries heart disease is still viewed as a "second epidemic," replacing infectious diseases as the leading cause of death (5)

According to WHO Cardiovascular Diseases are the number one cause of death globally: more people die annually from CVDs than from any other cause. An estimated 17.5 million people died from CVDs in 2012, representing 31% of all global deaths. Of these deaths, an estimated 7.4 million were due to coronary heart disease and 6.7 million were due to stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. Out of the 16 million deaths under the age of 70 due to non-communicable diseases, 82% are in low and middle income countries and 37% are caused by CVDs (13).

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management using counselling and medicines, as appropriate (13).

Types of cardiovascular disease includes: Angina, Heart Attack (myocardial infarction), atherosclerosis, heart failure, cardiovascular disease, and cardiac arrhythmias (abnormal heart rhythms). Other forms of cardiovascular disease include congenital heart defects, cardiomyopathy, and infections of the heart, coronary artery disease, heart valve disorders, myocarditis, and pericarditis (6).

Diagnosis of cardiovascular disease includes a complete and comprehensive medical evaluation involving history and physical examination. Tests may be used to diagnose cardiovascular disease or the risk of cardiovascular disease. Some of the common tests include blood tests, exercise stress testing, EKG, X-Ray, and imaging tests, such as heart

scan, ultrasound, and echocardiogram. A coronary angiogram may be done in certain cases that reveals which coronary arteries are narrowed or blocked.

There are high chances that a diagnosis of cardiovascular disease may be overlooked or delayed because there may be no symptoms in some forms of cardiovascular disease, especially in early stages. This happens in many disease such as hypertension, arteriosclerosis, etc.

To solve such problems and many others in the health sector related to heart disease diagnosis, a fast and effective detection method must be deployed. This method should cater to all developing countries where there is a shortage of specialists and misdiagnosed cases are high. Many experts have come up with a way to extract information from enormous datasets that are collected in the past. Data mining can be a solution and a convenient tool, to assist physician by obtaining knowledge and information from these enormous datasets. Through this knowledge rules can be generated which then can be used to predict heart disease well in advance (7).

Work on data mining started long time back in 90s. As the technology progressed and knowledge obtained by data has been efficiently utilized, numerous work related to heart disease diagnosis using data mining techniques have been done.

Many studies which have been reviewed in this report have been published in the International Journal of Science and Research where different data mining techniques have been compared and their performances have been studied.

T. Revathi et al. (29) in a Comparative Study on Heart Disease Prediction System Using Data Mining Techniques have used 14 parameters for predicting heart disease that include blood pressure, cholesterol, chest pain and heart rate. The parameters have been used to improve an accuracy level. By comparing all the techniques of data mining, it was shown that neural network works well in predicting heart disease and it provides 100% of accuracy.

A paper by Priti Wadal et al. (30) intended to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. The research had developed a prototype Heart Disease Prediction System (HDPS) using Decision Trees, Naïve Bayes, and Neural Network. Using 1000 records in the prediction system with 15 medicals attributes the author could extract patterns in response to the predictable state. The most effective model

to predict patients with heart disease was Naïve Bayes followed by Neural Network and Decision Trees (17).

A study conducted by S. Saravanakumar et al (31) in 2014 proposed a frequent feature selection method for Heart Disease Prediction found out that, good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral.

A study to detect Cardiovascular Disease risk's level for adults using Naive Bayes Classifier that could also identify its risk level achieved a good performance for risk level detection. Nidhi et al (32) did an Analysis of Heart Disease Prediction using Different Data Mining Techniques. With an aim to analyze the various data mining techniques introduced in recent years for heart disease prediction the observations reveal that Neural networks with 15 attributes has outperformed over all other data mining techniques. Neural network provided 100% of accuracy, compared to Naïve Bayes and Decision Tree which showed a prediction of 90.74% 99.62% respectively. In a study of Intelligent Heart Disease Prediction System Using Data Mining Techniques, Sellappan Palaniappan et al (33) used all the methods, Decision Trees, Naïve Bayes, and Neural Network concluded that Naïve Bayes technique showed an accuracy of 86.179 %.

Data mining have shown a promising result in prediction of heart disease. It is widely applied for prediction or classification of different types of heart disease. For example, different data mining techniques were applied for prediction of ischemic heart disease and diagnosis of coronary artery disease (Tsipouras and Fotiadis, 2008; Kangwanariyakul et al., 2010) (34, 35). These successful studies which are conducted abroad have motivated this study to tackle the underlying problem that exists in our country related to heart disease diagnosis because, to the best of the researcher's knowledge there are no researches conducted locally regarding heart disease diagnosis using data mining techniques.

9. Global Perspective of Data Mining in Healthcare

Many healthcare facilities now are switching to electronic medical records (EMRs), such change is making OEMs face difficulties in managing large volumes of healthcare data with high resolution. Also, healthcare vendors are now finding it difficult to manage and interpret large amount of patient data efficiently. The Global Healthcare Data Analytics Market is thus moving up to fill up the space of managing this enormous amount of healthcare data.

In the recent years, this segment has seen a lot of activity lately, with vendors providing solutions in many areas like, financial analytics, clinical control and administration and even managing R&D. The market for data analytics is expanding at a fast pace and is expected to grow at a CAGR of 25.04 percent over the period 2013-2018. To maintain a competitive edge, many such healthcare data companies are moving towards advanced technology-supported solutions to improve patient experiences of illness.

In this extremely competitive scenario, not only doctors, and hospitals as the biggest gainers but vendors are increasingly competing to outperform each other while providing better healthcare data and solutions to doctors and hospitals (21).

Figure 10.1 shows the five most sought after companies in the Global Healthcare Data Analytics Market performing extensive Data mining.

Figure 9.1: Big Players in Healthcare Data Mining: Global Prospective

Health Informatics at IBM

The Health Informatics is one a professionally acclaimed activity at IBM Research worldwide. Health Informatics is the study of how computer systems can manage and analyze the complex processes inherent in the field of medicine.

This includes knowledge discovery and mining of clinical data, and genomic data, together with the integration and presentation of the information in useful ways applied to patients in a clinical setting. These are mostly hybrid of biology, medicine, and computer science, and draws upon research at IBM in pattern recognition, data analysis, simulation science, databases, knowledge discovery, data mining, and statistics and information theory (22).

Advance Analytics at Optum Health

Optum Health Care Solutions is another leading vendor in the Global Healthcare Data Analytics market. Its care solutions include health management solutions such as wellness, complex medical support, decision support systems, and physical health programs.

Optum One uses an unmatched foundation of health care data that is cleaned, normalized, and validated. An integrated analytics platform that is built for the health care provider market (24).

Oracle Data Mining (ODM)

Oracle Data Mining (ODM), is a component of the Oracle Advanced Analytics Database Option which provides powerful data mining algorithms that enable data analysts to discover insights, make predictions. Oracle Data Miner GUI, an extension to Oracle SQL Developer, enables data analysts, and others to work directly with data inside the database.

Oracle Data Miner creates predictive models that application developers can integrate into applications to automate the discovery and distribution of new business intelligence-predictions, patterns, and discoveries—throughout the enterprise (25).

Enterprise Analytics at MedeAnalytics

MedeAnalytics Enterprise Analytics aggregates data from multiple sources, or from an enterprise data warehouse (EDW), to create meaningful business insights. It offers Integrated Analytics for the Entire Enterprise covering Revenue Cycle, Documentation and Coding,

Supply Chain and Labor Productivity, Clinical Quality, Operational Efficiency, Strategic planning, etc. (26).

McKesson Analytics Explorer™

McKesson Analytics Explorer™ is high-speed Health Data Visualization and Reporting Software which offers many benefits in terms of handling healthcare data. With the use of this software, users can interactively visualize data without requiring predefined dimensions or constraints (23).

9.1 Information Landscape

A recent study by IBM estimates that humanity creates 2.5 quintillion bytes of data every day. That total includes stored information—photos, videos, social-media posts, word-processing files, phone-call records, financial records, and results from science experiments—and data that normally exists for mere moments, such as phone-call content and Skype chats.

The global healthcare analytics market is estimated at USD 9.96 billion in 2016 and is projected to reach USD 31.75 billion by 2021, at a CAGR of 26.1% during the forecast period (27).

The increasing need to improve the efficiency of current healthcare institutions, medical professionals, and medical practices, need to improve quality, and demand to lower the ever-increasing healthcare costs are the primary drivers of this market.

The global healthcare analytics market can be segmented based on technology/platform (predictive analytics, prescriptive analytics, and descriptive analytics), application (clinical data analytics, financial data analytics, administrative data analytics, and research data analytics), component (software, hardware, and services), deployment (on premise model and cloud-based model), end-user (healthcare, pharmaceuticals, biotechnology, academia, research, and others), and geography (North America, Europe, Asia-Pacific, Middle East & Africa and Latin America) (Figure 7.5.2).

Currently, the descriptive analytics segment accounts for the largest market share in healthcare analytics technology but its share is expected to decrease in the coming years, and share of predictive and prescriptive analytics is expected to increase during the forecast

period. By application, the clinical analytics segment is forecasted to have the highest growth rate. Although the market growth of the cloud-based services segment is increasing substantially, the on premise delivery system segment still accounts for the largest share due to its multi-vendor framework (27).

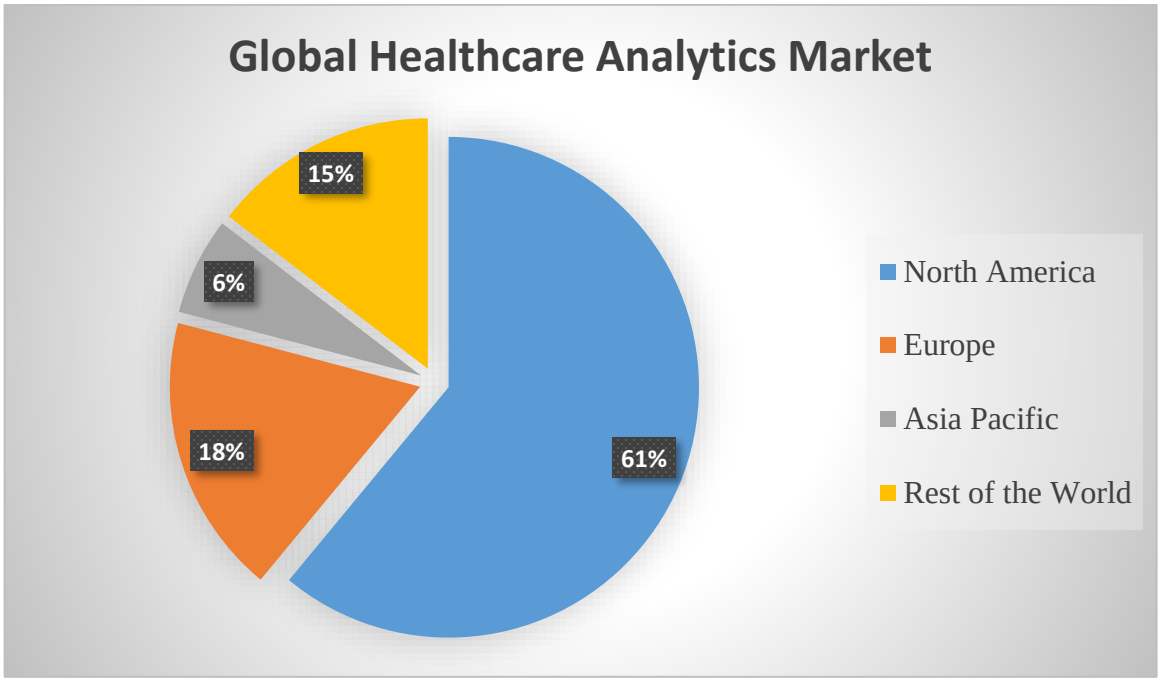


Figure 9.2: Global Healthcare Analytics Market: by Geography (USD million – 2015)

The global healthcare data has been doubling every two years, with emerging markets like Brazil, China and India producing 32 per cent off the global data in 2012. In 2019, it is forecasted that the same emerging economies like, India, China and Brazil will produce 62 per cent of the global data. In 2020, it is projected that the global data produce would be at a staggering level of 40 thousand exabytes, which will be 50 times greater than what is was in 2010 (Figure 10.2) (28).

Health data has shown greater projection in 2014-2016, with more technologies like EMR and EHR and better computer storage facilities. Healthcare data is expected to grow at a greater pace in the coming years. The Figure 10.3 shows the healthcare data in brown color. This huge amount of data which is expected to reach at a level of 6500 Exabyte would be available to be converted into useful information for various predictive reasons.

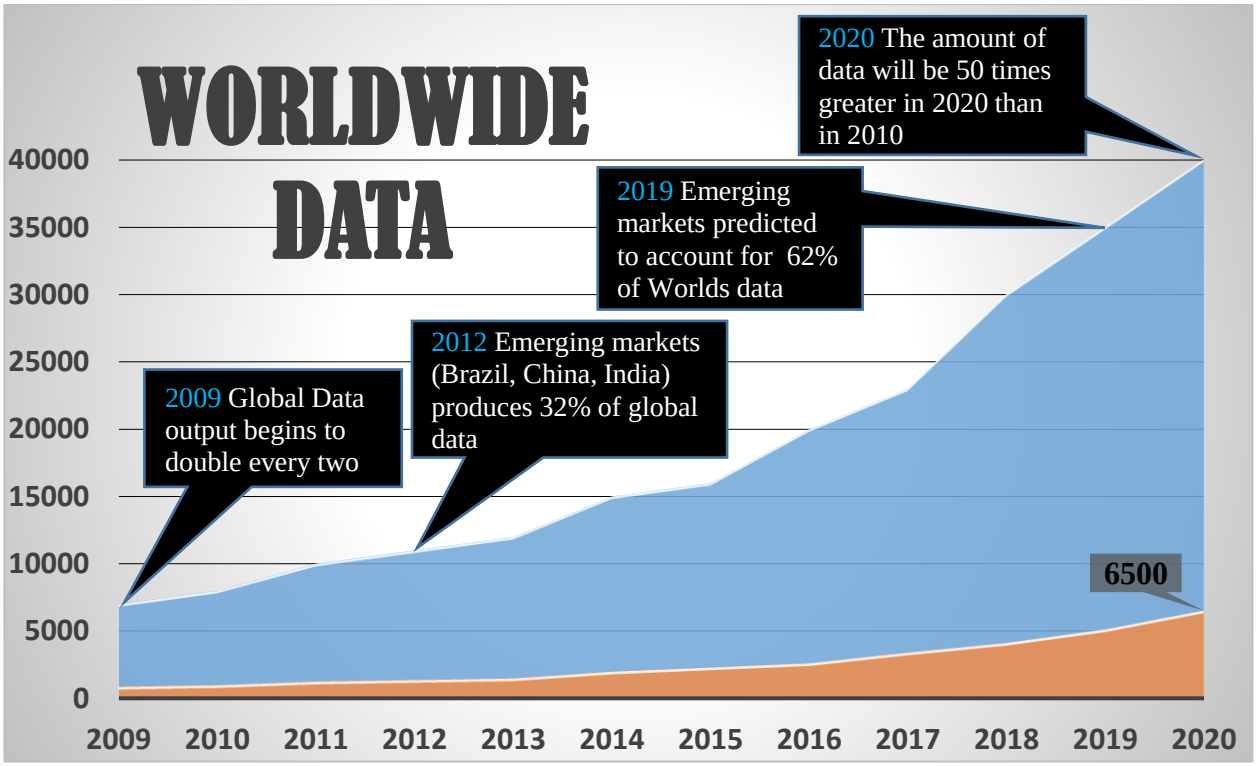


Figure 9.3: Worldwide Data Trend projected for 2020

10. Results

As one of the objective of this study was to study heart disease prediction using data mining techniques, three classification technique were studied based on the literature review of secondary data. The three techniques included are, Naïve Bayesian Classifier, Decision Tree Classification Algorithm, and Artificial Neural Network.

In a typical data mining application of heart disease prediction, experiments are conducted where multiple scenarios can be considered based on the study objective. In one scenario, all attributes may be contained and in other, only selected attributes might be contained. In the same way, two experiments with different attributes selection technique can be considered. With such different scenarios, models are developed and experiments are performed.

Results obtained from literature review have been divided into following parts.

10.1 Total Records/Instances:

Ina typical scenario, the experiments were first conducted on a full training dataset and Cross Validation was adopted for randomly sampling the training and test dataset.

49 studies were reviewed from different sources over a period of 10 years from 2008 to 2016. It was found that patient records/instances are taken maximum in the range of 200-400 (Figure 11.1).

Frequency distribution showed that, maximum frequency is seen in the range of 200-400 (Table 11.1).

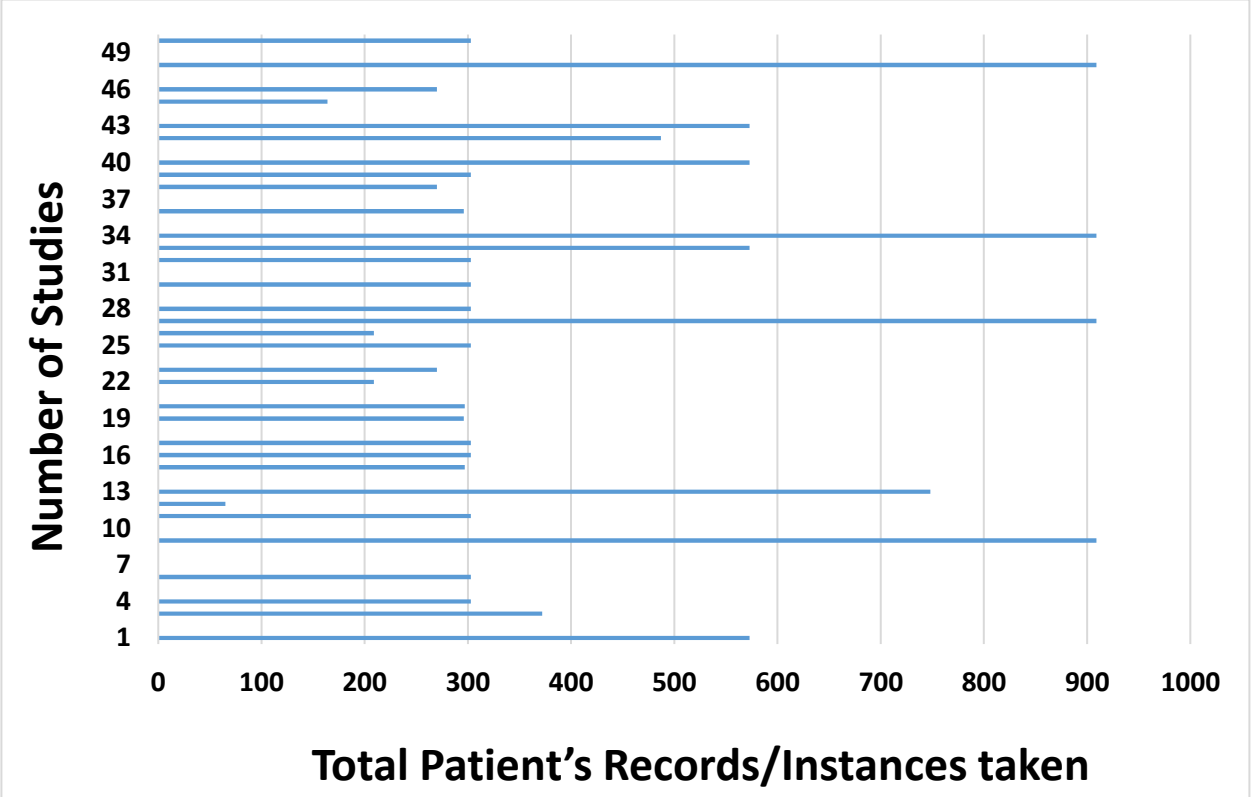


Figure 10.1: Distribution of Patient's Records among the number of studies reviewed

Table 10.1: Frequency Distribution of number of Patient's Records

Range of Records	
0-200	4
200-400	32
Above 400	13

10.2 Attributes Selected

Attribute selection is a data reduction techniques which is applied to the dataset. A domain expert selects the attributes that must be included on the dataset.

In the book, Han and Kamber (19) has stated that attribute selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions) and mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

Minimum set of attributes are found so that resulting probability distribution of the data classes will be as close as possible to the original distribution obtained using all attributes.

This can be done using Best First search method. It starts with an empty set of attributes as the reduced set, following which the best of the original attributes is determined and added to the reduced set. At each subsequent step, the best of the remaining original attributes is added to the set. This method selects the minimum number of attributes from a total number of attributes.

In this study, literature review was done to identify the most commonly used attributes. It was found that, thirteen attributes with one output attribute are most commonly used for heart disease prediction system. The selected attributes with their description have been listed in the Table 11.2. In some studies, few other attributes have also been used. These attributes were chosen to predict diagnosis through better attribute selection method. The most commonly used attributes are Age, Chest pain type, Resting Blood pressure, Serum Cholesterol, Rest ECG results, Heart Defect, Diagnosis. All other attributes have been listed in the Table 11.3

Table 10.2: List of most commonly used Attributes with their description

S. No.	Attribute	Description
1	Age	Age in years
2	Sex	Sex
3	Cpt	Chest pain type
4	Trestbps	Resting blood pressure
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar
7	Restecg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	CA	Number of major vessels coloured by fluoroscopy
13	Thal	Heart defect
14	num	The predicted attribute diagnosis of heart disease (angiographic disease status)

Table 10.3: List of other attributes found in studies

S. No.	Other Attributes
1	Pt. ID
2	Obesity
3	Smoking
4	Diabetes
5	Alcohol
6	Heart rate
7	Diet

10.3 Performance Measure

In classification problems, the primary source of performance measurements is a confusion matrix (Coincidence matrix, classification matrix or a contingency table). Given n classes, a confusion matrix is a table of at least size n X n (20).

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

If the instance is positive and it is classified as positive, it is counted as a true positive (TP); if it is classified as negative, it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as a true negative (TN); if it is classified as positive, it is counted as a false positive (FP). These terms are useful when analyzing a classifier's ability (Figure 11.2).

	Predicted Heart Disease: YES	**Predicted Heart Disease: NO**
Actual Heart Disease: YES	True Positive (TP)	False Negative (FN)
Actual Heart Disease: NO	False Positive (FP)	True Negative (TN)

Figure 10.2: Confusion Matrix Table

Sensitivity is also referred to as the true positive rate i.e., the proportion of positive instances that are correctly identified, while specificity is the true negative rate i.e., the proportion of negative instances that are correctly identified. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples.

$$Sensititivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

In this study, performances measurements have been reviewed from the literature and results of Sensitivity, Specificity and Accuracy of the studied techniques were extracted. All selected studies had the same number and type of attribute to avoid false selection. The overall average Sensitivity was calculated for each technique. Among the three, Neural Network was shown to have highest average Sensitivity at 92.8 per cent. Lowest was for Decision tree at 84.5 per cent (Figure 11.3)

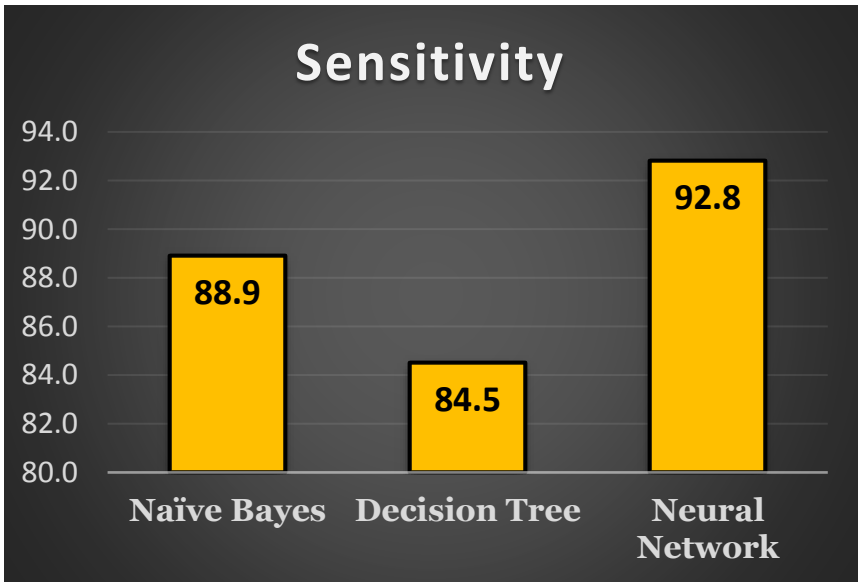


Figure 10.3: Comparison of Sensitivity between different techniques of Heart Disease Prediction

Specificity showed a different result than sensitivity. Naïve Bayes classifier at 91.3 per cent had an edge over Neural Network at 90.8 per cent. Decision tree was low at 84.2 per cent (Figure 11.4).

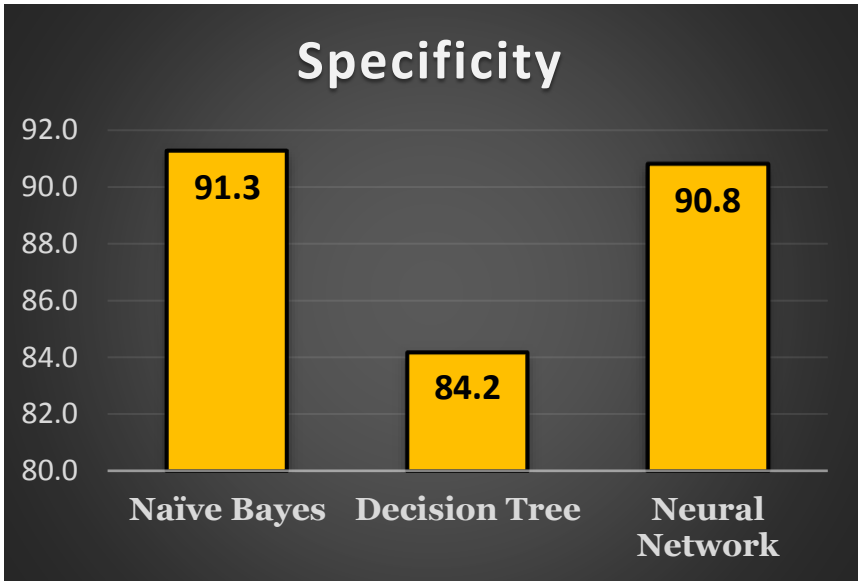


Figure 10.4: Comparison of Specificity between different techniques for Heart Disease Prediction

While comparing the three Classifiers on the grounds of accuracy, it was found that Neural Network and Naïve Bayes have almost equal value at 89.8 and 89.3 per cent respectively. This clearly justifies how Neural Network outperforms the other two in predictive diagnosis of Heart Disease (Figure 11.5)

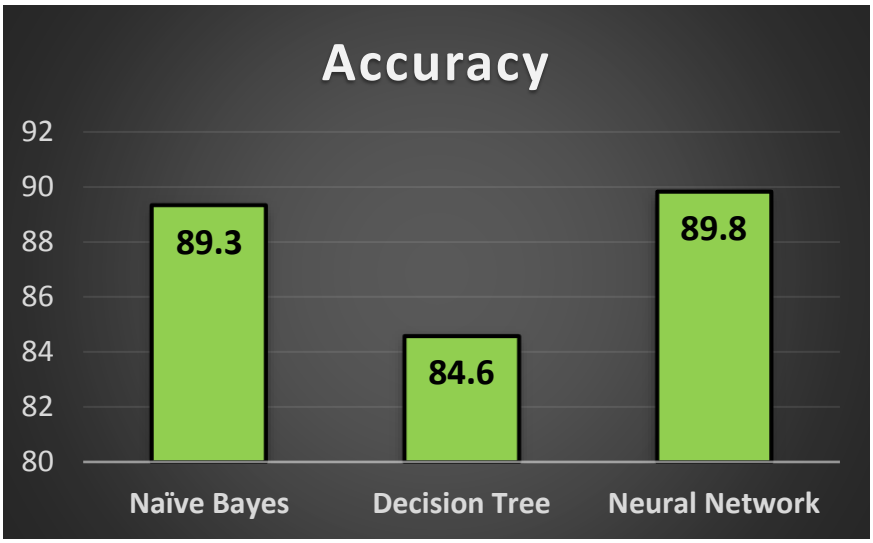


Figure 10.5: Comparison of Accuracy between different techniques for Heart Disease Prediction

11. Conclusion

In the report, the aim was to review predictive model for heart disease detection using three data mining classifiers which can enhance the reliability of heart disease diagnosis. These classifiers have been studied in detail to bring out more clarity on the subject and how these experiments work. Data Mining practice has been going on in different areas of healthcare sector as identified in the report by studying some of the big players / global leaders from this field who are taking it to the next level. The market for Predictive analytics in healthcare has been increasing many folds with emerging economies like Brazil, China and India predicted to contribute a greater share in this segment.

Heart disease is a fatal disease by its nature and misdiagnosis of this disease can cause serious, even life threatening complications such as cardiac arrest and death. This study reviewed how data mining techniques can be used efficiently to model and predict heart disease cases by identifying the common types of attribute and number of patient records required. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease. Naïve Bayes model has shown the highest specificity rate which makes it a handy tool for cardiologists to screen out patients who have a high probability of not having the disease and transfer those patients to for further analysis. The Neural Network model reviewed from the previous literature shows the maximum Accuracy, predicting cases with heart disease, with a classification accuracy of 89.8 per cent. Decision tree on the other hand could not exceed a classification accuracy of 84.6 per cent.

12. Recommendations

This study has reviewed three most commonly used data mining techniques applied in the prediction of heart disease. To improve the classification accuracy of the models further researches can be conducted using different classification algorithms like support vector machine (SVM) and Rule Induction.

Most of the experiments reviewed in this study were implemented with default parameters of the algorithms, further investigations can be performed with different parameter settings to enhance and expand the capabilities of the prediction models. In addition, the Neural

Network and Naïve Bayes classification algorithms can be thoroughly tested to get more accurate results.

The data mining methods used in this study were capable of extracting patterns and relationships hidden in the dataset. The patterns found can be used to design a Knowledge Based System (KBS) with the help of domain experts who have years of experience in the problem domain.

As a future work, a researcher can plan to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types and more accurate results.

13. Limitations

This study is a scientific review of literature which does not include comprehensive search of both published and unpublished literature. A systematic literature review would provide better results and more accurate prediction.