REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

USE OF DATA MINING TECHNIQUES FOR EARLY CARDIOVASCULAR DISEASE PREDICTION

Aws Nabeel AHMED

Master`s Thesis

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, 2022

USE OF DATA MINING TECHNIQUES FOR EARLY CARDIOVASCULAR DISEASE PREDICTION

Aws Nabeel AHMED

Information Technologies

Master`s Thesis

ALTINBAŞ UNIVERSITY

2022

The thesis titled USE OF DATA MINING TECHNIQUES FOR EARLY CARDIOVASCULAR DISEASE PREDICTION prepared by AWS NABEEL AHMED ALSALIM and submitted on 13/12/2022 has been **accepted unanimously** for the degree of Master of Information Technologies.

Asst. Prof. Dr. Sefer KURNAZ
Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Sefer KURNAZ	Department of Computer Engineering, Altınbaş University	_____
Asst. Prof. Dr. Oguz KARAN	Department of Software Engineering, Altınbaş University	_____
Assoc. Prof. Dr. Adil DENİZ DURU	Department of Sports and Health Sciences, Marmara University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master`s Thesis.

Submission date of the thesis to Institute of Graduate Studies: ____/____/____

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Aws Nabeel AHMED

Signature

DEDICATION

To the soul of my father and brother (may Allah have mercy on them). To my dear mother who taught me to be giving, may Allah grant her long life. To my wife, companion on the life and my daughter, my soul mate, who have been the best help to me throughout my life journey. To my brothers who are my support and a source of pride. To everyone whom I received advice and support. My regards to the above

PREFACE

First of all, I thank Allah Almighty for helping me complete this thesis.

I am pleased to extend my thanks, appreciation and gratitude to my honorable supervisor, Asst. Pro. Dr. Sefer Kurnaz, who I was pleased to his supervision on this research for his efforts to create a scientific environment and to overcome all the difficulties that I faced throughout the research period. I extend my thanks to the Altınbaş University, Department of Computer Engineerin**g /** Information Technologies, for supporting me in completing this work.I special thanks to my mother, who supported me with her prayers, and my wife, who was my support at this stage and did not leave me alone at all times with such circumstances. I appreciate my friends and everyone who encouraged me and contributed to the success of this study, especially my dear friend Dhulfiqar Aamer. this accomplishment would not be possible without them.

Finally, I would like to thank everyone who helped me with their advice and efforts.

ABSTRACT

USE OF DATA MINING TECHNIQUES FOR EARLY CARDIOVASCULAR DISEASE PREDICTION

Ahmed, Aws Nabeel

M.Sc. Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: December / 2022

Pages: 49

As healthcare collects such vast amounts of information, the medical field is seen as "too rich in important data" and "information poor". As dynamism in the context of healthcare is one of the most common issues that decision makers look at, the importance of data quality has risen to the top of the list of requirements for decision makers. It is possible to solve this problem by using advanced data mining tools. Data mining is particularly large and useful for examining and analyzing a large amount of information due to the various information flows and large assortments of the supporting data set. Our research hopes to help experts make the best decision to predict this deadly disease by using a variety of data techniques (selection trees, naïve Bayes, and neural networks), as well as comparisons to mining discoveries from previous experts. Several patient records were used, which were separated into 700 records with 300 males and 400 females, and 13 highly significant disease-related characteristics (age, smoking, genetics, glucose, etc.). The analysis is divided into two parts: the first part involves cleaning the excellent data using various appropriate procedures, while the second part involves applying the formulas described earlier to that data. The results show that each technique has a clear advantage when it comes to producing the best number. The results

were: individually, Bayesian technique (95.44%), decision tree approach (98.15%), neural networks (89.16%). Finally, we must achieve our goals while being absolutely certain of the soundness of the decision.

Keywords: Cardiovascular Disease, Decision, Decision Trees, Environment, Health Care, Naive Bayes, Neural Networks, Patient, Prediction, Technique

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

WHO	:	World Health Organization
IHDPS	:	Intelligent heart disease prediction system
CMAR	:	Classification based on Multiple Association Rules
CART	:	Classification and Regression Tree
EMA	:	Expectation-Maximization Algorithm
MLP	:	Multilayer Perceptron

LIST OF SYMBOLS

x_i : Independent scaling factors

1. INTRODUCTION

The healthcare sectors have many difficulties and challenges in finding diseases. Healthcare organizations are collecting bulk amount of patient data. The Data mining methods are utilized to decide covered data that is valuable to healthcare specialists with effective analytic decision making. Data mining strategies are utilized in the field of the healthcare industry for different purposes [1]. Therefore, Coronary artery disease is increasingly becoming one of the leading causes of death. Due to the vast amount of information generated by the healthcare environment and especially clinical diseases, it is necessary to formulate a few important and powerful developments in order to examine, focus on and access key information. Data mining is essential for extracting important data from massive data sets since it is entirely dependent on a computer. Data mining is particularly large and profitable to explore and study vast amounts of information due to the amount of information generated by this environment and the large assortments of supporting evidence.

There are not many information structures in these medical institutions to support patient billing, inventory management, and knowledge development. Medical offices with limited resources rarely provide useful medical data that is not scattered and can be utilized, as during the normal period for people with coronary vein infection, as well as different and simple data sets for other patients. Many different strategies have resulted in trips to psychotherapy centers lasting more than 10 days. How many unmarried women over the age of 30 have been diagnosed with a carcinoid tumor. Should the course of treatment consist of chemotherapy alone, radiotherapy alone, or both chemotherapy and radiotherapy given the patient's knowledge of the dangerous turn of events? How likely are people to develop cardiovascular disease, given their experience? Basic pre-scientific theories cannot adequately answer these questions due to their complexity. Most clinicians choose a course of treatment based on their knowledge and experience rather than a data set with a lot of data. Such planning can lead to biases, errors, and excessive clinical expenses that harm the patient's cognition. Using various clinical data from the huge datasets available in the industry, it is possible to identify a few distinct coding schemes used in the clinical sector. Such situations should be used for clinical purposes to deal with this disease accurately. On the other hand, big and diversified data are

the primary data types that have been widely disseminated. Before data can be extracted, it must be fully collected and formatted. Some researchers have used a model consisting of a cardiological profiling unit based on approximate groups and a classification unit based on ambiguous rules. The rules generated from fuzzy classifiers are improved by applying the adaptive genetic algorithm. [2].

The labeled data must then be combined to create a system containing relevant clinical information that can be retrieved using extraction techniques. The goal of the extraction cycle is to identify novel and confidential examples in the various educational files of each patient, as well as those examples themselves. Many academics are interested in using data mining techniques and various real-world courses to focus information from many educational lists and work on logical data that is relevant to this serious disease [3]. Accurate and distinct evidence of disease is the critical indicator of the success and robustness of the data mining algorithm. According to the Planet Wellbeing Society, coronary artery disease is said to have been the leading cause of death on our large planet for at least ten years. Researchers use quantitative analysis techniques to help medical professionals identify and predict cardiovascular disease. The coronary artery disease hypothesis architecture, which is based on clinical data from previous patients, can help medical professionals identify cardiovascular disease. Large amounts of information are collected by many clinical benefit providers and centers from the people who obtain it with the help of existing structures. There are a few devices on the market that use prediction calculations, but all of them have defects that may lead to patient death [4]. In order to isolate and analyze various data and important hidden information from these informational indexes, an AI approach is very necessary. It actually enhances speed and accuracy in forecasting.

Many companies, especially pharmaceutical companies, have used data mining techniques for further investigation necessary to determine if data mining and automated reasoning can be used to predict Cardiovascular disease. In order to determine whether a patient has cardiovascular disease or not, a few data mining methods are used. In addition, the assumption not only helps specialists reach the correct conclusion more quickly, but also validates the result, igniting future research that may be able to prevent or reduce coronary

disappointments. Hidden relationships can be found and diseases can be identified more accurately through a combination of data mining using modern methods and the experience of a doctor who is an accurate specialist in this disease [5]. The main objective of this investigation is to build predictive models of cardiovascular disease using decision trees, naive Bayes, and neural networks. It is possible to identify and retrieve cases and links in the field of cardiovascular disease using a reliable source of information. So he may be able to support professionals in making joint clinical decisions in ways that wouldn't help with regular decision making. Amazing thinking may enable people to use fewer prescribed medications in general. It displays the discoveries in an equal and graphical way for easier understanding.

Various legacy technologies, although with significant limitations, can handle noise data to some extent with many problems and drawbacks:

a. Filtering software is not perfect. Most writing relies on accurate information to get clear and reliable results.
b. The issue of information purification is rarely taken into account due to the claim that unrelated information (the clamor) often does not outperform the 40% for the minority class.

Since the information is so lopsided in this case, it is difficult to make calculations with great accuracy and make enough discoveries to predict this fatal condition.

2. RELATED WORK

It is very difficult for medical service institutions in a healthcare environment to provide high-quality and at the same time affordable treatment (clinics, clinical centers). Accurate diagnosis and treatment of the patient is one of the most important things necessary to provide the highest level of patient care. It is important to realize that making a combination of poor clinical decisions can have serious repercussions in the future. In addition, we must consider reducing the cost of various clinical tests. With the help of a group of networks that provide emotional support or computer-based big data, they may be able to achieve many of their goals. Currently, the vast majority of different clinics use emergency clinics data frameworks to track the medical information of their patients. These systems often provide a wealth of information, including facts, messages, frameworks, and other visual representations. Unfortunately, only a small portion of this data is used to support an important clinical decision. There is a lot of new information in this content. To assist medical practitioners in learning how to make clinical choices for the benefit of the patient [6].

A variety of tests have been used to identify and treat cardiovascular disease. Different information mining techniques are used to obtain and evaluate different levels of accuracy in a variety of ways. The positive aspects of evaluating the ownership of a photograph are assumed to be independent of those of other properties because the NB classifier technology uses the unexpected open door. Guileless Bayes are recommended when there is a high probability of developing cardiovascular disease. This application was provided as a component of electronic clinical thinking and important web-based clinical decisions.

The data obtained and organized was considered a ready-made set because the labor was separated into two basic stages: ordering and pre-processing, which carried out various activities such as data normalization, reduction and cleaning, among others. So the process of forecasting is to predict the disease and what will happen in the future. A different experimental group is then formed in response to the concerns raised by the disease depending on the type and characteristics of the condition. Send to the expert the results of the forecast at

the conclusion of the course. Several investigations into this case have been carried out through information mining. The topics of the article are briefly summarized below [7].

According to [8], predicting heart disease in its early stages may prevent potential deaths due to heart attacks. A good classification algorithm may help a doctor predict the presence of cardiovascular disease before it actually occurs. This research focuses on the prediction of potential heart disease by integrating the most recent data set available in the UCI and Convolutional Neural Networks (CNN) repository. This data set consists of some heart test standards as well as general human habits. The results showed that the proposed model is superior to the current techniques referenced in this paper. The overall accuracy of the proposed model is 97%.

According to [9], this paper describes a correlation rule mining algorithm for detecting intrusions on different networks. The KDD data set is used for experimentation. There are three input features categorized as core features, content features, and traffic features. Several attacks are in the Dataset Classified to Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). The proposed method gives a significant improvement in detection rates compared to other methods. A correlation rule-mining algorithm is proposed to evaluate the KDD dataset and dynamic data to improve efficiency, reduce false positive rate (FPR), and provide lower processing time.

The experts have employed a variety of information mining techniques to the Savvy Heart Expectation Framework (IHDPS) to estimate cardiovascular illness, including decision trees, Gullible Bayes, and brain networks. These methods served as a reliable source of knowledge. Following the application of these distinct methodologies, several collections of 909 unmistakable records with diverse features were obtained, including 15 specific clinical highlights from various data set collections with a focus on cardiovascular concerns in Cleveland. According to findings, ANN had a precision pace of 85.53, Decision Trees had an exactness pace of 89%, and naive Bayes had a precision pace of 86.53%. Naive Bayes is the best method in terms of performance since it comprises a substantial and crucial portion of the set of correct assumptions (86.53%) for the group of patients who have heart problems and

suffer the negative consequences of cardiovascular diseases. Decision trees are the final approach, followed by neural network in that order.

Decision trees, however, are the simplest and most reliable way in the field of cardiology and visualization for those who don't suffer from the negative consequences of cardiovascular disease (89%) when compared to other processes, despite combined discoveries.

According to [10], others were of the opinion that all problems should end here. The application of various data mining techniques led to the development of a highly accurate model (Decision Trees, Neural Network, and Naive Bayes). When it comes to reaching important and relevant exploration objectives for a variety of information, each of these significant developments in the healthcare environment has a certain strength that sets it apart from the others. Frameworks in IHDPS are able to support a range of simple and complex options. Regarding the suitability of different mining techniques, there is no doubt.

Dynamic frameworks have no questions left unanswered due to IHDPS. Using this approach, it is possible to find a range of different relationships and patterns in cardiovascular diseases that had not been recognized before. With this solution, you have reliability, simplicity, and room to expand as your business expands. Diabetes, circulatory stress and coronary artery disease were identified using neural network clusters. On the basis of the different set of information from the patient's records, given that a few different examinations were prepared. Age range, pulse, and other factors were among the thirteen variables used in this experiment. The innovation of the back spread was used to finish the preparation process. As a result, based on the physician's blurry knowledge and the new information he has gathered, the framework creates a list of potential diseases for which patients may be at risk.

Using two things: artificial intelligence (A.I) calculations and a digital stethoscope. Supported by the use of accurate heart sounds to assess persistent cardiovascular infections, according to [11]. Our method outperformed the latest classifiers with an accuracy rate of 96%. During the primer, 152 different sounds were captured using computerized stethoscopes. These good results came from a survey of 122 people. A different data set goes through several continuous rounds of partitioning, segregation, etc.

According to [12], it is often difficult for medical practitioners to predict cardiovascular disease because it requires experience and knowledge and is a complex task to accomplish. This health industry contains huge amounts of useful data to draw effective conclusions using its hidden information. Therefore, by using appropriate results and making effective decisions about the data, some superior data analysis techniques are used, for example Naive Bayes and Decision Tree. Using some characteristics (age, gender, blood pressure, etc.) chances of developing cardiovascular disease can be predicted. In this study, we collected 301 samples with 12 clinical features. Logistic regression, decision tree, SVM, and Naive bayes classification algorithms were applied to predict heart disease. In this case, logistic regression provided an accuracy of 86.25%. However, we also compared the results based on the UCI data set with our model.

According to [22], The researchers proposed creating an app that could predict a person's risk of developing Cardiovascular disease based on a number of variables, including age, circulatory stress, orientation, and more. According to this study, the best, most accurate and most reliable innovation for the development of a very important application that assesses chronic cardiovascular disease with great accuracy, correctly and reliably is the use of a neural network. The various Cleveland datasets from the UC Irvine library served as a very important basis for building and testing the model. Prepare and evaluate different datasets using multifaceted cognitive brain regulators.

According to [13], he claims to use CNN and NN, which are two artificial intelligence technologies, to predict very serious cardiovascular diseases. At first, the analysis shows that NN is more accurate than CNN. The different information layer, the second secret layer, and the final output layer are the three basic layers that are important for a standard mild neural network. The three ordered dimensions of a CNN nerve are first length, second width, and third depth. Since CNNs have three different and unique layers, they differ greatly from ordinary neural networks (ONN) in this way. CNNs are more accurate and reliable than conventional and regular NNs. When I used different hidden layers of NN 3, I was finally able to achieve an average accuracy level of 86.40 percent. By using three different and hidden layers, CNN was able to achieve a lower accuracy rate of 78.65%.

Four artificial intelligence models, according to [14], were used to predict very serious cardiovascular disease. SVM may choose an accuracy speed of 83 percent. An accuracy rate of 79 percent was achieved using decision trees. In 78 percent of the different cases, linear regression provided very accurate results. K-Nearest Neighbor was found with an accuracy rate of 87 percent. Look for situations where the model is either inappropriate or inappropriate in the option trees. Significant arithmetic is used to make a set of judgments based on the k points closest. UCI's Big Data Center is used to provide public data to build and test the accurate model as a tool for predicting cardiovascular disease. Python is the preferred and good programming language for doing such important work.

A critical problem is the consistency of clinical data. Several studies have examined the problem of poor data quality in the clinical field. Many different types of anomalies in the data have been examined by experts in the disciplines of data sets and data quality. Examinations, for example, treated anomalous data. By using informative groups to create useful right-handed dependency rules, data consistency concerns can be alleviated [15].

According to [16], SVM, Naive Bayes and Decision tree has been applied with and without using PCA on the dataset. We used PCA to reduce the number of attributes. After reducing the size of the dataset, SVM outperforms Naive Bayes and Decision tree. SVM can further be used to predict heart disease. A GUI desktop application can be built using SVM and this dataset to predict the possibility of cardiovascular disease in a patient.

A comparison was done in another publication. They promoted a method for identifying data links and identifying possible infringement to remedy information mistakes. In order to fill in any gaps in the data, the developers included an attribution model. To fill in the gaps in a dataset of certifiable medical services, they used three different calculations. The designers enabled information from a few sources to combat heterogeneity. A semantic-based framework may be in charge of managing clinical data, including free clinical notes, acoustic data, and clinical images [17].

Academic research was done on data standardization. When the ongoing electronic clinical consideration data doesn't adhere to a specific electronic clinical advantages record creating,

they set up a worldview display structure as well as a model supervisor and arranging device. We managed the duplicate archive issue throughout our request. They came up with a method for identifying duplicate evidence records that is simple to use.

According to [18],In this paper the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The focus is on using different algorithms and combinations of several target attributes for intelligent and effective heart attack prediction using data mining. For predicting heart attack, significantly 15 attributes are listed and with basic data mining technique other approaches e.g. ANN, Time Series, Clustering and Association Rules, soft computing approaches etc. can also be incorporated. The outcome of predictive data mining technique on the same dataset reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction.

Audits suggest that if clinical decision support and PC-based patient information are taken into account, patient outcomes may be enhanced. It is possible to envision a situation in which new techniques for displaying and analyzing information, such "data mining," are used to greatly improve the nature of therapeutic treatment decisions.

3. MOTIVATION

It is quite difficult for clinical benefit organizations to provide top-notch care at an affordable price (facilities, clinical core interests). For the greatest care, precise patient diagnosis and treatment are essential. It is crucial to understand that making bad clinical decisions might have detrimental effects. Additionally [19], the cost of clinical testing needs to be reduced. They may be able to achieve their goals with the support of organizations that provide in-depth assistance or computer-based knowledge. Currently, a staggering majority of centers use crisis facility information systems to track the medical data of their patients. These structures frequently provide enormous amounts of data, which take the shape of informational fragments, messages, frames, and other visual representations. Tragically, only a small portion of these data are used to support clinical judgment. There is a lot of fresh information in these materials. How could information be transformed into knowledge to support clinical decision-making by clinical experts? Remember that this is an important highlight. On this, we'll concentrate our request [20].

3.1 PROBLEM STATEMENT

There aren't many information structures in crisis centers that support patient billing, stock management, and the creation of essential experiences. What is the typical lifespan of those who have coronary artery disease? A few medical offices with restricted capacity offer ongoing encouragement services. How many different systems have resulted in trips to the ER that lasted longer than 10 days? How many older single women (over 30) were found to have malignant growths? Should the course of therapy consist of chemotherapy alone, radiation alone, or both chemotherapy and radiation given the patient's knowledge of the adverse turn of events? Furthermore [22], how likely are people to get a heart condition given their experience? Basic prescience models cannot handle these types of requests. When choosing a course of treatment, most clinicians depend on their experience and skill rather than a dataset with a lot of data. This training might lead to biases, errors, and irrational clinical costs that are detrimental to patient consideration [17].

According to Wu et al's review [18], a reconciliation between clinical choice assistance and PC-based patient records may enhance patients' results. It is feasible to establish an environment where the quality of medical treatment options is considerably enhanced by using new techniques for displaying and analyzing data, such as "information mining."

3.2 PROBLEM THE AIM AND OBJECTIVES OF THE STUDY

Before implementing the prediction process, this study attempts to organize and clean up the various medical data set in the healthcare and cardiovascular disease environment using a variety of important techniques and modern algorithms.

The following set of goals are achieved in order to reach the best prediction of this disease, and this is the goal:

a. To clean the different and unbalanced data and make a prediction very accurately, the unbalanced noise data set is removed through a set of traditional methods, after that a set of learning algorithms are applied and the results are compared with the results of previous researchers [23].

b. To find a set of important and relevant characteristics as a function to improve the prediction process, a set of techniques is used, the best of which is the genetic algorithm to increase the accuracy of the results of different algorithms applied to those different data.

c. In order to compare the results and the accuracy of this method, the set of results should be compared with the previous results from the work carried out by previous researchers.

4. DATA MINING REVIEW

Although information digging has been practiced for more than 20 years, its full potential is only now becoming apparent. By combining quantitative analysis, artificial intelligence, and data set innovation, information digging is a process for extracting hidden instances and relationships from large collections. By "the nontrivial cycle of extricating verifiable, previously obscure, and possibly useful data from the information preserved in a data collection," Fayyad defines information mining. Giudici describes it as "a process of choosing, investigation, and displaying of enormous volumes of information to discover normalities or links that are at first opaque intended to generate obvious and significant objectives for the proprietor of the set. [24]"

Information mining uses both guided and unguided learning. In managed learning, model bounds are obtained from a preparation set, but not in solo learning (e.g., kmeans grouping is unaided). Every information mining technique has a distinct capacity depending on the display's goal. The two most often used showing aims are categorization and forecast. While order models provide discrete, unordered outright names, expectation models predict continuous highly regarded capabilities. Decision Trees and Brain Organizations use arrangement techniques, whereas Relapse, Affiliation Rules, and Grouping are expectation computations.

Decision tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Two-Partition 3), and C4.5. These algorithms differ in choosing splits, when to stop a node from splitting, and assigning a class to a non-partitioned node. CART uses a Gini index to measure the inclusions of a department or group of training sets. It can handle high dimensional categorical data. Decision trees can also process continuous data (as in regression) but must be converted to categorical data.

The establishment for some AI and information mining procedures is Naive Bayes or Bayes' Standard. The standard is utilized to assemble models that are fit for expectation. It offers new ways to deal with information investigation and perception. By deciding the association between the objective (i.e., reliant) and other (i.e., free) factors, it gains from the "proof"

introduced. Input, stowed away, and yield units are the three levels that make up brain organizations (factors). The significance of the appointed worth (weight) of every individual information unit lays out the association between input units, stowed away units, and result units. more noteworthy weight demonstrates a more prominent significance. The exchange capabilities utilized by brain network techniques are straight and sigmoid. Huge volumes of information might be prepared with not many data sources utilizing brain organizations. At the point when different techniques are inadequate.

5. METHODOLOGY

5.1 MATERIALS

The main objectives of our research are to assist professionals in making the best possible decision to predict this disease (cardiovascular disease) using a different set of information mining methods (decision trees, Naive Bayes, and neural network), as well as comparing the different set of results with previous analysts to arrive at The best (the best possible prognosis for this disease). In this section, we will explain how it works and how to use a set of different algorithms on the medical data set of cardiovascular disease in a healthcare setting. We have strong confidence that we will achieve what we set out to do [27].

5.2 PREDICTION MODELS

The several methods used in this study to extract data and predict cardiovascular disease are fully summarized in this section.

5.2.1 Neural Network

It all boils down to having a fundamental knowledge of how the mind functions. For instance, it enables computers to contrast or compare two particular samples. Since neurons are made of solely numbers rather than qualities, they arrange traits and exclude them from instances. There are several layers of connected axons in the brain. The perceptron is a network of connected nodes that, as its name suggests, simulates several kinds of direct setbacks. In an indirect running task, this perceptive aftermath has seen multiple straight setbacks. As seen in Fig. 5-1, each seprtpetrep layer has a common link.

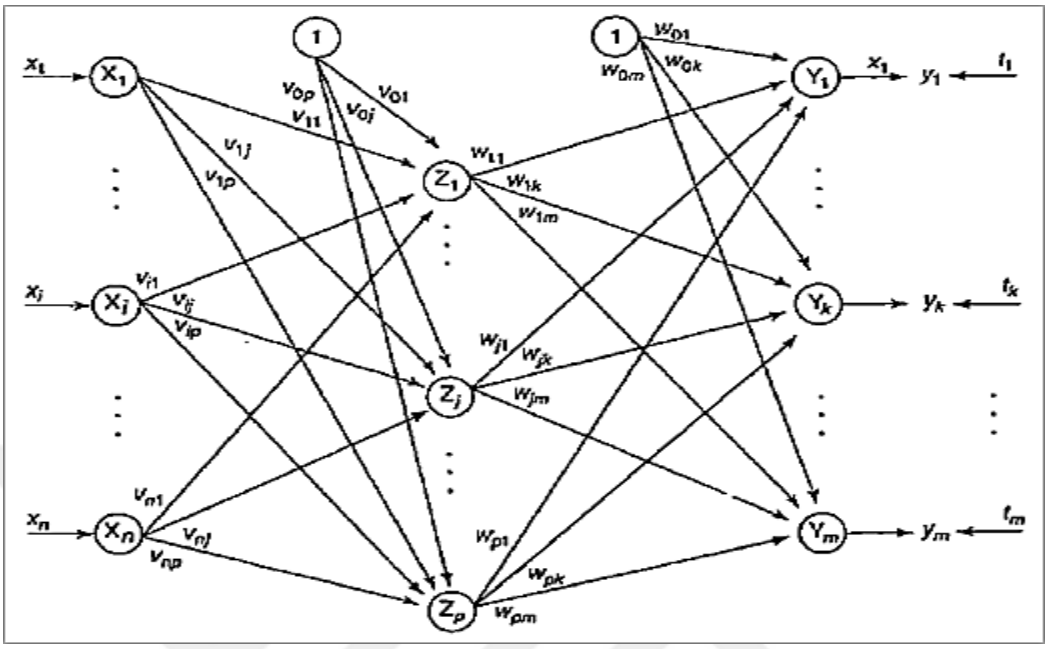

Figure 5.1: Multi-layer perceptron (MLP) feed forward Neural Network

Hidden layers change the information layer loads to lower fatal mistake rates. A network of brain cells serves as a classifier in managed learning [19].

There are three layers in the neural network: the input layer first, the hidden layer second, and the output layer third, and it also contains a group of 13 neurons, 2 neurons, and 6 neurons. The figure below shows these three layers and cells.

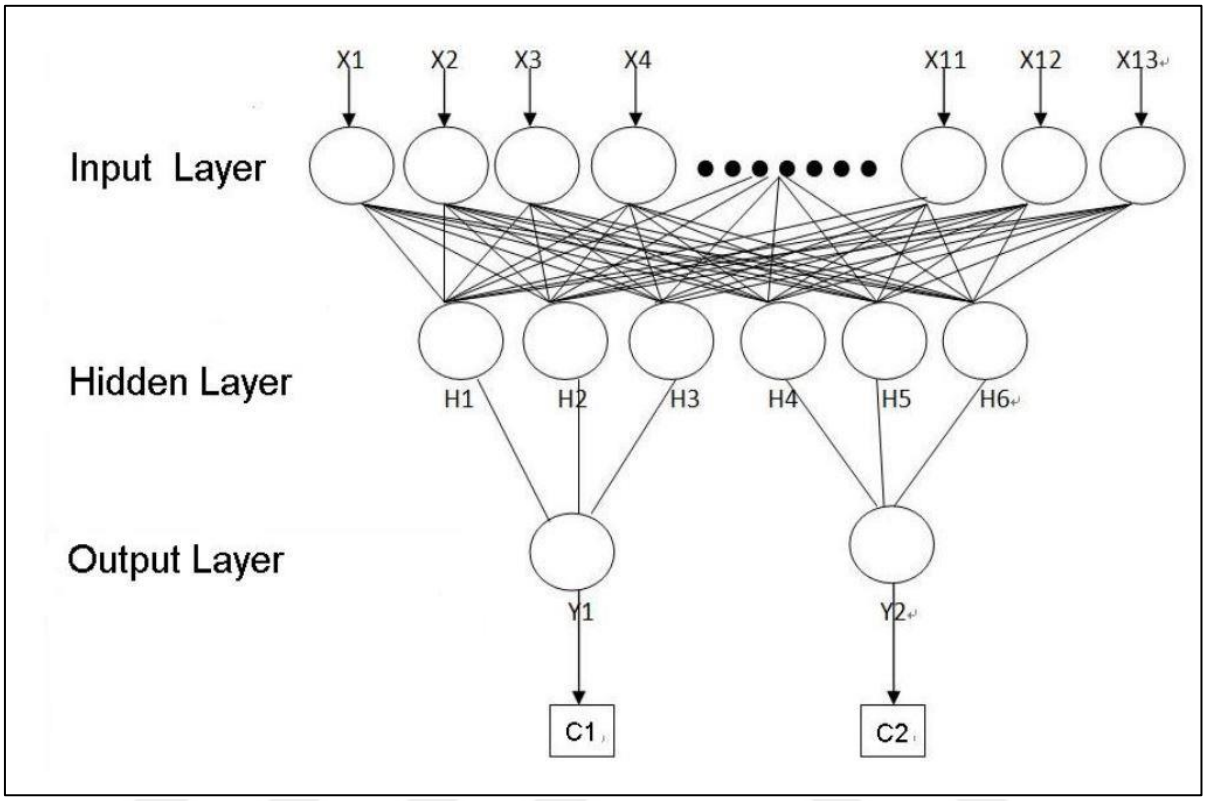

Figure 5.2: The model contains three layers.

5.2.2 Naive Bayes

Naïve Bayes is a statistical classifier that does not assume any dependency between attributes. It tries to maximize the subsequent probability of determining the class. In theory, this classifier has a minimum error rate but it may not always be the case. However, the inaccuracy results from assumptions due to conditional independence of the category and the lack of available probabilistic data. Observations show that Naïve Bayes performs consistently after reduction of number of attributes. According to Bayesian theorem

$$P(A|B) = P(A)*P(B/A)/P(B), \text{ Where } P(B|A)=P(A\cap B)/P(A) \quad (5.1)$$

Based on the above formula, the Bayesian classifier calculates the conditional probability of an instance belonging to each class, and based on this conditional probability data, the instance is classified as the class with the highest conditional probability. In cognitive expression, it has

excellent interpretability like a decision tree and is able to use past data to build an analysis model to predict or categorize in the future [27].

The next classifier used in this experiment is known as naive Bayes. For example, data may be requested using a model for controlled learning groups. The probability of each class is calculated, then the most likely class is assigned to the entire transaction. naive Bayes is a popular approach to season prediction for a variety of data sets, including educational and clinical information mining. This method can be used to coordinate a wide range of data sets, including general assessment assessments and geographic locations of diseases [28].

Each record belongs to a specific category, and each record is awarded to a person based on their own merits and assuming they fall under that category, forbidding the addition of any other components. This approach can be very useful when dealing with massive data sets. Despite its simplicity, Naive Bayes outperforms even the most complex assembly schemes. The approach is clear in some circumstances. The first and important step is to make a frequency table from the data set. Based on the various information provided, create a probability table. Finally, for each category, the naive Bayesian criteria for posterior probability are used. The prediction results in the class with the best return probability. The Naive Bayes formula can be used to calculate the probabilistic probability in a probabilistic model [29]:

$$(CK|x)=P(x|CK)P(CK)/\ P(x). \tag{5.2}$$

An alternative definition of "Back" is the total of all previous probabilities and current evidence. The limitation of the class variable C to the following may lead to the following condition:

$$(CK|x1, \dots . , xn) = 1\ Z\ ((CK) \prod (|x_i\ |CK)\ n\ i=1. \tag{5.3}$$

Evidence (Z)=P(x), with x1,..., xn being independent scaling factors. The Bayesian Classifier is frequently predicted using prior knowledge and previously available data on the original

distributions in the absence of such prior information, despite the fact that it has certain issues [32].

5.2.3 Decision Tree

The selection tree is made up of three main components: hubs that deal with quality checks, branches that display the results of those tests, and sheets (or final hubs) that deal with the results of a particular class. The pivot at the base of the tree is the most important. Selection tree classifiers are a good choice for locating exploratory data because they do not require any form of topographical knowledge or boundary preparation. Selection trees are capable of handling complex data. Therefore, people find it easier to understand their representation of the collected information as a tree structure, as shown in Figure 5-3.

It is easy to understand and prepare for the selection tree registration procedure. To ensure that each paper is free from defects, iterative characterization is frequently used in the assessments. As a result, the goal is to gradually increase the flexibility and accuracy of the decision tree while also ensuring that the need for data is high. This method was used to reduce the entropy of unstructured data [33].

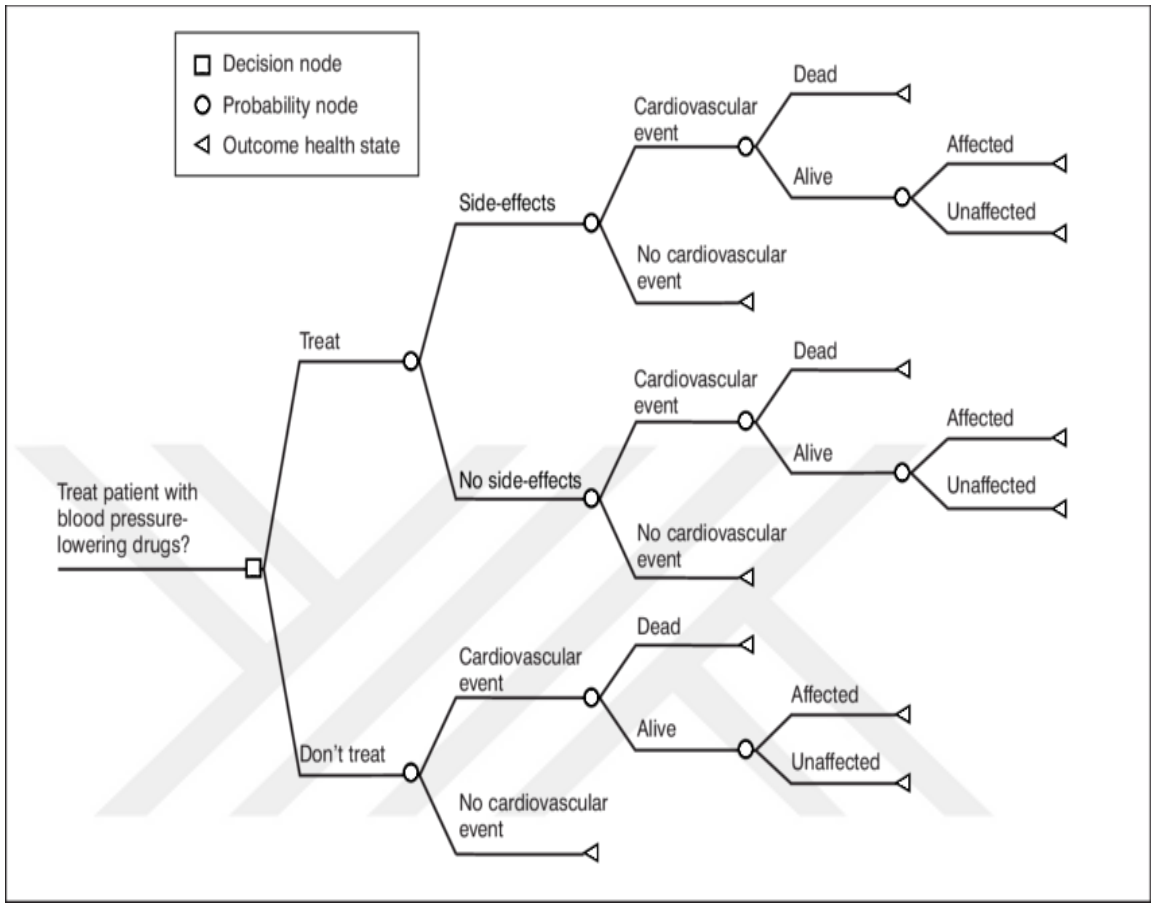

Figure 5.3: Decision tree to treat one of the causes of cardiovascular disease, high blood pressure.

The size of the dataset affects the branches and center points of the decision tree. The sending structure distributes a certain measure of values to each credit. Each trade has a unique set of regulations that each branch and center must follow in order to get the best results. The record will finally receive the grade suggested by the decision center. Therefore, for as long it is granted a class, the trade is remains the focus. This method divides the credit into branches and centers because it wants to pick one of the qualities to serve as the grade. The class mark in Fast Digger may be selected while bringing in the dataset [34].

5.3 ABOUT DATA CLEANING

The most common way of cleaning information, especially clinical information in the medical services climate, and especially in the records of heart patients, includes various huge and proficient techniques.

The objective is to make the information homogeneous with one another and partition it into related classifications to work with the use of critical and vital methods to create the best analysis and forecast of this illness that is both the most risky and common on the planet. Given the weakness of this infection and the risk it stances to individuals' lives, this cycle is done to genuinely introduce the aftereffects of the procedures utilized, analyze these methods, and distinguish the most ideal innovation that anyone could hope to find for diagnosing and anticipating coronary illness. Everything should be done precisely and carefully to help experts in arriving at the best choices for patients.

Table 5.1: Heart attack parameters

Parameters	Weightage	
Male and Female	Age < 30	0.1
	>30 to <50	0.3
	Age>50 and Age <70	0.7
	Age>70	0.8
Smoking	Never	0.1
	Past	0.3
	Current	0.6
Overweight	Yes	0.8
	No	0.1
Alcohol Intake	Never	0.1
	Past	0.3
	Current	0.6
High salt diet	Yes	0.9
	No	0.1
High saturated fat diet	Yes	0.9
	No	0.1
Exercise	Never	0.6
	Regular	0.1
	High If age < 30	0.1
	High If age > 50	0.6
Sedentary Lifestyle/inactivity	Yes	0.7
	No	0.1
Hereditary	Yes	0.7
	No	0.1
Bad cholesterol	Very High >200	0.9
	High 160 to 200	0.8
	Normal <160	0.1
Blood Pressure	Normal (130/89)	0.1
	Low (< 119/79)	0.8
	High (>200/160)	0.9
Blood sugar	High (>120&<400)	0.5
	Normal (>90&<120)	0.1
	Low (<90)	0.4
Heart Rate	Low (< 60bpm)	0.9
	Normal (60 to 100)	0.1
	High (>100bpm)	0.9

The table below shows the set of 25 traits selected for the trials by some researchers that were most relevant to serious cardiovascular disease. Some researchers have conducted experiments on a different real-world data set of some heart patients to study the effect of limitations on that data and to eliminate completely unreliable rules with careful validation of the test set on this important data.

Table 5.2: Characteristics set to predict Cardiovascular disease

Attribute name	Medical meaning	Neg	Constraints		
			itemFilter	group	ac
AGE	Patient age	0	0	0	1
LM	Left Main	1	1	0	2
LAD	Left Anter Desc	1	1	0	2
LCX	Left CircumfleX	1	1	0	2
RCA	Right Coronary	1	1	0	2
AL	Antero-Lateral	0	1	1	1
AS	Antero-Septal	0	1	1	1
SA	Septo-Anterior	0	1	1	1
SI	Septo-Inferior	0	1	1	1
IS	Infero-Septal	0	1	1	1
IL	Infero-Lateral	0	1	1	1
LI	Latero-Inferior	0	1	1	1
LA	Latero-Anterior	0	1	1	1
AP	Apical	0	1	1	1
SEX	Gender	0	0	0	1
HTA	Hypertension Y/N	0	1	0	1
DIAB	Diabetes Y/N	0	1	0	1
HYPLPD	Hyperloip Y/N	0	1	0	1
FHCAD	Faml hist dis Y/N	0	1	0	1
SMOKE	Smokes Y/N	0	1	0	1
CLAUDI	Claudication Y/N	0	1	0	1
PANGIO	Prev angina Y/N	0	1	0	1
PSTROKE	Prior stroke Y/N	0	1	0	1
PCARSUR	Prior surgery Y/N	0	1	0	1
CHOL	Cholesterol	0	0	0	1

Examples of heterogeneous data and noise data:

Table 5.3: Table showing missing data.

age	sex	Cp	rbp	chol	fbs
45	0	0	135	450	1
	1	1	124	500	0
66	1	1	152	388	1
62	1	1	295	251	1
68	1	0	313	285	1
52	1	0	313	595	
80	1	0	125	120	1
45	0		135	111	0
52	1	0	124	313	
55	0	1	152	141	1
41	1	0	295	350	1
	1	1	175	120	1
54	1	0	165	185	0
45	0	1	99	122	0
80	1	0	160	120	0
45	1	0	155	111	0
52	1	0	175	313	0
55	0	1	150	141	0

Table 5.4: Table showing noise data.

age	sex	Cp	rbp	chol	fbs
62	1	0	178	M	1
0	1	0	195	450	0
45	1	1	111	500	0
52	1	0	192	388	1
36	1	1	135	251	0
55	0	0	120	-565	0
D	1	0	122	285	400
50	1	1	135	251	-0
52	1	0	124	384	0
50	1	0	152	401	0
45	1	1	295	299	1
52	1	0	175	350	0
45	1	1	111	500	0
52	1	0	192	388	1
36	1	1	135	251	0
55	0	0	120	100	0
50-	1	0	122	285	1

Table 5.5: Table showing inconsistent data.

age	sex	Cp	rbp	chol	fbs
36	Female	0	Normal	255	0
65	1	0	195	355	0
Less than 30	1	0	111	384	1
45	1	0	192	295	1
44	1	0	135	313	0
66	1	0	122	255	0
62	1	1	144	High	0
Small	1	1	212	384	1
44	1	0	152	200	1
52	1	0	166	400	0
More than 60	1	0	125	389	1
45	One	1	145	295	1
52	1	1	136	313	0
36	Female	0	Normal	255	0
65	1	0	195	355	0
Less than 30	1	0	111	384	1
45	1	0	192	295	1
44	1	0	135	313	0
66	1	0	122	255	0
62	1	1	144	High	0
Small	1	1	212	384	1
44	1	0	152	200	1
66	1	1	136	366	0

6. DATA SOURCE

A patient's mental environment provides access to a variety of clinical data. The term "coronary artery disease" is used to refer to all diseases that directly or indirectly affect the heart. According to the World Health Organization, cardiovascular disease is the leading cause of death globally. The term "cardiovascular disease" generally refers to a variety of diseases that affect the circulatory system. Scientists will review the health records of a group of people with diverse clinical characteristics to determine if they can develop more accurate methods for predicting this deadly disease (see Figure 6-1).

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Figure 6.1: Description of attributes

There are two different types of records, each containing 13 related items that can be used to investigate and estimate injury. There are 300 male records and 400 female records. The accompanying table lists the main points we want to address [41].

Table 6.1: Table showing for some patients with (39) features.

id	age	sex	cp	rbp	chol	fbs	ecg	thalach	angina	oldpeak	slope	ca	concept	dm	thal	smoke	famhist	num
1	65	1	1	120	401	1	1	100	0	2.3	0	0	0	0	0	1	0	3
2	60	1	1	115	299	0	0	155	1	1.6	2	1	0	0	1	0	1	1
3	41	1	0	80	350	1	0	175	0	0	0	0	0	0	0	1	0	3
4	33	1	1	85	120	1	0	145	1	1.5	1	0	0	0	0	0	0	2
5	54	1	0	90	185	0	0	114	1	2.9	2	0	1	0	2	1	1	1
6	45	1	1	99	132	0	1	125	1	3.1	0	0	0	0	1	0	1	1
7	52	1	0	111	200	0	0	175	0	3	1	0	1	0	0	1	0	2
8	36	1	1	313	499	1	0	100	0	2.2	0	0	1	0	2	1	0	3
9	65	1	0	313	200	1	0	166	0	1.5	0	0	1	0	1	1	1	3
10	71	1	1	122	366	0	0	125	0	1.9	1	0	0	0	0	0	0	3
11	45	1	0	135	450	0	1	145	0	0.2	0	0	0	0	0	0	0	1
12	44	1	1	124	500	0	0	100	0	1.8	2	0	0	2	1	0	0	1
13	66	1	1	152	388	1	1	178	1	0.8	0	0	0	0	0	0	0	3
14	62	1	1	295	251	1	0	195	0	0.4	1	0	0	1	1	0	1	3
15	68	1	0	175	285	0	0	111	0	2.6	0	0	0	0	0	0	0	2
16	52	1	0	165	595	1	1	90	1	1.7	0	0	0	1	0	1	0	1
17	80	1	0	160	120	0	0	135	1	1.5	2	0	1	0	2	0	0	3

Table 6.1: Table showing for some patients with (39) features. “tables continued”

18	45	1	0	155	111	0	1	122	0	2.9	1	0	1	0	1	0	0	1
19	52	1	0	175	313	1	0	144	1	3.1	0	0	0	2	0	1	1	3
20	36	1	0	145	400	0	0	90	0	3	2	0	0	1	2	1	0	3
21	65	1	1	114	389	1	1	98	1	2.2	1	0	1	0	1	0	1	1
22	71	1	0	125	295	0	0	136	0	1.5	0	0	0	2	0	1	0	1
23	45	1	0	175	313	1	1	164	0	1.9	0	0	0	1	0	0	0	2
24	44	1	1	180	255	0	0	75	1	0.2	1	0	0	0	1	1	1	3
25	66	1	1	166	355	1	0	66	1	1.8	0	0	1	0	0	0	0	2
26	62	1	0	125	384	0	1	100	0	0.8	2	0	0	1	2	0	1	1
27	68	1	1	145	200	1	0	135	0	3.5	1	2	1	0	1	1	0	1
28	52	1	0	136	366	0	1	124	1	0.3	0	0	0	2	0	1	0	2
29	80	1	1	178	450	1	0	152	0	1.5	2	1	1	1	0	0	1	3
30	60	1	0	195	500	1	0	100	0	0.6	1	0	0	0	2	0	0	2
31	41	1	0	111	388	1	1	120	0	0.1	0	1	1	0	1	1	1	1
32	33	1	1	192	251	1	0	111	0	0.2	0	0	1	2	0	0	0	3
33	54	1	1	135	285	0	0	120	1	1.2	2	0	1	1	2	0	0	3
34	45	1	0	122	595	0	1	150	0	1.3	1	2	0	0	1	0	1	2
35	52	1	0	144	120	1	0	96	0	2.4	0	1	0	2	0	0	0	1
36	36	1	1	212	111	0	1	99	0	0.8	2	0	0	1	2	1	1	1
37	65	1	0	152	313	0	0	67	0	0.4	1	2	0	0	0	0	0	2

Table 6.1: Table showing for some patients with (39) features. “tables continued”

38	71	1	0	136	400	0	0	125	1	2.6	0	1	0	2	0	0	0	2
39	45	1	0	164	389	0	1	100	0	1.7	2	0	0	0	0	0	1	3
40	44	1	0	313	295	1	0	90	1	1.5	0	0	1	0	1	0	0	1
41	66	1	1	313	313	0	1	100	0	2.9	0	1	0	0	0	1	0	2
42	62	1	0	122	255	0	1	125	0	3.1	2	0	0	1	0	0	1	3
43	68	1	0	135	355	1	0	145	1	3	1	2	1	0	2	1	0	3
44	52	1	0	124	384	0	0	136	0	2.2	0	1	1	0	1	0	1	3
45	80	1	0	152	401	0	1	100	1	1.5	2	0	0	2	0	0	0	3
46	45	1	1	295	299	1	0	120	0	1.9	1	0	1	1	2	1	0	2
47	52	1	0	175	350	0	0	111	0	0.2	2	2	0	0	1	0	1	2
48	36	1	1	165	120	1	0	85	1	1.8	0	1	1	2	0	1	0	1
49	65	1	0	180	185	1	1	88	0	0.8	0	0	0	1	0	0	1	2
50	71	1	0	166	132	1	1	92	1	3.5	2	2	0	0	1	0	1	2

Table 6.2: Table showing for some patients with (33) features.

id	age	sex	cp	ecg	thalach	CA	slope	angina	smoke	Old peak	famhist	num
1	65	1	1	1	100	1	1	0	1	0	0	3
2	60	1	1	0	155	0	1	1	0	1	1	1
3	41	1	0	0	175	1	0	0	1	0	0	3
4	33	1	1	0	145	0	1	1	0	0	0	2
5	54	1	0	0	114	1	0	1	1	1	1	1
6	45	1	1	1	125	0	1	1	0	1	1	1
7	52	1	0	0	175	1	0	0	1	0	0	2
8	36	1	1	0	100	1	1	0	1	0	0	3
9	65	1	0	0	166	1	0	0	1	1	1	3
10	71	1	1	0	125	0	1	0	0	0	0	3
11	45	1	0	1	145	0	0	0	0	0	0	1
12	44	1	1	0	100	0	1	0	0	0	0	1
13	66	1	1	1	178	0	1	1	0	0	0	3
14	62	1	1	0	195	0	1	0	0	1	1	3
15	68	1	0	0	111	0	0	0	0	0	0	2
16	52	1	0	1	90	1	0	1	1	0	0	1
17	80	1	0	0	135	0	0	1	0	0	0	3
18	45	1	0	1	122	0	0	0	0	0	0	1
19	52	1	0	0	144	1	0	1	1	1	1	3
20	36	1	0	0	90	1	0	0	1	0	0	3
21	65	1	1	1	98	0	1	1	0	1	1	1
22	71	1	0	0	136	1	0	0	1	0	0	1
23	45	1	0	1	164	0	0	0	0	0	0	2
24	44	1	1	0	75	1	1	1	1	1	1	3
25	66	1	1	0	66	0	1	1	0	0	0	2
26	62	1	0	1	100	0	0	0	0	1	1	1

Table 6.2: Table showing for some patients with (33) features. “tables continued”

27	68	1	1	0	135	1	1	0	1	0	0	1
28	52	1	0	1	124	1	0	1	1	0	0	2
29	80	1	1	0	152	0	1	0	0	1	1	3
30	60	1	0	0	100	0	0	0	0	0	0	2
31	41	1	0	1	120	1	0	0	1	1	1	1
32	33	1	1	0	111	0	1	0	0	0	0	3
33	54	1	1	0	120	0	1	1	0	0	0	3
34	45	1	0	1	150	0	0	0	0	1	1	2
35	52	1	0	0	96	0	0	0	0	0	0	1
36	36	1	1	1	99	1	1	0	1	1	1	1
37	65	1	0	0	67	0	0	0	0	0	0	2
38	71	1	0	0	125	0	0	1	0	0	0	2
39	45	1	0	1	100	0	0	0	0	1	1	3
40	44	1	0	0	90	0	0	1	0	0	0	1
41	66	1	1	1	100	1	1	0	1	0	0	2
42	62	1	0	1	125	0	0	0	0	1	1	3
43	68	1	0	0	145	1	0	1	1	0	0	3
44	52	1	0	0	136	0	0	0	0	1	1	3
45	80	1	0	1	100	0	0	1	0	0	0	3
46	45	1	1	0	120	1	1	0	1	0	0	2
47	52	1	0	0	111	0	0	0	0	1	1	2
48	36	1	1	0	85	1	1	1	1	0	0	1
49	65	1	0	1	88	0	0	0	0	1	1	2

7. RESULTS

7.1 RESULTS OF LEARNING ALGORITHMS TO PREDICT DISEASE RISK

7.1.1 The Results of using Learning Algorithms Compared to Traditional Algorithms

The results show that clinical information is sensitive and has to be handled extremely carefully since the findings made after using the information cleaning approach are more precise than the results made before using the cleaning system. The review method entails doing three computations on various pieces of cleansed clinical data, comparing the results with those of previous exams, and repeating the process.

The critical drawbacks of this study are the clinical data's veracity, as the outcomes depend on it to make the best decision and accurately assess the condition. The clinical information purging cycle in the healthcare business may be aided by the results of this study, and the challenges associated with selecting the best course of action for a given condition may be addressed. Perhaps the biggest challenge facing the investigation is choosing the correct individual to incorporate clinical data, which establishes the accuracy of the results [42].

Consequently, specialists should take their work seriously and communicate reliable data in order to reach the highest level of precision.

In Table 7-1, the outcomes of three cleaning strategies are shown.

Table 7.1: Results for the classifiers

Techniques	Accuracy Before cleaning	Accuracy After cleaning
Decision tree	95.41	98.15
Naïve Bayes	94.49	95.44
Neural network	65.96	89.16

Information mining techniques and computations used to calculate cardiac disease are robust and sophisticated. The outcomes of the extraction interaction are displayed in Table 7-1. These results are more accurate in identifying this catastrophic illness when compared to those of other research since our study investigated uncorrelated information or irrational information and then incorporated a succession of important applications throughout the period spent information mining.

7.1.2 Results of Selecting the Characteristics Most Relevant to Cardiovascular Disease

Various information mining techniques are used to identify and research these 13 traits. Decision trees consistently perform superior in terms of exactness and execution when compared to other techniques with regard to coronary sickness expectancy. When applied to the option tree, Bayesian grouping can provide amazing results [40]. ID: The unique ID number assigned to the client. Between the ages of 1 and 0, men dwarf women. In the case of chest discomfort, respond as follows: RBP: If the answer is yes and the answer is no, take a break. Hypertension, a fasting blood glucose level more than or equal to 120 milligrams per deciliter, and the assessment of cholesterol per blood deciliter Effects of the resting ECG for THALACH: most noticeable pulse One (1) case of angina pectoris, none (0). Old pinnacle with low ST and a sloping surface. Number of important vessels as determined by fluoroscopy shaded: CA (esteem 0 3) Topical category: Vascular status of diabetic mellitus (DM): history of diabetes = 1, no history of diabetes = 0 in THAL, where 1 addresses the usual value and 0 represents the decent defect. There are just two possible answers for to: "yes" or "no." Yes,

FAMHIST and FAMHIST might be used to differentiate between cardiovascular problems. Generally speaking, a safe bet is similar to 1, a medium bet to 2, and a big bet to 3, individually. The outcomes are very comparable to those of the decision tree method. This method differs from others in that it is quick and efficient in addition to having a high level of precision. Decision trees may be subjected to a genetic computation in order to anticipate the outcomes of a cardiac condition while also reducing the amount of information. The necessity to concentrate on the method of organizing that crucial information and then applying those calculations to it, as shown in Table 1, is necessitated by the distinctions in scientific findings.

$$P(c|x) = P(x|c)\, P(c)\, /P(x) \tag{7.1}$$

Decision Tree,K- Nearest Neighbor (KNN), and Naive Bayes are three famous data mining techniques that stand out for their simplicity and drawbacks, respectively. A trustworthy, simple-to-understand, and effective technique for sorting is Guileless Bayes Innovation. In terms of the probability of P (c | x) from the preceding probabilities of P (c), P (x | c), and the probabilities of consistency, it is mostly based on Paez's conjecture, which is as per the following [39].

7.2 COMPARISON WITH PREVIOUS WORKS

The review interaction's findings demonstrate that, among the several computations, the decision tree approach is the best one for handling information that is highly uncorrelated. Against determine the validity of this technique; the decision tree computation was compared to a variety of earlier worker models that made use of comparable knowledge gathering. The method of foreseeing this severe disease serves as a demonstration of the accuracy of the computation [41].

8. DISCUSSION

The variety of findings demonstrates that significant barriers emerge when therapeutic decisions are based on clinical expertise rather than the techniques used to conceal facts.

The suggested method involved a number of information cleansing stages, a number of computations, and the production of findings that were acceptable when compared to the earlier discoveries.

This study has some flaws since controlling human existence is extremely dangerous. Therefore, it is possible for unlucky patient assistance to result in excessive errors and exorbitant costs. As we have already discovered, these errors are disregarded in the anticipation and analysis of this illness since they may result in the demise of several people. In this environment, it is harder to identify the most important concerns [37].

This investigation can improve the expectation cycle since the cleaning approach, heredity calculation-based selection of the most suitable qualities, and subsequent application of the calculations produced beneficial results.

The process utilized to gather clinical information for parties with relation to cardiac sickness may be advantageous to many qualified professionals and students since it may be used to make rational clinical judgments.

9. CONCLUSIONS

a. Only 13 parameters were chosen for further investigation because of their greater significance in identifying the onset of coronary supply pathway disease. This information may be obtained by using data development and processing.

b. Compared to other approaches, the Naive Bayes performs second and the Neural Network third in terms of precision and execution. In terms of execution, probable accuracy, and extreme judgment, decision trees come out on top. Because of the gravity of the situation, several concerns must be addressed before making a choice. In order to aid chiefs, you should improve your capacity to predict cardiovascular disease by using cost-effective diagnostic tests and genetic computations.

c. When compared to earlier studies that used comparable clinical data, our method-which makes use of the decision tree calculation—performs better and yields more accurate findings.

REFERENCES

[1] A. Kumar, P. Kumar, A. Srivastava, V. D. Ambeth Kumar, K. Vengatesan, and A. Singhal, "Comparative analysis of data mining techniques to predict heart disease for diabetic patients," in International Conference on Advances in Computing and Data Sciences, 2020, pp. 507–518.

[2] G. T. Reddy, M. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," Evol. Intell., vol. 13, no. 2, pp. 185–196, 2020.

[3] S. Caravita et al., "Haemodynamics to predict outcome in pulmonary hypertension due to left heart disease: a meta-analysis," Eur. Respir. J., vol. 51, no. 4, 2018.

[4] M. A. Khan and F. Algarni, "A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS," IEEE Access, vol. 8, pp. 122259–122269, 2020.

[5] D. Kinge and S. K. Gaikwad, "Survey on data mining techniques for disease prediction," *Int. Res. J. Eng. Technol.*, vol. 5, no. 01, pp. 630–636, 2018.

[6] 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.

[7] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison," Comput. Biol. Med., vol. 136, p. 104672, 2021.

[8] A. Mehmood et al., "Prediction of heart disease using deep convolutional neural networks," Arab. J. Sci. Eng., vol. 46, no. 4, pp. 3409–3422, 2021.

[9] D. Sellappan and R. Srinivasan, "Association rule-mining-based intrusion detection system with entropy-based feature selection: Intrusion detection system," in Handbook of Research on Intelligent Data Processing and Information Security Systems, IGI Global, 2020, pp. 1–24.

[11] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS international conference on computer systems and applications, 2008, pp. 108–115.

[12] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, "Comparing deep and classical

machine learning methods for human activity recognition using wrist accelerometer," in Proceedings of the IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence, New York, NY, USA, 2016, vol. 10, p. 970.

[13] S. Islam, N. Jahan, and M. E. Khatun, "Cardiovascular disease forecast using machine learning paradigms," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 487–490.

[14] Y. Pan et al., "Brain tumor grading based on neural networks and convolutional neural networks," in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 699–702.

[15] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *2010 5th International Conference on Computer Science & Education*, 2010, pp. 1344–1349.

[16] S. Archana and K. Elangovan, "Survey of classification techniques in data mining," Int. J. Comput. Sci. Mob. Appl., vol. 2, no. 2, pp. 65–71, 2014.

[17] A. Dey, J. Singh, and N. Singh, "Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis," *Int. J. Comput. Appl.*, vol. 140, no. 2, pp. 27–31, 2016.

[18] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 96–104, 2013.

[19] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[20] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019.

[21] A. Powar, S. Shilvant, V. Pawar, V. Parab, P. Shetgaonkar, and S. Aswale, "Data Mining & Artificial Intelligence Techniques for Prediction of Heart Disorders: A Survey," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1–7.

[22] H. G. Lee, K. Y. Noh, and K. H. Ryu, "Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2007, pp. 218–228.

[23] V. Chaurasia, S. Pal, and B. B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” *J. Algorithm. Comput. Technol.*, vol. 12, no. 2, pp. 119–126, 2018.

[24] V. Chaurasia and S. Pal, “A novel approach for breast cancer detection using data mining techniques,” *Int. J. Innov. Res. Comput. Commun. Eng. (An ISO 3297 2007 Certif. Organ. Vol*, vol. 2, 2017.

[25] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in *2008 IEEE/ACS international conference on computer systems and applications*, 2008, pp. 108–115.

[26] S. P. Rajamhoana, C. A. Devi, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, “Analysis of neural networks based heart disease prediction system,” in 2018 11th international conference on human system interaction (HSI), 2018, pp. 233–239.

[27] T. J. Peter and K. Somasundaram, “An empirical study on prediction of heart disease using classification data mining techniques,” in *IEEE-International conference on advances in engineering, science and management (ICAESM-2012)*, 2012, pp. 514–518.

[28] S. B. Patel, P. K. Yadav, and D. P. Shukla, “Predict the diagnosis of heart disease patients using classification mining techniques,” *IOSR J. Agric. Vet. Sci.*, vol. 4, no. 2, pp. 61–64, 2013.

[29] M. B. Priya, P. L. Juliet, and P. R. Tamilselvi, “Performance analysis of liver disease prediction using machine learning algorithms,” *Int. Res. J. Eng. Technol.*, vol. 5, no. 1, pp. 206–211, 2018.

[30] M. Anbarasi, E. Anupriya, and N. Iyengar, “Enhanced prediction of heart disease with feature subset selection using genetic algorithm,” Int. J. Eng. Sci. Technol., vol. 2, no. 10, pp. 5370–5376, 2010.

[31] X.-F. Zhang, J. Attia, C. D’Este, X.-H. Yu, and X.-G. Wu, “A risk score predicted coronary heart disease and stroke in a Chinese cohort,” J. Clin. Epidemiol., vol. 58, no. 9, pp. 951–958, 2005.

[32] M. Anbarasi, E. Anupriya, and N. Iyengar, “Enhanced prediction of heart disease with feature subset selection using genetic algorithm,” *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.

[33] N. A. Setiawan, P. A. Venkatachalam, and A. M. H. Fadzil, “Rule selection for coronary artery disease diagnosis based on rough set,” *Int. J. Recent Trends Eng.*, vol. 2, no. 5, p. 198, 2009.

[34] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in 2013 IEEE conference on information & communication technologies, 2013, pp. 1227–1231.

[35] K. Yang *et al.*, "HerGePred: heterogeneous network embedding representation for disease gene prediction," *IEEE J. Biomed. Heal. informatics*, vol. 23, no. 4, pp. 1805–1815, 2018.

[36] A. N. Arbain and B. Y. P. Balakrishnan, "A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data," *Int. J. Data Sci. Adv. Anal.*, vol. 1, no. 1, pp. 1–11, 2019.

[37] P. E. Puddu and A. Menotti, "Artificial neural network versus multiple logistic function to predict 25-year coronary heart disease mortality in the Seven Countries Study," Eur. J. Cardiovasc. Prev. Rehabil., vol. 16, no. 5, pp. 583–591, 2009.

[38] P. Ghosh et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021.

[39] M. C. Beyene and P. Kamat, "Survey on prediction and analysis the occurrence of heart disease using data mining techniques," *Int. J. Pure Appl. Math.*, vol. 118, no. 8, pp. 165–174, 2018.

[40] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, and P. Singh, "Latest trends on heart disease prediction using machine learning and image fusion," Mater. Today Proc., vol. 37, pp. 3213–3218, 2021.

[41] R. Shinde, S. Arjun, P. Patil, and J. Waghmare, "An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm," *IJCSIT) Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 637–639, 2015.

[42] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, and K. Li, "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing," *Inf. Sci. (Ny).*, vol. 435, pp. 124–149, 2018.

[43] M. J. A. Alkhafaji, A. F. Aljuboori, and A. A. Ibrahim, "Clean medical data and predict heart disease," 2020, doi: 10.1109/HORA49412.2020.9152870.

APPENDIX A

Decision Tree	**Accuracy and performance**
Decision Tree with 13 features	98.15%
Decision Tree with 19 features	97.50%