Heart Disease diagnosis using machine learning classification techniques

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted by:

SANJIB KUMAR SHAW

2k20/SWE/20

Under the supervision

of

Mr. SANJAY PATIDAR

(Assistant Professor)

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

JUNE, 2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, SANJIB KUMAR SHAW, Roll No. 2k20/SWE/20 student of M. Tech (SOFTWARE ENGINEERING), hereby declare that the project Dissertation titled "Heart Disease diagnosis using machine learning classification techniques" which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associate ship, Fellowship or other similar title orrecognition.

Sanjib Kr Shaw

SANJIB K SHAW

2k20/SWE/20

Place: Delhi

Date: 31/05/2022

i

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Heart Disease diagnosis using machine learning classification techniques"** which is submitted by SANJIB KUMAR SHAW, 2k20/SWE/20 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this Universityor elsewhere.

Place: Delhi

Date: 31/5/2022

Sanjay Patidar

Assistant Professor

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organization. I am grateful to everyone who helped in shaping the result of the project.

I express my sincere thanks to **Mr**. **Sanjay patidar**, my project guide, for providing me with the opportunity to undertake this project under his guidance. His constant support and encouragement have made me realize that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

Sanjib Kr Shaw
2k20/SWE/20

ABSTRACT

Predicting heart disease is difficult in medicine. In India, heart disease causes most deaths. In many nations, overwork, stress, and other factors cause cardiovascular disease deaths. It's linked to heart disease in adults. For identifying cardiac disease, a decision support system is needed. Our work uses data mining to better predict cardiac disease. Heart disease is a leading cause of mortality worldwide, notably in Bangladesh. Forecasting cardiac disease accurately is a difficult and time-consuming procedure, but machine learning (ML) methods may help. This article explains our preferred approach for predicting cardiac problems, which uses machine learning algorithms to discover key indicators and improve accuracy. The UCI Repository has 14 features from our dataset. We built our model by categorising the world using Maximum Entropy, Random Forest, and SVM. SVM delivered the best performance in our suggested system, with 92.67 percent accuracy for the threshold instances of the dataset. The new method has produced 20% more accurate results than before.

Keywords : Machine Learning , maximum entropy, random forest, support vector machine (SVM), UCI repository.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

LIST OF ABBREVIATIONS

µL	Micro liter
µm	Micrometer
0C	Celsius
Abs	Absorbance
ANN	Artificial Neural Networks
AVG	Average
BMI	Body Mass Index
BP	Back-Propagation
BP	Blood Pressure
BPNN	Back-Propagation Neural Network
CAD	Coronary Artery Disease
CBR	Case-Based Reasoning
CCBs	Calcium Channel Blockers
CHD	Coronary Heart Disease
CHF	Cardiac Heart Failure
CNS	Central Nervous System
DASH	Dietary Approach to Stop Hypertension
DBP	Diastolic Blood Pressure
DC of WBC	Differential Count of White Blood Corpuscles
DM	Diabetes mellitus
DOCA	Deoxycorticosterone Acetate
ECF	Extra Cellular Fluid
EDTA	Ethylene Di-amine Tetra Acetic Acid
ELISA	Enzyme Linked Immune sorbent Assay
EME	Established Market Economy
Hb	Hemoglobin
HCl.	Hydrochloride
IDH	Isolated Diastolic Hypertension

IHD	Ischemic Heart Disease
ISH	Isolated Systolic Hypertension
M	Molarity
MAE	Mean Absolute Error
MAP	Mean Arterial Pressure
MCH	Mean Corpuscular Hemoglobin
MCV	Mean Corpuscular Volume
Mg	Magnesium
Mg	Milligram
mg/dl	Milligram Per Deciliter
mm	Millimeter
mmHg	Millimeter of Mercury
mmol/L	Mill mole Per Liter
Na+	Sodium ion
NaCl	Sodium Chloride
NH-34	National Highway-34
NUS	Narrow Urinary Space
PA	Primary Aldosteronism
PCT	Proximal Convoluted Tubule
pH	Potential of Hydrogen
PP	Pulse Pressure
RAAS	Renin-Angiotensin-Aldosterone-System
RAS	Renin-Angiotensin-System
RBC	Red Blood Corpuscles
RN	Round Nucleus
RMES	Root Mean Square Error
SVM	Support Vector Machine
WBC	White Blood Corpuscle
WHO	World Health Organization
WHR	Waist Hip Ratio

xi

CHAPTER I

INTRODUCTION

CHAPTER I
INTRODUCTION

This chapter provides a summary of the study, starting with the need of researching heart disease and the difficulties that come along with it. The explanation of the issue and the need of the project are both discussed in the parts that follow.

1.1. Data Mining

The collection and analysis of information and data have emerged as genuine resources for most enterprises. The process of learning revelation in medicinal databases is highly defined, and information mining is an essential step forward in the field. Databases are aggregations of information that have an explicit and well-defined structure and justification for their existence. DBMS refers to the efforts that are undertaken to produce and manage such information. The process that is generally related with the revelation of learning from information is the technique that is associated with information revelation in databases. The process of computationally isolating hidden learning structures that are referenced in models and examples from large information storehouses is what information mining is concerned with.

The fields of databases, machine learning, and representation are the primary areas of focus in the field of information mining, which is an interdisciplinary subject. It does this by identifying models of successful therapeutic treatments for a variety of disorders, and it furthermore endeavours to unearth useful information from vast stores of data. Mining information is at the core of knowledge discovery and data mining (KDD), and it is used to extract interesting instances from large amounts of data that are not difficult to view, translate, or manipulate. It is the process of searching through vast stockpiles of information with the intention of locating patterns in order to generate useful data. The KDD process is comprised of a few steps that, beginning with the acquisition of raw data and leading to the production of new knowledge.

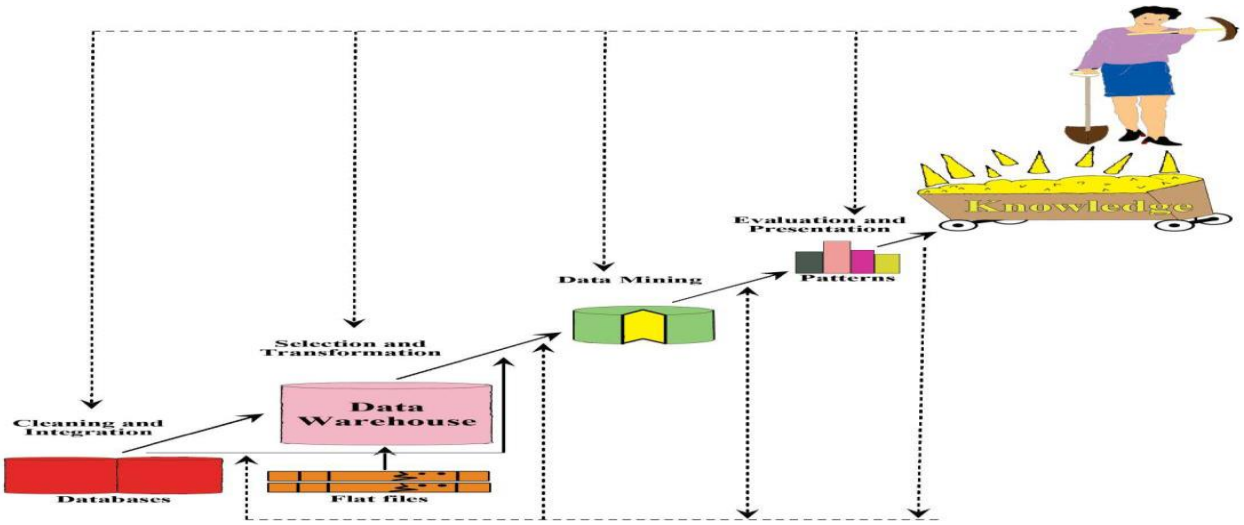


Figure 1.1 Knowledge discoveries as a process [1]

The process of learning revelation is shown as consisting of an iterative succession of information cleansing, information incorporation, information determination, information mining design acknowledgement, and learning introduction. This is shown in Figure 1.1. The process of "digging" for information entails searching through enormous databases in search of linkages and global instances that may be buried deep inside the mass of material. Before information mining calculations can be used, an objective informative index has to be compiled and compiled first. A common hotspot for information is a store or stockroom dedicated to the storage of information, and preliminary processing is required in order to partition these multivariate informative collections. The last step in the disclosure of information obtained from data is to verify that the instances identified by the data mining algorithms are present in the more comprehensive informative collection. The information that was obtained may comprise decisions that describe the features of the information, patterns that occur as often as possible, and things that are seen to be in bunches in the database, etc.,

In the realm of research on medical treatments, information mining serves as a testing zone. The extraction of useful information from databases and the provision of rational basic guidance in the diagnosis and treatment of illnesses is becoming more important. Information mining in the field of medicine may help solve this problem. Digging up medical material has an outstanding potential for uncovering hidden instances in the informative archives of the therapeutic field [1].

1.2 Data Mining Tasks

In machine learning, one of the common tasks is the grouping of information. In the middle of the 1950s, artificial awareness first achieved acknowledgement as a kind of control. Learning is one of the most important pre-requisites for any clever behaviour you may exhibit. The majority of scientists working in the modern era are of the opinion that knowledge cannot exist apart from education. In the field of research into artificial consciousness, machine learning has been an essential component of the field's development from the very beginning. The field of software engineering known as machine learning focuses on the development of algorithms that provide personal computers the ability to acquire knowledge. It is often used in the process of developing systems that may ensure increased effectiveness and appropriateness on the part of the system. The tasks associated with prescient exhibiting include the study of order, relapse, and temporal organisation, as shown in Figure 1.2. Clear showing necessitates a number of tasks, some of which include grouping, affiliation rules, and representation.

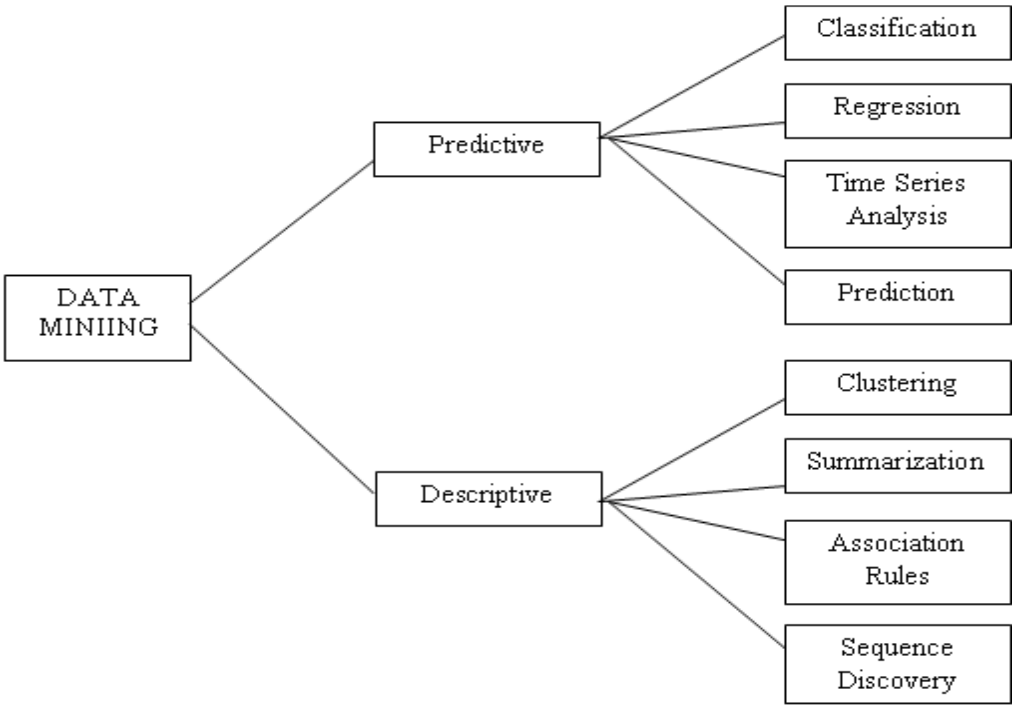


Figure 1.2 Data mining tasks [2]

In the field of machine learning, grouping refers to the process of determining which of a number of different categories another perception belongs to. The preparation of a large amount of information that contains perceptions or cases whose categorization involvement is known is

used as the basis for doing this. The process of establishing a model or capability that depicts and identifies information classes is referred to as grouping. The ultimate objective of this process is to enable the capacity to use the model to anticipate the class of objects whose class mark is unknown. It is a capability of learning that maps or organises an information item into one of a few predetermined groups or classes that fall under the category of controlled learning. Evaluating of informative index is enhanced the circumstance testing the order effectiveness, and the grouping model makes use of preparing informational gathering in order to manufacture an arrangement prescient model. Two distinct problems, such as parallel grouping and multiclass arrangement, may each be thought of as one of its two halves. In a twofold arrangement, just those two groups are taken into consideration, however in a multiclass grouping, one of the classes may be designated as the protest class [2].

The goal of expectation may be reached with the help of relapse. It is the means by which an examination of the existing and historical circumstances of the characteristic, as well as an anticipation of its upcoming state, may be carried out. A prediction of an esteem may be made with the use of an information mining technique called relapse. A mathematical dataset is used as the basis for developing a scientific formula that is tailored to the data. A dataset including achieved goal values is the starting point for a relapse project. Relapse analysis may be used to illustrate the link between at least one autonomous or indicator component and a ward or response variable. Direct relapse, multivariate straight relapse, nonlinear relapse, and multivariate nonlinear relapse are the several types of relapse procedures.

A collection of information foci, generally estimated at successive focuses in time spaced at regular time intervals, is what we mean when we talk about time organisation. Examining time arrangements requires using methods that include dissecting information about time arrangements in order to extract key insights and various properties associated with that information. Techniques for analysing time arrangements can be divided into two categories: recurrence area techniques, which include phantom analysis, and time-space techniques, which include auto-connection and cross-relationship analysis. Both of these categories can be subdivided further into a number of subcategories. When it comes to time arrangements, there are a few different sorts of inspiration and information analysis that are available, each of which is suitable for a different kind of goal. When it comes to information mining, pattern recognition, and machine learning, time arrangement analysis may be used for grouping, categorization, and inquiry by content, as well as for detecting and identifying abnormalities.

A group of objects that are similar to one another yet distinct from the items that belong to other groups is called a bunch. Bunching differentiates between collections of items that share specific properties and is an example of unsupervised learning since there are no specified classes in bunching. It analyses data structures without providing a class name as a counselling option. The articles are grouped together or bunched according to the guideline of increasing the level of familiarity within the class while decreasing the degree of familiarity between the classes. It is the first step in exploratory data mining and a common technique for factual information exploration that is used in a variety of domains, such as machine learning, design recognition, image analysis, data recovery, and bioinformatics, among others. In general, there are two types of grouping that can be referred to as "bunching": hard grouping and delicate grouping. Hard grouping determines whether or not each question fits into any of the groups, whereas delicate grouping determines whether or not each question fits into any of the groups to a certain degree. The K-implies grouping computation is one of the most common types of bunching calculations used.

Finding interesting connections between elements in respect to large datasets may be accomplished with the help of a method known as affiliation rule learning, which has garnered a lot of attention and is widely used. It is designed to identify reliable principles that may be discovered in databases by making use of a variety of levels of interest. There are a few computations based on the affiliation rules that are mostly helpful in condensing and identifying the instances. In addition to using connections, assistance, and trust, they do this so that they may find the appropriate instances. These are often necessary to satisfy a client's expressed need for the least amount of aid and a client's decided need for the least amount of assurance in the interim. It is broken down into two distinct stages, the first of which involves finding assistance for the purpose of locating all successive thing sets in a database. The second stage involves utilising these continuous thing sets along with the fundamental certainty requirement in order to formulate rules. In most cases, affiliation and connection are required in order to identify continuing item set discoveries amid significant informative indexes. The affiliation is different from the order in that it can forecast any characteristic, not only the class, and they may forecast the value of more than one property at the same time. There are a few different varieties of

affiliation rules, the most common of which are the staggered affiliation rule, the multidimensional affiliation rule, and the quantitative affiliation rule.

A classifier is the name given to the inferred capacity that is produced as a result of the managed learning calculation that breaks down the preparation information. When the yield consists of discrete or straightforward features, this phenomenon is referred to as characterisation; when the yield consists of numerical or ongoing characteristics, this phenomenon is referred to as relapse. The concept of unsupervised learning refers to the challenge of trying to unearth hidden structures within material that has not been labelled.

1.3 Statement of the Motivation

The prediction of coronary disease is one of the most significant challenges faced by the social insurance business. In response to the general rise in death rates among coronary disease patients, medical professionals are using a wide variety of information mining techniques in the process of diagnosing cardiovascular disease. Every method has a certain set of advantages as well as drawbacks. Each and every calculation that is used by each and every approach has particular capabilities that are helpful when analysing cardiac disease. When conducting an optimal inquiry of coronary disease, the results of each calculation should be combined and examined.

1.4 Statement of the Problem

The machine learning approach determines the degree to which new facts, such as instances or tenets derived from a large amount of information, are revealed. In the field of information mining, illness anticipation fills the role of an essential worker. In order to get at a diagnosis of a disease, it is necessary to conduct a number of different tests on the patient. The use of information mining technologies, on the other hand, has the potential to lessen the number of tests required. This reduced test set is responsible for an important task in terms of both timing and execution. The mining of data relating to coronary disease is important because it helps medical professionals to determine which aspects or characteristics are more important for

diagnosis. Some examples of these are age, weight, and other similar factors. Because of this, the doctors will be able to perform an even more accurate analysis of heart disease.

Although the medical services industry makes use of a variety of information mining systems, there is still research that needs to be conducted on the various methods of classification that are currently in use. This will allow for the selection of the information mining system that provides the highest quality results. It is anticipated that the investigation that is presented in this theory would address the challenge of improving the forecast model to anticipate coronary disease in individuals who already have coronary disease and to provide a favourable response when predicting an infection. Rapidly, the important research capabilities are afterwards described as questions such as,

- How distinct information mining tactics may be employed in the pharmaceutical services business and to separate their performance in expectation?"

- How can the use of an arrangement system contribute to the construction of a forecast display in order to correctly predict the risk of cardiovascular disease among diabetic patients?

1.5 Objectives of the Research

Exploring the existing relationships between elements may be accomplished in a satisfactory manner by using information mining as part of the process of dissecting therapeutic information. These days, the information that is stored in restorative databases is expanding at a pace that cannot be stopped by any means. The concept that restorative information analysis may lead to an improvement in social insurance has gained widespread acceptance in recent years.

The effective development of a forecast display by employing various characterisation approaches to foresee coronary disease and performance in anticipation is the primary objective of the research endeavour. It also indicates that data mining may be used to medical datasets in order to anticipate or organise the information with a level of accuracy that is reasonable.

Following that, we will discuss the objectives that need be accomplished in order to reach the primary objective mentioned earlier:

- To identify the greatest possible arrangement demonstration that may assist medical professionals in predicting the risk of heart disease by employing diabetes characteristics.

- The ability to recognise and categorise patterns among multivariate patient variables.
- To foresee the future outcomes relying on prior experiences and present situations.
- To identify patients who are in risk, with the goal of increasing the quality of care provided while simultaneously lowering the cost of care.
- To construct a prediction display making use of appropriate grouping algorithms such as Rendom Forest, SVM, and Decision tree.

1.6 Contribution ofthe Thesis

Within the scope of this postulation, we investigated coronary artery disease. Decision tree and credulous bayes calculation had been presented in previous work; nevertheless, we employ Maximum Entropy, Random Forest, and Support Vector Machine, all of which are thought about bases on accuracy offered in previous work.

1.7 Organization of the Thesis

In the second chapter, a literature review on data mining, its most important prediction approaches, applications, a survey of comparative analyses conducted by other researchers, and the criteria that will be utilised for model comparison in this study are discussed.

The Problem Statement and Data Sets for the Proposed Diagnosis System are Presented in Chapter Three.

In the fourth chapter, a summary of the findings is discussed, as well as a comparison of the findings about the approaches used to the planned work and the previous work.

The findings and recommendations for the future are presented in Chapter 5.

X

CHAPTER 2

LITERATURE REVIEW

CHAPTER 2
BACKGROUND STUDIES AND LITERATURE REVIEW

In this Chapter we depict the background of our research work and writing survey related issue. Learning revelation in databases and information mining is an interdisciplinary region concentrating on the procedures for separating helpful learning from information. Have express that the need of viable recognizable proof of data, relevant information, non clear and important for basic leadership from a substantial gathering of information has been on a consistent increment as of late.

2.1 Preliminaries

This is an intuitive and iterative process including a few subtasks and choices and is known as Knowledge Discovery from Data. The focal procedure of learning revelation is the change of information into information for basic leadership, known as information mining [1].

Today, information mining has developed so tremendous that they can be utilized in numerous applications and it is the way toward finding beforehand obscure examples and patterns in information, to assemble prescient models. It is another way to deal with information investigation and information disclosure and started from work of insights and machine learning as an interdisciplinary field. Insight is the capacity of getting the hang of, comprehension and discovering answers for issues in an explicit area and Artificial Intelligence is a part of Computer Science. Machine Intelligence is utilized for medicinal information mining and it extricates biomedical and social insurance learning for clinical basic leadership and creates logical speculation from huge restorative information. Therefore, KDD, which incorporates information mining systems, has turned into a prominent research device for social insurance scientists.

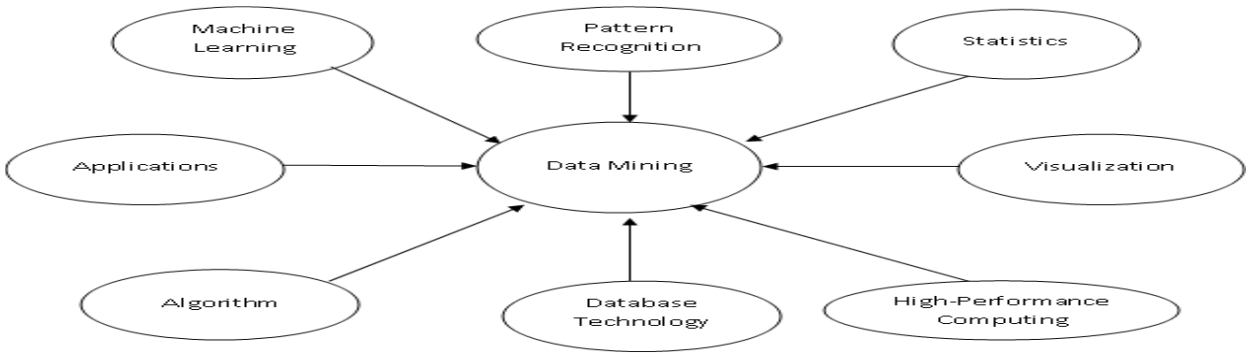


Figure 2.1 Origin of Data mining [1]

As appeared in Figure 2.1, information mining is a term that portrays diverse strategies utilized in a space of machine learning, measurable investigation, design acknowledgment, forecast, characterization, grouping, perception, displaying methods and database advancements that can be utilized in various ventures. It has wide applications in all parts of industry, for example, media communications, retail, generation, keeping money, instruction, and human services the executives. It includes different stages as business understanding, information understanding, information arrangement, displaying, assessment and organization.

The necessity for information mining arises from the fact that the quantity of information seems to increase rapidly on a daily basis throughout the majority of domains associated with data preparation, as does the need to learn how to mine and get insights from databases. Its applications also have the potential to benefit social insurance providers including doctor's facilities, centres, physicians, and patients by identifying effective treatments and best practises.

2.2 Data Mining Approaches in Medicine

Wellbeing is a typical topic in many societies. Among definitions still utilized, most likely the most seasoned is that wellbeing is the nonattendance of sickness. Wellbeing isn't fundamentally about issues of' specialists, social administrations and doctor's facilities. It is an issue of social equity. The broadly acknowledged meaning of wellbeing given by the WHO (1948) is "Wellbeing is a situation of complete physical, mental, and social prosperity; it is not only the absence of infection or illness," says the World Health Organization (WHO).

Present day prescription has developed instruments and strategies which might be utilized in different mixes for the evaluation of physical wellbeing. They incorporate self appraisal of by and large wellbeing, investigation into manifestations of sick wellbeing and hazard factors, investigation into the utilization of restorative administrations, institutionalized surveys for cardiovascular diseasees and clinical examination. The term wellbeing and family welfare administration covers a wide range of individual and network administrations for treatment of infection, aversion of ailment and advancement of wellbeing. The reason for wellbeing administrations is to enhance the wellbeing status of any populace.

There have been numerous endeavors to characterize ailment. The Oxford English Dictionary characterizes malady as a state of the body or some part or organ of the body in which its capacities are upset or unsettled'. Learning about human wellbeing and disease is whole of the commitments of a substantial number of orders.

As of late, information mining has been utilized generally in the zones of science and designing, bioinformatics, hereditary qualities and prescription. It is a gathering of algorithmic approaches to separate instructive examples from crude information and considered for instance in medicinal services. It assumes an essential job in handling the information over-burden in medicinal informatics. The invention of information mining provides a client-oriented technique to deal with innovative and concealed patterns in the information, and its applications may be developed to evaluate the efficacy of therapeutic drugs. The information created by social insurance exchanges is colossal. This restorative information about substantial patient populace is broke down to perform therapeutic research [2].

With the improvement of data innovation, broad restorative information is accessible. Restorative information grouping assumes a significant job in numerous therapeutic applications. It is the way toward changing depictions of therapeutic findings and strategies to general codes. Finding codes are utilized to follow infections and other wellbeing conditions, even unending ailments, for example, Heart Disease mellitus and coronary disease. Restorative order is broadly utilized in doctor's facilities for the measurable investigation of sicknesses and treatments. It tends to the issues of determination, examination and showing purposes in prescription. Therapeutic information has gained an extraordinary ground over the previous decades in the advancement and utilization of arrangement calculations. In medicinal services, these restorative information can be changed into accumulations, to compute normal qualities per patient and contrast and different qualities, to amass information into groups of comparable information and so forth., However, few difficulties incorporate information mining technique that are client connection, execution and adaptability.

In Data Mining, shrewd strategies are connected so as to separate information designs and are an enormous chance to help doctors manage this expansive measure of information. Therapeutic information mining has been connected to exact order and fast expectation for visualization and conclusion of patients in a particular restorative zone. The arrangement controls in the restorative information mining incorporate the utilization of the different classifiers on the informational indexes. The development of restorative databases is high. This quick development is the

primary inspiration for scientists to mine helpful data from these medicinal databases. As the volume of put away information builds, information mining procedures assume a critical job in discovering designs and extricating learning to give better patient consideration and compelling symptomatic abilities. Information mining strategies shape a gathering of heterogeneous apparatuses and systems and are utilized for various purposes. These processes and tactics rely on factual approaches, perception, machine learning, etc., and it very well may be related to process information in order to discover hidden instances and provide medical professionals with an additional source of learning for making decisions [3].

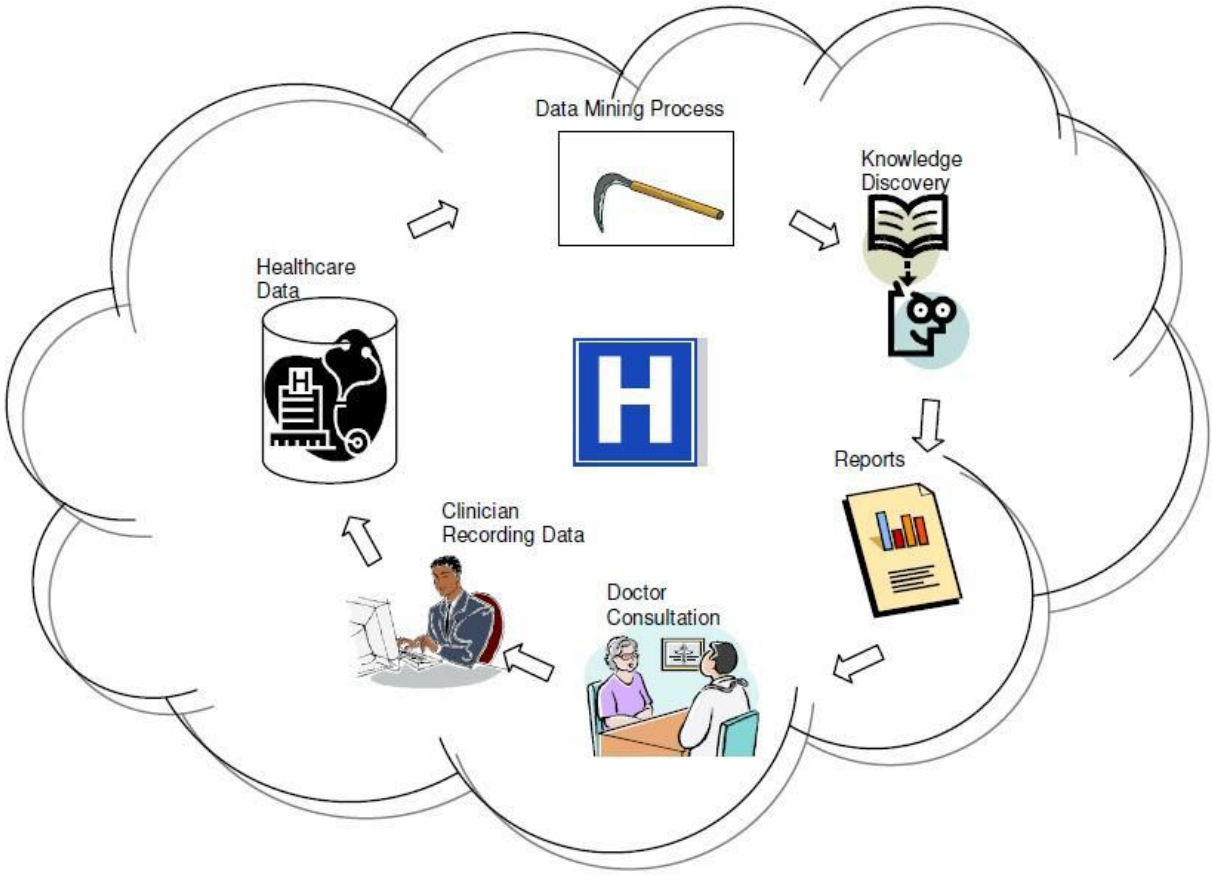


Figure 2.2 Mining of data for use in medicine and the design of health care systems [3]

A human service is responsible for managing the resources, devices, and procedures necessary for the perfect storage, recovery, and usage of data in the fields of health and biomedicine. Essential social insurance is another way to deal with wellbeing classification which incorporates at the network level every one of the elements required for enhancing the wellbeing status of the populace. The act of essential social insurance includes a decent arrangement of

deprofessionalizaion of medication. Laymen have come to assume a noticeable job in the conveyance of human services. The state of pharmaceutical services is still one of "data richness" but "learning poverty." Inside all the many systems that provide medical care, there is a wealth of information that may be accessed. As appeared in Figure 2.2, information mining has potential applications in a few fields, not the minimum of which is Health Care. Medicinal services associations are expanding uses on data innovation to enhance wellbeing, quality and access. All Health care associations hold nitty gritty and extensive records of patient information and these tremendous measures of information are put away as information stores organizing the information into segments and lines [4].

The reasons why human services will profit by the uses of information mining are:

- Health insurance agencies endeavor to diminish cash misfortune because of extortion by utilizing information mining strategies
- Huge measures of information are produced amid social insurance exchanges
- Data mining enhances basic leadership since it is utilized for the disclosure of patterns and examples in a lot of various sorts of information

This is because of the way that social insurance associations that perform information mining have better expectations about their mid and long hauling prerequisites. Calculations can naturally order the information dependent on likenesses of guidelines and examples got between the preparations on the testing informational index. They clarify the different prescient information mining procedures used to achieve the objectives and the strategies for looking at the execution of every one of the systems.

For the purpose of analysis, a variety of tests need to be carried out. However, via the usage of data mining techniques, it is possible to lessen the number of tests, and this reduced test set plays a significant part in both the timing and the completion of the project. Basic sicknesses, for example, type 2 Heart Disease and coronary disease result from a perplexing transaction of hereditary and natural variables. Coronary disease will establish a standout amongst the most essential human services issues worldwide throughout the following five to ten years. The pervasiveness of sort 2 Heart Disease and the weight of sickness caused by it have expanded quickly around the world. Ongoing examinations on the forecast of regular ailments depend on numerous hereditary variations alone or notwithstanding customary sickness hazard factors. Hazard models streamlined for Heart Disease are of uncommon significance, as type2 diabetic

patients have a two to four time higher cardio vascular infection chance than the non diabetic populace. Coronary disease and stroke are the primary driver of death and incapacity among individuals with sort 2 Heart Disease. Truth be told, no less than 65 percent of individuals with Heart Disease bite the dust from some type of coronary disease or stroke.

The term "information mining" refers to the process of analysing raw data with the aid of a personal computer in order to determine which aspects of the data are most relevant. Today's social insurance groups are well-equipped to generate and collect a large quantity of information. Due to the increase in the amount of information, a predetermined path must be specified for the extraction of data when it is necessary. The human services sector is loaded with data, and information mining is increasingly becoming a need, both of which have led to an increased interest in the use of data mining techniques in the social insurance industry in recent years. Information mining in restorative research begins with hypothesis, and the findings are balanced according to the requirements of the study. Digging out information is of utmost relevance in the field of medication, and it is a component of the whole procedure that calls for an in-depth knowledge of the requirements imposed by the social insurance organisations. Information mining will reach its full potential in the reveal of knowledge hidden in the medical information as a result of the continued development of data communication technologies in the future [2].

That restorative information mining has tremendous potential for researching the hidden instances in the informative indexes of the therapeutic field, so you should definitely give it a go. These cases are suitable for clinical examination and evaluation. The invention of information mining provides a client-organized technique to deal with innovative and hidden patterns in the data. The completion of therapeutic work is regarded as an important but challenging task that must be carried out in an accurate and competent manner [3].

Therapeutic science industry has colossal measure of information and propelled information mining order methods that have been broadly connected in the field of medicinal databases, especially in coronary disease forecast, and they have picked up a great deal of achievement.

2.3Naive Bayes Approach

In the first experiment, a process called the displayed Naive Bayes information mining classifier was used. This procedure generates an optimal forecast show by making use of the least amount of preparation set in order to predict the likelihood of a diabetic patient developing cardiovascular disease. The identification of diseases is an essential part of the work done in the medical sector. The suggested model makes use of diabetic's conclusion to make predictions about characteristics such as age, sex, blood pressure, and glucose, in addition to the likelihood of a diabetic patient developing cardiovascular disease. It is important to note that the features that are being used in our suggested technique are those that are being used for the detection of heart disease, and they are not immediate indicators of coronary disease.

2.3.1 Data set and Parameters used in Naïve Bayes

The risk of coronary disease increases with factors such as family history, smoking, poor eating habits, high blood pressure, high cholesterol levels, and being overweight or obese. The coronary disease dataset maintained by the Cleveland Clinic Foundation is available for use in calculating the accuracy rate in India. In any case, in order to carry out the tests for the current investigation, the records of around 500 diabetes patients who had been treated at the Seshiah Diabetic Research Institute in Chennai, India, were first obtained. The specific clinical information collection provides clear and concise definitions for a variety of topics related to heart disease. Table 2.1 presents the heart disease characteristics that were included in our naive Bayes analysis, along with their respective descriptions.

Table 2.1 HeartDisease attributes used in our Naïve Bayes experiments

Attribute	Description
Sex	A classification of the sex of the person
Age	Age of the patient
Family heredity	Previous history (Father / Mother)
Weight	Patient's weight
BP	Blood pressure
Fasting	Sugar level after fasting
PP	Post prandial blood glucose level
A1C	HbA1c level Glycosylated last 4 months sugar level
LP Tot cholesterol	Total cholesterol level

2.3.2 Data cleansing and selection are also included

Every computation has to accommodate the data in some kind of specified structure. The transformation of unstructured data into a form that can be reasonably organised by a computer is known as preprocessing. The stage of information arrangement encompasses all of the activities that are performed to produce the final dataset from the raw underlying information. These rudimentary records may be saved in a few different forms, such as content, surpass expectations, or other types of database records. After that, the raw data is transformed into informative collections that possess a number of attributes that are considered to be appropriate.

There are various factors that contribute to the fact that many datasets are lacking qualities. The crude data, for the most part, include a great deal of noise, which may be defined as a mistake made at random or a shift in the value of an intentional variable. It is not possible to use it especially to process, with the computations involved in machine learning. A connection may be made between information cleaning and the removal of noise and the correction of inconsistencies in the information. Its schedules make an effort to fill in any missing characteristics, calm down any agitation while still identifying any exceptions, and correct any anomalies in the material.

In order to prevent the development of confusing or inappropriate recommendations or examples, it is vital to clean and sort the information that will be used in the calculation for information mining. Altering the information so that it is more compatible with the mining process is something that has to be done.

Information joining is the process of combining data from several sources into a coherent information storage. This may be thought of as being analogous to a data warehouse or an information 3D square. A thorough and careful reconciliation of the information obtained from various sources is helpful in preventing and reducing instances of redundancy as well as anomalies in the following gathering of information. This contributes to improving the accuracy and speed of the mining process that ultimately takes place. Information reduction may lessen the amount of information required by gathering and removing unnecessary highlights. When using the techniques for information mining, the focus is placed on certain fields that allow for study of the information. This is accomplished by selecting and sorting a few fields as information, yield fields, and prescient fields.

All of the attributes that are used in Naive Bayes tests that are described in Table 2.1 have numeric qualities, with the exception of sex and familial heredity. When referring to males or females, the quality sex is contrasted with the characteristics "M" or "F," respectively. The trait of familial heredity competes against attributes that may be categorised as "Father," "Mother," or "Both." In the case that the patient does not have a history of heart disease, this attribute will be left empty. Because the mining computation requires that no property estimate be left blank for it to function correctly, the value "No" may be used for patients who had no history of heart disease in the past. In a similar fashion, there must be an absolute quality reliant on which the informative collections may be described in order to fulfil this need.

The purpose of the study being presented here is to determine the likelihood that a diabetic patient would develop coronary disease. Following that, the 'LP Tot Y/N' attribute was selected to serve as the defining feature of the class. Because the 'LP Tot Y/N' property is a number feature, the trait esteems have been categorised as either having an increased cholesterol esteem ('Yes') or a low cholesterol esteem ('No'). Within the framework of the information analysis mode, all of the distinctive choice modules that are pertinent to the data have been studied with

the purpose of accumulating an optimal subset of credits to forecast the risk factors of diabetic patients developing a variety of heart diseases.

2.4 Support Vector Machine Approach

In the second test in the current study, Support vector machine information mining classifier system has been applied with spiral premise work section to assess weakness of diabetic patients to heart disease. The great part of these frameworks have successfully employed SVM for the order rationale. On the demonstration of this, SVM classifier has been employed in the trials that figure in the present study. The repercussions of the suggested framework are quite large. The framework exhibits great accuracy in predicting the defenselessness of diabetic individuals to heart illnesses.

2.4.1 Data set and Parameters used in SVM

The method that is being shown here is one for determining the susceptibility of diabetic persons to heart diseases, and it is being applied to data of around 500 patients who have Heart Disease. With the exception of the powerlessness quality, all of the property tasks and Heart Disease characteristics employed in our SVM-suggested framework are very standard, save for the powerlessness quality.

The other 358 patients were determined to have a lower level of vulnerability to coronary disease. There were 500 records total, and 142 of them classified individuals as having a very high level of heart disease vulnerability. Because SVM generates only numeric qualities, the apparent attributes were converted into numeric characteristics by exchanging each incentive for a unique whole number. This was done because SVM generates only numeric qualities. For instance, the default sex values have been altered to read as follows: male = 1, and female = 0. Following this step, the estimates of the ascribes are scaled to fall somewhere between 0 and 1. After that, the SVM classifier was provided these records as a contribution, and the operation of calculating the help vector was scrutinised. Following that, this SVM model has the potential to be proposed for the order of the diabetes dataset.

2.5 Approaches Based on Support Vector Machines and Decision Trees

In the third trial of the current study, it is intended to determine whether method, help vector machine or choice tree acceptance, is more accurate in predicting the risk of coronary disease in diabetes individuals.

2.5.1 The data format that is used in support vector machines and decision trees

Here, the medical histories of approximately one thousand patients diagnosed with heart disease have been compiled, and the philosophy that has been portrayed is that the characteristics of the diabetic properties that make up each record are presented in Table 2.2. Aside from the hazard class characteristic, all of the trait jobs are standard.

Table 2.2HeartDiseaseattributes used in our SVM and Decision Tree experiments.

Role	Attribute Name	Attribute Type
Label	Risk Class	Binominal
Regular	Sex	Binominal
Regular	Age	Integer
Regular	Family/Heredity	Polynomial
Regular	Smoking	Numeric
Regular	BP	Polynomial

A mapping of occurrences between distinct classes or gathers is what we mean when we talk about an order display. It accurately forecasts the names of certain classes. In addition to this, it organises the information into groups according to the preparation set and the qualities that are used in defining the characteristics, and then it uses these groups to sort the new information. The models are structured on the basis of support vector machines and decision tree enlistment, and the goal is to properly predict the objective class for each example included in the data.

2.6 Naive Bayes, Support Vector Machine and Decision Tree Approach

In the last experimentation, a near examination on the classifiers which can arrange the danger of diabetic patients getting coronary disease from a machine learning viewpoint has been given. It intends to assess and look at utilizing three changed information mining grouping methods, for example, Naïve Bayes, Support Vector Machine and Decision Tree to decide the conceivable approaches to foresee the danger of coronary disease for diabetic patients dependent on their prescient precision.

2.6.1 Data structure used in Naïve Bayes, SVM and Decision Tree

Additionally, in this study, the medical histories of approximately one thousand patients diagnosed with heart disease were analysed and investigated with the assistance of three distinct information mining order systems. The findings were evaluated in terms of their affectability, particularity, F-score, and exactness.

In the chapter on Statistics, the terms "normal" and "standard deviation," which are represented by the symbols (+/-), are used to talk about numeric and numerical properties. The polynomial and apparent characteristics are referred to as the minimum and median qualities, respectively. The'mode' is the appearance that is the most severe, and the'slightest' is the appearance of information that is the most fundamental. The traits that are absent are the ones that are obvious, whereas the qualities that are present are the ones that have information.

Using stratified ten times cross-approval is the typical method for predicting the error rate of a learning system given an established example of knowledge. This method was developed in the 1960s. After extensive testing on a variety of unique datasets using a variety of learning methods, it has been shown that 10 folds is about the optimal number of folds to produce the best measure of error. The information is divided into preparation and testing information using 10-fold cross approval to evaluate the precision of our learning model so that we can quantify the consistency of the proposed model. This allows us to determine how reliable the suggested model is. In light of this circumstance, we will divide the dataset into ten distinct parts, and then train and test each segment separately.

In this investigation, three well-known information mining classification processes, namely the Naive Bayes hypothesis, Decision tree acceptance, and Support vector machine, are linked with one another and compared with one another based on how accurately they predict the future. Because of its learning administrators and administrator structure, which enable arrangement of almost self-assertive procedures, Powerless and Rapid excavator has been used as a device for assessing and contrasting various grouping strategies and given patient datasets. This has been done in order to better understand how to categorise patients.

In this section, a disease dataset is presented, together with information on pertinent topics such machine learning, its methods along with short descriptions, data preparation, performance evaluation measures, and data preprocessing. A area of artificial intelligence known as machine learning (ML) comprises the creation of algorithms that are capable of learning from their prior experiences. In order to work properly, machine learning algorithms must first search the input dataset for previously undetected patterns before utilising that knowledge to build models. After that, they will be able to provide accurate predictions for newly collected datasets that are wholly foreign to the algorithms. Learning allowed the computer to improve its level of intelligence in this way. It can now recognise patterns that would be very challenging or perhaps impossible for humans to discover on their own in the first place. Algorithms and methodologies used in machine learning are able to work on vast datasets and generate decisions and predictions using that data. A simplified representation of the many different types of machine learning algorithms is shown in Figure 1.

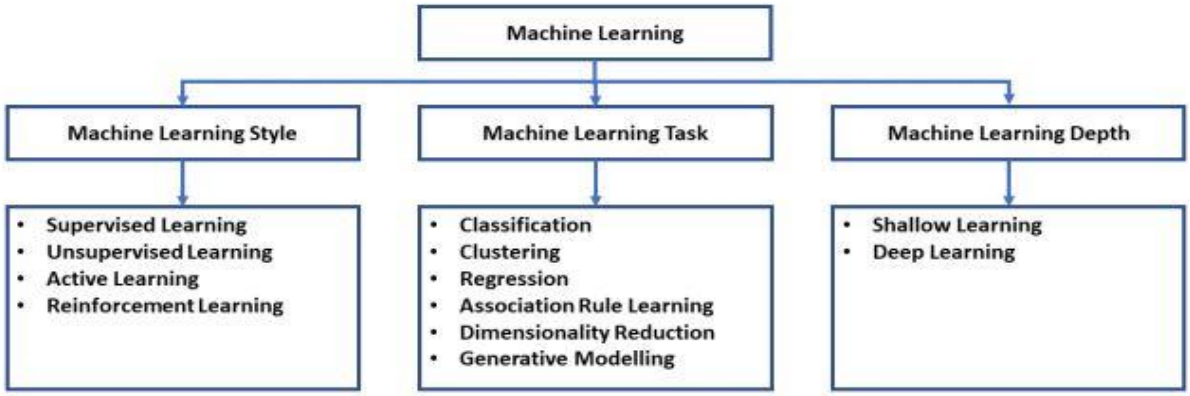


Figure 2.3. Machine Learning [6]

In this study, a Cleveland heart-disease dataset was one-hot encoded, standardized, and cleaned. Key heart disease traits, such as male gender, have been discovered, and additional study is planned. The researchers identified the most significant features in this work by displaying Random Forest classifier permutation feature significance and partial importance plots. These characteristics included reversible thallium stress test abnormalities, exercise-induced ST depression, and asymptomatic chest discomfort. [13]

During this period of rapid technical advancement, time, energy, and spirit are being put into the construction of an intelligent healthcare system. It's possible that the development of an automated system for forecasting cardiac disease may mark a turning point in the field of medical research. This study uses a dataset obtained from the machine learning repository at UC Irvine in order to evaluate the efficacy of a strategy for predicting cardiovascular disease. Typical machine learning algorithms benefit from feature selection since it increases their performance [14]. The Random Forest method in conjunction with PCA achieves a classification accuracy of 92.85 percent for heart illness. By incorporating actual data from both healthy and sick patients into the process of training the models, the purpose of this initiative is to improve the level of confidence and accuracy with which medical professionals make diagnoses. When it comes to patient classification, having several machine learning models is preferable than having just one. According to the findings of the hard voting ensemble model, our method attained an accuracy of 90 percent in the final outcomes. [15]

Researchers say model heart rate data might be used to predict a person's heart illness a year from now. Smart Mirrors, Smart Mice, Smart Phones, and Intelligent Chairs collect this information. Internet-based heart rate data was kept on a server for analysis and interpretation. This approach gathers a year's worth of data to create reliable future forecasts. Receiving a one-year prognosis of oncoming heart disease may increase individuals' and the community's understanding of heart disease. As projected, the approach will minimize heart disease patients and deaths in the future. [16]

This study links the data sets to medical parameters. Much information was extracted after running the datasets via the Python-created ML Algorithm and Random Forest Algorithm. Using current patient data, this technique may predict the creation of new ones early, minimizing the likelihood of life-threatening complications and mortality. The Random Forest algorithm is used

to create and implement a reliable cardiac disease prediction system. The application displays a CSV file containing patient record data. After accessing the dataset, the programme produces a heart attack level. The recommended strategy offers the following benefits: Software performance and accuracy haven't been criticized. It's also very adaptable, allowing for long-term success. [17]

Researchers use machine learning to identify and forecast sickness. Predicting cardiac illness based on available clinical data is a major issue for academics today. Combining clinical data with machine learning algorithms has produced state-of-the-art results, but there's potential for improvement. This research proposes using machine learning techniques to forecast cardiac disease. Next, we'll examine algorithms. Aims: This study uses correlational analysis to discover factors that may improve prediction results. Our findings are compared to a previously published study using the vascular heart disease dataset from UC Irvine. Our model showed an accuracy of 86.94%, which was higher than the previously reported Hoeffding tree technique's 85.43%. [18]

Several machine learning algorithms investigate and predict chronic diseases, including heart disease, cancer, and Alzheimer's. Data scientists and medics use machine learning to save lives and reduce population mortality. Data scientists may conduct predictive research on big data to do early analysis on the cardiac syndrome to save the lives of individuals suffering from the ailment, as indicated in the title. Many factors must be assessed when completing AN analysis and predicting cardiac disease. The author of this case study used machine learning to analyze the syndrome's performance in the field and confirm the predicted statistics. [19]

Increased alcohol use, cigarette smoking, and lack of exercise are three reasons heart disease has become a public health issue. Obesity and diabetes also contribute. This article explains how machine learning algorithms may detect cardiac disease. Both training and testing employ a data-gathering method that considers human health aspects. AI and machine learning are used to predict cardiovascular diseases. After deploying and testing the machine learning algorithm, its overall performance may be assessed. [20]

In the next sections of this article, we will demonstrate how both of the machine learning models were put to use in addressing the heart disease detection difficulties that were covered in a previous post. After the algorithms were developed and the parameters were fine-tuned, a number of tests were carried out to ensure that both the algorithms and the parameters were

operating as expected. The purpose of this research is to examine the accuracy of the predictions that are provided by the two models while using different parameter value combinations. The Cleveland database, which was collected from the learning dataset repository at the University of California, Irvine, was applied in order to diagnose heart disease (UCI learning dataset repository). It was shown that neural networks are superior than convolutional neural networks in terms of accuracy of prediction in the great majority of cases. [21]

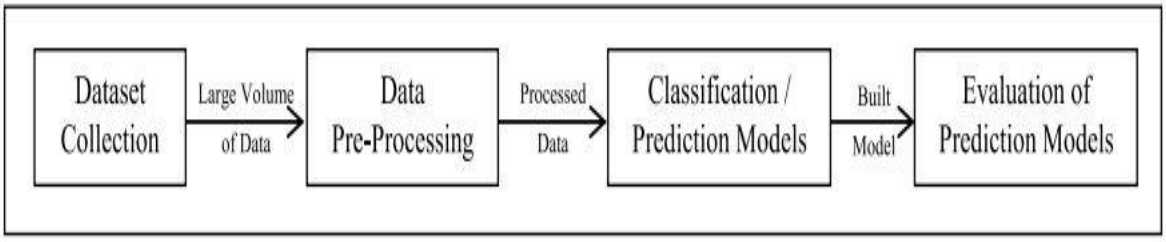


Figure 2.4. Steps of evaluation of prediction models

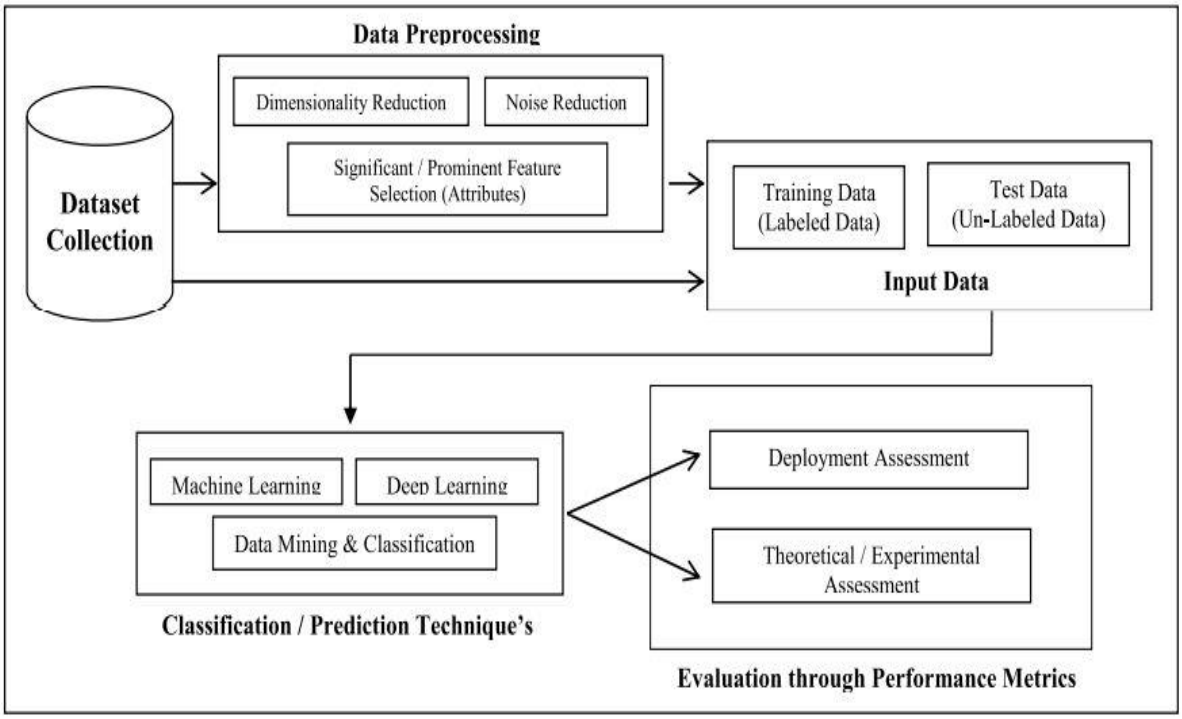


Figure 2.5. Steps of data collection to evaluation through performance metrics.

Mohammad and colleagues' [22] model predicts cardiovascular disease using several feature combinations. UCI's machine learning repository stores Cleveland database data. Decision Tree,

Logistic Regression, Support Vector Machine, Naive Bayes, and k-Nearest Neighbor were employed. These models were each tested on a distinct set of feature combinations. The suggested model predicts heart disease with 87.4% accuracy, according to the study.

Princy et al. [23] use decision trees, K nearest neighbours, logistic regression, naive Bayes, random forests and support vector machines to classify cardiovascular illness using machine learning (SVM). Data on cardiovascular disease was used to test the models [24]. The research results show that decision trees are more accurate than other classifiers. 78% of the time, the DT model accurately predicted the patient's status.

Prediction accuracy can be improved. Bashir et al [25] chose features to increase accuracy by testing various heart disease datasets. Logistic Regression, Decision Trees, Naive Bayes SVMs Classifiers include SVMs and RFs. Results demonstrate improved accuracy. Sabab et al. [26] found that data mining and feature selection may increase the accuracy of cardiac disease classification algorithms. The authors used the Naive Bayes model, the 87.8% accurate SVM model, and the 86.80% accurate C4.5 DT model to identify the cardiovascular disease (79.9 percent).

Zunaidi et al. [27] found that Multilayer Perceptron Neural Networks are best for early heart disease detection. [27] Needs citation 85.3% accuracy was achievable. Khemphila et al. [28] used MLP to identify the cardiac disease with 80.17 percent accuracy.

Random forests, decision trees, and hybrid models may predict cardiac disease. This investigation employed Cleveland heart disease data. Princy et al [30] employed Nave Bayes, ID3, and Neural Network to categorize heart disease risk. The authors observed that the more features a model includes, the more accurate it is. This inquiry obtained 80.6% accuracy. Gawali et al. [31] used Nave Bayes and K-Means to quantify heart disease risk. In this study, accuracy was 84%.

Maiga and colleagues [32] provide four machine learning techniques for cardiology. These algorithms employ cholesterol, BMI, and blood pressure as inputs. The recommended experiment uses Random Forest, KNN, and Naive Bayes algorithms. The most accurate random forest classifier predictions are 73% accuracy, 65% sensitivity, 80% specificity. Mohan et al. [33] use machine learning to uncover heart disease-related data. Hybrid HRFLM combines LM

and RF feature sets. 88.7% accuracy was achieved. Haq et al. [34] developed a strategy for diagnosing cardiac disease using machine learning. LASSO, mRMR, and LASSO feature selection algorithms were employed in the study. The suggested technique achieves 89% accuracy in 10-fold logistic regression cross-validation.

Using the Z Alizadeh Sani dataset, a neural network method was enhanced to identify coronary artery disease. Generic weight initialization increased network speed by 10%. This made the model 93.58 percent correct. Latha et al. [36] present a new strategy for improving classification model performance. "Ensembling characteristics" improve model predictions. Ensembling improved week classifier accuracy by 7%.

2.7 Literature Review

In their paper titled "Choice Support in Heart Disease Prediction System utilising Gullible Bayes," Dhanashree Medhekar and colleagues present the classifier method for the detection of coronary disease. Additionally, they demonstrate how gullible Bayes can be utilised for the purpose of characterization reason. Within this framework, they have categorised information about restorative practises into five distinct categories, namely No, Low, Average, High, and Very high. As a consequence of this, two basic capabilities, namely grouping (also known as preparation) and expectation (also known as testing), were carried out. The Cleveland centre setup served as the source for this information, which was acquired using 14 characteristics [37].

A. Mert, N. Kılıc, and A. Akan. When it came to the diagnosis of coronary disease, the sizeable dataset under consideration was put to use; all things considered, the RF classifier demonstrated better performance versus C4.5. During this interim period, the mix system was implemented in [38], which

J. Zhu, L. He, and Z. Gao are the three names. NN is a revolutionary classifier that is widely used in many different areas of medicinal research and learning methodologies. The neural system based on the k-nearest neighbour. Fuzzy principles Other approaches besides the support vector machine RF, C4.5, GMM, Clustering Classifiers of the current state of the workmanship are shown in figure 3. 4 Research and Practice in Cardiology because it is fundamental and not

very difficult to put into practise. NN is dependent on the structure and components of natural neural systems that are concerned with neurons. This computation records a response in a manner that is comparable to how the human mind operates. Through the use of writing, we were able to see that NN is successfully carried out in the presence of circulatory abnormalities [39].

H. R. Marateb and S.Goudarzi. In contrast to tactics that had been cited in the past, discretization approaches were presented as examples for the parameter interims. For the purpose of characterisation of coronary conduit disease parameters, discretization is the process of separating the constant parameter from among the many variables in a study. Nonetheless, this method had a direct impact on the execution of the arrangement procedure, which is often used for the analysis of data in [40].

S. Faziludeen and P. Sankaran. Have shown that the KNN classifier achieves a higher accuracy estimate than the SVM classifier when it comes to determining the location of cardiac peculiarities. Accordingly, KNN and SVM for highlight determination approach with improved F-score estimate, although during this procedure, the conclusion proclaimed that KNN figures better element choosing when contrasted with SVM [41].

Dinesh Kumar G et al. This project presents a prediction model that can determine whether or not a person has cardiac disease and may also give awareness or a diagnosis about that condition. To accomplish this, a comparison is made between the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier, and logistic regression on the dataset taken in a region in order to present an accurate model of predicting cardiovascular disease [42].

In order to increase the overall performance of models, Kanika Pahwa and colleagues suggested selecting characteristics first before moving on to classification. The SVM-RFE and gain ratio algorithms are used to the dataset for the purpose of feature selection. The end result is an assignment of weight to each feature. This strategy helps to enhance accuracy while also reducing the amount of time spent computing. The suggested method of picking features, as shown by the results of the experiments, leads in an improvement in accuracy for both models. [43].

A method for diagnosing coronary artery disease (CAD) using PCG and PPG data was developed by Banerjee et al. The authors first extracted unusual features from the signals, and then they used a support vector machine (SVM) classifier to determine which among the patients had CAD. After further investigation, it was discovered that their suggested method achieved an accuracy of categorization equal to eighty percent when using the SVM classifier [44].

Paradkar et al. made use of an SVM-based classifier in order to determine whether or not PPG signals were present in CAD patients. Important highlights are extracted from the morphological structure of PPG signals using the proposed procedure, which is then used in the administered learning process. SVM was given a better rating than other methods in terms of affectability and particularity [45].

The neural system classifier that Kim and Kang et al. devised was for the purpose of early identifying coronary supply channel illness hazard prediction from the dataset. However, the study discovered that their recommended technique using the neural system is better to everything that the Framingham risk score [46] has to provide.

Priyanka et al. Coronary artery disease is the leading cause of mortality worldwide, accounting for more deaths than any other illness or disease combined. Every year, there is a steady increase in the number of men and women who are diagnosed with coronary disease. Because of this, its diagnosis and therapy are started immediately. The anticipation of coronary disease might at times be a problem since there are insufficient resources available in the area of therapeutics. It is possible that the use of suitable technology support in such a way might wind up being very beneficial to both the medical community and patients. The provision of Data mining solutions is all that is required to resolve this problem. In this study, we plan to use both the Naive Bayes model and the Decision Tree model, which are both information gathering approaches for producing an accurate prognosis of heart disease. In order to determine which of the two systems is superior in terms of productivity and accuracy, it considers both of them [47].

2.5 Chapter Summary

In this section, a variety of audits and specialist papers on information mining arrangement systems that are tied to social insurance databases are discussed. The mining of information may result in a learning-rich environment, which, in turn, would make it possible to fundamentally improve the quality of healthcare decisions. In this instance, data restoration and information mining are related to one another through the databases. The development of tools such as Rapid Miner has made it possible to employ this technique with relative ease with large datasets including a large number of features. In addition to this, it makes intelligent predictions and organises the information in a logical manner.

Using stratified ten times cross-approval is the typical method for predicting the error rate of a learning system given an established example of knowledge. This method was developed in the 1960s. It has been shown via extensive testing on a variety of different datasets and through a variety of different learning techniques that 10 folds is about the optimal number of folds to acquire the best measure of error. The information is separated into preparation and testing information using 10-overlap cross approval so that the accuracy of our learning model can be evaluated. This is done so that the strength of the suggested model can be quantified. In light of this circumstance, we will divide the dataset into ten distinct parts, and then train and test each segment separately.

In this investigation, three common information mining grouping methods, namely the Naive Bayes hypothesis, Decision tree acceptance, and Support vector machine, are connected and contrasted with each other based on their predictive accuracy. These methods include the decision tree acceptance, the Naive Bayes hypothesis, and the support vector machine. Because of its learning administrators and administrator system, R and Rapid mineworker has been used as a tool for comparing and contrasting a variety of order methods and patient datasets. This is possible thanks to the fact that these two programmes enable the arrangement of almost subjective procedures.

CHAPTER 3

PROBLEM STATEMENT AND PROPOSED METHOD

PROBLEM STATEMENT AND PROPOSED METHOD

This chapter looks at the issue proclamation, the proposed strategy, and the related data set. There are many information mining systems currently in use in the human services industry; however, the examination that needs to be done is on the execution of the various order procedures, in order to enable the choice of which of them is the best.

3.1 Problem Statement

The machine learning approach determines if a large amount of information results in the discovery of new data pertaining to instances or tenets. In the field of information mining, one of the most important jobs is disease forecasting. In order to diagnose a patient with a disease, it is necessary to conduct a number of different tests on them. The use of information mining technologies, on the other hand, has the potential to lessen the number of tests required. This reduced sample size of the test now plays an important role in both timing and execution. The mining of data pertaining to coronary disease is very important since it helps experts to determine which aspects or characteristics are more important for determination. Some examples of these include age, weight, and other similar factors. Because of this, the doctors will be able to perform an even more accurate analysis of heart disease. The human services business makes use of a variety of information mining techniques; nevertheless, there is still research that has to be conducted on the efficacy of the various organisational approaches, so that the industry may choose the information mining approach that provides the greatest results. The examination that is outlined in this postulation is intended to answer the test of boosting the expectation model to anticipate cardiovascular disease in diabetic people and offering a positive reply when predicting the illness. In a flash, the fundamental research abilities are represented as such in this manner.

- How may various information mining techniques be used in the medical services business, and how can their expected execution be recognised?

- How exactly does the use of grouping systems contribute to the formation of an expectation display in order to properly forecast the risk of cardiovascular disease in diabetic patients?

3.2. Data Source

The Cleveland Heart Disease database provides the source for the acquisition of a dataset including information attributes. The expectations of a heart attack with essential instances can no longer be met thanks to the support provided by the record established. The attribute "Conclusion" with a value of "1" is recognised as a Heart Disease prediction, but the attribute "Conclusion" with a value of "0" is identified as their being no anticipation of Heart disease for patients. In this case, the "PatientId" property is the most important one, while the other attributes serve as sources of information.

1. Diagnosis (value 0: a diameter reduction of less than fifty percent, indicating the absence of heart illness; value 1: a diameter reduction of more than fifty percent, indicating the presence of heart disease) Predictable characteristic 1.

Essential quality

PatientId – The individual's unique identifier

The input characteristics

1. Sex (value 1: Male; value 0: Female)
2. Your age in years 3.
3. Decrease in the height of the old peak caused by a ST dip
4. In a state of rest - findings of the resting electrogram (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
6. Incline: the angle at which the peak workout ST segment is performed (value 1: unsloping; value 2: flat; value 3: down sloping)
7. Diagnostic procedure: exercise-induced angina (value 1: yes; value 0: no)
8. Serum Cholesterol (mg/dl)
9. Have your blood pressure checked (mm Hg on admission to the hospital)
10. The (value 3: normal; value 6: fixed defect; value 7: reversible defect)
11. CA represents the number of main vessels that can be seen via fluoroscopy (value 0-3)
12. The Latch - Achieving Your Highest Possible Heart Rate

13. Type of Chest Discomfort (with a value of 1 indicating typical type 1 angina, a value of 2 indicating typical type angina, a value of 3 indicating non-angina pain, and a value of 4 indicating asymptomatic)

3.3. Random Forest Method

There are two types of methods that fall under the umbrella term "supervised learning technique." It is also possible to use this technique to complete the classification challenge. The supplied enormous dataset is utilised to generate a large number of decision trees in a very efficient way. The average increase in the accuracy of the forecast is the formula for making the prediction. In our suggested technique, we are using this algorithm for the categorization of heart disease that has been widely publicised on Heart disease . Through the use of the random forest algorithm, the Heart Disease are divided into groups according to the patterns that can be seen in the Heart Disease. The algorithm that was used in this categorization may be found in the section below.

Input: Dataset (Heart Disease) for training (80%) and testing (20%)
Output: Classification result interns of Accuracy, Precision, Recall, and F-measure

begin
Preparation and standardisation of data; this is done for To complete the training data set
Calculation of characteristics

1. At the beginning of the provided training set, choose any K random data points to use as starting points.
2. For each of these K values, locate the decision trees that correspond to the dataset that was provided.
3. Following that, we choose the N-th number of trees that we wish to plant. Now, for the new datapoints.
4. We acquire the predictions from the decision trees that have been formed.
5. We allocate the specific points of data that were previously picked to each tree that has received the greatest number of votes.

Construct classifiers;
end

6. Make use of the worth of characteristics for each individual heart disease data points.
7. Should be completed for all entries in the testing data collection
8. Verify the correctness of the model;

Training and Testing at an end
end

3.4 SVM

It is another another classification strategy for the classification issue that we are now dealing with. It is as a consequence of this that decision boundaries are formed, which split the n-D area into subclasses. When selecting the next data points that will be valuable in the future, this

technique is used, and it has shown to be highly successful. A hyperplane is a term that refers to the decision boundaries that have been found in the field of mathematics and statistics. We have used the UCI dataset, and the SVM proved to be superior for text classification. We are now working on a similar project for Heart Disease see below. The algorithm is shown in the next section:

I/P: Dataset (Heart Disease) for training (80%) and testing (20%)
Output: Classified result intern of Accuracy, Precision, Recall, and F-measure

---

begin
Preparation and standardisation of data; this is done for To complete the training data set
Calculation of characteristics

1. The four functions of the kernel are implemented here by SVM, which stands for support vector machine.
2. Linear, polynomials, Sigmoid, and Radial Based Function (RBF) are examples of linear functions.
3. Construct classifiers.

end

4. Make use of the worth of characteristics for each individual heart disease data points.
5. Should be completed for all entries in the testing data collection.
6. Verify the correctness of the model,

Training and Testing at an end
end

3.5 Maximum Entropy

The maximum entropy classifier is capable of successfully classifying texts by using mutually dependent properties. Using this classifier, we are operating under the assumption that we should model everything that is known and make no assumptions about what is unknown. The classifier that optimises entropy is chosen from among all classifiers that are empirically compatible with a given set of training data in order to achieve this aim. We utilised the UCI dataset, and we found that the maximum entropy method worked best for text categorization. We are now working on the similar assignment for Heart Disease. The algorithm is shown in the next section:

I/P: Dataset (Heart Disease) for training (80%) and testing (20%)
Output: Classified result intern of Accuracy, Precision, Recall, and F-measure

begin
Preparation and standardisation of data; this is done for To complete the training data set
Calculation of characteristics

1. The link was effectively used via categorization.
2. Among emotive keywords, part of speech tags, and other similar terms
3. The use of negation in the study of sentiment

end

4. Make use of the worth of characteristics for each individual heart disease data points;
5. Should be completed for all entries in the testing data collection
6. Verify the correctness of the model;

Training and Testing at an end
end

3.6Flow Heart of Project Simulation

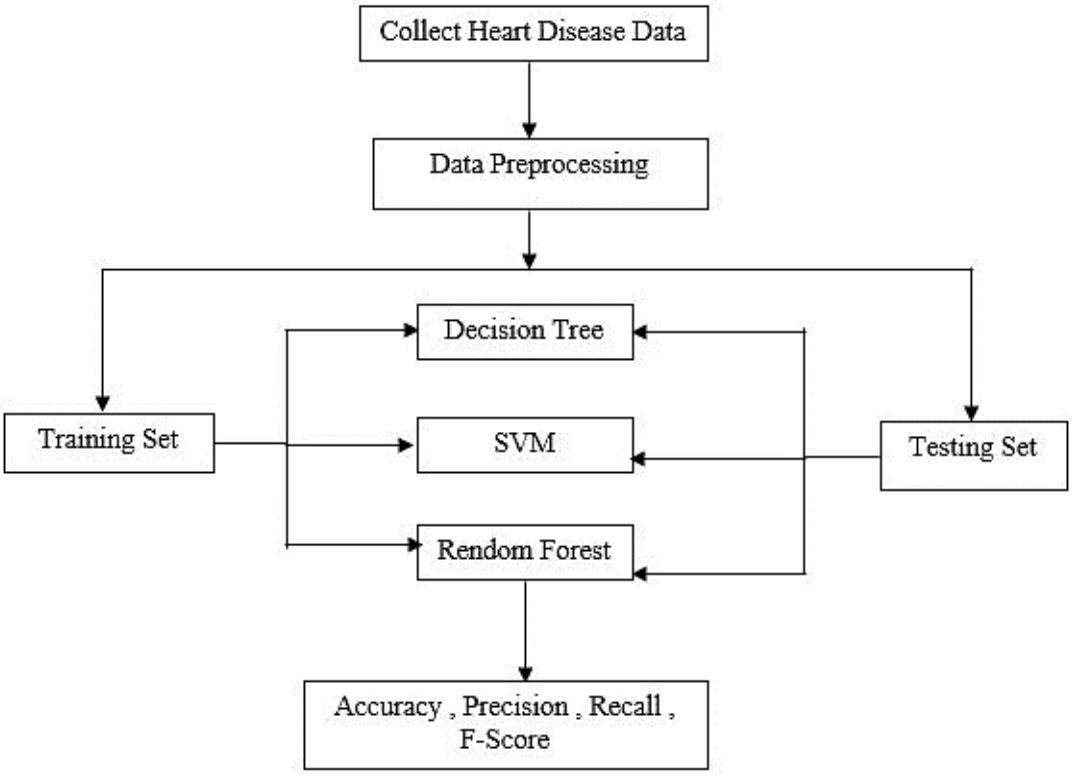


Figure 3.1 Flow Heart of Project Simulation

Figure 3.1 shows Flow Heart of Project Simulation, in which first of all we import data and then preprocess the data and apply Maximum Entropy, Random Forest, and Support Vector Machine Algorithm then compare and analyze both of results for getting accuracy.

3.7 Algorithm The overview of the Proposed System

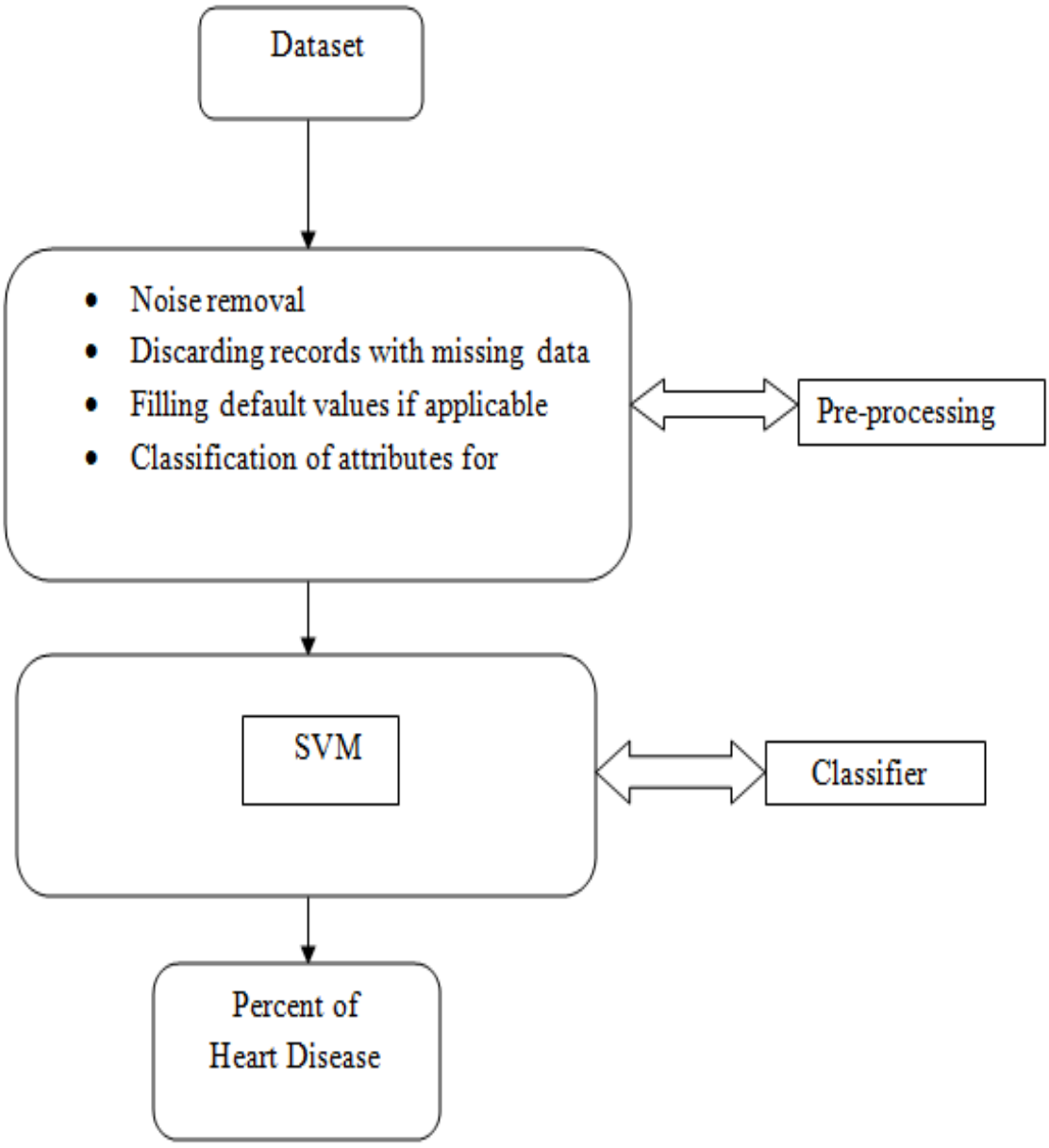


Figure 3.2. AlgorithmThe overview of the Proposed System.

Shows Figure 3.2 apply Maximum Entropy, Random Forest, and Support Vector Machine Algorithm on the given dataset then finding the percent of heart disease.

3.8 Overview of the Proposed System

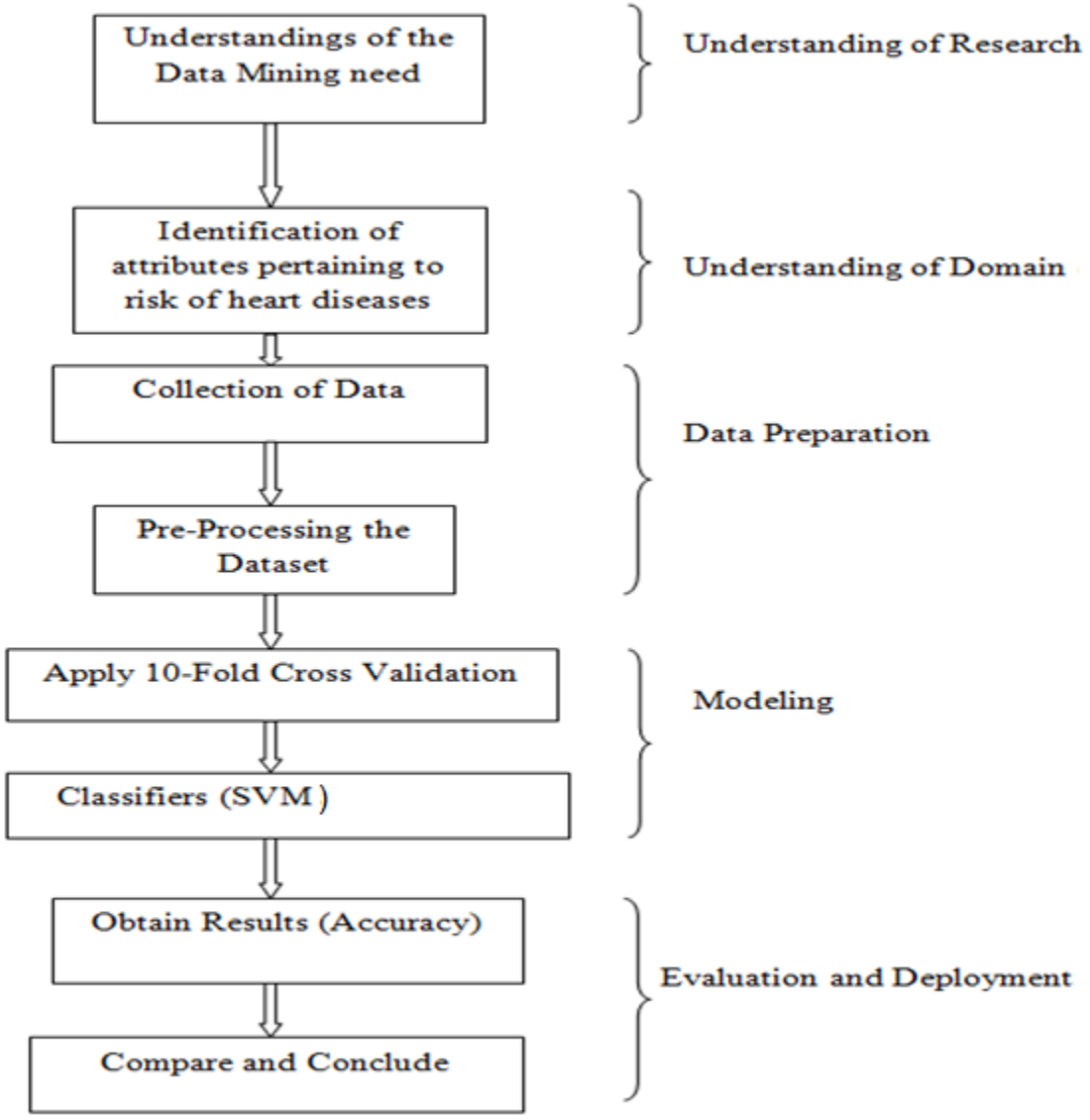


Figure 3.3 Overview of the Proposed System

Shows Figure 3.3 divided into five parts like Understanding of Research, Understanding of Domain, Data Preparation, Modeling and Evaluation and Deployment. First Understanding of the Data Mining Need, second Identification of attributes pertaining to risk of heart diseases , third collection of data and pre-processing the dataset , Fourth Apply 10-fold cross validation and classifiers (Maximum Entropy, Random Forest, and Support Vector Machine) and Five Obtain Results (Accuracy) and Compare and Conclude.

CHAPTER 4

IMPLEMENTATION AND RESULT

CHAPTER 4
IMPLEMENTATION AND RESULT

In this chapter, we discussed the implementation, and the results, which included using a total of 308 records from the Cleveland Heart database, each of which had 14 characteristics. [3] User inputs values for various medical parameters, such as gender, age, and so on. According to the value of this parameter, this model makes a prediction as to whether or not the patient has heart disease; in either case, the physicians will advise the patient to undergo more cardiac testing.

.

4.1 Hardware and Software :General. Brand. HP. Model.Display. Size. 15.60-inch. Resolution.Processor. Processor. Intel Core i5 8250U. Base Clock Speed. Memory. RAM. 8GB.Graphics. Graphics Processor. Intel HD Graphics 620. Storage. Hard disk. No. Connectivity. Wi-Fi standards supported. 802.11 ac. Operating system : windows 10. library like NumPy , SciPy , Scikit-learn, Theano, Pandas, Matplotlib , Plotly.

4.2 Dataset : The benchmark dataset of UCI Heart disease prediction was utilised for the purp ose of this research endeavour. This dataset includes 14 different factors that are associated with Heart Disease, including 'age,"sex,' 'chest pain type,"resting blood pressure,' 'cholestoral,' 'fasting blood sugar,"resting electrocardiographic,"maximum heart rate,' 'exercis

The data set was divided into two phases: training and testing; eighty percent of the data were us ed for training purposes, while twenty percent were used for testing.

4.3 Simulation

In this section, we have to explain the implementation of our work on google colab , and write a library related to implementation.

```
+ Code  + Text
%matplotlib inline

[ ]  from sklearn.metrics import classification_report,confusion_matrix,accuracy_score

[ ]  from sklearn.preprocessing import StandardScaler
     from sklearn.model_selection import RandomizedSearchCV,train_test_split

[ ]  from xgboost import XGBClassifier
     #rom catboost import CatBoostClassifier
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.svm import SVC

[ ]  data = pd.read_csv("heart.csv")
     data.head(6)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1

Figure 4.1 Dataset load and display

Figure 4.1 shows the dataset attributes link age, sex,cp,treatbps,chol,fbs,restecg, thalach,exang,oldpeak, slope,ca,thal,target.

```
sns.set_context('notebook',font_scale = 2.3)
data.drop('target', axis=1).corrwith(data.target).plot(kind='bar', grid=True, figsize=(20, 10),
                                                       title="Correlation with the target feature")
plt.tight_layout()
```

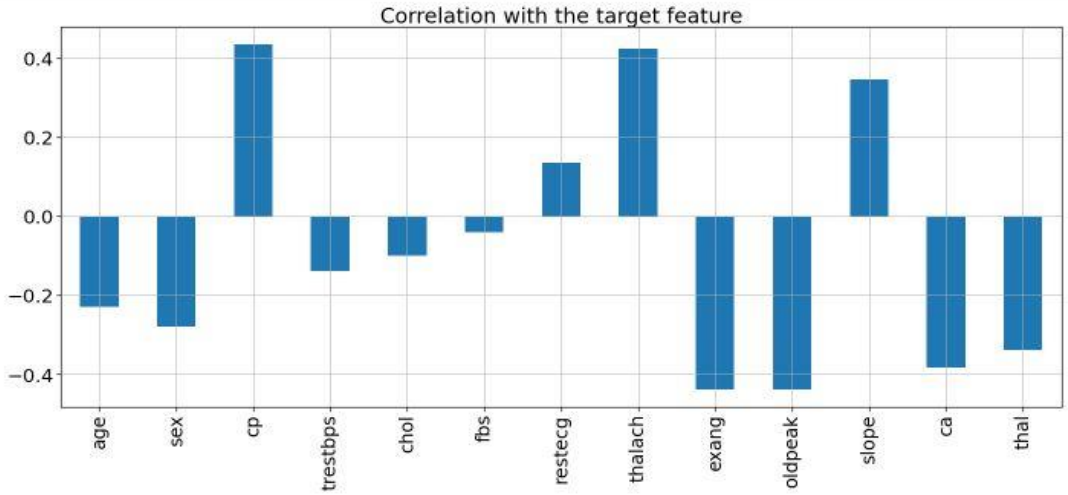


Figure 4.2 Correlation with the target feature

Figure 4.2 shows the correlation with the target feature of dataset attributes link age, sex,cp,treatbps,chol,fbs,restecg, thalach,exang,oldpeak, slope,ca,thal,target.

Figure 4.3 Tight_Layoutthe target feature

Figure 4.3 shows the Tight_Layout the target feature of dataset attributes link age, sex,cp,treatbps,chol,fbs,restecg, thalach,exang,oldpeak, slope,ca,thal,target

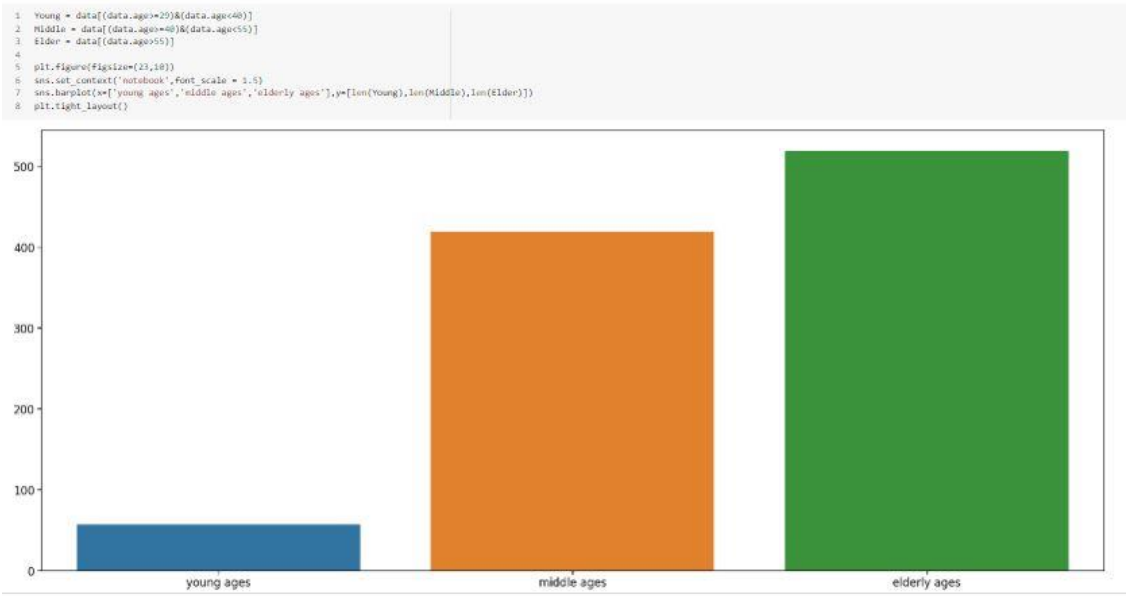


Figure 4.4 Tight_Layout the target feature of ‘young ages','middle ages','elderly ages

Figure 4.4 shows the Tight_Layout target feature of ‘young ages', 'middle ages', and 'elderly ages. Elderly ages peoples are maximum facing issues related to the heart.

4.4. Result

The execution of each approach is performed a certain number of intervals, and the outcomes are achieved by a range of grouping algorithms on a Heart disease dataset to get the final findings. On the basis of the numerous results that have been calculated, graphs have been constructed. The effectiveness of our proposed approach was evaluated using the characteristics stated below:

Table 4.1 Evaluate Metric with Contingency Table

		Prediction	
		Predicted Negative	Predicted Positive
Reality	**Actually Negative**	True Negative (TN)	False Positive (FP)
	Actually Positive	False Negative (FN)	True Positive (TP)

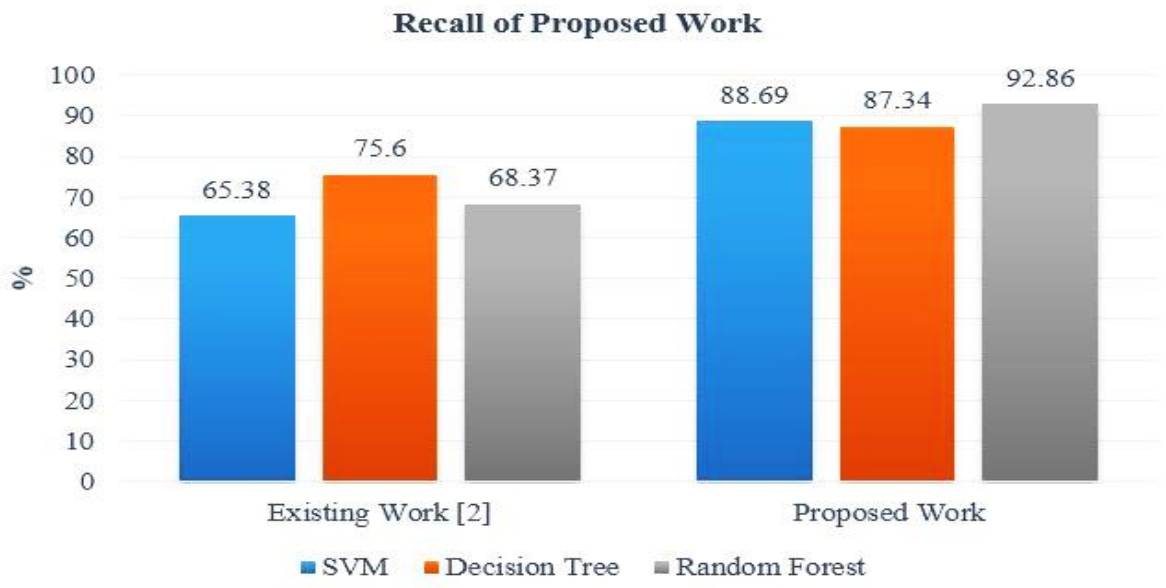
$$recall = \frac{tp}{tp+fn} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$precision = \frac{tp}{tp+fp} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Graph 4.1, graph 4.2, graph 4.3, and graph 4.4 are labeled x-direction are existing and proposed work and y-direction in terms of %.

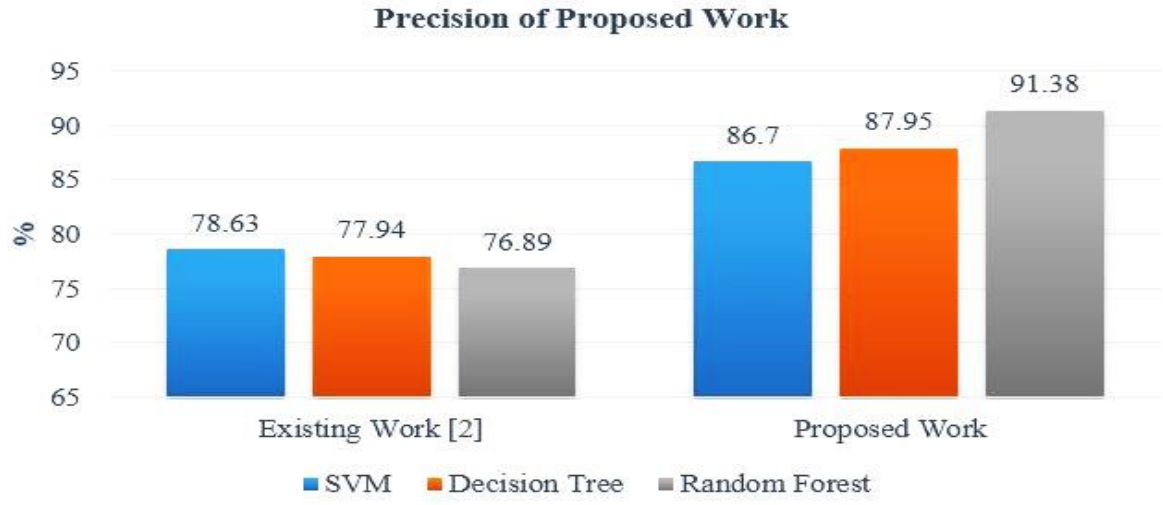
4.4.1 Recall: Recall is a statistic that quantifies the number of positive class expectations that are created from each and every positive model in the dataset, and it is used in machine learning. The recall is determined with the help of equation 1.



Graph 4.1 Recall graph between Existing Work and Proposed Work

For all approaches, including both known and new algorithms, the recall value is computed in Graph 4.1. In addition, a graphic is used to illustrate the results. According to our research, when compared to collaborative and content-based strategies, the suggested strategy has a greater accuracy positive rate than the other two options.

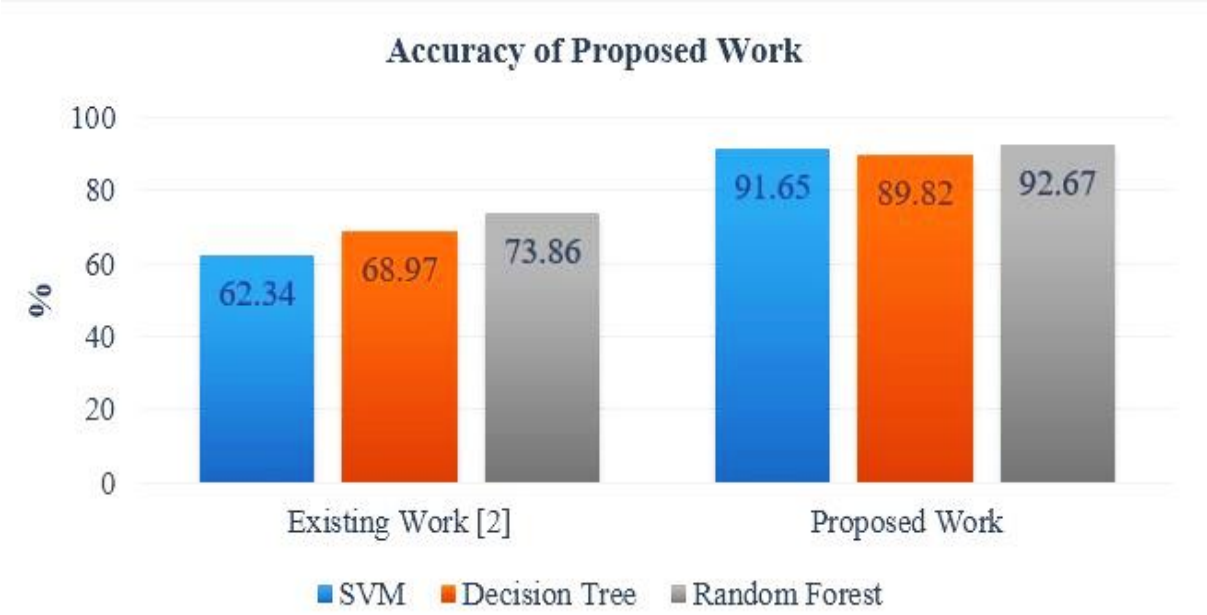
4.4.2 Precision: Precision is the numeral of optimistic class potentials connected with a optimistic class that are really related with the positive class under consideration. The precision of the capacity is resoluteby using Calculation 2 see below.



Graph 4.2 Precision graphs between Existing Work and Proposed Work

On the graph 4.2, you can see how to calculate the accuracy value for all algorithms, including current and new work. In addition, a graphic is used to illustrate the outcomes, the optional strategy has a higher accuracy optimistic rate, according to our data.

4.4.3 Accuracy: True Positives and True Negatives are both included in the accuracy calculation, which is a percentage of the correctly expected categories to the total Test Dataset. In order to determine the accuracy, equation 3 is used.



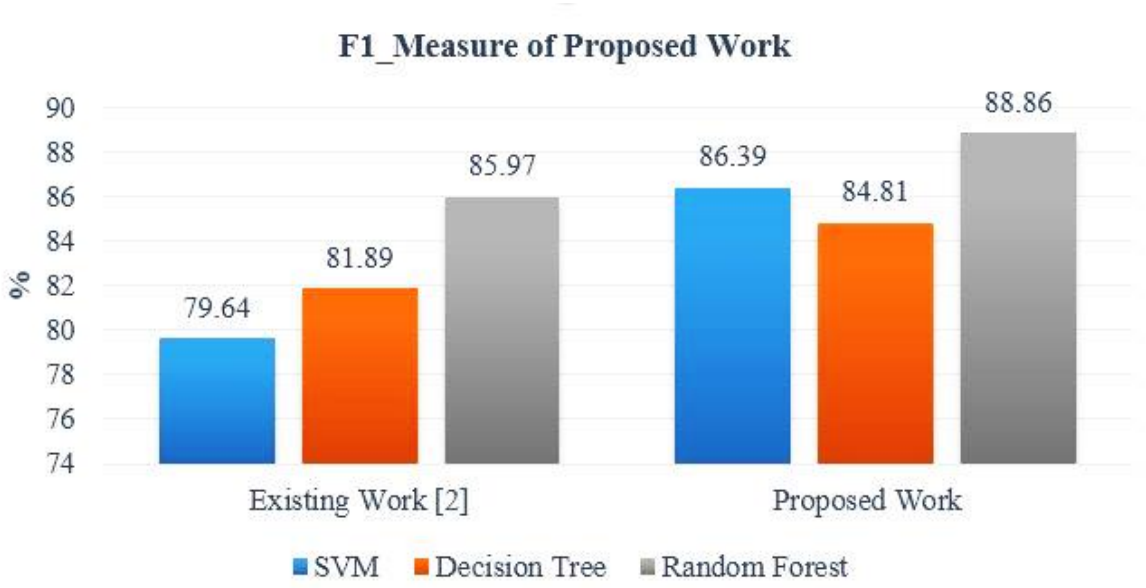
Graph 4.3Accuracy graphs between Existing Work and Proposed Work

Using the available research, we can obtain a precision value for each approach, both existent and suggested, as shown in graph 4.3. In addition, a graphic is used to illustrate the outcomes, the optional strategy has a higher accuracy optimistic rate, according to our data.

4.4.4 F1_Measure

The F1 measure, which is specifically employed in binary classification, is used to determine the correctness of the test results. For the computation of the F1 measure, the accuracy and recall are taken into consideration.

In the case of all samples that should have been classified as positive on graph 4.4, the F1 Score = 2 * ((precision * Recall) / (precision * Recall)) was calculated.



Graph 4.4 F1_Score graphs between Existing Work and Proposed Work

Based on the provided data, we may conclude that the proposed strategy works. The entire running time is also lowered while maintaining the highest possible suggestion quality. As illustrated above, the suggested technique is scalable and can be used to big datasets.

4.5 Chapter Summary

In this chapter we discuss about implementation and result about our work like Rendom Forest, SVM, and Decision tree, were investigated. According to our findings, utilising performance indices, Rendom Forest achieves an accuracy of 92.67 percent, which is higher than SVM's accuracy of 91.65 percent and Decision Tree's accuracy of 89.82 percent, respectively. The sensitivity and specificity of the Rendom Forest are higher than those of SVM and Decision tree, respectively. The proposed working approach has achieved 20% more accurate results as compared to existing results.

CHAPTER 5

CONCLUSION AND FUTUREWORK

CHAPTER 5

CONCLUSION AND FUTUREWORK

5.1 Conclusion

Based on the findings of this study, it was recommended that a hybrid intelligent machine-learning-based prediction system be created to aid in the diagnosis of heart disease. A dataset consisting of instances of coronary disease in Cleveland served as the testing ground for the system. Two well-known classifiers—Random Forest and SVM, respectively—were explored using feature selection as a research tool. Decision trees were also used. According to our results, which make use of performance indicators, Random Forest obtains an accuracy of 92.67 percent, which is greater than the accuracy achieved by SVM, which is 91.65 percent, and the accuracy achieved by Decision Tree, which is 89.82 percent, respectively. Both the sensitivity and specificity of the Rendom Forest have been shown to be superior to those of the SVM and Decision tree, respectively. The findings obtained using the suggested working technique are twenty percent more accurate than those obtained using the present approach. The performance of these predictive classifiers for heart disease diagnosis will be improved in future research by the use of other feature selection techniques and optimization methodologies. This will allow for improvements to be made to the classification accuracy.

5.2 Future Work

This system has other applications in upcoming development, such as, for example. It is possible for it to combine various therapeutic characteristics in addition to those listed above. Text mining is a tool that can be exploited, and is available in the database of the medical services business. It may be used to mine large amounts of unstructured information.

Reference

[1] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.

[2] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik and A. U. Zaman. Heart Disease Prognosis Using Machine Learning Classification Techniques. In:6th International Conference for Convergence in Technology (I2CT). pp.1-6 (2021). http://doi.org/10.1109/I2CT51068.2021.9418181.

[3] A. Kumari and A. K. Mehta, "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544544.

[4] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[5] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.242.

[6] Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy and M. Omair Shafiq, " The Threat of Adversarial Attacks Against Machine Learning inNetwork Security: A Survey,", 6 Nov 2019 (v1), last revised 4 Oct 2020.

[7] S. Farzana and D. Veeraiah, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-5, doi: 10.1109/ICCCS49678.2020.9277165.

[8] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2019, pp. 1-5, doi: 10.1109/WITS.2019.8723839.

[9] A. Lakshmanarao, A. Srisaila and T. S. R. Kiran, "Heart Disease Prediction using Feature

Selection and Ensemble Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 994-998, doi: 10.1109/ICICV50876.2021.9388482.

[10] A. Erdoğan and S. Güney, "Heart Disease Prediction by Using Machine Learning Algorithms," 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1-4, doi: 10.1109/SIU49456.2020.9302468.

[11] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[12] M. Chakarverti, S. Yadav and R. Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019, pp. 1578-1582, doi: 10.1109/ICICICT46008.2019.8993191.

[13] Y. Lin, "Prediction and Analysis of Heart Disease Using Machine Learning," 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI), 2021, pp. 53-58, doi: 10.1109/RAAI52226.2021.9507928.

[14] F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 338-341, doi: 10.1109/ICREST51555.2021.9331158.

[15] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019, pp. 1-6, doi: 10.1109/ICTCS.2019.8923053.

[16] R. Wijaya, A. S. Prihatmanto and Kuspriyanto, "Preliminary design of estimation heart disease by using machine learning ANN within one year," 2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T), 2013, pp. 1-4, doi: 10.1109/rICT-ICeVT.2013.6741541.

[17] M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala, "Machine Learning Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402653.

[18] M. A. Alim, S. Habib, Y. Farooq and A. Rafay, "Robust Heart Disease Prediction: A

Novel Approach based on Significant Feature and Ensemble learning Model," 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020, pp. 1-5, doi: 10.1109/iCoMET48670.2020.9074135.

[19] N. Basha, A. K. P.S., G. K. C. and V. P., "Early Detection of Heart Syndrome Using Machine Learning Technique," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019, pp. 387-391, doi: 10.1109/ICEECCOT46775.2019.9114651.

[20] N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-3, doi: 10.1109/ISCON52037.2021.9702314.

[21] C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020, pp. 158-161, doi: 10.1109/ECBIOS50299.2020.9203614.

[22] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telemat. Inform., vol. 36, pp. 82–93, 2019.

[23] R. J. P. Princy, S. Parthasarathy, P. S. H. Jose, A. R. Lakshminarayanan, and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570–575.

[24] KAGGEL, Cardiovascular Disease dataset. 2020.

[25] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," in 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623. doi: 10.1109/IBCAST.2019.8667106.

[26] S. A. Sabab, M. A. R. Munshi, A. I. Pritom, and others, "Cardiovascular disease prognosis using effective classification and feature selection technique," in 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), 2016, pp. 1–6.

[27] G. Biau et al., "Performances Analysis of Heart Disease Dataset using Different Data

Mining Classifications," Test, vol. 8, no. 6, pp. 2677– 2682.

[28] A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," in 2011 21st International Conference on Systems Engineering, 2011, pp. 406–409.

[29] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329–1333.

[30] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in 2016 international conference on circuit, power and computing technologies (ICCPCT), 2016, pp. 1–5

[31] M. Gawali, N. Shirwalkar, and A. Kalshetti, "Heart disease prediction system using data mining techniques," Int. J. Pure Appl. Math., vol. 120, no. 6, pp. 499–506, 2018.

[32] J. Maiga, G. G. Hungilo, and others, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," in 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2019, pp. 45–48.

[33] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.

[34] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mob. Inf. Syst., vol. 2018, 2018.

[35] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Comput. Methods Programs Biomed., vol. 141, pp. 19–26, 2017.

[36] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Inform. Med. Unlocked, vol. 16, p. 100203, 2019.

[37] Dhanashree Medhekar, S, Mayur Bote, P & Shruti Deshmukh, D 2013, “Decision Support in Heart Disease Prediction System using Naive Bayes”, International Journal of Enhanced Research in Science Technology & Engineering, vol. 2, no. 3, pp. 1-5.

[38] A. Mert, N. Kılıç, and A. Akan, “Evaluation of bagging ensemble method with time-

domain feature extraction for diagnosing of arrhythmia beats," Neural Computing and Applications, vol. 24, no. 2, pp. 317–326, 2014.

[39] J. Zhu, L. He, and Z. Gao, "Feature extraction from a novel ECG model for arrhythmia diagnosis," Bio-Medical Materials and Engineering, vol. 24, no. 6, pp. 2883–2891, 2014.

[40] H. R. Marateb and S. Goudarzi, "A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system," Journal of Research in Medical Sciences, vol. 20, no. 3, pp. 214–223, 2015.

[41] S. Faziludeen and P. Sankaran, "ECG beat classification using evidential K-nearest neighbours," Procedia Computer Science, vol. 89, pp. 499–505, 2016.

[42] Dinesh Kumar G , Santhosh Kumar D , Arumugaraj K , Mareeswari V, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India , 2018, pp. 1-7.

[43] Kanika Pahwa , Ravinder Kumar, "Prediction of Heart Disease Using Hybrid Technique For Selecting Features," 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) , 2017 , pp. 500-504.

[44] R. Banerjee, A. Dutta Choudhury, S. Datta, A. Pal, and K. Mandana, Non Invasive Detection of Coronary Artery Disease Using PCG and PPG, vol. 181, Springer International Publishing, Germany, 2017 , pp. 234-242.

[45] N. Paradkar and S. R. Chowdhury, "Coronary artery disease detection using photoplethysmography," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, Republicof Korea, July 2017, pp. 100–103.

[46] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," Journal of Healthcare Engineering, vol. 2017, Article ID 2780501, 2017 ,pp. 1-13.

[47] Priyanka , Dr.Pushpa RaviKumar , "Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree", IEEE, International Conference on circuits Power and Computing Technologies [ICCPCT], 2017, pp. 1-7.

PUBLISHED PAPER

[1] **Heart Disease diagnosis using machine learning classification techniques(accepted)**

[2] **Review of Heart Disease Diagnosis using Machine Learning Classification Techniques(Communicated)**

PLAGIARISM REPORT